JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Novel E- Focused Crawler and Enhanced k- mean (n-gram) clustering technique for Automatic classification of attribute level customer healthcare sentiments

Saroj kushwah
Noida International University
sarojkushwahsiem@gmail.com

Neeta Sharma
Noida International University
neeta.sharma@niu.edu.in

Sanjoy Das
Dept. of Computer Science
Indira Gandhi National Tribal University, Imphal, India
sdas.jnu@gmail.com

ABSTRACT

Nowadays, there is a big trend in e-commerce websites where a large volume of valuable consumer sentiment data increases at an astonishing rate. Analyzing, summarizing, and explaining these big data (sentiments) quickly and efficiently is critical for data scientists. Many works exist in the public domain, such as hospitality, movies, hotel, electronic, and political reviews. Medical data analysis is a less explored area, and lots of research possibilities exist. Many people search for outsiders' opinions or views on different healthcare web portals, blogs and social sites related to a healthcare product, method, and services but cannot decide either sentiment is positive or negative. In this paper, we focus on attribute-level healthcare sentiment analysis. Firstly, design a Novel E-focused crawler and then pre-processing gathered reviews to remove noise content/outlier and improve storage capacity, efficiency, and Accuracy of the proposed system's relevant document. Secondly, corpus, Part of speech, thesaurus, and SentiWordNet dictionary are used to find out the implicit, explicit attributes and the polarity of sentiments.

At last, unsupervised clustering as Enhance K-Means (n-gram)-a machine learning approach is proposed. This proposed methodology can apply to several healthcare products and domains. The results obtained show our novel proposed approach outperformed all the existing methods in terms of Accuracy of 92%precision of 90 %, recall of 96%, F-Measure of 93%, G-Mean of 91%.

Keyword - Clustering, Enhance K-means (n-gram), sentiment analysis, focused crawler, healthcare.

1 Introduction

Disease surveillance is an information-based activity comprising the gathering, analysis, and interpretation of a considerable amount of data (sentiments) growing at the different social sites and other sources. This information gathered is then accustomed in distinct ways to calculate the effectiveness and efficiency of preventative health measures. To achieve this purpose, surveillance for a disease or other health problems should have clear objectives. Social media have drastically changed how we communicate, and reviews are accessible in minimum time, speedily, with men and women worldwide providing status updates with sentiments/reviews, videos, pictures, etc.(Denecke & Deng, 2015a).For men and women, health surveillance social media has come to be a moderately appropriate tool. It has been said for ages that awareness is the key to solve any problem from its root. In this era of Data Science, social media had played a crucial role in creating awareness to deal with diseases, unlike in the past when people were scarce of proper medical aid leading to a lack of time in getting treatment. In modern times, People can easily adapt to these platforms and get in touch with their doctors without a physical appointment where we are supposed to stand in a long queue for our turn. Social media is

also a cost-effective medium for public health and provides valuable information much faster than old age (Lim et al., 2017).

Monitoring any pandemic disease is the most important because it spreads on a large scale and can create an unfavorable and dangerous situation in society (Dai & Bikdash, 2016).
The equipment and other related material should be perfect in quantity and quality. For example, laboratories, clinics, health care systems, surgeons, and emergency departments through which the perfect information can be gathered and provided(*Overview of Influenza Surveillance in the United States*, 2016). This information should be gathered regularly and monitored (Dai & Bikdash, 2017)(Ginsberg et al., 2009).

Social media like Facebook, Twitter, Pinterest and Instagram, is the most trending platform through which a vast volume of data is shared quickly among millions of users. For example, Twitter is the best social media platform where 284 million people use per day, and more than 500 million Twitter are shared(Twitter, 2016). Through social media, users share their feelings, opinions, and activities(B. Singh et al., 2015). The study of social media shows that it is also helpful for public healthcare information.

The collection of information nature has revolved around people's thinking. Today, many new challenges are coming every day for data scientists while reviews are becoming popular and creating astonishing rates like online feedback, personal blogs, microblogging platforms, forums, etc. (Cunha et al., 2015). Data science is the most helpful (like an umbrella); it includes ML, data analytics, data mining, and other disciplines. Data scientists collect information through many different sources and use sentiment analysis and predictive analytics to find accurate information that helps retail industries, a trade organization, etc.

Data Science has enhanced to analyze the vast amount of opinions from various establishments. These sentiments are posted by different people on different sites related to various domains like product material, machinery, and estate building construction, day to day personal products, consumer maintenance services, hospitality, movies, T.V broadcast, and many more(X. Yu et al., 2012)(Cruz et al., 2010). Producing a summary of product features from collected reviews based on their sentiments is one of the most critical problems the researcher faces (Zhai et al., 2011).

The public health care domain is less explored. Much research scope is there on public health care attributes, especially on the oral and contraceptive methods for their effectiveness, cost, maintenance, STDs, and side effect analysis. The users' positive or negative comments and feedback on the market's product help new users decide the right things.

The rest of the paper is organized as; in section-2 Motivation of the work is discussed. In Section-3, we have presented various related works. Section-4 elaborates discussion on Attribute-based contraceptive method categorization. Section-5 includes a brief dataset description. Section-6 evaluation and results. Section -7 shows the result discussion. Finally, the paper is concluded in section-8 and in section-9 shows future direction.

2 Motivation for work

Most of the researcher mainly focuses on document level opinion extraction. They extract objective information from the opinion of the review. (Nguyen et al., 2020), examine the difference in contraceptive adoption by a group of obesity specified undesirable pregnancy.
(Deschênes et al., 2019)analyze the collaborative opinion of texts reporting functional psychological and physical attributes of recovery executive exploration examination of sensuality with distinct in addition to moderate to TBI, was organized. (LORRAINE, GOEURIOT JIN-CHEON, et al., 2011)describes the outcomes of the begging study of data prepared and linguistic features of posts on various sites related to drug – with attention opinion along with sentiments posted. This comprehensive opinion recommends an ecological framework to analyze research prescribing multilevel providers to health disparities among Latina delivery womanhood in rural U.S. regions. A dataset (Schminkey et al., 2019) contained 660 journals and pre-processed for scaling and fill out missing principles. To consider health level(Halim &

Khan, 2019) and analyze anomalies in complicated core router system. Conventional methods fail to determine unusual/ suspicious patterns when the examined data includes temporary measurements (Jin et al., 2019).
In (S. Singh et al., 2018) notice the frequency of adoption of contraceptives, contraceptives choose, or also undesirable demand for birth control method from non-pregnant 15 to 49 age group married women using manipulated as well as pre-tested utensils by the representative survey.(Benson et al., 2016),notice the method of contraception acceptance by 319,385 females searching abortion in 2326 publicly-owned health services from 2011 to 2013. Health-related Ministry services of abortion and contraception methods, in addition to scientific collaborators from Ipas. (Al Riyami et al., 2004) explain nationwide research in 2000 that basic data on evermore family woman female's authorization in Oman, analysis relates to women's empowerment or the response of authorization on unsatisfied demand for birth control. (Chaurasia, 2014), examine the pattern of contraception methods in our nation by competing use of contraception among females based on various demographics, cultures, economies.(Simmons et al., 2017),consider notwithstanding the interaction between amidst rifamycins and HC outcome in each therapy's pointed efficacy or elevated poisonousness.(Alabi et al., 2019),review the job of women's independence in using birth control methods amongst recently married females in northern Nigeria.(Caetano et al., 2019) study the rate of undesirable pregnancy is incredibly high in Rates that can be related to their predisposition not to recommend accurate birth control regularly.
In big data, OMGA is accustomed to an invaluable way of partitioning the reviews into various sentiments and analyzing people's attitudes (Shayaa et al., 2018). To summarize the outcomes of the questionnaire Fuzzy co-clustering is evaluated individual questions.(Honda et al., 2019). In (Y. Yu et al., 2015) proposed a Cludoop algorithm, a distributed density-based clustering technique used in big data using Hadoop. (Orkphol & Yang, 2019), studies on sentiment analysis using microblogging with K-Mean and Artificial Bee colony, (Chung et al., 2019)and (Khalid & Prieto-alhambra, 2019) study sentiment analysis using the machine learning approach while Some of the researchers work on sentence-level opinion mining and a knowledge-based approach. (Tg et al., 2019) studied only the long-acting contraception method. (Kumari & Savita, 2018) used the contraceptive attitude scale (Questionnaire) to analyze the opinion related to each statement and response of contraceptives. The number of researchers who have been studied the aspect or attribute-based opinion mining is very less. The researcher studied the same using contraceptive methods related to the health domain in which different attributes and corresponding sentiment are extracted from the user-generated opinion is even less by the researcher.

Many authors only consider stopwords and stemming in S. Das et al. (Das et al., 2016),(Gopalakrishnan & Ramaswamy, 2017) proposed a PNN RBFN for opinion classification. Adverse drug reactions (ADRs) is analyzed from filtering big data on Social Medias.(Yang et al., 2015). This approach(Samriya et al., 2016)based on a single organization and language (English) or could not apply to another organization that usage various documentation strategies, methods, or documentation system commanding to distinct documentation pattern (Tang et al., 2018), (Bharat Singh et al., 2016) removed stopwords and suffixes from the root word. (Mohanadevi, 2017) also, extract modest words(stem & stop words).(Shobana et al., 2019) pre-processing of raw data and considers the standard dictionary, whitespace, and tokenization.(Krishnan & Amarthaluri, 2019). Some author does not consider spell-checker and stemming (Dai et al., 2017)and eliminate words which do not begin with a lower or upper case letter (Aa, Bb, Cc,…Zz) and not use stemming (Chandra Pandey et al., 2017) and not convert numbers as well as all lowercase, etc. and some other author also not consider -,{},/,\,[],(), etc. respectively(Coletta et al., 2014), (Pattani, 2016).
However, it can also be said that the literature alone does not contribute much to filter the noise content or outlier, and it is one of the significant challenges in pre-processing. So it is needed to address the limitation of their methods and to find out a solution. Our proposed work's major contributions are given below: we have collected online and offline reviews of consumers from developed and developing countries through the feedback form of contraception methods and other multimedia sources. The main improvement of pre-processing to keep at minimum time, improve the storage capacity, and maximize Accuracy. We also proposed a Novel E-focused crawler to retrieve only topic-specific document pages to reduce the network traffic and download and remove error or noise opinion. Using a corpus, context information is used to identify the relevant and irrelevant documents to improve the relevant document's relevancy and efficiency. Identifying the implicit feature is a challenging task; for this purpose, we use the corpus or seed list of the attributes and match the sentiment word to the available document reviews and find corresponding parameters. We can get a more accurate result based on the expansion of the seed list of polarity identification. We have proposed an Enhance K- Means clustering algorithm (n-gram), which is highly efficient and effective in categorizing attributes, which ultimately improves the Accuracy and efficiency of clustering results of healthcare products.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

3 Related work

In (Jain & Cherikkallil, 2018) analyze and retrieve different information as diseases, sentiment, symptoms, treatment, location, etc., from healthcare data collected from Twitter. Medinsights is a twitter-based systematic or scientific healthcare platform to retrieve and analyze sentiments of healthcare data. A gradient boosting classifier is used to classify or analyze tweets sentiments into the medical or non-medical domain. Word2vec word embedding with a feed-forward neural network is used. Medical entities are extracted by conditional random field (CRF) from the tweets. This methodology's advantage is that the inference system helps increase the consumer's knowledge base on their stated question(Valsamidis et al., 2013) elaborate on a framework that retrieves valuable recorded knowledge (positive and negative sentiments) from agriculture blogs. It includes the creation of a weblog, opinion accumulation, and text cleaning processing of data. The advantages of a framework to the analysis of farmer perception by agriculture attributes utilizing opinion mining tools will improve their care by the establishment. This framework was not used in other agriculture blogs to get a better benchmark. The Accuracy could increase if the author comprises this approach with Word Sense Disambiguation (WSD) program.

