Determining PolyCystic Ovarian Syndrome Severity from Reddit Posts using Topic Modelling and Association Rule Mining

Santhi Selvaraj
Department of Computer Science and Engineering
Mepco Schlenk Engineering College, India
santhicse@mepcoeng.ac.in

Selva Nidhyananthan Sundaradhas
Department of Electronics and Communication Engineering
Mepco Schlenk Engineering College, India
nidhyan@mepcoeng.ac.in

Abstract: *Nowadays social media plays a vital role in various real-time applications, especially in healthcare applications. PolyCystic Ovarian Syndrome (PCOS) is a condition that affects females between the ages of 15 and 35 who are of reproductive potential. The symptoms of PCOS are hormonal issues, irregular periods, weight gain, follicles, infertility, excessive hair growth in the skin, hair loss, acne, pimples, dark scars, and depression. Most of the earlier researchers analyzed the PCOS based on clinical text and health records using a machine learning approach. The main motivation of this proposed work is to predict the upcoming PCOS symptoms based on current symptoms and find the severity of the PCOS from Reddit users. This is done by collecting head symptoms from Gynecologists, gathering present symptoms from Reddit users, collecting unstructured data is pre-processed and PCOS sub symptoms are extracted using Bag of Words. The sub symptoms are mapped into head symptoms using Latent Dirichlet Allocation (LDA) for dimension reduction. The major issue in that approach is a single user has experienced the same type of symptom multiple times. This issue is solved by implementing a novel method called Symptom Segmentation and grouping Labeled Latent Dirichlet Allocation (SSG_LLDA) is designed to reduce the dimensionality and map the social media users sub symptoms into head symptoms. Association Rule Mining (ARM) with Apriori is employed to produce the frequent symptoms, and effective rule sets, and form the distinctive symptom patterns. Among several mini-mum support and confidence metrics, 0.02 and 0.1 delivers the best rule sets and symptom patterns. Based on rulesets of symptom patterns and combinations, the severity of PCOS is determined for Reddit users. The novelty of this work is the construction of PCOS symptom patterns from topic modelling results instead of original data so the dimensionality of the features is reduced and more scalable.*

Keywords: *Association rule mining, bag of words, frequent symptoms set, PCOS, topic modelling.*

Received September 27, 2023; accepted March 31, 2024
https://doi.org/10.34028/iajit/21/3/8

1. Introduction

Social media is a collection of web-based applications that create Web 2.0 which is the conceptual and technological frameworks that enable users to share their ideas, opinions, and thoughts through virtual or social networks and communities [26]. According to data and applications, various forms of social media are used like social networks, media sharing networks, discussion forums, consumer review networks, blogs and microblogs, and social shopping networks example [17]. Based on global statistics [22] report, YouTube, Facebook, Instagram, and Twitter have the largest percentage of users and access rates.

Millions of individuals use these types of social media to share their information, images, news, live audio, and videos [1]. Trillions of bytes of data are shared every second on social media, which can be in structured or unstructured data formats [61] so big data in social media could be analyzed [55]. Most researchers are now concentrating their efforts on social networks analysis, which involves capturing relational or structured data in the form of social entities such as persons, groups, and organizations with some interactions between them [62]. Social media analytics is used to convert large volumes of unstructured data into meaningful information by applying various Natural Language Processing (NLP) techniques and tools [37].

Currently, social media plays a vital role in healthcare applications compared to other commercial applications. The significance of social media in the healthcare context in establishing interactions between patients and doctors, as well as connecting various health informatics systems has been investigated by healthcare organizations and medical management [20]. The relationship between patients and healthcare professionals is described in terms of perceptions of patients and facts by professionals [58]. Best practices, benefits, and risks of social media in health care professionals [65] were suggested in various social networking sites (blogs, microblogs, wikis) and their usage in health care. Social media and health policy [9] described how health policies are in-spired in social media and listed the dangers when users handle social media carelessly. Some social benchmark datasets [52] were used to classify the diseases from various social media posts and comments. Flu-related tweets [3] were classified by applying Keyword features and FastText classifier techniques and flu relevant and irrelevant posts were separated. An influenza analysis and prediction [5] system was used to extract the features and classify the Arabic tweets into reporting and non-reporting groups.

PolyCystic Ovarian Syndrome (PCOS) is a condition that leads to the growth of ovarian cysts and prevalent endocrine disorders in childbearing women which can lead to infertility. The growing inclination of PCOS is mostly perceived in the age group 15 to 45 years [54]. Stein and Leventhal [12] first found PCOS in 1935 in Chicago and they observed some similar symptoms among groups of people like lack of menstruation, infertility, excess body hair, and hormonal abnormalities. Common symptoms [54] of PCOS as infertility, hyperandrogenism, obesity, abdominal fat, baldness, hirsutism, acne, deep voice, insulin resistance, sugar graving, frequent urination, anxiety, mood swings, depression, and hypertension. Today, the prevalence of PCOS is increasing in different countries, and it ranges from 4%-20% worldwide. The prevalence ranges will increase year by year and it will be reached as hidden Epidemic diseases in the future. A disease study in 2017 [34] suggested measuring the burden of PCOS among 194 countries.

In earlier days, most people were unaware of this disease and only went to the hospital if they had severe symptoms of PCOS. Several ultrasound tests and puncture or surgical treatment are required at this time to heal from PCOS. Following that, researchers focused on data mining, and machine and deep learning strategies for diagnosing and predicting PCOS from medical datasets like ultrasound images [60] and symptom parameters [13]. Today most women use social media to share PCOS-related symptoms, medications, diets, and activities. According to patients' tweets and posts, the system assessed the prevalence of PCOS among women and obtained information on how many women are affected by specific symptoms.

The proposed work is implemented by using two significant approaches topic modelling and Association Rule Mining (AMR). Topic modelling is used to reduce the dimensionality of the features, works well in a large corpus, and can be fitted in any other models like clustering, AMR, etc. Latent Dirichlet Allocation (LDA) is an unsupervised machine learning algorithm that is used to group related words into topics and related topics into documents. Reddit users have provided their PCOS symptoms in their own words, so this topic modelling is utilized here to reduce the dimensions by grouping the users' sub symptoms into the primary symptoms using a modified LDA approach Symptom Segmentation Grouping_Labelled Latent Dirichlet Allocation (SSG_LLDA). ARM is commonly used to uncover significant relationships in datasets, identify frequent objects present in each record, and predict future behavior based on association rules. ARM is utilized in the proposed work to identify the symptom pattern from head symptoms, i.e., to find the subsequent symptoms based on the current symptoms by employing more efficient rule-building.

In this regard, we propose a system for predicting the upcoming PCOS symptoms based on current symptoms and finding the severity of the PCOS from Reddit users. The main contributions of the proposed work are given below:

1. PCOS head symptoms gathering from Gynaecologist.
2. Collect the PCOS-related posts from the Reddit users.
3. Pre-process the collected data by applying various pre-processing techniques.
4. Extract the PCOS related sub symptoms by using the feature extraction technique.
5. Map the relevant sub-symptoms into head symptoms using topic modelling SSG_LLDA.
6. Frequent Head Symptom and consequent symptom identification from topic modelling results using AMR.
7. Selection of effective Rules and construction of the unique symptom pattern for predicting the upcoming symptoms and finding the severity of the PCOS by taking various symptom combinations.

The remaining sections of this paper are outlined as follows: Section 2 describes the related works. Section 3 describes the overall architecture and methodologies of the proposed work, while section 4 describes the experiments and results of the proposed work. Section 5 concludes the proposed work.

2. Related Work

We have arranged the literature works as per the order like disease prediction from social media, different topic modelling techniques, and ARM for various applications, and finally included the works related to PCOS Disease Diagnosis based on the Kaggle dataset and ultrasound images. Most researchers have concentrated disease-level research using social media tweets and posts in the last ten to fifteen years. Zhang *et al*. [67] proposed a Support Vector Machine (SVM) which was used to detect and evaluate influenza epidemics in China via social media. Alkouz *et al*. [6] developed the Tweetfluenza model which was used to anticipate the spread of influenza in the United Arab Emirates in real time using cross-lingual data obtained from Twitter data streams. Zhang *et al*. [68], predicted the opinion for coronavirus pandemic from real-time tweets using Machine Learning Techniques like decision tree, linear regression, k-nearest neighbor, random forest, and SVM. Amin *et al*. [8] implemented the Long Short-Term Memory (LSTM) and Word2vec method for detecting Dengue or Flu from social media tweets.

Liu *et al*. [35] described the overview of topic modelling and its applications in the bioinformatics field using clustering results. Alga *et al*. [4] analyzed the temporal changes of research topics in scientific publications during the COVID-19 pandemic using LDA. Lossio-Ventura *et al*. [36] evaluated the different clustering and topic modelling techniques for health-related emails and tweets. Garbhapu and Bodapati [18] did a comparative analysis of Latent Semantic Analysis (LSA) and LDA in Bible data, LDA achieves 60 to 70% superior performance and better coherence score compared to LSA. Mohammed and Al-Augby [39] implemented LSA and LDA topic modelling classification and its comparison study on e-books. Various topic modelling algorithms [7, 19, 21, 28, 66, 69] were used to identify the highest probability risk factors in coronary heart disease, infectious disease, lipoprotein(a), diabetes, COVID-19, etc., Many researchers have done a clustering-based topic model in Arabic texts [24, 27] and topic discovery in social networks using the sparse topic model [57].

Lau *et al*. [33] proposed pattern mining for Dyspepsia symptoms based on Apriori with a subgroup constraint framework and formed the symptom clusters over time. Dogan and Turkoglu [14] diagnosed Hyperlipidemia using association rules and constructed the decision support system. Patil and Kumaraswamy [45] implemented the heart attack diagnosis by extracting the efficient pattern using K-means clustering and the maximal frequent itemset algorithm. McCormick *et al*. [38] proposed a hierarchical model with ARM for automatic prediction of future symptoms based on current and past symptoms. Kumar and Arumugaperumal [31] surveyed much literature related to AMR, rule generation algorithms, and its measures for medical applications. Three ARM and generation algorithms Apriori, Predictive Apriori, and Tertius were implemented by Nahar *et al*. [41] for detecting the factors of heart disease in both men and women. The risk factors of Early Childhood Caries [23], Diabetes Mellitus [25], and patterns of heart diseases [59] were found using AMR. Khare and Gupta [30] implemented association rule analysis for cardiovascular disease from the UCI dataset. Nguyen *et al*. [43] found the toxicities of cancer treatment using temporal AMR. Ramasamy and Nirmala [53] predicted the disease using ARM for extracting the information from the hospital database and keyword-based clustering to find which disease was affected by the patient. Authors surveyed the various algorithms for frequent item-set or pattern mining [10] in different models and datasets [40]. Nandhini *et al*. [42] implemented predictive association rule classifiers in healthcare datasets and compared the performance with other association classification algorithms. Domadiya and Rao [15] implemented ARM with distributed privacy preservation for electronic healthcare data. Tandan *et al*. [63] discovered symptom patterns of COVID-19 using ARM Dahmani *et al*. [11], improved the big data in a cloud environment using association rules using Ontology.

Pradeepa *et al*. [51] proposed a Detection and prEvention of polycystic Ovary synDrome using assOciation rule hypeRgrAph and domiNating set properTy (DEODORANT) system for early detection and prevention of PCOS from blogs, posts, and social media chats using frequent items in hypergraphs with the dominating set symptoms and form the clusters. Nsugbe [44] implemented machine learning-based decision-making systems for classifying the PCOS from the Kaggle dataset. Elmannai *et al*. [16] predicted PCOS disease from the Kaggle dataset using different feature selection methods. Khanna *et al*. [29] proposed an explainable artificial intelligence framework for predicting PCOS in the Kaggle dataset. Alamoudi *et al*. [2] proposed a deep-learning fusion approach for predicting PCOS from ultrasound images. Tiwari *et al*. proposed [64] a SPOSDS system for diagnosing PCOS using a machine learning algorithm in the Kaggle dataset.

