AN IMPROVED TOPIC MODELLING FRAMEWORK FOR DISCOVERING DOMINANT PCOS SYMPTOM FROM REDDIT POSTS

Santhi Selvaraj[1], Selva Nidhyananthan Sundaradhas[2] and Umakanth Nagendran[3]
[1,3]Department of Computer Science and Engineering, Mepco Schlenk Engineering College, India
[2]Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, India

Abstract

Nowadays social media plays a vital role in health care applications. A disorder known as PolyCystic Ovarian Syndrome (PCOS) affects females who are capable of reproducing between the ages of 15 and 35. The symptoms of PCOS are hormonal issues, irregular periods, weight gain, ovaries follicles, infertility, excessive hair growth in skin, hair loss, acne, pimples and dark scars in skin and depression. The main scope of this proposed work is to discover the dominant PCOS symptom based on current symptoms given by the Reddit users. The collected unstructured data from Reddit users are pre-processed and PCOS symptoms are extracted using Bag of Words and TF-IDF. A novel and improved topic modelling methods called Symptom Segmentation and Grouping (SSG) of Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation and BERTopic is designed to reduce the dimensionality of the features and map the sub symptoms of social media users into the head symptoms of Gynecologists. Finally, using maximum likelihood probabilities of these algorithms, the dominant head and sub symptoms are identified within the less time compared to traditional algorithms. Periods issues achieved the highest probabilities and dominant symptom with the value of 0.706 rather than other symptoms.

Keywords:

PCOS, Bag of Words, TF-IDF, Social Media, LDA, LSA, BERT

1. INTRODUCTION

The term social media refers to a grouping of web-based programmes that together form Web 2.0, the conceptual and technical frameworks that allow people to express their ideas, opinions, and thoughts via online communities and virtual networks [1]. Millions of people contribute their information, photos, news, live audio, and videos on various social media platforms [2]. The diseases were categorized from a variety of social media posts and comments using some social benchmark datasets [3]. Through the use of Keyword features and FastText classifier algorithms, flu-related tweets [4] were categorised, allowing relevant and irrelevant posts to be distinguished.

PCOS was initially identified in 1935 by Stein and Leventhal [5] in Chicago. Since then, it has become more common worldwide, with prevalence ranging from 4% to 20%. Numerous techniques in data mining, machine learning, and deep learning are employed to diagnose and forecast PCOS based on medical datasets such as ultrasound pictures [6] and symptom characteristics [7]. Nowadays, the majority of women communicate PCOS-related symptoms, drugs, diets, and hobbies on social media.

In this regard, the proposed system is developed by using modified and improved topic modelling methods to determine the dominant PCOS symptom from social media users. The remaining sections of this paper are outlined as follows: Section 2 describes the related works for the current system. Section 3 describes the architecture and methodologies of proposed system. The outcomes and conclusion of the suggested method are covered in Section 4 and 5.

2. BACKGROUND

SVM was developed by Fang et al. [8] and used to monitor and assess influenza outbreaks in China through social media. The overview of topic modelling and its uses in the field of bioinformatics were provided by Lin Liu et al. [9] utilizing clustering results. Latent Dirichlet Allocation was used by Andreas et al. [10] to evaluate the temporal shifts in study topics in scientific publications during the COVID-19 pandemic.

Juan et al. [11] assessed the various topic modelling and clustering approaches for emails and tweets pertaining to health. According to a comparison analysis of LSA and LDA in Bible data conducted by Vasantha et al. [12], LDA performs 60–70% better than LSA and has a higher coherence score. Shaymaa et al. [13] conducted comparison research on E-Books using LSA and LDA topic modelling categorization.

Various topic modelling algorithms [14-19] were used to identify the highest probability risk factors in Coronary Heart Disease, Infectious Disease, Lipoprotein(a), Diabetes, COVID-19, etc. Gethsiya et al. [20] implemented topic modelling-based sentiment analysis for PCOS from online forum using Naïve Bayes, SVM and LSTM algorithms.

3. PROPOSED METHOD

This system is used to analyze the major symptoms of PCOS from social media users and their posts by mapping the sub symptoms into head symptoms using modified topic modelling approaches which is shown in Fig.1.

3.1 PCOS SYMPTOM COLLECTION AND DATA COLLECTION FROM REDDIT

We consult with a doctor at the Lakshmi Fertility Centre in Sivakasi, Tamil Nadu, before beginning data gathering on social media. Gynecologists have mentioned more important symptoms of PCOS and its severity in today's world. Reddit is the best social media site for healthcare, and the r/PCOS tag offers more details regarding PCOS symptoms in people all over the world. Reddit has 160k users, and between January 2020 and December 2022, we collected 25000 posts from among them. In reality, Reddit users described their symptoms on their own words rather than providing specific symptoms so symptom mapping is essential.