In (Nithya et al., 2013)describe the different clustering tactics in medical science and the method of improving the design of clustering methods in ahead augmentation. Clustering is not perfect that is accustomed to collect the data points, not including any extra information of group definition. Clustering is used to find the internal grouping of unbalanced data. These methods are used to find the exact beneficial data from a variety of databases. To get the exact data is the primary purpose of clustering in the medical field.

In (Mohanadevi, 2017) use the auditable manifestation or symptoms name, which is attached with the word as three views from clinical notes, and after that, applied multi-views NMF and K-means clustering of documents. Use of any idea or techniques for analyzing the different datasets to check and calculate the correctness through parallel views NMF and K-means NMI as check the metrics and observe results. This whole technique gives the knowledge that after using pharmaceutical and manifestation names, clustering performances may be increased, and in the Comparison of K-means techniques, parallel view NMF gives a better result.

In (Shobana et al., 2019) discuss independent patient profiles to get the correctness of pharmaceutical related to every manifestation clinical notes. The proposed method consists of 7 phases; section annotator, word/sentence annotator, negation annotator, pharmacy annotator, manifestation or symptoms annotator, age, and climate annotator.

In (Garima et al., 2015) discuss and compare several clustering algorithms such as K-means, CURE, CLARINS, CLARA, DBSCAN, etc. The partitioned clustering algorithms are decidedly beneficial when convex shape clusters coming to have equivalent size. Hierarchical clustering algorithms are useful for sizable datasets. DBSCAN is also decidedly beneficial in excavating sizable datasets as they can effortlessly recognize noise.

In (Reddy et al., 2019) is solving the problem and complications in recovering datasets which is associated with the various clusters. In other words, clustering of healthcare big data using fuzzy c-means algorithms. In this fuzzy implementation, the data of each patient is collected and stored based on their disease. This algorithm will provide methodical results in the comparison of other clustering algorithms.

In (Bigorra et al., 2019) focuses to glean a procedure in the case of classification of the set of extracted features in terms of one-dimensional and enthrall kano-categories. Kano suggested three basic categories. These are (1) Must be (M) (2)One-dimensional (3)Attraction. With this procedure's help, the target setting process can be improved for the design team for objectivity, correctness, and efficiency. With this procedure's help, the target setting process can be improved for the design team for objectivity, correctness, and efficiency.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

In (Rajesh & Rao, 2019) noted that the methodology that uses TSC to cluster TS (Time Series) data are (1) Prepare the standard clustering approaches (2) Convert Time Series data into entity viz given categorization algorithm. (3) Multiple-step clustering methodology. The data after the research is compared with that works, which consider inter-patient categorization based on AAMI. It is necessary to recognize the unsupervised method for a long time ECG data monitoring and poured to find the cardiovascular disease as soon as possible to avoid premature deaths.

In (Primpeli et al., 2019) explore the advantages of the semantic web indicates by the WDC training dataset. To find a suitable product offer from 79 thousand e-shops is not possible but not easy too. The researchers determined objects originating from different sources and considered that the WDC training dataset and the gold standard were helpful. For the task of goods coordinating, both artifacts give to evolve an extra considering of the intensity of latent semantic representation and deep neural N/W.

In (K. Shyam Sunder Reddy & Bindu, 2017)elaborate a report is prepared and presented regarding various density-based clustering for maximum data examination, presenting a comprehension differentiation among the various techniques based upon variant evaluation metrics. Calculating the dataset as various new challenges are constituted by static datasets, clustering data streams like minimal time and memory, removing noisy content, and controlling varying and high-dimensional data. It is observed that none of the algorithms can handle all these challenging issues.
In (Denecke & Deng, 2015b) discusses the medical sentiment concerns the patient's health status, medical conditions, and treatment.
(Shivaprasad & Shetty, 2017) describe the classification of different types of sentiments analysis methods are presented. They discussed lexicon-based and machine learning methodologies. Due to the complexity involved in human language and communication, it becomes a complex process. The communication and understanding way of human and machine is different. It is observed that in the case of a suitable decision for a particular service or a product, sentiment analysis has to lead the leading role.
TF–IDF(Bafna et al., 2016) is accustomed to fuzzy K-means as well as hierarchical clustering. The experiment considers in two stages the 1st stage analysis the most appropriate approach, and in the 2nd stage, it is adapted to increased data set. To identify data duplication available in the corpus etc. Outcomes attained after processing several datasets represent the efficacy and effectiveness of the methodology. Domain-specific better semantic relativity concepts should be used to achieve better outcomes.

In (Samriya et al., 2016) have worked on the clustering of healthcare data which is employed by various migraine algorithms under the Weka tool. They found that migraine works better for better medical services and must think about how the maximum data can be stored, included, and mined. In these possible orders, for the increase, the profit, distribution of data is included in the organization. Instead of the medical-related data does not distribute with quantitative data like doctor's data.
In (Ogbuabor & F. N, 2018) noted the clustering method's performance such as DBSCAN & K-Mean available in medical services. The clustering algorithm performance is evaluated, datasets taken from the "myhealthavtar" domain. The result analysis shows that the k-means algorithm performs better than DBSCAN techniques in terms of execution time and Accuracy of clustering.

4 Methodology

Attribute-based contraceptive method categorization including five steps:

- Fetch the web pages using a Novel E-focused crawler (tf-idf as well as pre-processing).
- Pre-processing of collected reviews.
- Distinguish the implicit and explicit parameters and sentiments (opinion words) from sentences.
- Recognize the orientation of sentiment words along with attributes.
- Cluster, the attributes of the contraceptive method using Enhance K-mean clustering (n-Gram).

4.1 Collection of review from different sources online as well as offline

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Online reviews played their role as a trend and passion. Approximately 84% of consumers said they consult reviews online, read the positive and negative comments, and star ratings through software.
The information is gathered online and offline through the feedback form of contraceptive method or E-commerce sites. We have collected online reviews through the web such as healthcare blogs or sites, E-commerce sites (amazon, Flipkart, etc.), and approximately 49% of consumers reviews have collected through their network as Instagram, Facebook, gamil.com, etc. The rest of the reviews are gathered online as well as offline through the feedback form of the contraceptive method from the patient of private and government hospitals and a couple of eastern or western areas of U.P. shown in Fig.2 and makes a report when they satisfy; they take their right decision.

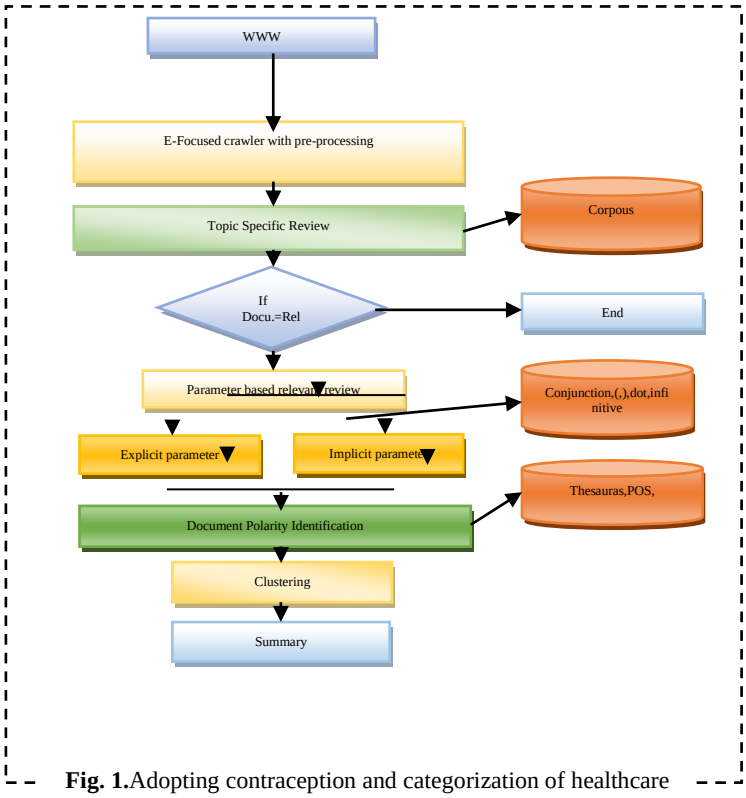


Fig. 1. Adopting contraception and categorization of healthcare sentiments- A machine learning approach

The demonstration of the workflow diagram of our proposed method in Fig. 1. Where Douc is the total number of document reviews, and Rel is the relevant document review.

We use the interview schedule, which consists of a series of audio and video interviews of family planning and their success. The essential components needed for the success of family planning are (1) Conducting an evidence-based interview, (2)Effective leadership and management,(3)Ensuring effective communication,(4)Proper protection in using Contraceptives, (5)Deploying trained supporting staff

Through this form, we have collected the reviews online and offline. Through the best communication, staff can explain their views to the illiterate or uneducated person who is so far from the developing area.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

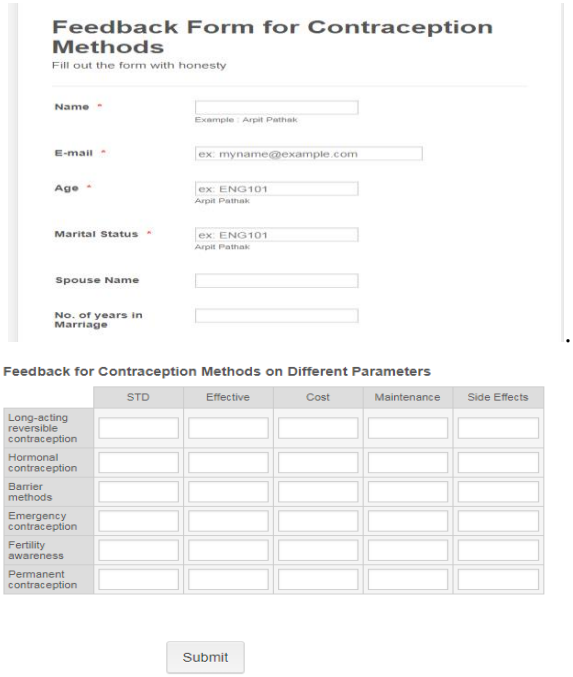


Fig. 2. Feedback form for contraception method on different parameters.

4.1.1 Collect the reviews from social media using Novel E-Focused crawler

Based on the previous reviews and crawled reviews, the corpus is extended for words, creating problems in sentiment analysis. The web is being constantly monitored to gather reviews. A focused crawler is being used where the download of the documents is being followed by the stage of verification of the documents' domain.