In all previous relevant works, the main research gap was exclusively focused on the Kaggle dataset or ultrasound pictures for predicting and detecting PCOS using machine and deep learning algorithms. However, using the Kaggle dataset’s symptoms and recommendations from Gynaecologists, we chose the PCOS symptoms for our suggested work from social media. Many PCOS-related works have solely focused on categorizing patients based on their clinical symptoms, but this work also analyses the current and any subsequent symptoms that may be present in the women.

3. Proposed Work

This work proposes an innovative model for determining the symptom pattern and analyzing the severity of PCOS from social media users. This proposed work benefits predicting the PCOS symptoms earlier based on our symptom pattern construction and finding the prevalence of PCOS according to the association rules. The primary benefit of this work is the efficient integration of ARM with topic modelling. The sequence of proposed work is shown in Figure 1, a mathematical illustration of each step is given in Algorithm (1), and workflow is given in Figure 2:

1. PCOS head symptom collection.
2. Data collection from Reddit.
3. Data pre-processing.
4. PCOS sub symptoms extraction.
5. Sub symptoms to head symptoms extraction.
6. Association rule construction from head symptoms.

Algorithm 1: Flow of Proposed Work

```
Input:    Dinput: Reddit PCOS Posts input Document
Output:   SPSumm[]: Best Symptom Pattern Summary
BEGIN
                    //Step 1: Data Pre_Processing using NLTK
      Function Pre_Proc($D_{input}$)
          $P_{Tot}$:[$p_1$, $p_2$ ..., $p_n$] ⟵ $D_{input}$                          //n - number of posts in the $D_{input}$
          $P_{lower}$ ⟵toLower(PTot)                              //Lowercase Conversion
          $P_{sent}$  ⟵SentenceTokenizer($P_{lower}$)                 //Sentence Tokenization
          $P_{words}$ ⟵WordTokenizer($P_{sent}$)                     //Word Tokenization
          $P_{punc}$  ⟵removePunctuation($P_{words}$)                //Punctuation Removal
          $P_{stop}$  ⟵stopwordsRemoval($P_{punc}$)                 //Stop words Removal
          $P_{lemma}$ ⟵WordNetLammatizer.lemmatize($P_{stop}$)      //Stem Word Extraction
          $P_{BG}$    ⟵N_grams($P_{lemma}$, 2)                   //Bigram Modelling
          $P_{Neg}$   ⟵negationHandling($P_{BG}$)               //Negation Word finding
          PP      ⟵normalization($P_{Neg}$)               //Normalize and pre-processed word
          return PP                                                 //return pre-processed data
        END Function
                    //Step 2: Sub Symptom Extraction using BoW
      Function Sym_Ext(PP, S)         //PP – { $PP_1$ ... $PP_n$} & S – {$S_1$, $S_2$, ... $S_m$}– Sub Symptoms Dictionary as per Table 1
         for each $PP_i$ in PP
            for each symptom $s_j$ in S
                if $PP_i$ == $s_j$
                  $SS_{i,j}$ ⟵ 1
                else
                   $SS_{i,j}$ ⟵ 0
               end if
             end for
         end for
         return SS                                       //return extracted sub symptoms as Bow
       END Function
                    //Step 3: Head Symptom Mapping
      Function SSG_LLDA(SS, HS, α, β, PP)
                              //SS–sub symptoms, HS–Head Symptoms, α, β–Sparsity parameter, PP–Pre-Processed posts
     $HS\epsilon 0,1^{SS}$, $PP\epsilon 0,1^{HS}$                           //Initialization
     for tϵ[1.....|HS|] do
         $SS_t \sim Dir(\beta)$                            //Dirichlet distribution of SS on HS as Eq.10
     end for
      SS ← [$SS_1, SS_2$ .....$SS_t$]                        //Collection of all SS distributions
      for jϵ[1.....|PP|] do
         $PP_j \sim Dir(\alpha)$                            //Dirichlet distribution of HS on P as  Eq.9
       end for
       for each HS in PP
         for each sub symptom segment ($SS_s$) in SS
            if |$SS_s$| = = 1
              Pr(SSs) ⟵ Pr(HS|$SS_S$, PP)            // as per Eq. 14
           else
               Pr(SSs) ⟵ Pr(HS|$SS_S$, SS, PP)       // as per Eq. 15
           end for
      end for
      //Labelling and grouping Head Symptoms
       NHS=10        //Total number of head symptoms as 10 as per Table 1
       Label[SSs]={}
       for each SS in PP
       for each sub symptom segment ($SS_s$) in SS
            Label[SSs]=Label[SSs] ∪HS[SSs]      //label and group the same type of symptoms in a single pre-processed post
       end for
  end for
return Label
    END Function
    //Step 4: PCOS Symptom pattern generation
    Function ARM(Label, PP)                              //HS - Head Symptoms, PP–Pre-Processed Posts
      import apriori functions from frequent_patterns package
      min_support ⟵ 0.02, min_confidence ⟵ 0.1
      for each Label i in PP
            $FHS_i$ ⟵ apriori(i, min_support)                    //frequent Head Symptoms
     $RS_i$ ⟵ association_rules (FHSi, min_confidence)          //Rule sets
       end for
       compute Support of $FHS_i$
 if support > min_support
      return $FHS_i$
   compute confidence and lift for $RS_i$
    if confidence > min_confidence
        order the $RS_i$ using lift
         return $RS_i$
     end if
end if
$SP_{Summ}$ ⟵ RSi
  END Function
END
```

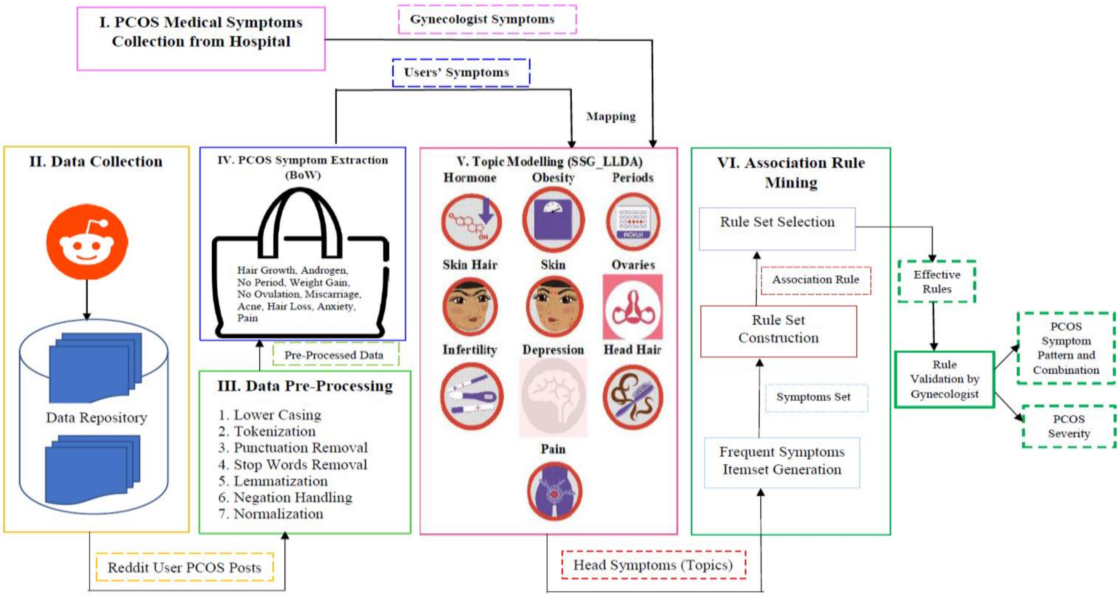


Figure1. Schematic diagram of the proposed work.

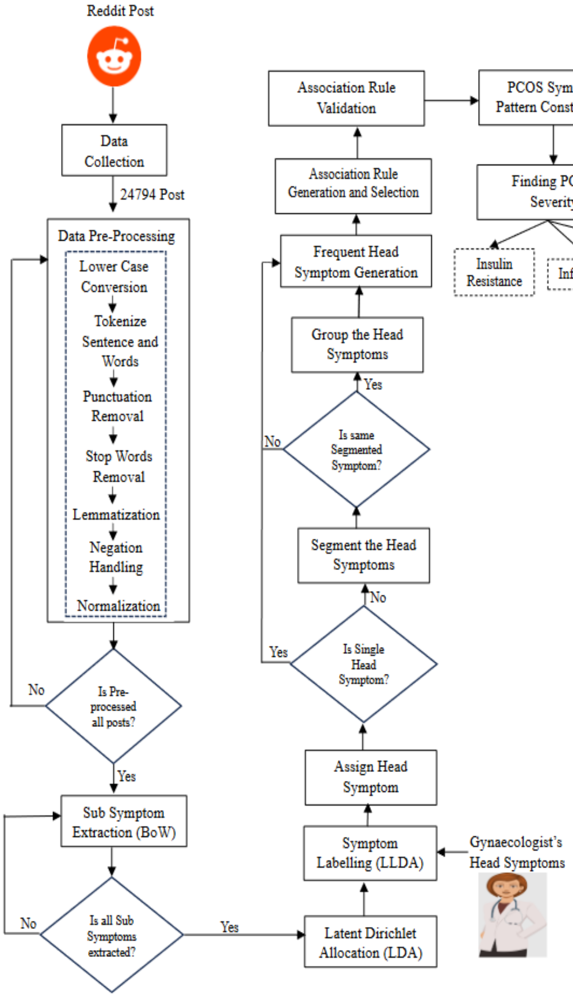

Reddit Post
Data Collection
24794 Post
Data Pre-Processing
Lower Case Conversion
Tokenize Sentence and Words
Punctuation Removal
Stop Words Removal
Lemmatization
Negation Handling
Normalization
Is Pre-processed all posts?
No
Yes
Sub Symptom Extraction (BoW)
Is all Sub Symptoms extracted?
No
Yes
Latent Dirichlet Allocation (LDA)
Symptom Labelling (LLDA)
Gynaecologist's Head Symptoms
Assign Head Symptom
Is Single Head Symptom?
Yes
No
Segment the Head Symptoms
Is same Segmented Symptom?
No
Yes
Group the Head Symptoms
Frequent Head Symptom Generation
Association Rule Generation and Selection
Association Rule Validation
PCOS Symptom Pattern Construction
Finding PCOS Severity
Insulin Resistance
Inflammatory
Adrenal

Figure 2. Flow chart for proposed workflow.

3.1. PCOS Symptom Collection

Before starting data collection on social media, we consult with a Gynaecologist from Lakshmi Hospital and Fertility Centre in Sivakasi, Tamil Nadu [32]. The doctor has explained more information about PCOS as well as a list of the most common PCOS symptoms. In reality, users on social media did not provide exact symptoms, instead describing their symptoms in their terms. As a result, mapped the doctor's symptoms to the symptoms described by social media users, as shown in Table 1.

Table 1. Gynaecologists mentioned symptoms vs. social media users' symptoms.

Sl. No.	**Gynaecologist (head) symptoms**	**Social media users' (sub) symptoms**
1	Hormone problem	hormonal abnormality, excess androgen, breast change, testosterone, hormonal disruption, sugar craving, low libido, appetite, frequent urination, hormonal imbalance, endometriosis, diabetic, insulin resistance, adrenal hyperplasia, osteopenia, progesterone, estrogen, hyperinsulinemia, hyper and rogegism, hunger.
2	Obesity	Weight gain, obesity, inflammation, fat, belly fat, overweight, abdomen fat, excess weight, bloating.
3	Periods problem	Irregular period, skip menstrual period, skip period, heavy bleeding, irregular cycle, abnormal period, heavy period.
4	SkinHair problem	excessive hair growth, hair growth, dark hair, hirsutism, facial hair, excess body hair, upper lip hair, stomach hair, chest hair, chin hair, mustache, unwanted hair growth.
5	Skin problems	oily skin, acne, patch, darkening skin, skin tag, dark patch, dark spot, pimple, irritability, redness skin, stretch mark.
6	Ovaries problem	No ovulation, follicle, cyst ovary, polycystic ovary, cyst, ovarian cramp, cystic acne, oligoovulation, anovulation.
7	Infertility	Not pregnant, never get pregnant, no pregnant, infertility, miscarriage.
8	Depression	Fatigue, sleep problem, depression, anxiety, low energy, mental hell, mood swing, brain fog, insomnia, high blood pressure, stressful, anger, tired.
9	HeadHair problems	Thinning hair, hair loss, baldness, hair fall.
10	Pains	Headache, lower abdominal pain, painful period, breast pain, pelvic pain.