3.2 DATA PRE-PROCESSING

Data pre-processing is a method for transforming unstructured data into meaningful information and maintaining quality of the data. It is done by applying various NLP techniques like URL removal, punctuations removal, lowercase conversion, spell checking, stop words removal, tokenization, stemming, lemmatization and normalization. The dataset is transformed into data frames prior to pre-processing, and the results of each step are stored in a column of data frames.

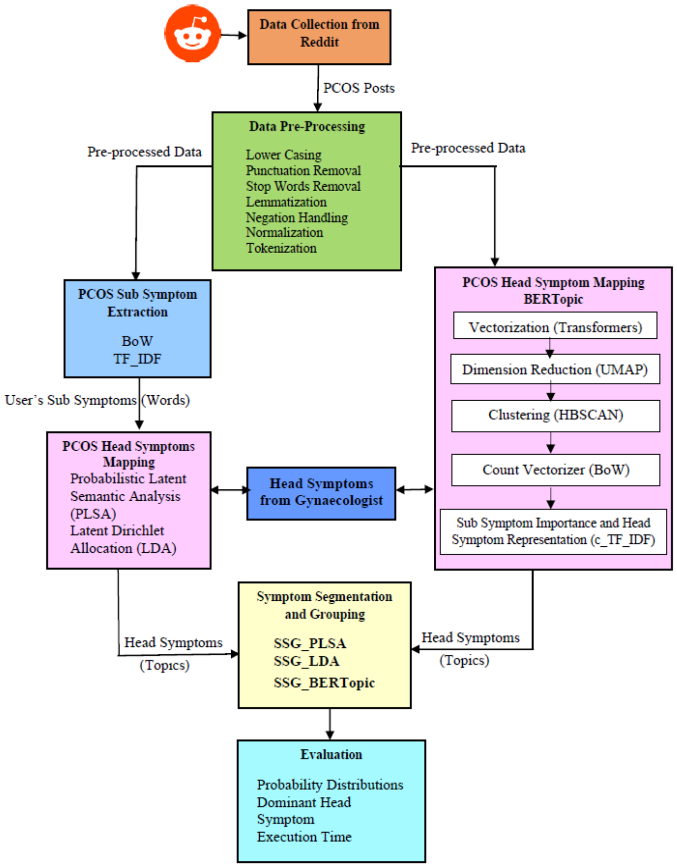
Data Collection from Reddit
PCOS Posts
Data Pre-Processing
Lower Casing
Punctuation Removal
Stop Words Removal
Lemmatization
Negation Handling
Normalization
Tokenization
Pre-processed Data
Pre-processed Data
PCOS Sub Symptom Extraction
BoW
TF_IDF
User's Sub Symptoms (Words)
PCOS Head Symptom Mapping BERTopic
Vectorization (Transformers)
Dimension Reduction (UMAP)
Clustering (HBSCAN)
Count Vectorizer (BoW)
Sub Symptom Importance and Head Symptom Representation (c_TF_IDF)
PCOS Head Symptoms Mapping
Probabilistic Latent Semantic Analysis (PLSA)
Latent Dirichlet Allocation (LDA)
Head Symptoms from Gynaecologist
Head Symptoms (Topics)
Head Symptoms (Topics)
Symptom Segmentation and Grouping
SSG_PLSA
SSG_LDA
SSG_BERTopic
Evaluation
Probability Distributions
Dominant Head Symptom
Execution Time

Fig.1. Proposed System Design

3.3 PCOS SUB SYMPTOM EXTRACTION

The pre-processed textual raw data is converted into numerical form using BoW and TF-IDF extraction to improve results during the topic modelling step. Bag of Words, which stands for count the number of times a word appeared in the document, whereas TF/IDF uses a computing method to ascertain how crucial a word is to the document. A single user may also experience more symptoms, whether they are similar or distinct symptoms. Let consider dataset D is the set of posts $P=\{p_1,p_2,p_3,\cdots,p_n\}$, the unique symptoms are given as the features $S=(s_1,s_2,s_3,\cdots,s_m)$. Each post p_i, $1\leq i\leq n$ is represented by symptom vector. The weight of each symptom, which is taken from the lexicon of the symptoms are represented by a value in the symptom vector and it is denoted as $<w_{i1},w_{i2},w_{i3},\ldots,w_{im}>$, where w_{ij} represents the weight value of symptom s_j in the post p_i. Symptom Extraction by BoW is calculated as follows:

$$w=\begin{cases}1, & \text{if } s \text{ occurs in } p\\ 0, & \text{if } s \text{ not occurs in } p\end{cases} \tag{1}$$

$$\text{BOW}(p_i,s_j,D)=\sum_{i=1}^{n}\sum_{j=1}^{m}\begin{cases}1, & \text{if } w_{ij}=1\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