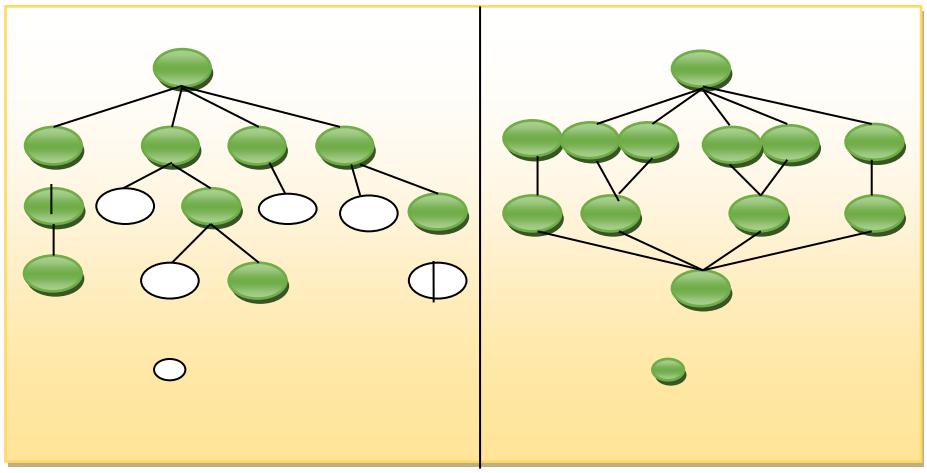


Fig. 3. Universal crawler in the left-hand side and focused crawler on the right-hand side

Web or spider crawler is a program that regularly transmits the web's hyperlink then saves this linked page to local storage. Two types of crawler: simple or universal crawler and focused crawler. Fig. 3 shows that a universal crawler compiles as many pages to form the specific set of URLs. On the other hand, focused crawler compiles the document on a particular topic to reduce the network traffic and download.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

In focused crawling research, the prioritization of unfrequented URLs (webpages) in the crawling frontier during proceeding these URLs, followed by very close to their priority, is the main issue or difficulty. The utmost frequent feature, took in different focused crawling studies, prioritizing unvisited Urls is the relevance of its origin Urls, as, its in-linked URLs. Although it is a limited feature, we cannot consider the frequency of the unvisited Urls accurately if we have some origin URLs. Solving this problem or issue and increase the efficiency of a focused crawler, we propose a new focused crawler, namely Novel E-Focused crawler, which is consists of a combination of two (1) Implemented pre-processing (shown in Fig.2) (2) tf-idf.

The Novel E-focused crawler can predict the probability that any unvisited page will be taken into consideration before it is downloaded. It means to find the good nodes out of the whole web page or graph. This crawler is not interested in all URLs at the same time in a particular domain. It downloads the whole page of domain-specific Urls, collects these links afterward. Then, such links remain saved within layers. It is concluded which page to move next from the sub-layer. A high-quality retrieval Novel E-focused crawler is different from a general crawler, which judges whether the document pointed by URLs is relevant for a particular domain.

4.2 Sentiments Cleaning Process (pre-processing)

In the next phase, the crawled data is going through implemented pre-processing. The extended words/abbreviations/slangs are selected using the n-gram model, which works for defining the probability of occurrence of the very next word or character in the string, and on detection, the particular word includes in the corpus. Generally, the reviews on social sites are different features, (1) the size of a huge number of opinions. (2) Several opinions accommodate noisy data. (3) There is a huge volume of opinion which provide valuable information for consumer as well as providers. (4) Several opinions include parameters along with the sentiment of the product. The process of implementing pre-processing is shown in Fig. 4.

There are some widely known techniques: Step: 1 Noisy Element Removal, Step: 2 Normalization, Step: 3 Word Standardization.

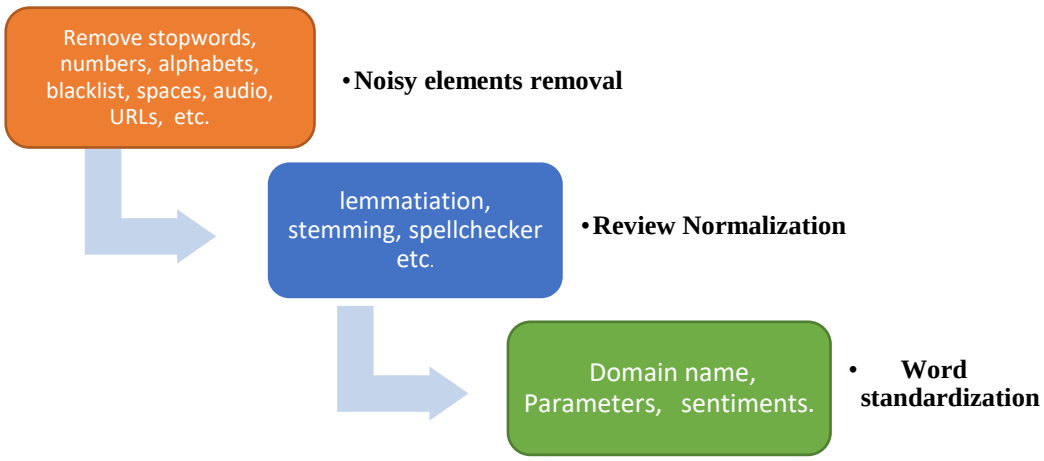


Fig. 4. Step by step process of sentiments cleaning process

Step: 1 Noisy Element Removal
In this phase, during the pre-processing, we have to eliminate all the unwanted noisy entities to improve the precision and Accuracy of the opinion collected from the social web. Under this process, we remove those words and numbers, helping verbs, determiners, Wh-family words and articles, prepositions, etc., which do not change the definition of the text or which cannot add is called the stopwords and numbers list. Stopwords and numbers are restricted for searching as they repeatedly occur in language for which the indexing engine has been turned.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Steps for Noisy Elements Removal

1. To remove the URLs using the list of rational expression matching.
2. To eliminate all the special characters and symbols.
3. To replace multiple spaces between or before the words and sentences with a single space.
4. To eliminate all the stopwords by comparing them with the list of stop words. The frequency of occurrence of stopwords in opinions present on social media is very high. They are neither related to any domain nor contain meaningful information about any product.
5. To eliminate all the numeric values or numbers.
6. To eliminate blacklist and non-sensical opinions from a large number of reviews.
7. To eliminate unnecessary pictures/photos/ images or videos.
8. To eliminate dash, forward/ backward slash (/,\), etc. from the reviews.
9. To eliminate punctuation marks.

Step: 2 Review Normalization

Data science's fundamental component is the review normalization that the researcher mostly ignores during the noisy data removal processing. Converting of data, called transforming the origin review into other patterns, permits processing reviews effectively minimize and also eliminate reviews is the primary motive of reviews normalization.

The review normalization steps are given below:

1. To convert each word of opinion into lowercase.
2. Sometimes, “hash-tag(#)” contributes some meaningful data, for this reason, to eliminate the only # and to keep sentiment word. E.g., "#good" is change with "good". Stemming of sentiment words: To reduce insufficiency, a stemming algorithm was executed.
3. The sequence of three or more repeated characters in the opinion word is replaced with one character. E.g. “cheapppppp” is replaced with “cheap”.
4. Some opinion words do not begin with an alphabet and provide helpful knowledge, therefore eliminating only symbol or number 234567900-=/.,;][= and to keep the word. E.g. “*-99t321 AMAZING” is replaced with “amazing”.
5. Spell-checker: to correct the maximum awkward mistakes to improve the Accuracy and speed. The grammar and spelling of opinion words are checked by this tool.
6. To eliminate all the suffixes to sentiment like s,ss,ies->y,ly, ful,es,sses->ss, etc. This process can reduce word strength by removing suffixes which further optimizes the search.

Step 3: Word Standardization

A data processing progress that transforms the composition of a different set of data into an ordering data arrangement is called data standardization. In the data transform operations, it can work as a transform work unit. It allows data users to investigate and utilize data in an irregular manner.

4.3 Identification of Attributes

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

After detecting the relevant reviews, the same is being considered for differentiated as explicit and implicit attributes (Cost, STDs disease, effect, side effect, and maintenance) from the reviews. The identification procedure of the noun and pronoun word points out all the procedure's attributes, but the system found the lack of protection of frequent attributes.

4.3.1 Identification of Explicit attributes

Users use different words to present their opinions. The system manually annotates these attributes will reject them to recognize. Moreover, the individuals will suggest many reviews and comments related to the attribute of the contraceptive method. The features are manually explained in other procedures. The recommended (proposed) methodology retrieve noun phrase, adverb, the noun from the collected large volume of sentiments. Retrieving words can be known as explicit attributes. Extracting noun forms has automatically determined the attribute of birth control method using wordnet and the POS from given in the dataset.

4.3.2 Implicit attribute identification

The process by which the implicit attribute to identify by the people is used as phrases, dialogues speaking, and omission of sentences or para. As such, the parameters cannot directly define in any sentence or in speaking. For example, She has headaches, allergy and not fit. This example review sentence indirectly defines the side effect and maintenance attributes of the contraception method.

In this kind of review sentence, context to all information is given in the sentences and use of a corpus to identify the information given in the context.

For all implicit reviews, the attributes detected as verbs/adjectives (thesaurus is being used for detecting the synonyms and other related aspects of the attributes) are considered for matching with a corpus. This also works for the extended words, slang, and abbreviations, or if found, the same is considered the implicit attributes.

4.4 Determination of sentiment and polarity/orientation of reviews

The polarity of the attributes like cost, STD, etc., is being counted based on the negative/positive reviews, as for every sentiment of a specific attribute, the polarity of added as +1/-1 for positive/negative sentiments.

4.4.1 The recognition of word statement

As it is known, desirable and undesirable are two emotional stocks of words opinion. The word that has favored (desirable) meaning, concluded in possible attitude orientation, emphasizes other undesirable sentiments in a negative orientation.

This is the authenticated study to classify the sentiments given in the text.
For example:

- I am ***satisfied*** with the contraception method- positive
- This product is ***not good***-negative

The frequent words used to describe the meaning of the given text and its word used to make a sentence.

4.4.2 Evaluation of the polarity (orientation) of reviews

Negation rules strictly followed while determining the sentiments of collected reviews. There are some negation words viz no, not, never as well as other terms in the same pattern viz cheap and stop can also convert the polarity in the way given below: to determine the polarity of sentiment by Negation rules taking advantage of negation word for instance never, not and no, etc.

Some parts of the dictionary approach are used to identify the adjective, the adverb, & verb to prepare the sentiment list, and all the negation words are stored in the negation list at the sentence level. Sentiment word is matched to the Thesaurus one by one.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

4.5 Clustering

The data collected is considered for clustering, the clustering process's major usage to present the data for further processing and usage as for any medical clinical study, diagnoses, market study, etc.
Clustering is similar to the classification through which we can categorize the data into different groups(Shivaprasad & Shetty, 2017)(Bafna et al., 2016). K-mean is almost the widely-known unsupervised machine learning algorithm in data science. For clustering, extended k-means come to existing, which uses similarity of the data point with the centroid and the Euclidian distance, as it is not entirely clear to match the similarity between the grouped data Euclidian distance for grouping. The k-means algorithm is used for clustering the dataset of reviews generated. In the present study, the similarity index concerning the attributes is also being considered for the group the data. The clustering of the data is being based on parameters like cost, side effects, STDs, etc., by use of the proposed Enhanced K-mean (n-gram) clustering algorithm.