3.2. Data Collection

Nowadays, most social media sites, such as Twitter, Tumblr, Facebook, and Reddit, are being used to share PCOS-related opinions among people. Many PCOS-affected women have shared their symptoms, clinical reports, feelings, scan images, medicines, treatments, diets, exercises, and awareness about PCOS. The tags on Twitter [50] @pcosdiva, @PCOSnutrition, @pcoschallenge, @AwarenessPCOS, @PCOSA, @PCOSDiet is mainly focused on spreading the PCOS awareness and challenges in terms of images and videos. The tags on Facebook [47] #PCOSIndia, #PCOSsup-portgroup, #PCOSTips, #PCOSAwareness, and #PCOSSociety mainly concentrate on awareness, diets, and treatments in terms of videos and images. On Tumblr [49], the #pcos tag describes the images related to PCOS symptoms. The majority of tweets on Twitter, Facebook, and Tumblr are images and videos rather than text and do not focus on symptoms-related posts of the people.

Reddit is the best healthcare social media platform and r/PCOS [48] tag provides more information about PCOS symptoms among people all around the world (112k members). PCOS symptoms such as weight gain, irregular periods, hirsutism, mental health, fertility, hair loss, hair thinning, and aches have been discussed by members of this group. This platform supports a greater number of texts rather than images and videos so Reddit posts are more helpful for this proposed model. 24794 Reddit PCOS posts were gathered between January 2020 and September 2022 [46], but the individual profiles were not accessible on the Reddit page. The collected user's posts are saved into a .csv file and it is converted into a data frame during the process. Table 2 shows the sample posts and what type of symptoms collected from the PCOS community on the Reddit page as mentioned in Table 1.

Table 2. Sample posts and its relevant symptoms.

Sample posts and sub symptoms	**Reddit URL**	**Relevant head symptoms**
Yes, I'm **fat**. I have **hormonal abnormalities**.	https://www.reddit.com/r/PCOS/comments/fvyg2n/i_hate_having_the_pcos_belly/	Obesity, Hormone Problem
Facial hair was the first and big sign for me. I went to doctor who diagnosed me with **hirsutism, hair loss** on my scalp	https://www.reddit.com/r/PCOS/comments/fuwv06/just_had_the_best_doctor_experience_of_my_life/	SkinHair and HeadHair Problem
Frequent **headaches**, and the worst symptoms are **obesity**	https://www.reddit.com/r/PCOS/comments/qb8eko/what_were_your_first_signs_of_pcos/	Pain, Obesity
Luckily my only symptoms are **irregular periods**. But I Can feel as my period gets more irregular, my **anxiety** levels increase	https://www.reddit.com/r/PCOS/comments/qw6zpj/anyone_with_pcos_a_regular_period/	Periods problem, Depression
I had been told for many years I would **never get pregnant.**	https://www.reddit.com/r/PCOS/comments/plorbp/it_just_hit_me_that_i_might_never_be_able_to_get/	Infertility
I worry about all the **scars** and **acne**	https://www.reddit.com/r/PCOS/comments/qw6zpj/anyone_with_pcos_a_regular_period/	Skin Problem
No ovulation and I have the **follicles** on both ovaries	https://www.reddit.com/r/PCOS/comments/13je4sc/pcos_no_symptomps_of_ovulation_and_risk_of/	Ovaries Problem
When I was diagnosed 18 years ago, there was uncertainty about whether Insulin Resistance and excess androgens caused weight gain, caused the hormonal disruption	https://www.reddit.com/r/PCOS/comments/q6zyd1/what_causes_pcos_exactly/	Hormone Problem, Obesity

3.3. Data Pre-Processing

The social media resources and text tweets do not adhere to conventional English and contain acronyms, hashtags, URLs, punctuation, special characters, numerals, and misspelled words which are not needed for the learning and analysis process. With the growth of data generation and increasing heterogeneity of data in social media, the rate of noisy, incomplete, and inconsistent data has risen dramatically. This unstructured and inconsistent data will reduce the accuracy of the model so data pre-processing is an essential step for model construction. Data pre-processing is a technique for converting raw data into relevant and intelligible information, as well as ensuring data quality [56].

Text pre-processing is a required step that is carried out with the help of Natural Language Processing (NLP) pre-processing techniques. It includes cleaning and removing noise from the data, lower-case conversion, punctuation removal, URL removal, spell checking, tokenization, removing stop words, stemming and lemmatization, normalization, etc., This data pre-processing helps to enhance the corpus by removing unrelated content and increasing the accuracy of the learning algorithms by extracting essential features.

Before doing the pre-processing, we saved our dataset in data frames and each step of the pre-processing result is kept in a column of data frames. The following data pre-processing algorithms are implemented in our dataset using Python and the Natural Language Tool Kit (NLTK).

1. Lowercasing

All the text or tweets should be converted into lowercase because of the uniformity of the text, the same entity of the text, and the consistent string format. For example, PCOS, Pcos, PcoS ought to be converted into pcos.

2. Tokenization

It is the process of splitting the paragraph into sentences in turn sentences into words and returns the list of tokens. The length of the list is equivalent to the number of words in the original text. This token separation is dependent on whitespace and punctuation to determine when one word is finished and the next one begins. Punctuation occurred for the terms don't, didn't, and it has a tokenization issue, so it can be dealt with separately.

3. Punctuation Removal

Using a regular expression in the string module, we delete the extra punctuations from the tokenized input. This module performs three phases to replace the 32 punctuations with space: find punctuations, change punctuation, and give space.

4. Stop Words Removal

Stop words are words that appear frequently in languages but do not provide any useful information. Stop words include prepositions, articles, auxiliary verbs, conjunctions, and pronouns. NLTK library listed 180 stop words in the English language and it is removed from the original texts.

5. Stemming or Lemmatization

It is the process of deleting affixes from the original posts to reduce the stem/lemma/root word. The stemming method does not check the word's part of speech or context, but lemmatization uses the WordNet dictionary to check the word's context. It is used to maintain the various forms of the words into a single base word. For example, the words looked, looking and looks are reduced to the base word 'look'.

6. Negation Handling

It is one of the important steps in NLP since the negated words affect the scope of the given text. Not, nor, never, no, didn't, don't, doesn't, etc., are negation words, but they also stop words. We deal with this negation by combining two words using Bigram modelling rather than unigram because the negative scope is missed in the unigram. It means combining current words and previous words at a time. In our dataset, the phrases not pregnant, never getting pregnant, and no period are merged to show the negative scope which is handled using the Bigram method and normalization technique. For example, Unigram and Bigram of the word "never getting pregnant" as <never>, <getting>, <pregnant>, <never getting>, <getting pregnant> and then normalization into a single word <nevergettingpregnant>.

7. Normalization

It is used to convert the raw text into canonical representation and maintain the uniformity of the words. The social media posts contain shortened words, variously spelled words, numbers used instead of text, non-standard English or neologism words, etc. Text normalization algorithms are used to manage these words and combine many words into a single word. For example, weight gain, irregular periods, hair loss, belly fat, heavy bleeding, and so on. Those terms are normalized, resulting in a single word such as weightgain, irregularperiod, bellyfat, hairloss, and heavybleeding, which may be used to extract features or symptoms quickly in the next step. The above all pre-processing steps outputs are discussed in the experimental results section

3.4. PCOS Sub Symptoms Extraction

Feature extraction entails converting pre-processed textual input into numerical information. In this work, PCOS symptoms are extracted from pre-processed data using the Bag of Words model together with a single hot representation and the count vectorizer method, which makes it simple to produce binary outputs. In this method, the unique symptoms are maintained in the dictionary as per Table 1, and more than one symptom is also present in the single user's post.

Let us consider dataset P as the set of posts $P=\{p_1, p_2, p_3,\dots, p_n\}$, where n is the total number of posts in the dataset. The unique sub symptoms are given as the features $S=(s_1, s_2, s_3,\dots, s_m)$, where m is the total number of symptoms. Each post p_i, $1\leq i\leq n$ is represented by a symptom vector. Each value of the symptom vector represents the weight of the symptom which is extracted from the vocabulary of the symptoms in the given dataset and it is denoted as $<wt_{i1}, wt_{i2}, wt_{i3},\dots,wt_{im}>$, where wt_{ij}

represents the weight value of symptom s_j in the post p_i. The post and symptom extraction matrix are shown in Table 3 and bag of words is given in Equations (1) and (2).

$$wt_{ij} = \begin{cases} 1, & if\ s_j\ occurs\ in\ p_i \\ 0, & if\ s_j\ not\ occurs\ in\ p_i \end{cases} \tag{1}$$

$$BOW\ (p_i, s_j) = \begin{cases} 1, & if\ wt_{ij} = 1 \\ 0, & otherwise \end{cases} \tag{2}$$

Table 3. Post and symptom extraction matrix.

Post	s_1	s_2	$\ldots s_j \ldots$	s_m
p_1	wt_{11}	wt_{12}	wt_{1j}	wt_{1m}
p_2	wt_{21}	wt_{22}	wt_{2j}	wt_{2m}
⋮				
p_i	wt_{i1}	wt_{i2}	wt_{ij}	wt_{im}
⋮				
p_n	wt_{n1}	wt_{n2}	wt_{nj}	wt_{nm}

3.5. PCOS Head Symptom Mapping (SSG_LLDA)

Reddit users described their symptoms in their own words, which meant that they did not match the primary symptoms indicated by the Gynaecologist. This raised the dimension of the sub symptom features. Use the different topic modelling strategies for reducing the symptom features by mapping the users' sub symptoms into head symptoms specified by the Gynaecologist. Topic modelling is an unsupervised learning algorithm that is used to identify the different topics from a set of documents based on statistics of the words. Topic modelling is a dimension reduction technique that uses probabilistic algorithms to find latent topics rather than words in a collection of documents. Topic model and document model construction are used to implement basic topic modelling. Topic model means the probability of the term occurring in how many topics (Sub Symptom to Head Symptoms) and document model means the probability of the topic occurring in how many documents (head symptoms to post) in the corpus.

Let us Consider 'm' words as extracted user's sub symptoms *SS*={w_1, w_2, …, w_m}, '*k*' topics as head symptoms *HS*={t_1, t_2, …, t_k} and '*n*' posts *P*={p_1, p_2, p_3,…, p_n} in the document. The basic topic modelling is given in Equation (3).

$$Pr(SS|P) = Pr(SS|HS)\, Pr(HS|P) \tag{3}$$

Where, *Pr(SS\HS)* is topic model, *Pr(HS\P)* is the document (post) model and *Pr(SS\P)* is the number of extracted sub symptoms and head symptoms occurred across the document.

Change the general LDA as in Equation (3) for single head symptoms into all head symptoms are given in Equation (4).

$$Pr(SS|P) = \sum_{HS} Pr(SS|HS, P)\, Pr(HS|P) \tag{4}$$

Apply conditional independence *Pr(SS|HS,P)=Pr(SS|HS)* and change Equation (4) into Equations (5) and (6) as follows:

$$\Pr(SS|P) = \sum_{HS} \Pr(SS|HS) \Pr(HS|P) \tag{5}$$

$$\Pr(SS|P) = \theta_{HSP}\ *\ \varphi_{SSHS} \tag{6}$$

Where θ_{HSP} is the probability of head symptoms *HS* occurring in posts *P* and φ_{SSHS} is the probability of sub symptoms *SS* occurring in head symptoms HS.