Symptom Extraction by TF-IDF is calculated as follows:

$$tf(s,p)=\frac{f_{(s,p)}}{w_p} \tag{3}$$

$$idf(s,D)=\frac{|D|}{|p\in D, s\in p|} \tag{4}$$

$$w_{ij}=tf\text{-}idf(s_j,p_i,D)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left[tf(s_j,p_i)\times idf(s_j,D)\right] \tag{5}$$

Each distinct symptom is retrieved from each post using either approach, and the distinct symptom for the entire post or document is then summarized.

3.4 PCOS HEAD SYMPTOMS MAPPING

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. Three of the most important topics modelling techniques, including Probabilistic Latent Semantic Analysis, Latent Dirichlet Algorithm, and BERTopic are used for mapping.

3.4.1 Probabilistic Latent Semantic Analysis:

It is a statistical method for finding the relationship between a collection of posts and the sub symptoms present those posts by obtaining the semantic relationship between those symptoms. Singular value decomposition is used, which entails creating symptoms and posts matrices for the entire dataset. Let consider sub symptoms or terms as S, head symptoms or topics as H and posts as P, then SVD is represented as follows:

$$S\times P=(S\times H)\times(H\times H)\times(H\times P) \tag{6}$$

$$M=L\sum R^{T} \tag{7}$$

where M is Sub Symptoms to each Post mapping matrix, L is a left singular matrix i.e. Sub Symptoms to Head Symptom mapping matrix, $\sum$ is diagonal matrix i.e Head Symptom importance, R^T is a right singular matrix and take the transpose i.e head symptoms to each post mapping.

The LSA of Eq.(6) and Eq.(7) is modified into Probabilistic LSA as follows:

$$P(P,S)=\sum_{H}\left[P(S|H)P(H)P(P|H)\right] \tag{8}$$

$$M=L\sum R^{T}$$

where $P(P,S)$ is the probability of sub symptoms or terms into the posts or documents, $P(S|H)$ is the probability of sub symptoms into the head symptoms or topics, $P(H)$ is the probability of the head symptoms and $P(P|H)$ is the probability of head symptoms into the posts.

3.4.2 Latent Dirichlet Allocation:

It is a statistical generative model for finding the hidden head symptoms present in the post by applying Dirichlet distributions of sub symptoms and head symptoms since posts are mixture of head symptoms and head symptoms are mixture of sub symptoms. LDA provides the better results and visualization of topics compared to pLSA. General LDA for mapping sub symptoms into head symptom and all head symptoms for the entire post as follows:

$$P(S|P)=\sum_{H}\left[P(S|H,P)P(H|P)\right] \tag{9}$$

Apply conditional independence P(S|H, P)=*P*(*S*|*H*) and change the Eq.(9) into Eq.(10) and Eq.(11) as follows:

$$P(S|P)=\sum_{H}\left[P(S|H)P(H|P)\right] \tag{10}$$

$$P(S|P)=\theta_{HP}*\phi_{SH} \tag{11}$$

where θ_{HP} is the probability of head symptom *H* occurred in post *P* and φ_{SH} is the probability of sub symptom *S* occurred in head symptom *H*.

Consider the Dirichlet Distributions of head symptom and sub symptom is given in Eq.(12) and Eq.(13).

$$\theta^{(P)}\sim Dir(\alpha) \tag{12}$$

$$\phi^{(H)}\sim Dir(\beta) \tag{13}$$

where, α and β are sparsity parameter which controls per-post and head symptom distribution and per-head symptom and sub symptom distribution and these values are less than one. The head symptom assignment and sub symptom assignment are given in Eq.(14) and Eq.(15):

$$H_i\sim Multinomial\ \theta^{(P)} \tag{14}$$

$$S_i\sim Multinomial\ \phi^{(H)} \tag{15}$$

The joint distribution of the head symptom assignments *H*, sub symptom assignments *S* for all '*k*' head symptoms φ and topic mixture Θ is given as Eq.(16):

$$P(S,H,\Theta,\phi|\alpha,\beta)=\prod_{i=1}^{k}\left[P(\phi_k|\beta)\prod_{j=1}^{P}P(\theta_j|\alpha)\right]\prod_{t=1}^{N_j}\left[P(H_{(j,t)}|\theta_j)P(S_{(j,t)}|\phi_{H_{(j,t)}})\right] \tag{16}$$