Enhanced K-mean (n-gram) clustering algorithm
The use of the algorithm to identify the different data points (parameters) and sentiments and clustering to review the given parameters as follows:

INPUT: Opinion Texts posted by Reviewer on various websites and other sources (hospitals or a couple of backward areas.).
OUTPUT: Clustering of Products based on positive and negative sentiments of review.
Step 1: Reviews are being initiated from different sources as online (E-focused crawler) and the rest of the online and offline using (feedback form of contraceptive).

(1) To parse individual URLs of initial page (z)
(2) Compute TF-IDF as well as Step 2 (Implemented pre-processing)

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

Where:
$tf_{i,j}$= Total no. of occurrence of the relevant term (i) in collected reviews (j)
df_i=Total no. of reviews containing the relevant term (i)
N = total no. of reviews

Classifier← sorting of URLs (P=0, 1,3, 4 …n)
Extract←URLs
If (webpage = =Relevant)
` Insert← RQ
Continue this URL
Check←All hyperlinks (at sub-layers)
Else if (webpage = =Relevant) (sub-layer)
Continue this Url up to n-sub-layer
Else if (web page≠ relevant)
Discard it
end
end

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Step 2: Step-by-step sentiments cleaning process shown in Fig.4 as removal of noisy contents, review normalization, and word standardization.

```
Array [] ←retrieved reviews
Specialchar []= string {“\n”, “/”, “@”, “&”, “-”, “^”, “$”};
number[]= int {0,1,2,3,4,5,6,7,8,9}:
wordsuff []= newstring{“s”, “es”, “ies”, “ves”, “ly”, “ful”, “ss”, “ssess”};
hashtags [] = string {#\S+’,’}
htmltags []= string {<.*?>;}
Extraspaces [] = string {+ ‘, ’}
Alphabet []> 2 = string {aaa…., bbb…, ccc…, …….zzz….}
Retweets []/opinions and cc= string {rt/cc ‘,’}
Reviews [] = string {lower ( )}
Url[] = string { “https/ http://”, “www”, “ .com”, “./n”, ? : \/ ( \ W / \ / ~ | \ = | \ % )* \ b ;” }

(1)Remove Images If Review_ text Contains->“. Jpg”, “. png”, “ gif”
(2) Review number If Review _ text Contains-> number from number []
(3) Remove word suffix using for each loop.
For each statement, word replace [] word suffix by “ ” (blank space)
```

Step 4: Break reviews into the separate statement

```
# Separate statement store in an array
Array [] separate opinion← opinions
```

Step 5: Find attributes (from POS Tagging, sentiwordnet2, corpus, thesaurus)

```
If (word== “NN”, “Adverb”)
Array [Index] “NN”, “Adverb”= Explicit Attribute;
Else if (word== “Adjective”, “verb”, “Phrases”, “Idioms”, sentiment word)
Array [Index] “Adjective”, “verb”, “Phrases”, “Idioms”, sentiment word= Implicit Attribute;
Else if (word≠ “POS Tagging”, “SentiWordNet2”, “Corpus”, “Thesaurus”)
Array [Index] word= Neutral;
Discard this word
end
end
```

Step 6: Find out sentiment word (“Thesaurus”, “SentiwordNet2”, “Sentiment Dictionary”)

```
PSW=Positive Sentiment Word, NSW= Negative Sentiment Word, SW= Sentiment Word;
If (word==sentiment)
Sentiment [] = SW;
Else if (SW= =Positive)
Positive [] = PSW;
Else if (SW= =Negative)
Negative [] = NSW;
end
end
```

Step 7: Find out Polarity (from list of negation words)

```
Array []← negation words
N= Negation word, -ve= Negative, +ve= Positive;
If (N← NSW)
 +ve orientation [] = + ve score;
Else if (N← PSW)
 – ve orientation[] = -ve score;
```

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Else if (w= Negative)
Negative [] = -ve score;
Else if (w= positive)
Positive [] = +ve score;
Else (w=neutral)
Neutral = 0 score;
end
end
end
end

Step 8: Make Data set for store information (Data) Attribute-Sentiment polarity wise
// Collect Data into Dataset based on Attribute-sentiment pair
Count all no. of Array
Column 1= Attribute []
Column 2= +ve score []
Column 3 = –ve Polarity []
Column 4 = Neutral []

Step 9: Applied Enhanced K-Mean (n-gram) algorithm.
Clustering: Group Attributes (cost, maintenance, effectiveness etc.)
Define C= 3 (create three clusters)
For every dataset in order to contain a row (Attribute-sentiment pair)
Find out the minimum cost
Parameter – x1,x2,x3,x4,x5

i. Determined the cluster's Centroids.
ii. We employ the arithmetical formula of the K-mean algorithm expressed in step (a) for computing the Euclidian distance between the centroid and distinct points.
iii. Before including the selected point to the cluster, it is being examined for similarity, employing the similarity function.

$$simMsr(C_i) = \sum_{c_{ci}c_{pj}} simFunc\,(Cc_i, Cp_j) \qquad (2)$$

Where:
$C_{C_i,}$Represent the i_{th} cluster's centroid.
C_{p_j}Represent included cluster point
.
C Represents the total number of clusters.

Step 9.1: Examine the data (Attributes) for similarity checking.
i. Go to step 9.2 if similarity (nearest point) is high.
ii. Else evaluating the next off (nearest point)

Step 9.2: If the data points correctly describe the cluster similarity, the data points (attributes) are combined to the clusters.

Steps involved in clustering using K-means Algorithm:

Step (a):- Assignment of data (Consideration of precleaned or pre-processed data)
The clusters are expressed by the center data points referred to as the cluster's centroid.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

i. Categorize the given dataset of 5 data points (attributes) into clusters.
ii. Initially, a random data point is selected as the center of the cluster.
iii. To calculate the distance of each attribute from the center of the cluster.
iv. Find the minimum cost of each attribute to its nearest center using the Euclidean Distance Matrix.
v. The attributes whose distance from the center of the cluster has a minimum is assigned to the cluster center.
vi. Assigned data points to cluster using:

$$argmindist(C_i, x)^2 \quad (3)$$

Where:
$C_i \epsilon C$ (set of central points (C_i) belong to total no. of C),
dist(C_i,x) (represent the Euclidean distance)

Step (b):- Updating of the centroid
i. The new center of the cluster is recalculated by using:

$$C_i = \frac{1}{s_i} \sum_{x_i s_i} x_i \quad (4)$$

s_i => Number of parameter (data) in i^{th}cluster (C_i)

ii. To recalculate the distance of each parameter from the newly obtained center of the cluster.
iii. Stop if no parameter was reassigned, otherwise repeat step v. [in step (a)].

Considering the working example of the n-gram model working,
S (News): I I am not,

Training dataset:
S_1: I am a human,
S_2: I am not a stone,
S_3: I I live in Delhi,

- The model applied is a unigram

P(S) = P(I I am not) = P(I / <S>) * P(I / <S>) * P(am / <S>) * P(not / <S>)
= 3/3*3/3*2/3*1/3
=1*1*.6667*.3333
=0.22219, is the probability of the occurrence of these defined words as combination.

- Model applied is bigram

P(S) = P(I I am not) = P(I / <S>) * P(I / I) * P(am / I) * P(not / am)
= 3/3 * ¼ * 2/4 * ½
=0.0625, is the probability of occurrence of these words as combination using the bigram model.

- Model applied is bigram

P(S) = P(I I am not) = P(I I / <S>) * P(am / I I) * P(not / I am)
= 1/3 * 1/1 * ½
= 0.1666, is the probability of occurrence of these words as combination using trigram.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

5 Dataset

The dataset contains 10,156 reviews and 20,138 sentences for the contraceptive healthcare products displayed in Table 1. It also contains written sentiments without the ranking of concerned products. Different products have textual reviews indicating by the sentences.
All the reviews have been collected from the health care sites, twitter, Instagram, interview scheduling sites, healthcare blogs like www.drfameg.com,www.webmd.com, www.gmail.com, www.grouponehealthsource.com, www.facebook.com, www.healthnewsreview.org, www.reviewtrackers.com/doctor-review-sites. Further, Whatsapp, cell phones, use of feedback forms of contraceptive methods from peoples and hospitals of eastern/western area of Uttar pradesh are included. Each of the experiments has been processed on the dataset given below for clustering the contraceptive product's attribute by using the proposed methodology.

Table 1. Dataset of reviews

Data set	STDs	Effective	Cost	Maintenance	Side-Effect
Reviews	2000	2050	2080	2020	2006
Sentence	3008	5950	4050	4070	3060

Table 1 depicts the total no. of Reviews (sentiment) and sentences gathered from the different sources (social media and interview schedule). We determine the sentiments and various kinds of sentences from the data set (reviews) based on different attributes such as STDs, effectiveness, cost, maintenance, and side effect.

6 Evaluation and analysis

The experiments have been processed on the dataset shown in Table1 to extract and group the attribute of contraceptive methods. The dataset contains 10,156 reviews and 20,138 sentences. It also contains different attributes and corresponding sentiments to calculate the value precision, average precision, and Accuracy, etc.
We tested seven different methods to proposed adopting contraception and categorization of the attributes based healthcare data approach as follows:

1. The DBSCAN method clustering algorithm for categorization.
2. [(ST,SWR)+DBSCAN] method
3. SC+K-Mean methods we also represent as a crawler or retrieving the web page and clustering for categorization.
4. SC+ Enhance K-Mean method
5. [SC+(ST,SWR)+ Enhance K-Mean]
6. [FC+(ST,SWR)+Enhance K-Mean]
7. Proposed methods represent a Novel E-focused crawler (tf-idf and implemented pre-processing) and Enhance the K-mean (n-gram) clustering algorithm for categorizing parameters of the healthcare products.

The DBSCAN represents Density-based spatial clustering of applications with noise, SC represents a simple crawler, ST is stemming and SWR is the stopwords removing algorithm, FW is a focused crawler and the K-Mean algorithm is used for clustering the parameters of the contraception method.
The efficacy and efficiency of each parameter-based healthcare product categorization methodology are considered by the precision, average precision, recall, F-Measure, G-Mean, TNR (True Negative Rate), and Accuracy. The performance

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

matrix determines the performance of clustering the different parameters/features and sentiment using different methods represented in Table 2 to Table 8.