Now, we have transformed from traditional LDA to SSG-LLDA. *HS*, SS∈N, where *HS* is the total number of head symptoms and SS is the size of the sub symptoms. The discrete distributions of these vectors' representation are given as Equation (7).

$$\theta \epsilon R^{HS} \text{ and } \varphi \epsilon R^{SS} \tag{7}$$

One hot representation of a particular sub symptom in the head symptom and head symptom in the post is given as Equation (8):

$$HS\epsilon\{0,1\}^{SS}\ and\ P\epsilon\{0,1\}^{HS} \tag{8}$$

Consider the dirichlet distributions of head symptoms and sub symptom given in Equations (9) and (10):

$$\theta^{(P)} \sim Dir(\alpha) \tag{9}$$

$$\varphi^{(HS)} \sim Dir(\beta) \tag{10}$$

Where, α and β are sparsity parameter that controls per-post and head symptom distribution and per-head symptom and sub symptoms distribution and these values are less than one.

The head symptom and sub symptom assignment are given in Equations (11) and (12):

$$HS \sim Multinomial\ \theta^{(P)} \tag{11}$$

$$SS \sim Multinomial\ \varphi^{(HS)} \tag{12}$$

The joint distribution of the head symptom assignments *HS*, sub symptom assignments *SS* for ∀k∈ℕ head symptoms Pr(SS, HS, θ, *φ*, |αβ) is given as Equation (13).

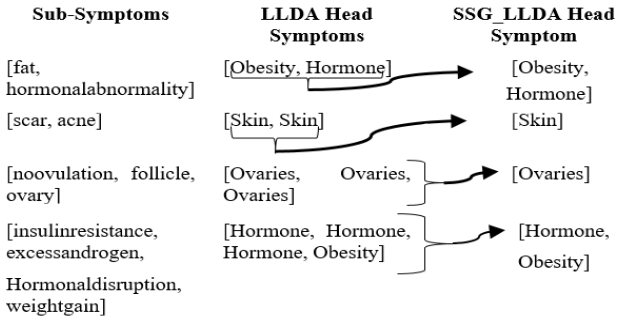
$$\Pr(SS, HS, \theta, \varphi|\alpha, \beta) = \prod_{j=1}^{P} \Pr(\theta_j|\alpha) \prod_{i=1}^{k} \Pr(\varphi_k|\beta) \prod_{l=1}^{N_j} \Pr(HS_{j,l}|\theta_j) \Pr\left(SS_{j,l}\middle|\varphi_{HS_{j,l}}\right) \tag{13}$$

Where, *Pr(*θ_j*|α)* is the probability distribution of head symptoms into posts per sub symptom-head symptom distribution parameter α, *Pr(φk|β)* is the probability distribution of sub symptoms into head symptoms per head symptom-post distribution parameter *β*, *Pr(*$HS_{j,l}$*|*θ_j*)* *Pr(*$SS_{j,l}$*|*$\varphi_{HS_{j,l}}$*)* is the probability distribution of specific head symptom and sub symptom. When choosing a new head symptom *HS* for sub symptom *SS* in posts *P*, Gibbs sampling, a technique used in LDA, is employed to select the maximum likelihood values. It is written in Equation (14) and modified into sub Symptom Segmentation (SS_S) LDA is written as Equation (15):

$$Pr(HS|SS,P) = \left(\begin{matrix} SS \\ in\ P\ that \\ assigned\ to\ HS \end{matrix} + \alpha\right) * \frac{SS\ in\ HS\ +\ \beta}{Total\ No.of\ SS\ in\ HS\ +\beta_{tot}} \quad (14)$$

$$P(HS|SS_S,SS,P) = \left(\begin{matrix} No.of\ SS_S \\ with\ different\ SS\ in\ P + \alpha \\ that\ assigned\ to\ HS \end{matrix}\right) * \frac{SS\ in\ HS\ +\ \beta}{Total\ No.of\ SS\ in\ HS\ +\beta_{tot}} \quad (15)$$

Currently, the topic model is crucial for identifying disease-related trends in tweets and news reports. Comparing LDA to LSA and Probabilistic Latent Semantic Analysis (pLSA), which assigns the likelihood to new posts, produces better results, performance, and topic visualization. As a result, this proposed implemented modified LDA like SSG-LLDA to map and reduce the user's several extracted sub symptoms into head symptoms specified by the Gynaecologists. If a single post contains many sub symptoms related to the same head symptom, categorize the topics together; otherwise, keep them separate. The following example shows the PCOS head symptom mapping using SSG_LLDA and detailed mapping is given in the Results Section.


Sub-Symptoms
LLDA Head Symptoms
SSG_LLDA Head Symptom
[fat, hormonalabnormality]
[Obesity, Hormone]
[Obesity, Hormone]
[scar, acne]
[Skin, Skin]
[Skin]
[noovulation, follicle, ovary]
[Ovaries, Ovaries, Ovaries]
[Ovaries]
[insulinresistance, excessandrogen, Hormonaldisruption, weightgain]
[Hormone, Hormone, Hormone, Obesity]
[Hormone, Obesity]

3.6. Association Rule Mining

3.6.1. Frequent Symptoms Itemset Generation

PCOS head symptoms among social media users are found and grouped by using topic modelling. ARM is one of the machine learning approaches used to generate the frequent head symptoms (items) and create the association among the symptoms. Frequent head symptom means, that head symptoms is occurred at least a minimum number of times in the dataset. In association analysis, the collection of more head symptoms is called a symptom itemset. For example, 4-symptoms itemset is denoted as {hormone, infertility, periods problem, skin}.

3.6.2. Rule Set Construction

ARM is a brute-force method and generates set of rules that would forecast the occurrence of one PCOS symptom based on occurrence of other symptoms. The association rules or patterns are formed by grouping the combination of frequent items or events that occurred at the same time. The association rule between PCOS symptoms is expressed in the form of A→B, Where A and B are disjoint sets of symptoms like A ∩ B=Ҩ. Here, A is named as antecedent of the rule and B is named as consequent of the rule. It is also known as "if→then" rules if denotes the antecedent and then denotes the consequent. For example, two PCOS symptoms are represented as Hormone→Periods. It means if patient has Hormone problem, then it leads to Irregular periods.

3.6.3. Rule Set Selection

The effectiveness of the rule set is determined by using various measurements like support, confidence, and lift. Consider rule A→B, Where A is the antecedent symptoms and B is the consequent symptoms, and apply the following measurements for constructing the rules:

3.6.3.1. Support

Support means both antecedent and consequent symptoms affected users and a total number of users. It means how frequently both symptoms occurred together as a percentage of all users and its range is [0,1].

$$\text{Support}(\text{A} \rightarrow \text{B}) = \frac{\text{Users Affected by both A and B}}{\text{Total Number of Users}} \quad (16)$$

3.6.3.2. Confidence

Confidence means both antecedent and consequent symptoms affected users and the number of users affected antecedent symptom only. It determines how frequently consequent symptom occurs those who already affected by antecedent symptom and it's range is [0,1].

$$Confidence\ (A \rightarrow B) = \frac{Users\ Affected\ by\ both\ A\ and\ B}{Users\ Affected\ by\ A} \quad (17)$$

3.6.3.3. Lift

Lift is calculated as the number of users affected by both antecedent and consequent symptoms and divided by the number of users affected by antecedent symptom only and take the division of fraction of users affected by consequent symptom. The fraction means the number of patients having consequent symptom is divided by the total number of patients and it's range is [0, ∞].

$$Lift\ (A \rightarrow B) = \frac{Usr.Affected\ by\ both\ A\ and\ B/Usr.Affected\ by\ A}{Fraction\ of\ Usr.Affected\ by\ B} \quad (18)$$

This proposed work, first generate the frequent itemset and construct the association rules by applying Apriori algorithms and also using the above measures for selecting the effective ruleset.

4. Experimental Results

The dataset considered for this work is a collection of 24794 PCOS posts from Reddit users. The user posts in the dataset are taken from https://www.reddit.com/r/PCOS/, which is a popular healthcare forum in the world. This section discusses the outcomes of pre-processing, symptom extraction, topic modelling, and ARM on the collected data.

4.1. Data Pre-Processing

Figure 3 shows the output of all pre-processing steps, Bag of Words, and SSG_LLDA approach on some example PCOS patient postings from Reddit.

4.2. Topic Modelling

Reddit users have described their symptoms in their own words so the symptom extraction module extracts their symptoms and it is denoted as words. The processing of words takes more dimension, time and it would be a tedious process for performing the next module AMR. In this regard, the users' extracted symptoms (words) are mapped into head symptoms (topics), and the dimensions are reduced using SSG-LLDA. We identified ten topics and its related words from BoW extracted symptoms using LDA, LLDA and SSG-LLDA are represented in Table 4. Topic modelling methods map the keywords into topics and topics into documents based on probabilities of the words occurred in the topic and document. Table 4 also describes the topic number, topic label and keywords with its frequencies of SSG_LLDA.