When choosing a new head symptom H for sub symptom S in Post *P*, Gibbs sampling, a technique used in LDA, is employed to select the maximum likelihood values. It is written in Equation 17.

$$P(H|S,P)=\frac{\begin{pmatrix}\text{No. of } S \text{ in}\\ P \text{ that assigned to } H+\alpha\end{pmatrix}\begin{pmatrix}\text{No. of } S \text{ in}\\ H+\beta\end{pmatrix}}{\text{Total No. of } S \text{ in } H+\beta_{tot}} \tag{17}$$

3.4.3 BERTopic:

It is one of the deep learning-based methods for topic modelling and it is derived from Bidirectional Encoder Representations from Transformers (BERT) method. When comparing BERTopic with LSA and LDA, it doesn't need sub symptom extraction phase as one of the processes in the BERTopic comprises the extraction. It contains five processes: 1) Vectorization is used to convert the pre-processed data into numerical representation then form the vectors and it is done by SentenceTransformer. 2) After vectorization the dimension of the vectors are increased so Uniform Manifold Approximation and Projection (UMAP) is applied for reducing the higher dimensional structure into lower dimensional structure. 3) The related sub symptoms are grouped into head symptoms by forming the similar clusters using HBSCAN algorithm. 4) The clusters are formed based on number of posts so determining the frequency of each sub symptoms present in each cluster using count vectorizer method. 5) Next, form each cluster into head symptoms according to sub symptom importance by calculating the TF-IDF of each cluster which is given in Equation 18.

$$c\text{-}tf\text{-}idf(s,h)=\square\, tf_{(s,h)}\,\square\times\log\left(1+\frac{A}{f_s}\right) \tag{18}$$

where $tf_{s,h}$ means frequency of sub symptom *s* in head symptom or cluster *h*, f_s means frequency of sub symptom *s* across all cluster *h*, *A* is the average number of sub symptoms per cluster *h*.

3.5 SYMPTOM SEGMENTATION AND GROUPING

As PLSA, LDA and BERTopic only provides the topic number, the head symptoms resulting from each method are labelled in accordance with Table 1 head symptoms. A single post has multiple sub symptoms that have been mapped and generates multiple labelled head symptoms. As a result, this proposed study applied the innovative idea of symptom segmentation and grouping using modified versions of the algorithms like SSG_PLSA, SSG_LDA, and SSG_ BERTopic to segment the sub symptoms, map the head symptoms, and group the related head symptoms that were present in a single post. These modified algorithms are used to reduce the dimensionality of the features by combining similar type of symptoms which is represented in Algorithm 1.

Algorithm 1: Symptom Segmentation Grouping

Input: P- Post, SS – Sub Symptoms, HS – Head Symptoms
Output: Label – Grouped Labelled Head Symptoms

```
for each post P in corpus
    map the SS into HS using Eq.(8), Eq.(17) and Eq.(18)
    //Labelling and grouping Head Symptoms
      NHS=10        //Total number of head symptoms as 10
      Label[SSs]={}
      for each SS in PP
         for each sub symptom segment (SSs) 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 for
END
```

3.6 EVALUATION

Using a variety of topic modelling techniques, the secondary symptoms provided by PCOS patients on Reddit are translated into the primary symptoms indicated by gynecologists. If more than one head symptom appears in a post and they are similar, these mapped head symptoms are also segmented and grouped; otherwise, they are segmented separately. This is done after getting topic modelling results by using symptom segmentation and grouping. These modified algorithms are evaluated using sub symptoms distribution, dominant symptom finding and execution time of each algorithm.

4. RESULTS AND DISCUSSION

The dataset considered for this work is a collection of 25000 PCOS posts from Reddit users. The user posts in the dataset are taken from https://www.reddit.com/r/PCOS/ and analyze the following results.

4.1 SUB SYMPTOMS AND HEAD SYMPTOMS MAPPING USING SSG

The Table.1 shows the ten head symptoms given by the Gynecologist, sub symptoms given by the Reddit users, sample posts and its relevant outputs of each step.