6.1 Performance matrix

To calculate the attributes classification of the contraceptive/healthcare product using the given proposed methodology. There are different evaluation matrix PPV, TPR, F-score, G-Mean, TNR, average precision, Accuracy, etc. are used to evaluate the result. A large number of review documents extracted are used to calculate the value of precision, average precision, and Accuracy, etc. can be shown mathematically as under:

(1) Positive predictive value (PPV): It is also referred to as precision. It is described as the ratio of total no. of true extracted relevant opinion and all the positive, relevant opinions. The computation of precision is represented by Equation (5).

$$\text{Precision}=\sum_{i=1}^{N}\frac{\text{TPRi}}{TPRi+FPRi} \quad (5)$$

(2) TPR (True Positive Rate): It is also known as recall. TPR is calculated by Equation (6) and may be described as the ratio of the total number of true extracted relevant opinions and the sample of the total number of true, relevant opinions.

$$\text{Recall}=\sum_{i=1}^{N}\frac{\text{TPRi}}{\text{TPRi+FNRi}} \quad (6)$$

(3) F- score: F- score is also called an F-measure. It contains both PPV and TPR. Suppose it is the case that PPV or TPR is minimum F- the measure will be minimum. The calculation of the F-measure is represented by the given Equation (7).

$$\text{F}=\frac{2*\text{Precision}*\text{Recall}}{Precision+Recall} \quad (7)$$

(4) G-Mean: For enumerating classifier G-Mean is employed on the dataset. This metric denotes equivalence b/w performance on the majority and minority and also considers both TNR and TPR. The calculation of G-Mean shows in Equation (8).

$$\text{G-Mean}=\sqrt{\text{TPR+TNR}} \quad (8)$$

(5) True Negative Rate (TNR): TNR is computed as divide the no. of correct irrelevant opinion by the total number of irrelevant opinion negatives.

$$\text{TNR}=\sum_{i=1}^{N}\frac{\text{TNRi}}{\text{TNRi+FPRi}} \quad (9)$$

(6) Average Precision: Average precision is the ratio of the sum of precision to the total no. of parameters of opinion.

$$\text{Average Precision}=\frac{\text{Sum of precision}}{\text{Total number of parameter}} \quad (10)$$

(7) Accuracy: The performance of the proposed system is measured by Accuracy. It is measured as a ratio of correctly extracted relevant opinions to the total number of opinions. The higher Accuracy represents a better outcome. Accuracy is calculated by Equation (11).

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

$$\text{Accuracy} = \sum_{i=1}^{N} \frac{\text{TPRi+TNRi}}{\text{TPRi+FPRi}} \qquad (11)$$

Where,
TPRi= Total no. of correctly extracted as a relevant review. FPRi= Total no. of false extracted as relevant opinion.
TNRi=Total no. of correctly extracted as an irrelevant review. FNRi=total no. of false extracted as an irrelevant review.

We have conducted different experiments on the dataset includes 10,156 reviews and 20,138 sentences.

Precision:
The precision computed through the experiment listed in Table 2.We have compared the proposed approach methodology's performance with other methods; the results show that our proposed approach gives better results than other methods for retrieving health reviews and clustering the contraceptive method based on attributes. Table 2 displays the precision of the proposed approach as clustering the attribute/feature-based contraceptive healthcare products.

Table 2. Precision Value of healthcare attributes using different clustering methods

Factor	STD	Effective	Cost	Maintenance	Side Effect
Method					
DBSCAN	0.41	0.38	0.40	0.45	0.42
(ST,SWR)+DBSCAN	0.44	0.40	0.42	0.48	0.45
SC+ k-Mean	0.46	0.41	0.45	0.51	0.47
SC+ Enhance k-Mean	0.51	0.49	0.57	0.56	0.54
[SC+(ST,SWR)+Enhance K-	0.62	0.60	0.66	0.61	0.65
[FC+(ST,SWR)+Enhance K-	0.71	0.68	0.75	0.70	0.78
Proposed approach	0.89	0.85	0.90	0.82	0.86

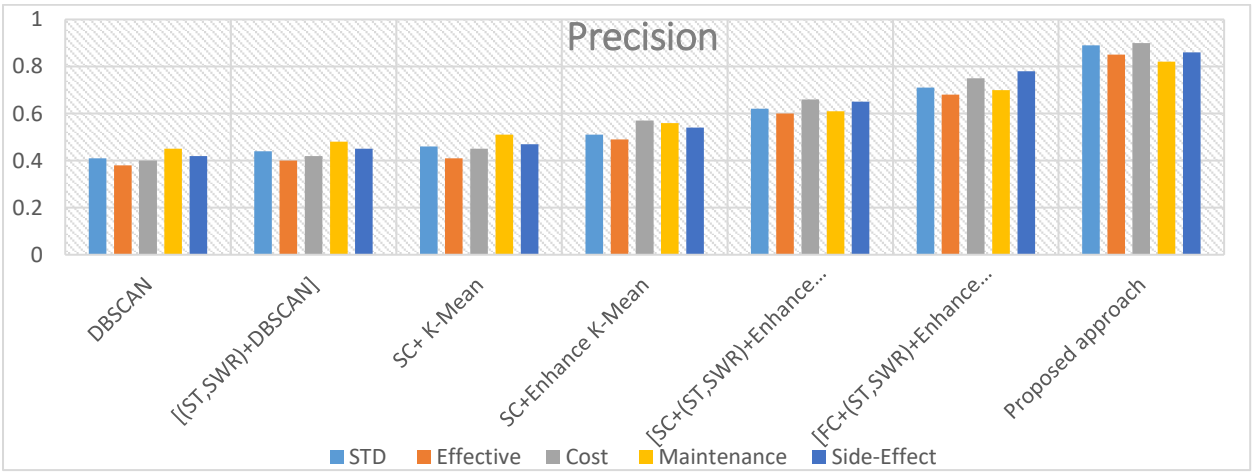


Fig.5. Precision Value of healthcare attributes using different clustering methods.

Recall:
Table 3 displays the recall of the proposed approach as clustering the attributes/feature-based contraceptive healthcare product.
We compare proposed Adopting contraception performance and categorize healthcare data methodology to other methods to show that our proposed approach obtains a better outcome than other given methods for retrieving documents and clustering.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Table 3. Recall value of healthcare attributes using different clustering methods.

Factor Method	STD	Effective	Cost	Maintenance	Side Effect
DBSCAN	0.47	0.43	0.46	0.51	0.48
(ST,SWR)+DBSCAN	0.50	0.46	0.47	0.55	0.51
SC+ k-Mean	0.52	0.47	0.53	0.56	0.54
SC+ Enhance k-Mean	0.57	0.55	0.64	0.62	0.61
[SC+(ST,SWR)+Enhance K-Mean]	0.69	0.66	0.70	0.67	0.71
[FC+(ST,SWR)+Enhance K-Mean]	0.86	0.75	0.82	0.77	0.84
Proposed approach	0.95	0.93	0.96	0.89	0.92

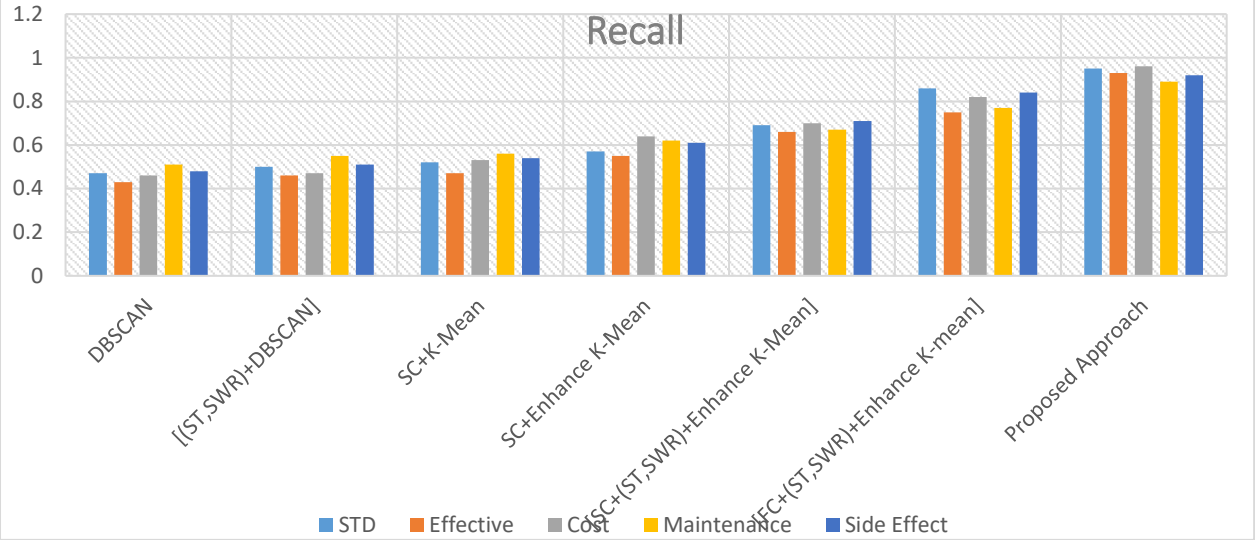


Fig.6. Recall value of healthcare attributes using different clustering method

F-Measure:

Table 4 displays F-Measure of the proposed approach as clustering the attribute/feature-based contraceptive healthcare product. We compare the Adopting proposed contraception performance and categorize healthcare data methodology to other methods to show that our proposed approach obtains better outcomes than other given methods for retrieving documents and clustering.

Table 4. F-Measure the value of healthcare attributes using different clustering methods.

Factor Method	STD	Effective	Cost	Maintenance	Side Effect
DBSCAN	0.44	0.41	0.43	0.47	0.44
(ST,SWR)+DBSCAN	0.46	0.43	0.44	0.51	0.47
SC+ k-Mean	0.48	0.44	0.48	0.53	0.51
SC+ Enhance k-Mean	0.54	0.52	0.61	0.58	0.57
[SC+(ST,SWR)+Enhance K-Mean]	0.65	0.63	0.67	0.64	0.67
[FC+(ST,SWR)+Enhance K-Mean]	0.77	0.72	0.78	0.73	0.81
Proposed approach	0.92	0.89	0.93	0.85	0.89

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

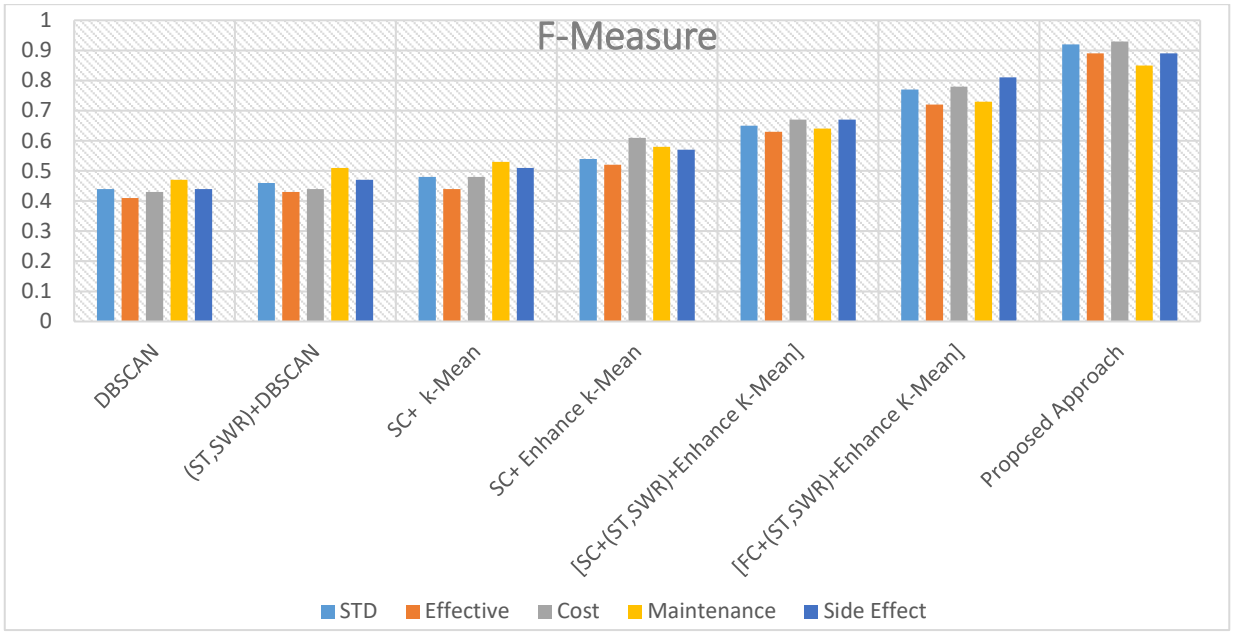


Fig.7.F-Measure value of healthcare attributes using different clustering methods.