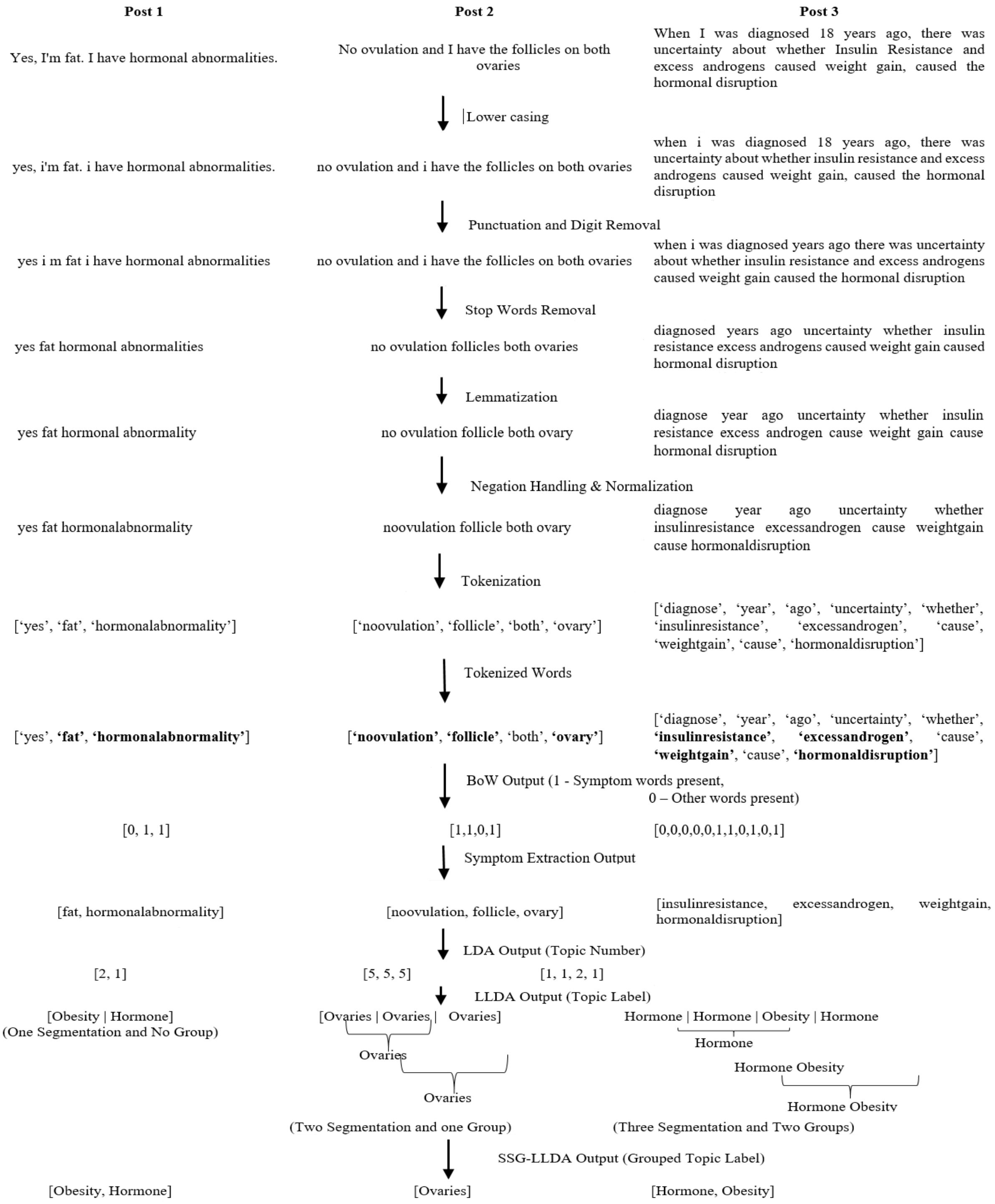

Post 1
Post 2
Post 3
Yes, I'm fat. I have hormonal abnormalities.
No ovulation and I have the follicles on both ovaries
When I was diagnosed 18 years ago, there was uncertainty about whether Insulin Resistance and excess androgens caused weight gain, caused the hormonal disruption
Lower casing
yes, i'm fat. i have hormonal abnormalities.
no ovulation and i have the follicles on both ovaries
when i was diagnosed 18 years ago, there was uncertainty about whether insulin resistance and excess androgens caused weight gain, caused the hormonal disruption
Punctuation and Digit Removal
yes i m fat i have hormonal abnormalities
no ovulation and i have the follicles on both ovaries
when i was diagnosed years ago there was uncertainty about whether insulin resistance and excess androgens caused weight gain caused the hormonal disruption
Stop Words Removal
yes fat hormonal abnormalities
no ovulation follicles both ovaries
diagnosed years ago uncertainty whether insulin resistance excess androgens caused weight gain caused hormonal disruption
Lemmatization
yes fat hormonal abnormality
no ovulation follicle both ovary
diagnose year ago uncertainty whether insulin resistance excess androgen cause weight gain cause hormonal disruption
Negation Handling & Normalization
yes fat hormonalabnormality
noovulation follicle both ovary
diagnose year ago uncertainty whether insulinresistance excessandrogen cause weightgain cause hormonaldisruption
Tokenization
['yes', 'fat', 'hormonalabnormality']
['noovulation', 'follicle', 'both', 'ovary']
['diagnose', 'year', 'ago', 'uncertainty', 'whether', 'insulinresistance', 'excessandrogen', 'cause', 'weightgain', 'cause', 'hormonaldisruption']
Tokenized Words
['yes', **'fat'**, **'hormonalabnormality'**]
[**'noovulation'**, **'follicle'**, 'both', **'ovary'**]
['diagnose', 'year', 'ago', 'uncertainty', 'whether', **'insulinresistance'**, **'excessandrogen'**, 'cause', **'weightgain'**, 'cause', **'hormonaldisruption'**]
BoW Output (1 - Symptom words present, 0 – Other words present)
[0, 1, 1]
[1,1,0,1]
[0,0,0,0,0,1,1,0,1,0,1]
Symptom Extraction Output
[fat, hormonalabnormality]
[noovulation, follicle, ovary]
[insulinresistance, excessandrogen, weightgain, hormonaldisruption]
LDA Output (Topic Number)
[2, 1]
[5, 5, 5]
[1, 1, 2, 1]
LLDA Output (Topic Label)
[Obesity | Hormone]
(One Segmentation and No Group)
[Ovaries | Ovaries | Ovaries]
Ovaries
Ovaries
(Two Segmentation and one Group)
Hormone | Hormone | Obesity | Hormone
Hormone
Hormone Obesity
Hormone Obesitv
(Three Segmentation and Two Groups)
SSG-LLDA Output (Grouped Topic Label)
[Obesity, Hormone]
[Ovaries]
[Hormone, Obesity]

Figure 3. Output of pre-processing, BoW and SSG_LLDA.

According to Table 4 results, the symptoms (topics) that social media users are affected in that order are: Periods Problem, Obesity, Skin Hair, Skin, Head Hair, Depression, Hormone Problems, Ovaries Problem, Infertility, and Pains. SSG_LLDA results are also used to find number of posts in social media users affected the single and more than one combination of symptoms which is represented in Figure 4.

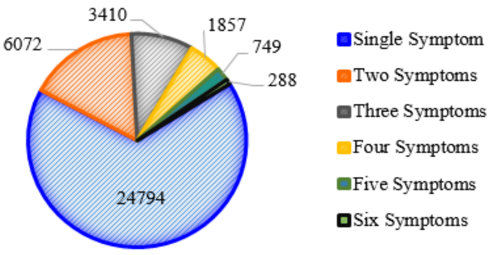


Figure 4. Number of Reddit users affected with different symptoms combination.

Table 4. Topics, keywords and probabilities of SSG_LLDA.

Topic No.	Topic label	Keywords and probabilities
1	Hormone problem	**insulinresistance (0.23)**, testosterone (0.158), hormonalimbalance (0.119), hormonaldisruption (0.015), hormonalabnormality (0.012) excessandrogen (0.107), estrogen (0.104), hyperinsulinemia (0.101), hyperandrogegism (0.094), hunger (0.0065), breastchange (0.0059), diabetic (0.0055), adrenalhyperplasia (0.0036), progesterone (0.0034), appetite (0.014), sugarcraving (0.012), frequenturination (0.005), endometriosis (0.003) osteopenia (0.002).
2	Obesity	**weightgain (0.750)**, obesity (0.147), overweight (0.086), bellyfat (0.006) inflammation (0.0012), fat (0.004), abdomenfat (0.003), excessweight (0.002), bloating (0.001).
3	Period problem	**irregularperiod (0.756)**, irregularcycle (0.070), heavybleeding (0.069), heavyperiod (0.057), skipperiod (0.042), abnormalperiod (0.005), skipmenstrualperiod (0.001).
4	SkinHair problem	**hairgrowth (0.631)**, facialhair (0.24), hirsutism (0.025), excessivehairgrowth (0.017), chinhair (0.016), excessbodyhair (0.014) darkhair (0.013), upperliphair (0.012), stomachhair (0.01), chesthair (0.009), mustache (0.007), unwantedhairgrowth (0.006)
5	Skin Problem	**Acne (0.555)**, skintag (0.01), oilyskin (0.150), darkpatch (0.04), irritability (0.038), patch (0.05), darkeningskin (0.03), stretchmark (0.02), pimple (0.01) darkspot (0.09), rednessskin (0.007).
6	Ovaries problem	**noovulation (0.22)**, cyst (0.19), cystovary (0.17), polycysticovary (0.15), follicle (0.08) ovariancramp (0.07) cysticacne (0.06), oligoovulation (0.04), anovulation (0.02).
7	Infertility	**infertility (0.21)**, notpregnant (0.2), miscarriage (0.2), nevergetpregnant (0.2), nopregnant (0.19)
8	Depression	**Fatigue (0.245)**, anxiety (0.194), moodswing (0.17), depression (0.16), sleepproblem (0.11), brainfog (0.006), anger (0.03), highbloodpressure (0.014), lowenergy (0.013), tired (0.011), stressful (0.04) mentalhell (0.004), insomnia (0.003).
9	HeadHair problem	**hairfall (0.5)**, thinninghair (0.06), baldness (0.02), hairloss (0.42).
10	Pain	**painfulperiod (0.2)**, headache (0.2), pelvicpain (0.2) lowerabdominalpain (0.2), breastpain(0.2).

Table 5. Web plot results for different symptom combinations.

Symptom combinations	**No. of symptom combinations**	**Level of symptom permutations and No. of users affected**					
		Highest	**Count**	**Middle**	**Count**	**Lowest**	**Count**
Single symptom	10	Period	3949	Head-hair	2882	Pain	638
Two symptom	41	Period and obesity	528	Hormone, depression	121	Ovaries and pain	31
Three symptom	39	Skin-hair, period and skin	209	Obesity, head-hair and depression	60	Infertility, ovaries and depression	14
Four symptoms	18	Skin, hormone, period and obesity	198	Ovaries, skin-hair, period and obesity	95	Ovaries, infertility, depression and pain	20
Five symptoms	9	Skin, depression, hormone, period and obesity	176	Skin-hair, infertility, period, obesity and skin	64	Head-hair, infertility, hormone, pain and depression	18
Six symptoms	6	Skin, depression, skin-hair, head-hair, period and obesity	89	Skin obesity, hormone ovaries, head-hair period	46	Depression, ovaries, hormone, pain, infertility and obesity	11

The same permutations of symptoms are regarded as a single combination of symptoms. For example, the permutation of three symptoms Period, Obesity and Skin are PeriodObesitySkin, PeriodSkinObesity, ObesityPeriodSkin, ObesitySkinPeriod, SkinPeriodObesity, SkinObesityPeriod and considered as a single symptom combination as PeriodObesitySkin. The same rule is followed in any combination of symptoms. Figure 5 shows sample web plot the proportion of women who have each symptom. According to this finding, a large number of users have Period-related disorders, a medium portion deal with Depression and Head-Hair problems, and a small number of users have Pain. The same plot is done for two to six symptom combinations which is represented in Table 5. According to these findings, the majority of women who have Period, Obesity, SkinHair, and Skin problems are followed by those who have Hormone, Depression, and HeadHair problems, while fewer women experience Infertility, Ovarian problems, and Pain.

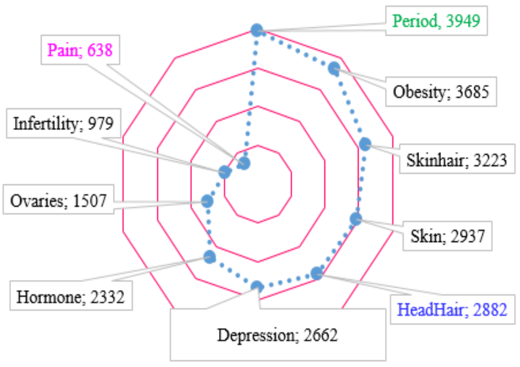


Figure 5. Web plot for single symptom.

4.3. Association Rule Mining

Women with PCOS are not only affected by just one symptom, and the state of their body dictates which of the many symptoms they will suffer during a short period. In this regard, this work identifies the frequent symptom sets and create association rules of symptoms to infer the symptom pattern of PCOS. The frequent symptom sets and association rules are produced using the Apriori algorithm using support, confidence, and lift measures in accordance with the topic modelling symptoms results. Minimum Support value is first established for creating frequent symptom sets, and then fix the confidence and lift metric for generating and choosing significant association rules for the produced symptom sets. The effective rules are selected based on confidence and lift quality measures. Table 6 describes the number of frequent symptoms set and association rules for different support and confidence values.

Table 6. Frequent symptom sets and rule sets for ARM.

Min. support	**No. of frequent symptom set**	**No. of association rules**									
		Confidence metric									
		0.05	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**
0.005	181	1089	843	663	533	439	369	292	229	177	138
0.01	112	521	409	333	274	216	183	139	107	74	51
0.015	85	337	259	214	180	141	119	91	68	42	24
0.02	**59**	174	**158**	118	98	87	72	52	37	22	11
0.025	49	130	128	96	77	68	59	41	30	16	7
0.03	44	104	104	83	64	56	47	32	22	14	5
0.035	38	80	80	69	53	46	37	23	15	10	4
0.04	32	56	56	51	42	35	26	14	7	6	1

Table 6 results lead us to fix that a minimal support value of 0.02 and a confidence value of 0.1 are appropriate since these values provided an adequate number of the frequent symptoms set and association rules in comparison to alternative metric values. For example, 59 frequent symptom sets and 158 association rules are generated for minimum support 0.02 and confidence 0.1. Additionally, it offers a useful criterion for identifying unique patterns to each PCOS symptom. Table 7 lists the top five significant rules for each subsequent symptom based on the antecedent symptoms, along with the support, confidence, and lift ratings for each rule. All top five rules of each consequent symptom are considered as the individual rule which means one rule of the consequent is not depending on another rule of consequent. At the same time, one issue with those results is that the sub-rules are formed with same support, confidence, and lift values in single consequent for example in Table 7 periods consequent rule 2 {‘Depression’, ‘Hormone’, ‘Obesity’}is a subset of Rule 1 {‘Depression’, ‘Hormone’, ‘Ovaries’, ‘Obesity’}. This issue can be solved by selecting the sub-rules with other measures like leverage, conviction and applying the hybrid optimization-based approach for avoiding the sub-rules with same metrics which would be our future contribution.