4.2 PROBABILITY DISTRIBUTIONS OF SUB SYMPTOMS IN HEAD SYMPTOM

The sub symptoms are extracted from the Reddit data using BoW and TF-IDF methods. Table 2 shows the sub symptom extraction and head symptoms mapping results. According to BoW and TF-IDF summation finding, the sub symptoms Ovaries, Infertility, and Pain are the least common, while Periods, Depression, and Obesity are the most common. Skin Hair, Head Hair, Skin, and Hormone are in the middle. Each SSG method has offered its own probability for head symptom mapping based on the distributions of the sub symptoms, and adding these probabilities together will result in a probability that is very close to one.

Table.1. PCOS Sub Symptoms, Head Symptoms, Sample Posts and Results of each step

PCOS Symptom Collection from Gynecologist	
Reddit Users' Sub Symptoms	**Head Symptoms**
adrenal hyperplasia, appetite, breast change, diabetic, endometriosis, estrogen, excess androgen, frequent urination, hormonal abnormality, hormonal disruption, hormonal imbalance, hunger, hyper androgegism, hyperinsulinemia insulin resistance, metformin, osteopenia, progesterone, sugar craving, testosterone	Hormone
abdomen fat, belly fat, bloating, excess weight, fat, inflammation, obesity, overweight, weight gain	Obesity
abnormal period, heavy bleeding, heavy period, irregular cycle, irregular period, skip menstrual period, skip period	Periods
chest hair, chin hair, dark hair, excess body hair, excessive hair growth, facial hair hair growth, hirsutism, mustache, stomach hair, unwanted hair growth, upper lip hair	SkinHair
acne, cramp, darkening skin, dark spot, irritability, oily skin, patch, pimple, redness skin, skintag, stretch mark	Skin
anovulation, cyst, cystic acne, cyst ovary, follicle, no ovulation, oligo ovulation, ovarian cramp, poly cystic ovary	Ovaries
infertility, miscarriage, never get pregnant, not pregnant	Infertility
anger, anxiety, brain fog, depression, fatigue, high blood pressure, insomnia, low energy, mental hell, mood swing, sleep problem, stressful, tired	Depression
baldness, hair fall, hair loss, thinning hair	HeadHair
breast pain, headache, lower abdominal pain, painful period, pelvic pain	Pains
PCOS Post Collection from Reddit	
Sample Posts	**Mapping**
I **miscarried** a couple times years back and **never get pregnant**	Infertility
Hi guys.I suffer from pcos. Luckily my only symptoms are **irregular periods**. my **anxiety** levels increase.	Periods, Depression
I should suffer with uncontrollable **weight gain**, **sugar cravings**, **pimples** and **acne**	Obesity, Hormone, Skin, Skin

I was diagnosed with PCOS. I suffer with the so called Lean PCOS, with **hirsutism** and **hair loss** as my main problems.	SkinHair, HeadHair
No ovulation and I have the **follicles** on both **ovaries**	Ovaries
I have **pelvic pain** and **headache**	Pain

Extracted Sub Symptoms	**Head Symptoms Mapping**	**Process of Symptom Segmentation and Grouping (SSG) after Mapping**	**Head Symptoms Output**
miscarried nevergetpregnant	Infertility Infertility	Infertility | Infertility (Two Segmentation and One Group)	Infertility
irregularperiod anxiety	Periods Depression	Periods | Depression (Two Segmentation and No Group)	Periods Depression
Weightgain sugarcraving pimple acne	Obesity Hormone Skin Skin	Obesity | Hormone | Skin | Skin Obesity, Hormone Obesity, Hormone, Skin Obesity, Hormone, Skin (Four Segmentation and Three Group)	Obesity Hormone Skin
noovulation follicle ovary	Ovaries Ovaries Ovaries	Ovaries | Ovaries | Ovaries Ovaries Ovaries (Three Segmentation and One Group)	Ovaries

4.3 DOMINANT SYMPTOMS AND EXECUTION TIME

The dominant sub and head symptoms are found using maximum likelihood values of sub symptoms. As per Table.2, insulin resistance is the highest for Hormone, weight gain for obesity, irregular period for Periods, facial hair for Skin Hair, acne for skin, cyst for Ovaries, infertility sub symptom for infertility, fatigue for depression, hair fall for Head Hair and headache for Pain. All mapped dominant sub symptoms from each head symptom are taken and find dominant head symptom using maximum probability value. The dominant symptom of SSG_PLSA is Periods, SSG_LDA is Head Hair and SSG_BERT is Skin Hair problem which is shown in Fig.2. These three modified algorithms have taken less execution time compared to traditional algorithms which is shown in Fig.3.