TNR:
Table 5 displays the TNR of the proposed approach as clustering the attribute/feature-based contraceptive healthcare product. We compare the performance of the proposed approach methodology to other methods to show that our proposed approach gives better results than other given methods for retrieving health reviews and clustering the contraceptive method based on parameters.

Table 5.TNR values of healthcare parameters of different clustering methods.

Factor	STD	Effective	Cost	Maintenance	Side Effect
Method					
DBSCAN	0.36	0.34	0.35	0.40	0.38
(ST,SWR)+DBSCAN	0.40	0.35	0.37	0.43	0.40
SC+ k-Mean	0.41	0.37	0.41	0.46	0.43
SC+ Enhance k-Mean	0.46	0.45	0.50	0.51	0.50
[SC+(ST,SWR)+Enhance K-Mean]	0.57	0.55	0.61	0.56	0.60
[FC+(ST,SWR)+Enhance K-Mean]	0.66	0.64	0.70	0.65	0.74
Proposed approach	0.84	0.80	0.86	0.78	0.80

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

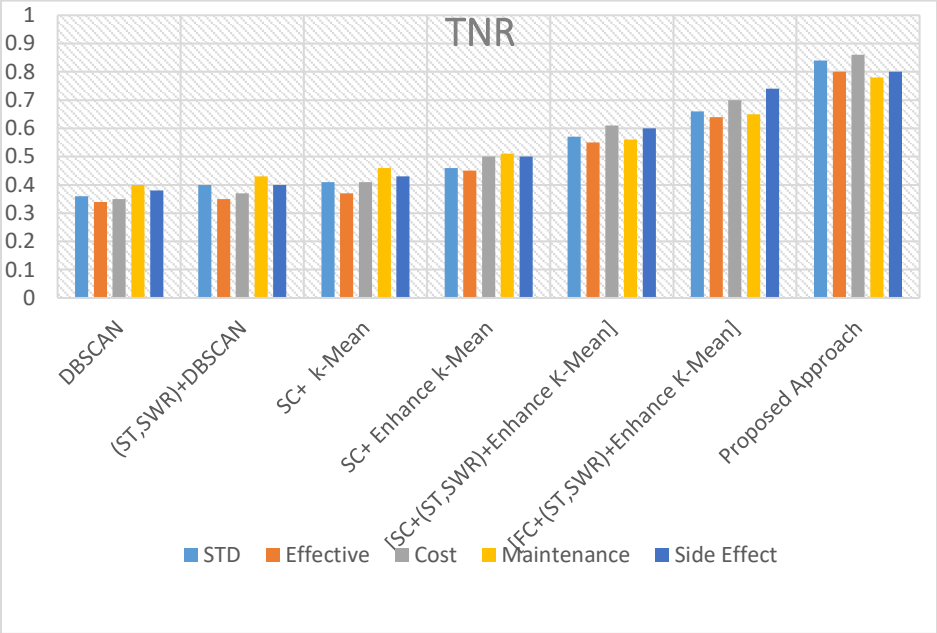


Fig.8. TNR Value of healthcare attributes using different clustering methods.

G-Mean:

In Table 6, comparing the performance of the proposed approach methodology to other methods show that our proposed approach gives better results than other given methods for retrieving health reviews and clustering the contraceptive methods based on attributes.

Table 6. G-Mean value of healthcare attributes of different clustering methods.

Factor Method	STD	Effective	Cost	Maintenance	Side Effect
DBSCAN	0.39	0.35	0.37	0.42	0.38
(ST,SWR)+DBSCAN	0.45	0.40	0.41	0.48	0.42
SC+ k-Mean	0.46	0.42	0.46	0.51	0.48
SC+ Enhance k-Mean	0.51	0.49	0.57	0.56	0.55
[SC+(ST,SWR)+Enhance K-Mean]	0.63	0.61	0.65	0.62	0.65
[FC+(ST,SWR)+Enhance K-Mean]	0.75	0.69	0.70	0.76	0.79
Proposed approach	0.89	0.86	0.91	0.83	0.86

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

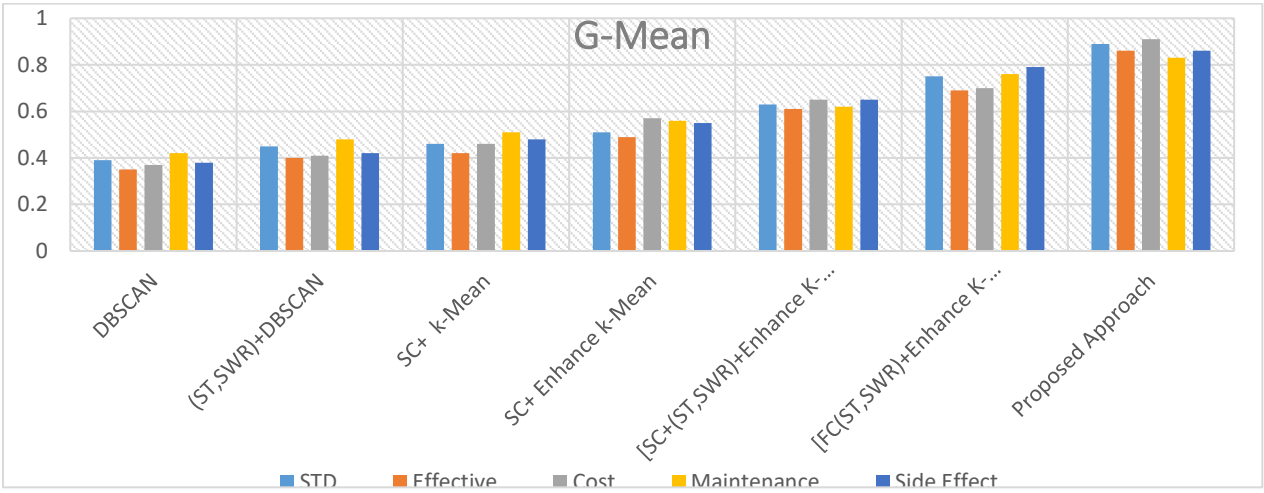


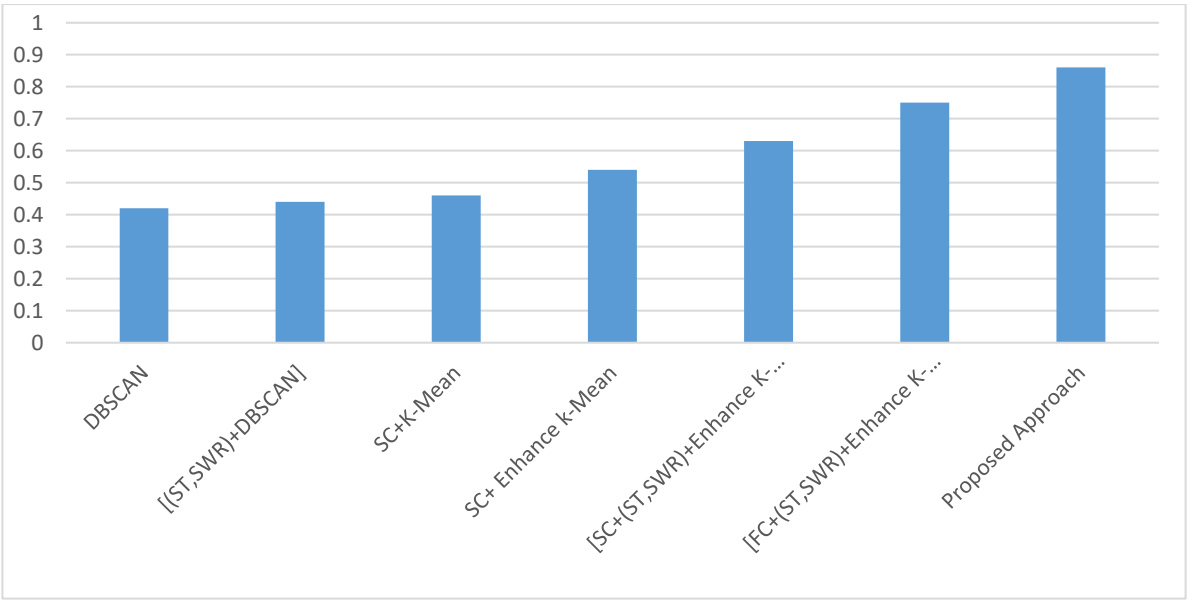
Fig.9. G-Mean value calculated by different clustering methods.

Average Precision:

Table-7 shows the Average precision values and comparison of the proposed approach methodology's performance to other methods. The results show that our proposed approach gives better results than other methods for retrieving health reviews and clustering the contraceptive methods based on parameters.

Table 7. The average precision of healthcare parameters of the different clustering algorithms.

Average precision	DBSCAN	0.42
	[(ST,SWR)+DBSCAN]	0.44
	SC+K-Mean	0.46
	SC+ Enhance k-Mean	0.54
	[SC+(ST,SWR)+Enhance K-Mean]	0.63
	[FC+(ST,SWR)+Enhance K-Mean]	0.75
	Proposed approach	0.86



JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

Fig. 10. The comparison of the performance of different parameter classification methods based on Average precision value.

Accuracy:
Table 8 displays the Accuracy of the proposed approach as clustering the parameters/features-based contraceptive healthcare products and experiments performed for collating the performance of the proposed methodology to other methods. It shows that our proposed approach produces preferable results compared to other given methods for retrieving reviews then clustering the healthcare products based on their attributes.

Table 8. Accuracy value of healthcare attributes of the various clustering methods.