4.4. Rule Validation and PCOS Severity

ARM is the important technique for discovering the patterns and relationships in huge datasets. All rules are validated and the top rules are selected by Gynaecologist and who suggested the PCOS unique symptom patterns for each symptom and the severity of PCOS according to the rules. When the size of the dataset and the number of rules is high, the computational cost and time of rule evaluation may grow. In this work, 158 rules are developed using minimal support and confidence values of 0.02 and 0.1 in ten separate consequent symptoms. Thus, optimize the performance and scalability of all rules can be improved by choosing the other metrics like lift and pruning the frequent symptoms set. There are three categories of PCOS severity: Insulin Resistance PCOS (INS), Inflammatory PCOS (INF), and Adrenal PCOS (ADR). The term Insulin Resistance PCOS refers to a condition in which the body’s own insulin is not properly utilized, leading to diabetes and the accompanying symptoms of Hormone Imbalance, Obesity, HeadHair and SkinHair. Inflammatory PCOS forms the chronic inflammation, causes the ovaries to generate excess testosterone and also affect the ovulation. This PCOS includes the symptoms such as irregular periods, ovaries problem, infertility, skin problems and pain. Adrenal PCOS is caused due to abnormal depression and pains, which includes depression and pain symptoms. Figure 6 describes the web chart for PCOS unique pattern of consequent symptoms based on ante-cedent symptoms and severity of the PCOS based on both antecedent and consequent symptoms. According to this chart, each consequent symptom has its own distinctive Antecedent symptom pattern as follows:

1. Hormone ← Obesity, Period, Depression.
2. Obesity ← Hormone, Period, Ovaries.
3. Period← Hormone, Obesity, Ovaries, Depression.
4. SkinHair ← Hormone, Obesity, Period, Skin.
5. Skin ← Hormone, Obesity, Period, SkinHair.
6. Ovaries, HeadHair ← Hormone, Depression.
7. Infertility ← Hormone, Obesity, Period, Ovaries.
8. Depression ← Period, Infertility, Pain.
9. Pain ← Hormone, Obesity, Period.

Table 7. Top five significant rules for each consequent symptom.

Rules	Antecedents	Consequents	Support	Confidence	Lift
SkinHair					
Rule 1	{‘Hormone’, ‘Obesity’, ‘Period’, ‘Infertility’}	{‘SkinHair’}	0.002	1	3.91
Rule 2	{‘Hormone’, ‘Obesity’, ‘Ovaries’, ‘Infertility’}	{‘SkinHair’}	0.001	1	3.91
Rule 3	{‘Hormone’, ‘Obesity’, ‘Ovaries’, ‘Skin’}	{‘SkinHair’}	0.003	1	3.30
Rule 4	{‘Hormone’, ‘Infertility’, ‘Skin’}	{‘SkinHair’}	0.002	1	3.12
Rule 5	{‘Obesity’, ‘Period’, ‘Skin’}	{‘SkinHair’}	0.008	0.75	2.79
Hormone problem					
Rule 1	{‘Depression’, ‘Obesity’, ‘Period’}	{‘Hormone’}	0.002	1	5.34
Rule 2	{‘Depression’, ‘Period’}	{‘Hormone’}	0.001	0.5	2.67
Rule 3	{‘Obesity’, ‘Period’}	{‘Hormone’}	0.009	0.30	1.58
Rule 4	{‘Period’}	{‘Hormone’}	0.004	0.15	1
Rule 5	{‘Obesity’}	{‘Hormone’}	0.003	0.12	1
Periods problem					
Rule 1	{‘Depression’, ‘Hormone’, ‘Ovaries’, ‘Obesity’}	{‘Period’}	0.001	1	3.15
Rule 2	{‘Depression’, ‘Hormone’, ‘Obesity’}	{‘Period’}	0.001	1	3.15
Rule 3	{‘Ovaries’, ‘Obesity’}	{‘Period’}	0.002	0.77	2.45
Rule 4	{‘Hormone’, ‘Obesity’}	{‘Period’}	0.001	0.32	1.13
Rule 5	{‘Obesity’}	{‘Period’}	0.003	0.35	1.01
Obesity					
Rule 1	{‘Hormone’, ‘Ovaries’, ‘Period’}	{‘Obesity’}	0.001	1	3.39
Rule 2	{‘Ovaries’, ‘Period’,}	{‘Obesity’}	0.001	0.37	1.26
Rule 3	{‘Hormone’, ‘Period’}	{‘Obesity’}	0.001	0.36	1.23
Rule 4	{‘Hormone’, ‘Ovaries’}	{‘Obesity’}	0.009	0.33	1.13
Rule 5	{‘Period’}	{‘Obesity’}	0.005	0.29	1
Ovaries problem					
Rule 1	{‘Depression’, ‘Hormone’, ‘Obesity’}	{‘Ovaries’}	0.001	1	8.27
Rule 2	{‘Depression’, ‘Hormone’}	{‘Ovaries’}	0.001	0.5	4.13
Rule 3	{‘Obesity’}	{‘Ovaries’}	0.001	0.175	1.45
Rule 4	{‘Hormone’}	{‘Ovaries’}	0.003	0.16	1.32
Rule 5	{‘Depression’}	{‘Ovaries’}	0.003	0.12	1
Infertility					
Rule 1	{‘Hormone’, ‘Obesity’, ‘Obesity’, ‘Period’}	{‘Infertility’}	0.001	1	12.73
Rule 2	{‘Obesity’, ‘Obesity’, ‘Period’}	{‘Infertility’}	0.001	1	10.25
Rule 3	{‘Obesity’, ‘Period’}	{‘Infertility’}	0.001	0.33	4.24
Rule 4	{‘Obesity’, ‘Period’}	{‘Infertility’}	0.007	0.17	2.12
Rule 5	{‘Hormone’, ‘Obesity’}	{‘Infertility’}	0.001	0.12	1.62
Skin problems					
Rule 1	{‘SkinHair’, ‘Hormone’, ‘Obesity’, ‘Period’}	{‘Skin’}	0.008	1	4.34
Rule 2	{‘Obesity’, ‘Period’, ‘SkinHair'}	{‘Skin’}	0.006	0.78	3.38
Rule 3	{‘Obesity’, ‘Hormone’, ‘Period’}	{‘Skin’}	0.006	0.64	2.76
Rule 4	{‘Period’, ‘Hormone’, ‘SkinHair’}	{‘Skin’}	0.009	0.30	1.32
Rule 5	{‘Obesity’}	{‘Skin’}	0.007	0.24	1.06
Head hair problems					
Rule 1	{‘Depression’, ‘Hormone’, ‘Obesity’}	{‘HeadHair’}	0.001	1	4.70
Rule 2	{‘Depression’, ‘Hormone’, ‘Period’}	{‘HeadHair’}	0.004	0.83	3.92
Rule 3	{‘Depression’, ‘Hormone’}	{‘HeadHair’}	0.002	0.57	2.69
Rule 4	{‘Depression’}	{‘HeadHair’}	0.002	0.375	1.76
Rule 5	{‘Hormone’}	{‘HeadHair’}	0.003	0.19	1
Depression					
Rule 1	{‘Period’, ‘Pain’, ‘Infertility’}	{‘Depression’}	0.001	1	4.24
Rule 2	{‘Obesity’, ‘Period’, ‘Infertility’}	{‘Depression’}	0.001	1	3.03
Rule 3	{‘Obesity’, ‘Period’, ‘Pain’}	{‘Depression’}	0.004	0.6	2.54
Rule 4	{‘Period’, ‘Pain’}	{‘Depression’}	0.004	0.36	1.52
Rule 5	{‘Period’, ‘Infertility’}	{‘Depression’}	0.007	0.22	1
Pain					
Rule 1	{‘Hormone’, ‘Obesity’, ‘Period’}	{‘Pain’}	0.001	1	5.53
Rule 2	{‘Hormone’, ‘Period’}	{‘Pain’}	0.002	0.5	3.66
Rule 3	{‘Obesity’, ‘Period’}	{‘Pain’}	0.004	0.4	2.08
Rule 4	{‘Period’}	{‘Pain’}	0.003	0.2	2

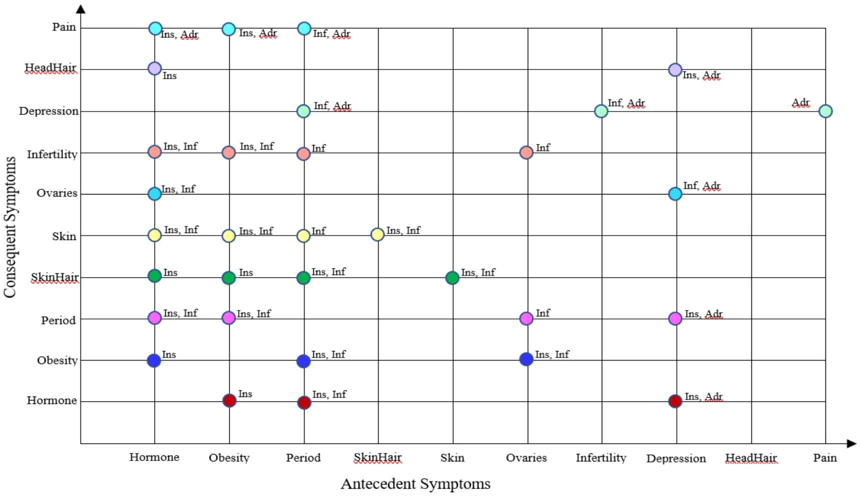


Figure 6. Web chart for severity of PCOS.

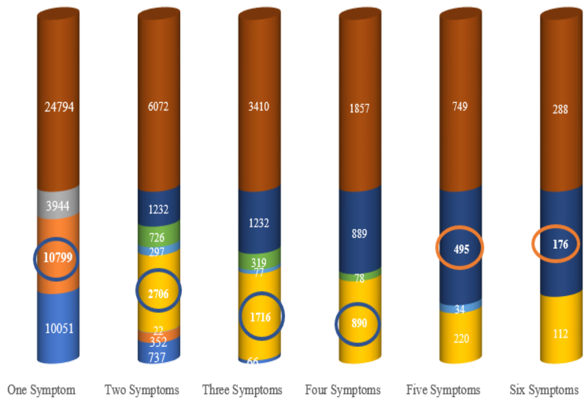


Figure 7. Symptom combinations vs. severity of PCOS.

The same process is applied for all other consequent symptoms and the severity of PCOS is calculated by all 24794 Posts from Reddit users based on symptom combinations which is represented in Figure 7. According to this statistic, Inflammatory PCOS in one symptom, insulin resistance and inflammatory PCOS in combinations of two, three, and four symptoms, and all insulin resistance, inflammatory and adrenal PCOS in combinations of five and six symptoms were relatively prevalent among Reddit users. This shows that women have high severity of PCOS if multiple symptoms are present, as opposed to PCOS that is primarily influenced by insulin resistance or inflammation. Women are also less impacted by adrenal PCOS on an individual basis, although it is more affected in women who also have insulin resistance, inflammatory or both type of PCOS.

Table 8 compares the proposed work with the state-of-art methods for various diseases with different risk factors and different topic modelling technique. As seen in Table 8, the suggested work is implemented using segmentation and grouping relevant symptoms of PCOS as opposed to the built-in topic modelling methodologies that were used for all previous models. While all previous approaches identified a variety of illness risk factors, our proposed work solely considers PCOS risk variables, making it simple to extract the association rules needed to create PCOS symptom patterns.

Table 8. Comparison with the topic modelling state-of-art methods for disease analysis through highest probability risk factors.