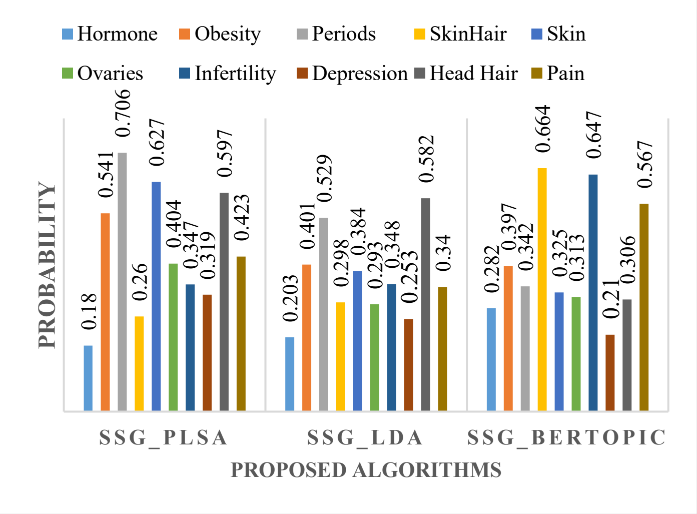

Fig.2. Dominant Symptoms

Table.2. Symptom Extraction and Probability Distribution of Modified Topic Modelling

Sub Symptoms	BoW	TF-TDF	SSG_PLSA	SSG_LDA	SSG_BERTopic
Hormone					
adrenal hyperplasia	8	5.5	0.036	0.019	0.016
appetite	4	1.53	0.015	0.039	0.034
breast change	20	9.7	0.059	0.019	0.023
diabetic	23	13.24	0.056	0.128	0.03
endometriosis	10	9.2	0.025	0.024	0.009
estrogen	7	4.41	0.040	0.021	0.027
excess androgen	22	15.5	0.107	0.054	0.014
frequent urination	3	2.41	0.014	0.028	0.018
hormonal abnormality	3	2.12	0.060	0.100	0.012
hormonal disruption	3	1.34	0.060	0.013	0.011
hormonal imbalance	34	21.62	0.100	0.014	0.025
hunger	10	10.34	0.063	0.061	0.172
hyper androgegism	2	1.41	0.020	0.012	0.02
hyper insulinemia	2	2.89	0.006	0.028	0.011
insulin resistance	**64**	**38.86**	**0.180**	**0.203**	**0.282**
metformin	11	8.95	0.047	0.036	0.029
osteopenia	4	1.73	0.001	0.011	0.007
progesterone	7	5.15	0.034	0.015	0.024
sugar craving	5	2	0.004	0.019	0.105
testosterone	35	23.81	0.128	0.185	0.136
Summation	277	179.34	1.055	1.029	1.005
Obesity					
abdomen fat	12	8.71	0.017	0.053	0.114
belly fat	25	14.91	0.051	0.074	0.097
bloating	16	8.65	0.137	0.015	0.102
excess weight	6	2.87	0.100	0.310	0.171
fat	11	6.24	0.022	0.041	0.081
inflammation	11	8.53	0.031	0.029	0.086
obesity	19	12.64	0.086	0.071	0.125
overweight	24	14.2	0.066	0.061	0.121
weight gain	**256**	**155.2**	**0.541**	**0.401**	**0.397**
Summation	380	231.95	1.051	1.055	1.2
Periods					
abnormal period	2	1.41	0.014	0.026	0.026
heavy bleeding	15	10.91	0.055	0.022	0.182
heavy period	1	1	0.047	0.024	0.063
irregular cycle	133	81.18	0.032	0.040	0.075
irregular period	**207**	**136**	**0.706**	**0.529**	**0.342**
skip menstrual period	13	2	0.002	0.387	0.048
skip period	3	2	0.204	0.014	0.243
Summation	367	234.5	1.06	1.042	1.0
Skin Hair					
chest hair	3	26.11	0.060	0.018	0.037
chin hair	23	13.33	0.127	0.034	0.047
dark hair	10	8.57	0.010	0.059	0.011
excess body hair	40	1	0.023	0.080	0.029
excessive hair growth	12	9.1	0.158	0.190	0.05
facial hair	**80**	**47.19**	**0.260**	**0.298**	**0.664**
hair growth	78	44.59	0.240	0.125	0.076
hirsutism	80	45.99	0.140	0.202	0.108
mustache	3	2.12	0.004	0.014	0.003
stomach hair	2	0.56	0.003	0.005	0.052
unwanted hair growth	1	1	0.012	0.003	0.02
upper lip hair	5	3.36	0.009	0.002	0.006
Summation	337	199.92	1.046	1.03	1.103
Skin					
acne	**200**	**112.77**	**0.627**	**0.384**	**0.325**
cramp	17	8.22	0.038	0.039	0.050
darkening skin	15	9.56	0.028	0.100	0.142
dark spot	4	2.12	0.038	0.035	0.071
irritability	12	4.92	0.038	0.013	0.068
oily skin	19	10.68	0.050	0.273	0.145
patch	18	10.45	0.036	0.038	0.099
pimple	9	4.95	0.012	0.022	0.089
redness skin	28	12.59	0.047	0.041	0.003
skintag	3	10.48	0.094	0.054	0.211
stretchmark	2	9	0.017	0.027	0.044
Summation	334	187.59	1.025	1.026	1.24
Ovaries					
anovulation	4	2.7	0.012	0.024	0.043
cyst	**60**	**36.49**	**0.404**	**0.293**	**0.313**
cystic acne	27	14.36	0.063	0.188	0.307
cyst ovary	52	29.21	0.293	0.233	0.267
follicle	10	4.87	0.032	0.110	0.052
no ovulation	13	7.72	0.037	0.032	0.048
oligo ovulation	3	2.52	0.008	0.019	0.035
ovarian cramp	5	1.89	0.021	0.056	0.028
poly cystic ovary	18	9.64	0.166	0.058	0.109
Summation	192	111.77	1.036	1.014	1.203
Infertility					
infertility	**44**	**30.26**	**0.347**	**0.348**	**0.647**
miscarriage	5	2.4	0.299	0.242	0.137
never get pregnant	4	2	0.281	0.239	0.218
not pregnant	1	2	0.111	0.203	0.056
Summation	54	36.66	1.039	1.033	1.058