Factor Method	STD	Effective	Cost	Maintenance	Side Effect
DBSCAN	0.42	0.40	0.42	0.45	0.43
(ST,SWR)+DBSCAN	0.45	0.42	0.43	0.50	0.45
SC+ k-Mean	0.47	0.43	0.46	0.52	0.50
SC+ Enhance k-Mean	0.52	0.51	0.59	0.56	0.55
[SC+(ST,SWR)+Enhance K-	0.64	0.62	0.65	0.63	0.66
[FC+(ST,SWR)+Enhance K-	0.76	0.70	0.76	0.72	0.80
Proposed approach	0.91	0.87	0.92	0.84	0.88

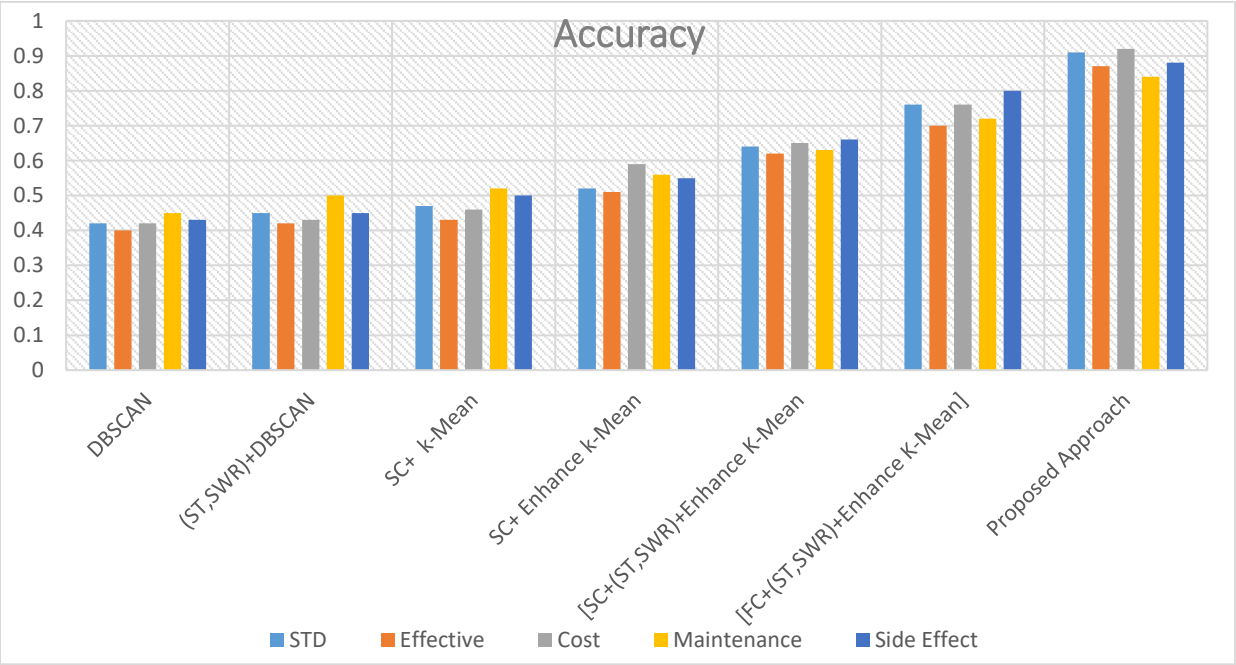


Fig. 11. Accuracy is calculated by different clustering methods.

7 Result and discussion

Clustering is an unsupervised machine learning technique used here to categorize the healthcare product based on its attributes/parameters. This study utilizes seven clustering methods as DBSCAN, [(ST, SWR)+DBSCAN], SC+K-Mean, SC+ Enhance K-Mean, [SC+(ST,SWR)+ Enhance K-Mean], [FC+(ST,SWR)+Enhance K-mean] and Proposed

Approach. The best clustering formation obtained from the optimum clustering formation section is utilized to classify the product. This results in six clustering methods for DBSCAN and K-Mean, one for clustering each of the parameters/attributes of the healthcare product. The given clustering method such as DBSCAN, [(ST, SWR)+DBSCAN], SC+K-Mean, SC+ Enhance K-Mean, [SC+(ST,SWR)+ Enhance K-Mean], [FC+(ST,SWR)+Enhance K-mean] get the useful result but not efficiently or effectively.

We use attribute-based clustering in place of referring prescription bottomed on subjective/objective clustering of healthcare products. The Accuracy and recall in indicating prescription via the proposed approach are illustrated in the above graphs and tables. The above graphs show Accuracy, recall, precision calculated by different clustering methods. The above graph shows the Accuracy of the clustering symptoms of the user. The given different graph gets better Accuracy, but including different attributes and considering is a compulsory one. Not focusing on pre-processing, clustering, and other attributes leads to not accurate results but in the sequence of calculation of the given proposed approach's efficiency and effectiveness considering it on parameter-based product categorization to cluster each attribute. Therefore our proposed approach could attain all these.

The implementation of the proposed methodology achieved a better result in efficient recommendations regarding attribute-based healthcare product clustering. Therefore Accuracy improved and attains effective and efficient results by the use of the proposed approach. Hence keep in touch with the proposed objective, and it attains attribute-based healthcare product categorization. Table from 2 to 8 lists the precision, recall, TNR, average precision, G-Mean, F-Measure, accuracy.

Table 2, Table 3, Table 4, Table 5, Table 6, Table 7, Table 8 lists the results of clustering the different classification approaches by using the methodology from 1-7 to extract documents (review) using crawler and categorization algorithm. When do clustering with the highest similarity of parameter review given in a cluster which after the uses of Enhance k-Mean clustering algorithm along with tfidf & pre-processing, corpus, and POS to extract the parameter? These parameters can be seen that the use of the proposed approach [E-focused crawler (tfidf as well as pre-processing) and Enhance K-Mean (n-gram)] to achieve the highest Accuracy, F-Measure, precision, and recall.

Our proposed approach obtains better performance in the Accuracy of 0.92% precision of 0.90 %, recall of 0.96%, F-Measure of 0.93%, G-Mean of 0.91%. These results suggest the best performance shown in Table 3 in terms of recall of 0.96% and Table 8 in the Accuracy of 0.92% using our proposed methodology. Therefore, the proposed methodology produces better results than other methodology used for extracting and clustering the parameters of contraceptives.

As displayed in Fig. 5, the performance of contraceptive adoption and categorization of health care data can obtain recall of 0.96%. In Fig.5, the result of the proposed approach illustrates the best Accuracy with 0.92% when we include Novel E-Focused crawler (tfidf &implemented pre-processing) and Enhance K-Mean (n-gram) for attribute categorization; this method obtains the better performance for extracting topic-specific healthcare reviews and clustering often involve effectiveness, STDs, side effect, cost, and maintenance attributes. Based on the above evaluation and comparative analysis, the proposed method gives a better result than the other methodology used for extracting and clustering the attributes of the contraception method.

8 Conclusion

Our paper has proposed a methodology that categorizes the adoption of contraceptives or healthcare data based on its parameters/attributes to produce a more relevant and useful summarization of each attribute. Novel E-Focused crawler's main improvement instead of simple Focused crawler, pre-processing of reviews collected from different sources, and use Enhance K-Mean (n-gram) instead of K-Mean. The proposed approach used a focused crawler with pre-processing and Enhanced K-Mean to improve the downloaded document's relevancy and efficiency. With corpus, POS, Thesaurus, and expansion of the seed list of polarity identification to get more accurate results. Removal of the irrelevant or noisy data improves the system's storing capacity and uses the Enhanced k-Mean algorithm to increase the precision, the average precision, recall, G-Mean, TNR, F-Measure and Accuracy of the cluster. The attribute’s to improve the performance of attributes classification and extract the high quality of information with precision, average precision, and Accuracy. The proposed method provides many satisfactory results and has achieved Accuracy with 92% and recall with 96% for categorizing healthcare products based on the attribute.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 68-94
https://publishoa.com
ISSN: 1309-3452

9 Future prospects

In our future work based on the proposed methodology, we will concentrate on M–Attribute segmentation for the segment the different attribute from the enormous amount of opinion where opinion about no. of attributes describe in a sentence or a statement freely available on social media and this methodology use to other area or field for their different attribute identification.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] A Al Riyami, M Afifi, and RM Mabry, Women's autonomy, education and employment in Oman and their influence on contraceptive use. *Reproductive Health Matters*, 2004, *12*(23), 144–154, DOI:10.1016/S0968-8080(04)23113-5.

[2] O Alabi, CO Odimegwu, N De-wet and JO Akinyemi, Does Female Autonomy Affect Contraceptive Use among Women in Northern Nigeria ? *African Journal of Reproductive Health*, 2019, *23*(2), 92–100, DOI: 10.29063/ajrh2019/v23i2.9.

[3] P Bafna, D Pramod and A Vaidya, Document Clustering: TF-IDF approach. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016*, 2016, 61–68, DOI: 10.4018/978-1-5225-0536-5.ch013.

[4] J Benson, K Andersen, D Brahmi, J Healy, A Mark, A Ajode, R Griffin, J Benson, K Andersen, D Brahmi and J Healy, What contraception do women use after abortion ? An analysis of 319, 385 cases from eight countries. *An International Journal for Research, Policy and Practice*, 2016, *1692*, 0–16, DOI: 10.1080/17441692.2016.1174280.

[5] A M Bigorra, O Isaksson and M Karlberg, Aspect-based Kano categorization. *International Journal of Information Management*, 2019, *46*(October 2018), 163–172, DOI: 10.1016/j.ijinfomgt.2018.11.004.

[6] C Caetano, T Peers, L Papadopoulos, K Wiggers, H Grant, C Caetano, T Peers, L Papadopoulos and K Wiggers and Millennials, contraception : why do they forget ? An international survey exploring the impact of lifestyles and stress levels on adherence to a daily contraceptive regimen exploring the impact of lifestyles and stress levels on adherence to a daily. *The European Journal of Contraception & Reproductive Health Care*, 2019, *0*(0), 1–9, DOI: 10.1080/13625187.2018.1563065.

[7] AC Pandey, DS Rajpoot and M Saraswat, Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing and Management*, 2017, *53*(4), 764–779, DOI: 10.1016/j.ipm.2017.02.004.

[8] AR Chaurasia, Contraceptive Use in India: A Data Mining Approach. *International Journal of Population Research*, *2014*, 1–11, DOI: 10.1155/2014/821436.

[9] S Chung, M Chong, JS Chua and JC Na, Evolution of corporate reputation during an evolving controversy. *Journal of Communication Management*, 2019, *23*(1), 52–71, DOI: 10.1108/JCOM-08-2018-0072.

[10] LFS Coletta, NFF De Silva and ER Hruschka, Combining classification and clustering for tweet sentiment analysis. *Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014*, 2014, 210–215, DOI: 10.1109/BRACIS.2014.46.

[11] FL Cruz, F Enríquez, FJ Ortega and CG Vallejo, A Knowledge-Rich Approach to Feature-Based Opinion Extraction from Product Reviews. *SMUC'10*, 2010, 13–20.

[12] J Cunha, C Silva and M Antunes, Health Twitter Big Bata Management with Hadoop Framework. *Procedia Computer Science*, 2015, *64*, 425–431, DOI: 10.1016/j.procs.2015.08.536.