Authors	Disease Considered	Data set considered	Year	Methods used	# Risk factors found	Risk factors discovered	Highest probability risk factors
Huang *et al*. [21]	Coronary heart disease	Electronic health records	2015	Probabilistic topic model	4	angina, diabetes, hypertension, cholesterol	Diabetes
Ghosh *et al*. [19]	Infectious disease outbreaks from USA, China, India	Media and health agencies	2017	Temporal topic modelling	9	cough, rabies, salmonellosis, e. coli infection, h7n9, hfmd, dengue, add, malaria	Rabies, H7N9, malaria
Zhao *et al*. [69]	Lipoprotein(a)	Electronic health records	2019	NMF	6	pneumonia, general pains, cardio vascular disease, lung cancer, diabetes, renal and liver diseases	Cardio vascular disease
Ki *et al*. [28]	Diabetes	PubMed data	2021	LDA, LSA, pLSA	4	maturity-onset diabetes, gestational diabetes mellitus, risk gene, pancreas	Maturity-onset diabetes
Amara *et al*. [7]	Covid-19	Facebook data	2021	Multilingual -LDA	12	coronavirus, covid, china, health, positive, mask, quarantine, outbreak, death, police, drug, pandemic	Quarantine
Yochum *et al*. [66]	Samut prakan city diseases	Web application	2022	LDA	10	flu, gastroesophageal reflux, common cold, gastritis, diarrhea, appendicitis, pneumonia, pharyngitis, bronchitis, and irritable bowel syndrome	Flu
Proposed work	PolyCystic ovarian syndrome	Reddit data	2023	**SSG_LLDA**	**10**	hormone, obesity, periods, skinhair, skin, ovaries, infertility, depression, headhair, pains	**Periods**

Table 9. Comparison with the ARM state-of-art methods for disease analysis through top rules support, confidence and lift values.

Authors	Disease Considered	Data Set Considered	Year	Risk Factors Considered	# of Rules	Top Rules		
						Support	Confidence	Lift
Nahar *et al*. [41]	Heart disease	UCI cleveland dataset	2013	Chest pain, bp, cholesterol, heart rate, blood sugar	25	-	0.95	-
Ivancevic *et al*. [23]	Early childhood caries	ECC data from preschool children	2015	Male gender, breastfeeding, high birth order, language, low body weight	44	0.07	0.72	2.36
Khare and Gupta [30]	Cardiovascular disease	UCI repository	2016	BP, cholesterol, heart rate, chest pain	26	0.16	0.94	2.12
Sonet *et al*. [59]	Heart diseases	NICVD data	2017	Unstable angina, myocardial infarction, coronary heart disease	9	0.25	0.99	3.66
Tandan *et al*. [63]	Covid-19	Wolfram data repository	2021	Covid symptoms, chronic conditions	107	0.004	1	2.7
Pradeepa *et al*. [51]	PCOS	Media sources	2020	Period, obesity, excess hair, hormone, etc.,	44	0..005	0.71	4.3
Proposed work **(SSG_LLDA + ARM)**	PCOS	Reddit data	2023	Hormone, obesity, periods, skinhair, skin, ovaries, infertility, depression, headhair, pains	**50**	**0.009**	**1**	**12.73**

Table 9 compares the proposed work with the state-of-art methods for various diseases with different risk factors and ARM technique. As shown in Table 9, the suggested method uses topic modelling-based AMR, as opposed to other systems that merely built association rules based on risk variables. One of the states of art DEODORANT system was developed by Pradeepa *et al.* [51] was used to generate frequent itemset and form the cluster based on itemset. They didn't create the symptom patterns in accordance with the antecedent and consequent rules, and they didn't develop an integrated technique like topic modelling with AMR. In our suggested method, the data was gathered from social media and developed SSG_LLDA to minimize the dimensionality of the symptom features. The top rules of each symptom set contain support, confidence, and lift measures, which are constructed from the reduced feature sets to create the association rules and symptom patterns. Finally, the suggested approach achieves good levels of support, confidence and lift for top 50 rules of all ten risk factors.

5. Conclusions and Future Work

In the modern world, PCOS affects the majority of adolescent girls and young women as a result of their food and lifestyle. Earlier, this disease was unknown, but now medical professionals and PCOS sufferers have raised awareness of it via social media. People become more knowledgeable about the numerous PCOS symptoms and causes by reading posts on social media instead of seeing the doctors in person. Therefore, this proposed work helps in recognizing the various symptoms of PCOS, understanding how one symptom might trigger another, and determining the severity of PCOS for a woman using social media posts.

This aim is achieved by implementing the sequence of steps starting from symptom collection up to ARM and finally forming the unique symptom pattern and finding the severity of PCOS from associated symptom rules. The SSG_LLDA result gives various symptom combinations along with the most and least prevalent impacted symptoms for each combination. Apriori produces the top five effective PCOS symptom rules with a minimum support value of 0.02 and a confidence value of 0.1. Based on symptom combinations and association rules, the severity of PCOS is determined initially for 1500 Reddit posts between January 2020 and April 2021 and has subsequently expanded to 17,000 posts between January 2020 and April 2022, and later eventually handled 24794 posts between January 2020 and September 2022. These findings show that women with insulin resistance and inflammatory PCOS are more affected than women with adrenal PCOS. Thus, the proposed work is well suitable when the data is added more or updated. The proposed approach is then contrasted with several state-of-the-art approaches for various diseases to highlight its special features. Apriori with SSG_LLDA has taken the less time to generate the frequent symptom sets and association rules and also produces best outcome for determining symptom patterns and consequences of PCOS based on Social Media posts.

In the future, we expand our work to include BERT-based SSs and grouping rather than LDA and evaluate the model's effectiveness. Using a variety of ARM algorithms, including FP Growth and Apriori, the symptom patterns will be created from BERT topics and comparing the execution times of each approach. Furthermore, a variety of optimisation methods will be employed to optimise association rules. Therefore, in subsequent work, we use optimisation approaches to obtain high-quality findings.

Acknowledgment

The authors thank the Mepco Schlenk Engineering college management for their support in carrying out the research work.

References

[1] Akram W. and Kuma R., "A Study on Positive and Negative Effects of Social Media on Society," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 10, pp. 347-354, 2017. https://doi.org/10.26438/ijcse/v5i10.351354

[2] Alamoudi A., Khan I., Aslam N., Alqahtani N., Alsaif H., Al Dandan M., Al Gadeeb M., and Al Bahrani R., "A Deep Learning Fusion Approach to Diagnosis the Polycystic Ovary Syndrome," *Applied Computational Intelligence and Soft Computing*, vol. 2023, pp. 1-15, 2023. https://doi.org/10.1155/2023/9686697

[3] Alessa A., Faezipour M., and Alhassan Z., "Text Classification of Flu-related Tweets Using FastText with Sentiment and Keyword Features," *in Proceedings of the IEEE International Conference on Healthcare Informatics*, New York, pp. 366-367, 2018. DOI:10.1109/ICHI.2018.00058

[4] Alga A, Eriksson O., and Nordberg M., "Analysis of Scientific Publications during the early Phase of the COVID-19 Pandemic: Topic Modeling Study," *Journal of Medical Internet Research*, vol. 22, no. 11, pp. 1-11, 2020. https://www.jmir.org/2020/11/e21559/

[5] Alkouz B. and Al Aghbari Z., "Analysis and Prediction of Influenza in the UAE based on Arabic Tweets," *in Proceedings of the IEEE 3rd International Conference on Big Data Analysis*, Shanghai, pp. 61-66, 2018. DOI:10.1109/ICBDA.2018.8367652

[6] Alkouz B., Al Aghbari Z., and Abawajy J., "Tweetluenza: Predicting Flu Trends from Twitter Data," *IEEE Transactions on Big Data Mining and Analytics*, vol. 2, no. 4, pp. 273-287, 2019. DOI:10.26599/BDMA.2019.9020012

[7] Amara A., Taieb M., and Ben Aouicha M., "Multilingual Topic Modeling for Tracking COVID-19 Trends Based on Facebook Data Analysis," *Applied Intelligence*, vol. 51, no. 5, pp. 3052-3073, 2021. https://doi.org/10.1007/s10489-020-02033-3

[8] Amin S., Irfan Uddin M., Zeb M., Alarood A., Mahmoud M., and Alkinani M., "Detecting Dengue/Flu Infections Based on Tweets Using LSTM and Word Embedding," *IEEE Access*, vol. 8, pp. 189054-189068, 2020. DOI:10.1109/ACCESS.2020.3031174

[9] Charalambous A., "Social Media and Health Policy," *Asia-Pacific Journal of Oncology Nursing*, vol. 6, no. 1, pp. 24-27, 2019. DOI:10.4103/apjon.apjon_60_18

[10] Chee C., Jaafar J., Aziz I., Hasan M., and Yeoh W., "Algorithms for Frequent Itemset Mining: A Literature Review," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 2603-2621, 2019. https://link.springer.com/article/10.1007/s10462-018-9629-z

[11] Dahmani D., Rahal S., and Belalem G., "A New Approach to Improve Association Rules for Big Data in Cloud Environment," *The International Arab Journal of Information Technology*, vol. 16, no. 6, pp. 1013-1020, 2019. https://www.iajit.org/portal/PDF/November%202019,%20No.%206/13038.pdf

[12] Darby L., "The Stein-Leventhal Syndrome: A Curable Form of Sterility" (1958), by Irving Freiler Stein Sr.," *Embryo Project Encyclopedia*, 2017. https://hdl.handle.net/10776/11884

[13] Denny A., Raj A., Ashok A., Ram C., and George R., "i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," *in Proceedings of the IEEE Region 10th Conference*, Kochi, pp. 673-678, 2019. DOI:10.1109/TENCON.2019.8929674

[14] Dogan S. and Turkoglu I., "Diagnosing Hyperlipidemia Using Association Rules," *Mathematical and Computational Applications*, vol. 13, no. 3, pp. 193-202, 2008. https://www.mdpi.com/2297-8747/13/3/193

[15] Domadiya N. and Rao U., "Privacy-Preserving Association Rule Mining for Horizontally Partitioned Healthcare Data: A Case Study on the Heart Diseases," *Sadhana*, vol. 43, no. 127, pp. 1-9, 2018. https://doi.org/10.1007/s12046-018-0916-9

[16] Elmannai H., El-Rashidy N., Mashal I., Alohali M., Farag S., El-Sappagh S., and Saleh H., "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," *Diagnostics*, vol. 13, no. 8, pp. 1-21, 2023. DOI: 10.3390/diagnostics13081506

[17] Gancho S., "Social Media: A Literature Review," *e-Revista LOGO*, vol. 6, no. 2, pp. 1-20, 2017. DOI:10.26771/e-Revista.LOGO/2017.2.01

[18] Garbhapu V. and Bodapati P., "A Comparative Analysis of Latent Semantic Analysis and Latent Dirichlet Allocation Topic Modeling Methods Using Bible Data," *Indian Journal of Science and Technology*, vol. 13, no. 44, pp. 4474-4482, 2020. DOI:10.17485/IJST/v13i44.1479

[19] Ghosh S., Chakraborty P., Nsoesie E., Cohn E., Mekaru S., Brownstein J., and Ramakrishnan N., "Temporal Topic Modeling to Assess Associations between News Trends and Infectious Disease Outbreaks," *Scientific Reports*, vol. 7, pp. 1-12, 2017. https://doi.org/10.1038/srep40841

[20] How to Use Social Media in Healthcare: A Guide for Health Professionals, https://blog.hootsuite.com/social-media-health-care/, Last Visited, 2024.

[21] Huang Z., Dong W., and Duan H., "A Probabilistic Topic Model for Clinical Risk Stratification from Electronic Health Records," *Journal of Biomedical Informatics*, vol. 58, pp. 28-36, 2015. https://doi.org/10.1016/j.jbi.2015.09.005

[22] India Social Media Statistics, https://www.theglobalstatistics.com/india-social-media-statistics/, Last Visited, 2024.