Depression					
anger	7	2.86	0.025	0.011	0.023
anxiety	85	47.18	0.249	0.216	0.103
brain fog	17	8.64	0.003	0.061	0.112
depression	71	37.79	0.128	0.166	0.095
fatigue	**100**	**57.72**	**0.319**	**0.253**	**0.210**
high blood pressure	8	4.75	0.024	0.055	0.160
insomnia	6	2.13	0.005	0.008	0.021
low energy	8	6.41	0.004	0.003	0.130
mental hell	3	1.05	0.009	0.003	0.108
mood swing	70	35.83	0.163	0.190	0.086
sleep problem	23	16.61	0.108	0.010	0.059
stressful	3	3	0.003	0.005	0.013
tired	10	6.66	0.014	0.069	0.034
Summation	411	230.63	1.054	1.049	1.156
Head Hair					
baldness	12	4.14	0.036	0.036	0.142
hair fall	**186**	**105.34**	**0.597**	**0.582**	**0.306**
hair loss	116	70.31	0.317	0.281	0.271
thinning hair	23	13.97	0.065	0.120	0.224
Summation	337	193.76	1.015	1.018	1.0
Pain					
breast pain	2	1.15	0.020	0.115	0.068
headache	**23**	**13.71**	**0.423**	**0.340**	**0.567**
lower abdominal pain	2	2	0.117	0.238	0.037
painful period	16	8.67	0.328	0.037	0.347
pelvic pain	10	7	0.131	0.275	0.067
Summation	53	32.53	1.019	1.005	1.087

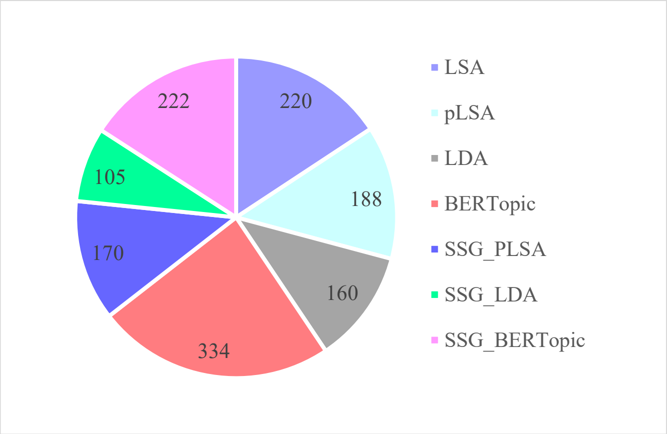

Fig.3. Processing Time (10^3 Seconds)

5. CONCLUSIONS

The main contributions of the paper could be summarized to following: PCOS symptom and posts collection from Gynaecologist and Reddit users, Pre-process the collected posts by applying various pre-processing technique, Extract the PCOS sub symptoms using BoW and TF_IDF technique, Map and group the sub symptoms into head symptoms using modified topic modelling algorithms like SSG_LDA, SSG_LSA and SSG_BERT and Evaluate the proposed algorithms with its own performance measures. Based on sub-symptom extraction and probability distributions, this proposed analysis indicates that each approach yields prominent symptoms. Compared to other methods, TF-IDF with SSG_LDA finds the dominating symptom and maximum likelihood more quickly. This approach can be expanded to analyse PCOS symptom patterns by integrating association rule mining and topic modelling.