[13] X Dai and M Bikdash, Trend Analysis of Fragmented Time Series for mHealth Apps : Hypothesis Testing Based Adaptive Spline Filtering Method with Importance Weighting. *IEEE Access*, *5*, 2017, 27767–27776, DOI: 10.1109/ACCESS.2017.2696502

[14] X Dai and M Bikdash, Distance-based Outliers Method for Detecting Disease Outbreaks using Social Media. *SoutheastCon 2016*, 1–8.

[15] X Dai, and M Bikdash and B Meyer, From social media to public health surveillance: Word embedding based clustering method for twitter classification. *Conference Proceedings - IEEE SOUTHEASTCON*, 2017, *Table I*. DOI: 10.1109/SECON.2017.7925400.

[16] S Das, B Singh, S Kushwah and P Johri, Opinion based on Polarity and Clustering for Product Feature Extraction. *International Journal of Information Engineering and Electronic Business*, 2016, *8*(5), 36–43, DOI: 10.5815/ijieeb.2016.05.05.

[17] K Denecke and Y Deng, Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*, 2015(a), *64*(1), 17–27, DOI: 10.1016/j.artmed.2015.03.006.

[18] K Denecke and Y Deng, Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence In Medicine*, (2015b), *64*(1), 17–27, DOI: 10.1016/j.artmed.2015.03.006.

[19] PM Deschênes, ME, Lamontagne and MP Gagnon, Talking About Sexuality in the Context of Rehabilitation Following Traumatic Brain Injury : An Integrative Review of Operational Aspects. *Sexuality and Disability*, *0123456789*, 2019, 2019, 10.1007/s11195-019-09576-5.

[20] Garima, H Gulati and PK Singh, Clustering techniques in data mining: A comparison. *2015 International Conference on Computing for Sustainable Global Development, INDIACom 2015*, *2015*, 410–415.

[21] J Ginsberg, MH Mohebbi, RS Patel, L Brammer, MS Smolinski and L BrillianL, Detecting influenza epidemics using search engine query data. *Nature*, 2009, *457*(7232), 1012–1014, DOI: 10.1038/nature07634.

[22] V Gopalakrishnan and C Ramaswamy, Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of Applied Research and Technology*, 2017, *15*(4), 311–319. DOI: 10.1016/j.jart.2017.02.005.

[23] Z Halim and S Khan, A data science-based framework to categorize academic journals. In *Scientometrics*, 2019, Vol. 119, Issue 1, 393–423, DOI: 10.1007/s11192-019-03035-w.

[24] K HOnda, R Yang, S Ubukata and A Notsu, Fuzzy Co-clustering for Categorization of Subjects in Questionnaire Considering Responsibility of Each Question. In I. M. Seki H., Nguyen C., Huynh VN. (Ed.), *Integrated Uncertainty in Knowledge Modelling and Decision Making*, 2019, 370–379, Springer London, DOI: 10.1007/978-3-030-14815-7_31.

[25] A Jain and S Cherikkallil, Medinsights: Twitter Based Platform for Health Care Analytics. *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, 2018, 1104–1109, DOI: 10.1109/ICIRCA.2018.8597360.

[26] S Jin, Z Zhang, K Chakrabarty and X Gu, Anomaly Detection and Health-Status Analysis in a Core Router System. *IEEE Design and Test*, 2019, *36*(5), 7–17, DOI: 10.1109/MDAT.2019.2906108.

[27] KSS Reddy and CS Bindu, A Review on Density-Based Clustering Algorithms for Big Data Analysis. *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017)*, 2017, 123–130.

[28] S Khalid and D Prieto-alhambra, Machine Learning for Feature Selection and Cluster Analysis in Drug Utilisation Research. *Current Epidemiology Reports*, 2019, *6*, 364–372.

[29] A Krishnan and A Amarthaluri, Large Scale Product Categorization using Structured and Unstructured Attributes. *KDD '19: ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, *2019*, http://arxiv.org/abs/1903.04254

[30] V Kumari and Savita, PERCEPTION AND PRACTICES OF MAN AND WOMEN ADOPTING METHODS OF CONTRACEPTION : AN EXPLORATORY STUDY ON GOVERNMENT. *International Journal of Engineering & Science Research*, 2018, *51*(51), 278–283.

[31] S Lim, S Tucker and S Kumara, An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics*, 2017, *66*, 82–94. DOI: 10.1016/j.jbi.2016.12.007.

[32] Lorraine, N Goeuriot Jin-Cheon, WYMKS Foo, C KHO, T YIN-LENG and C YUN-KE, Textual and Informational Characteristics of Health-Related Social Media Content: A Study of Drug Review Forums. *Asia Pacific Conference Library & Information Education & Practice, 2011 Textual*, 548–557.

[33] M Mohanadevi, A Clustering Based Collaborative Approaches for Health Care System Using Clinical Document. *International Research Journal of Engineering and Technology(IRJET)*, 2017, *4*(4), 1990–1995. https://www.irjet.net/archives/V4/i4/IRJET-V4I4514.pdf.

[34] BT Nguyen, JL Elia, CY Ha and BE Kaneshiro, Pregnancy Intention and Contraceptive Use among Women by Class of Obesity : Results from the 2006 – 2010 and 2011 – 2013 National Survey of Family Growth. *Women's Health Issues*, 2020, *28*(1), 51–58, DOI: 10.1016/j.whi.2017.09.010.

[35] NS Nithya, K Duraiswamy and PGomathy, A Survey on Clustering Techniques in Medical Diagnosis. *International Journal of Computer Science Trends and Technology*, 2013, *1*(2), 17–22. www.ijcstjournal.org

[36] G Ogbuabor and F. N, U, Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology*, 2018, *10*(2), 27–37, DOI: 10.5121/ijcsit.2018.10203

[37] K Orkphol and W Yang, Sentiment Analysis on Microblogging with K -Means Clustering and Artificial Bee Colony. *International Journal of Computational Intelligence and Applications*, 2019, *18*(3), 1–22, DOI: 10.1142/S1469026819500172

[38] Overview of Influenza Surveillance in the United States, Available at: http://www.cdc.gov/flu/weekly/overview.htm, 2016.

[39] RV Pattani, Efficient Density-Based Clustering of Tweets and Sentimental Analysis Based on Segmentation. *International Journal of Computer Techniques* -–, 2016, *3*(3), 53–57, http://www.ijctjournal.org.

[40] A Primpeli, R Peeters and C Bizer, The WDC training dataset and gold standard for large-scale product matching. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 2019,381–386. DOI: 10.1145/3308560.3316609.

[41] T Rajesh and KVG Rao, Time series clustering-introduction to healthcare system. *International Journal of Innovative Technology and Exploring Engineering*, 2019, *9*(1), 2958–2963. DOI: 10.35940/ijitee.A9115.119119.

[42] BR Reddy, Y V Kumar and M Prabhakar, Clustering large amounts of healthcare datasets using fuzzy c-means algorithm. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019*, 93–97, 2019, DOI: 10.1109/ICACCS.2019.8728503.

[43] J Samriya, S Kumar and S Singh, Efficient K-Means Clustering for Healthcare Data. *Advanced Journal of Computer Science and Engineering*, 2016, *4*(2), https://www.researchgate.net/publication/330412404.

[44] C Scaffidi, K Bierhoff, E Chang, M Felker, H Ng and C Jin, Red Opal : Product-Feature Scoring from Reviews. *EC '07: Proceedings of the 8th ACM Conference on Electronic Commerce*, 2007, 182–191.

[45] DL Schminkey, X Liu, S Annan and EM Sawin, Contributors to Health Inequities in Rural Latinas of Childbearing Age : An Integrative Review Using an Ecological Framework. *Community-Based Participation*, 2019,1–20, DOI: 10.1177/2158244018823077.

[46] S Shayaa, NI Jaafar, S Bahri, A Sulaiman, PS Wai, YW Chung, AZ Piprani and MA Al-Garadi, Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*, 2019, 37807–37827. https://doi.org/10.1109/ACCESS.2018.2851311

[47] TK Shivaprasad and J Shetty, Sentiment analysis of product reviews: A review. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2017*, *Icicct*, 2017, 298–303, DOI: 10.1109/ICICCT.2017.7975207

[48] G Shobana, B Vigneshwara and AM Sai, Twitter sentimental analysis. *International Journal of Recent Technology and Engineering*, 2019, *7*(4), 343–346, DOI: 10.5373/jardcs/v12sp5/20201828.

[49] KB Simmons, LB Haddad, K Nanda and KM Curtis, Drug interactions between rifamycin antibiotics and hormonal contraception : a systematic review. *An International Journal Of Obstetrics & Gynaecology*, 2017, *125*(7), 804–811, DOI: 10.1111/1471-0528.15027.

[50] B Singh, S Kushwah, S Das and P Johri, Issue and challenges of online user generated reviews across social media and e-commerce website. *International Conference on Computing, Communication and Automation, ICCCA 2015*, DOI: 10.1109/CCAA.2015.7148486.

[51] B Singh, S Kushwah and S Das, Multi-Feature Segmentation and Cluster based Approach for Product Feature Categorization. *International Journal of Information Technology and Computer Science*, 2016, *8*(3), 33–42. https://doi.org/10.5815/ijitcs.2016.03.04.

[52] S Singh, N Priya, D Roy, A Srivastava and S Kishore, Trends in contraceptive demands and unmet need for family planning in migrant population of Uttarakhand. *International Journal of Community Medicine and Public Health*, 2018, *5*(2), 590–595.

[53] C Tang, JM Plasek, Y Xiong, DW Bates and Zhou, Clustering Similar Clinical Documents in Electronic Health Records. *International Conference on Data Science*, 2018, *00*, DOI: 10.13140/RG.2.2.16490.41920.

[54] M Tg, K Y, A Shibru and B A, Utilization of Reversible Long Acting Contraceptive Methods and Associated Utilization of Reversible Long Acting Contraceptive Methods and Associated Factors among Women Getting Family Planning Service in Governmental Health Institutions of Gondar City, *Austin Journal of Public Health and Epidemiology*, 2019, *5*(1), 1–7.

[55] Twitter, *No Title*. Twitter Usage, Available at : https://about.twitter.com/company, 2016.

[56] S Valsamidis, T Theodosiou, I Kazanidis and M Nikolaidis, A Framework for Opinion Mining in Blogs for Agriculture. *6th International Conference on Information and Communication Technologies in Agriculture, Food and Environment (HAICTA 2013) A*, 2013, *8*, 264–274, DOI: 10.1016/j.protcy.2013.11.036.

[57] M Yang, M Kiang and W Shang, Filtering big data from social media - Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, 2015, *54*, 230–240, DOI: 10.1016/j.jbi.2015.01.011.

[58] X Yu, Y Liu, X Huang and A An, Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 2012, *24*(4), 720–734, DOI: 10.1109/TKDE.2010.269.

[59] Y Yu, J Zhao, Q Wang and Y Zhang, Cludoop: An efficient distributed density-based clustering for big data using hadoop. *International Journal of Distributed Sensor Networks*, *2015*, DOI: 10.1155/2015/579391.

[60] Z Zhai, B Liu, H Xu and P Jia, Clustering product features for opinion mining. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, 2011, 347–354, DOI: 10.1145/1935826.1935884