[23] Ivancevic V., Tusek I., Tusek J., Knezevic M., Elheshk S., and Lukovic I., "Using Association Rule Mining to Identify Risk Factors for early Childhood Caries," *Computer Methods and Programs in Biomedicine*, vol. 122, no. 2, pp. 175-181, 2015. DOI: 10.1016/j.cmpb.2015.07.008

[24] Jafar A., Fakhr M., and Farouk M., "Enhanced Clustering-based Topic Identification of Transcribed Arabic Broadcast News," *The International Arab Journal of Information Technology*, vol. 14, no. 5, pp. 721-728, 2017. https://iajit.org/PDF/vol%2014,%20no.%205%20sep/9013.pdf

[25] Kamalesh M., Prasanna K., Bharathi B., Dhanalakshmi R., and Canessane R., *Predicting the Risk of Diabetes Mellitus to Subpopulations Using Association Rule Mining*, Springer, 2016. https://link.springer.com/chapter/10.1007/978-81-322-2671-0_6

[26] Kaplan A. and Haenlein M., "Users of the World, Unite! The Challenges and Opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59-68, 2010. https://doi.org/10.1016/j.bushor.2009.09.003

[27] Kelaiaia A. and Merouani H., "Clustering with Probabilistic Topic Models on Arabic Texts: A Comparative Study of LDA and K-Means," *The International Arab Journal of Information Technology*, vol. 13, no. 2, pp. 332-338, 2016. https://www.iajit.org/portal/PDF/Vol.13,%20No.2/6146.pdf

[28] Ki C., Hosseinian-Far A., Daneshkhah A., and Salari N., "Topic Modelling in Precision Medicine with its Applications in Personalized Diabetes Management," *Expert Systems*, vol. 39, no. 4, pp. 1-21, 2021. https://doi.org/10.1111/exsy.12774

[29] Khanna V., Chadaga K., Sampathila N., Prabhu S., Bhandage V., and Hegde G., "A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome," *Applied System Innovation*, vol. 6, no. 2, pp. 1-26, 2023. https://doi.org/10.3390/asi6020032

[30] Khare S. and Gupta D., "Association Rule Analysis in Cardiovascular Disease," *in Proceedings of the 2nd International Conference on Cognitive Computing and Information Processing*, Mysur, pp. 1-6, 2016. DOI: 10.1109/CCIP.2016.7802881

[31] Kumar K. and Arumugaperumal S., "Association Rule Mining and Medical Application: A Detailed Survey," *International Journal of Computer Applications*, vol. 80, no. 17, pp. 10-19, 2013. DOI:10.5120/13967-1698

[32] Lakshmi Hospital, https://lakshmifertilitycentre.com/hormone-analysis/#/, Last Visited, 2024.

[33] Lau A., Ong S., Mahidadia A., Hoffmann A., Westbrook J., and Zrimec T., "Mining Patterns of Dyspepsia Symptoms across Time Points Using Constraint Association Rules," *in Proceedings of the Pacific-7th Asia Conference on Advances in Knowledge Discovery and Data Mining*, Seoul, pp. 124-135, 2003. https://doi.org/10.1007/3-540-36175-8_13

[34] Liu J., Wu Q., Hao Y., Jiao M., Wang X., Jiang S., and Han L., "Measuring the Global Disease Burden of Polycystic Ovary Syndrome in 194 Countries: Global Burden of Disease Study 2017," *Human Reproduction*, vol. 36, no. 4, pp. 1108-111, 2021. DOI:10.1093/humrep/deaa371

[35] Liu L., Tang L., Dong W., Yao S., and Zhou W., "An Overview of Topic Modelling and its current Applications in Bioinformatics," *Springer Plus*, vol. 5, no. 1608, pp. 1-22, 2016. https://doi.org/10.1186/s40064-016-3252-8

[36] Lossio-Ventura J., Gonzales S., Morzan J., Alatrista-Salas H., Hernandez-Boussard T., and Bian J., "Evaluation of Clustering and Topic Modeling Methods over Health-Related Tweets and Emails," *Artificial Intelligence in Medicine*, vol. 117, pp. 102096, 2021. https://doi.org/10.1016/j.artmed.2021.102096

[37] Madila S., Dida M., and Kaijage S., "A Review of Usage and Applications of Social Media Analytics," *Journal of Information Systems Engineering and Management*, vol. 6, no. 3, pp. 1-10, 2021. http://repository.mocu.ac.tz/xmlui/handle/123456789/583

[38] McCormick T., Rudin C., and Madigan D., "A Hierarchical Model for Association Rule Mining of Sequential Events: An Approach to Automated Medical Symptom Prediction," *Annals of Applied Statistics*, vol. 1, pp. 1-19, 2011. DOI:10.2139/ssrn.1736062

[39] Mohammed S. and Al-Augby S., "LSA and LDA Topic Modeling Classification: Comparison Study on E-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353-362, 2020. http://doi.org/10.11591/ijeecs.v19.i1.pp353-362

[40] Muliono R., Muhathir., Khairina N., and Harahap M., "Analysis of Frequent Itemsets Mining Algorithm against Models of Different Datasets," *in Proceedings of the 1st International Conference of SNIKOM*, Medan, pp. 1-9, 2019. DOI:10.1088/1742-6596/1361/1/012036

[41] Nahar J., Imam T., Tickle K., and Chen Y., "Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1086-1093, 2013. https://doi.org/10.1016/j.eswa.2012.08.028

[42] Nandhini M., Rajalakshmi M., and Sivanandam S., "Performance Analysis of Predictive Association Rule Classifiers Using Healthcare Datasets," *IETE Technical Review*, vol. 39, no. 1, pp. 143-156, 2022. https://doi.org/10.1080/02564602.2020.1827988

[43] Nguyen D., Luo W., Phung D., and Venkatesh S., "LTARM: A Novel Temporal Association Rule Mining Method to Understand Toxicities in a Routine Cancer Treatment," *Knowledge-Based Systems*, vol. 161, pp. 313-328, 2018. https://doi.org/10.1016/j.knosys.2018.07.031

[44] Nsugbe E., "An Artificial Intelligence-based Decision Support System for early Diagnosis of Polycystic Ovaries Syndrome," *Healthcare Analytics*, vol. 3, no. 2, pp. 1-7, 2023. https://doi.org/10.1016/j.health.2023.100164

[45] Patil S. and Kumaraswamy Y., "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," *International Journal of Computer Science and Network Security*, vol. 9, no. 2, pp. 228-235, 2009. http://paper.ijcsns.org/07_book/200902/20090230.pdf

[46] PCOS Comments in Reddit, https://www.reddit.com/r/PCOS/comments/, Last Visited, 2024.

[47] PCOS in Facebook, https://www.facebook.com/search/top?q=pcos, Last Visited, 2024.

[48] PCOS in Reddit, https://www.reddit.com/r/PCOS/, Last Visited, 2024.

[49] PCOS in Tumblr, https://www.tumblr.com/search/pcos, Last Visited, 2024.

[50] PCOS in Twitter, https://twitter.com/search?q=pcos&src=typed_query&f=user, Last Visited, 2024.

[51] Pradeepa S., Geetha K., Kannan K., and Manjula

K., "DEODORANT: A Novel Approach for early Detection and Prevention of Polycystic Ovary Syndrome Using Association Rule in Hypergraph with the Dominating Set Property," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 5421-5437 2023. https://link.springer.com/article/10.1007/s12652-020-01990-4

[52] Quwaider M. and Alfaqeeh M., "Social Networks Benchmark Dataset for Diseases Classification," *in Proceedings of the 4th International Conference on Future Internet of Things and Cloud Workshops*, Vienna, pp. 234-239, 2016. DOI:10.1109/W-FiCloud.2016.56

[53] Ramasamy S. and Nirmala K., "Disease Prediction in Data Mining Using Association Rule Mining and Keyword-based Clustering Algorithms," *International Journal of Computers and Applications*, vol. 42, no. 1, pp. 1-8, 2017. https://www.tandfonline.com/doi/pdf/10.1080/1206212X.2017.1396415

[54] Rani R., Hajam Y., Kumar R., Bhat R., Rai S., and Rather M., "A Landscape Analysis of the Potential Role of Polyphenols for the Treatment of Polycystic Ovarian Syndrome," *Phytomedicine Plus*, vol. 2, no. 1, pp. 1-21, 2021. https://doi.org/10.1016/j.phyplu.2021.100161

[55] Sahatiya P., "Big Data Analytics on Social Media Data: A Literature Review," *International Research Journal of Engineering and Technology*, vol. 5, no. 2 pp. 189-192, 2018. https://www.irjet.net/archives/V5/i2/IRJET-V5I245.pdf

[56] Sapountzi A. and Psannis K., *Principles of Data Science*, Springer, 2020. https://doi.org/10.1007/978-3-030-43981-1_4

[57] Shi L., Du J., and Kou F., "A Sparse Topic Model for Bursty Topic Discovery in Social Networks," *The International Arab Journal of Information Technology*, vol. 17, no. 5, pp. 816-824, 2020. https://iajit.org/PDF/September%202020,%20No.%205/16576.pdf

[58] Smailhodzic E., Hooijsma W., Boonstra A., and Langley D., "Social Media Use in Healthcare: A Systematic Review of Effects on Patients and on their Relationship with Healthcare Professionals," *BMC Health Services Research*, vol. 16, no. 442, pp. 1-14, 2016. https://doi.org/10.1186/s12913-016-1691-0

[59] Sonet K., Rahman M., Mazumder P., Reza A., and Rahman R., "Analyzing Patterns of Numerously Occurring Heart Diseases Using Association Rule Mining," *in Proceedings of the 12th International Conference on Digital Information Management*, Fukuoka, pp. 38-45, 2017. DOI:10.1109/ICDIM.2017.8244690

[60] Soni P. and Vashisht S., "Image Segmentation for Detecting Polycystic Ovarian Disease using Deep Neural Networks," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 3, pp. 534-537, 2019. https://doi.org/10.26438/ijcse/v7i3.534537

[61] Stieglitza S., Mirbabaiea M., Rossa B., and Neuberger C., "Social Media Analytics-Challenges in Topic Discovery, Data Collection, and Data Preparation," *International Journal of Information Management*, vol. 39, pp. 156-168, 2018. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

[62] Tabassum S., Pereira F., Fernandes S., and Gama J., "Social Network Analysis: An Overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 5, pp. 1-30, 2018. https://doi.org/10.1002/widm.1256

[63] Tandan M., Acharya Y., Pokharel S., and Timilsina M., "Discovering Symptom Patterns of COVID-19 Patients Using Association Rule Mining," *Computers in Biology and Medicine*, vol. 131, pp. 1-12, 2021. https://doi.org/10.1016/j.compbiomed.2021.104249

[64] Tiwari S., Kane L., Koundal D., Jain A., Alhudhaif A., Pola K., Zaguia A., Alenezi F., and Althubiti S., "SPOSDS: A Smart Polycystic Ovary Syndrome Diagnostic System Using Machine Learning," *Expert Systems with Applications*, vol. 203, pp. 117592, 2022. https://doi.org/10.1016/j.eswa.2022.117592

[65] Ventola C., "Social Media and Health Care Professionals: Benefits, Risks, and Best Practices," *Pharmacy and Therapeutics*, vol. 39, no. 7, pp. 491-499, 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103576/

[66] Yochum P., Nisamaneewong P., Karnchanapimonkul P., and Chomanan P., "Automated Disease Detection Based on Clinical Text Using Topic Modeling," *in Proceedings of the 10th International Conference on Information Technology: IoT and Smart City*, Shanghai, pp. 74-79, 2022. https://doi.org/10.1145/3582197.3582209

[67] Zhang F., Luo J., Li C., Wang X., and Zhao Z., "Detecting and Analyzing Influenza Epidemics with Social Media in China," *in Proceedings of the 18th Pacific-Asia Conference: Lecture Notes in Computer Science*, Tainan, pp. 90-101, 2014. https://link.springer.com/chapter/10.1007/978-3-319-06608-0_8

[68] Zhang X., Saleh H., Younis E., Sahal R., and Ali A., "Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System," *Complexity*, vol. 2020, pp. 1-10, 2020. https://doi.org/10.1155/2020/6688912

[69] Zhao J., Feng Q., Wu P., Warner J., Denny J., and Wei W., "Using Topic Modeling via Non-Negative Matrix Factorization to Identify

Relationships between Genetic Variants and Disease Phenotypes: A Case Study of Lipoprotein(a) (LPA),” *PLoS One*, vol. 14, no. 2, pp. 1-15, 2019. https://doi.org/10.1371/journal.pone.0212112

Santhi Selvaraj received her post graduate from Anna University, Chennai, Tamil Nadu, India. She is currently working as Assistant Professor, Selection Grade with the Department of Computer Science and Engineering. Her current research interest includes Text Mining, Sentiment Analysis, and Recommendation Systems.

Selva Nidhyananthan Sundaradhas received his PhD Degree from Anna University, Chennai, Tamil Nadu, India. He is currently working as Associate Professor, Senior Grade with the Department of Electronics and Communication Engineering. His current research interest includes Signal, Speech Processing, Image Processing, and IoT.