REFERENCES

[1] A.M. Kaplan and M. Haenlein, “Users of the World, Unite! The Challenges and Opportunities of Social Media”, *Business Horizons*, Vol. 53, No. 1, pp. 59-68, 2010.

[2] Waseem Akram, “A Study on Positive and Negative Effects of Social Media on Society”, *International Journal of Computer Science and Engineering*, Vol. 5, No. 10, pp. 347-354, 2018.

[3] Muhannad Quwaider and Mosab Alfaqeeh, “Social Networks Benchmark Dataset for Diseases Classification”, *Proceedings of International Conference on Future IoT and Cloud Work*, pp. 234-239, 2016.

[4] Ali Alessa, Miad Faezipour and Zakhriya Alhassan, “Text Classification of Flu-related Tweets using FastText with Sentiment and Keyword Features”, *IEEE International Conference on Healthcare Informatics*, pp. 1-6, 2018.

[5] Irving, “The Stein-Leventhal Syndrome”, Available at: https://embryo.asu.edu/home, Accessed in 2017.

[6] Palvi Soni and Sheveta Vashisht, “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.

[7] Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram and Remya George, “i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) using Machine Learning Techniques”, *Proceedings of IEEE International Conference on TENCON*, pp. 673-678, 2019.

[8] Fang Zhang, Jun Luo, Chao Li, Xin Wang and Zhongying Zhao, “Detecting and Analyzing Influenza Epidemics with Social Media in China”, *Lecture Notes in Computer Science*, Vol. 8443, pp. 90-101, 2014.

[9] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao and Wei Zhou, “An Overview of Topic Modelling and its Current Applications in Bioinformatics”, *Springer Plus*, Vol. 5, No. 1608, pp. 1-22, 2016.

[10] Andreas Alga, Oskar Eriksson and Martin Nordberg, “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. 1-22, 2016.

[11] Juan Antonio Lossio-Ventura, Sergio Gonzales, Juandiego Morzan, Hugo Alatrista-Salas and Tina Hernandez-Boussard, “Evaluation of Clustering and Topic Modeling methods over Health-Related Tweets and Emails”, *Artificial Intelligence in Medicine*, Vol. 117, pp. 1-27, 2021.

[12] Vasantha Kumari Garbhapu and Prajna Bodapati, “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.

[13] Shaymaa H. Mohammed and Salam Al-Augby, “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.

[14] Zhengxing Huang, Wei Dong and Huilong Duan, “A Probabilistic Topic Model for Clinical Risk Stratification from Electronic Health Records”, *Journal of Biomedical Informatics*, Vol. 50, pp. 28-36, 2015.

[15] Juan Zhao, QiPing Feng, Patrick Wu, Jeremy L. Warner, Joshua C. Denny and Wei-Qi Wei, “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.

[16] Chong Ni Ki, Amin Hosseinian-Far, Alireza Daneshkhah and Nader Salari, “Topic Modelling in Precision Medicine with its Applications in Personalized Diabetes Management”, *Expert Systems*, Vol. 45, pp. 1-21, 2021.

[17] Amina Amara, Mohamed Ali Hadj Taieb and Mohamed Ben Aouicha, “Topic Modelling in Precision Medicine with its Applications in Personalized Diabetes Management”, *Applied Intelligence*, Vol. 51, pp. 1-13, 2021.

[18] Phatpicha Yochum, Thanapoom Nisamaneewong, Phathaphol Karnchanapimonkul and Bhanusich Chomanan, “Automated Disease Detection Based on Clinical Text using Topic Modeling”, *Proceedings of International Conference on Information Technology, IoT and Smart City*, pp. 74-79, 2022.

[19] K. Gethsiya Raagel, M. Bagavandas, K. Sathya Narayana Sharma and P. Manikandan, “Sentiment Analysis and Topic Modeling on Polycystic Ovary Syndrome from Online Forum using Deep Learning Approach”, *Wireless Personal Communications*, Vol. 133, No. 1, pp. 1-20, 2023.