International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

Breaking the Taboos: Deploying Knowledge Differentiation to study COVID-19 ramifications on Women's Menstrual Health

Sarabjeet Kaur Kochhar[1] **and Rumjot Kaur**[2]

[1] *Department of Computer Science,Indraprastha College for Women, New Delhi, India*
[2] *Department of Computer Science,Indraprastha College for Women, New Delhi, India*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: The social media platforms serve as forums where people from the different strata of society, across the world, voice their unhindered opinions on a variety of subjects. Such platforms have been utilized to mine patterns that reflect the viewpoints, thought processes, and attitudes of the people. Sometimes the knowledge attained from social media messages highlights the emergent, actionable needs of the people. The rising anxiety and several problems, showcased by the tweets related to menstruation during the pandemic, serve as a ready reckoner for the same.

Psychological studies have established that people worldwide were exposed to grave and persistent psychosocial stressors due to the COVID-19 pandemic. It has also been established by studies that women are more affected by stress. Fear, anxiety, depression, and emotional instability caused due to stress affect the hypothalamic-pituitary-gonadal (HPG) and hypothalamic–adrenal axes, leading to menstrual cycle irregularities, dysmenorrhea, pre-menstrual symptoms, and menorrhagia. Menstrual health is a significant determinant of a woman's overall health and quality of life and is a contributor to the socio-economic burden on women, their families, and society, in general. Research to study the effect of the COVID-19 pandemic on women's menstrual health is an imperative public health initiative and is need of the hour. It is disappointing that only a handful of medical, survey-based studies have been conducted so far in this important area that affects almost half of the world's population. The work presented in this paper is intended as a novel, small step in the direction of the stated aspiration and urges the machine learning community to break the taboo, and start talking about menstrual health to complement the efforts of the medical fraternity. Technologies like artificial intelligence, data mining, and machine learning have the prowess of algorithmically harnessing gigantic datasets, sophisticated pattern detection, and presenting the knowledge discovered in an understandable form. This ready-to-be-leveraged knowledge can be used by domain experts for drawing conclusions and inferences. The framework presented in this paper uses a blend of supervised and unsupervised data mining techniques to uncover knowledge. The framework is based on the principle of knowledge differentiation and uncovers knowledge at two different levels of abstraction. This allows for analysis of the discovered knowledge from multiple perspectives, enabling its consolidation, comprehension, and actionability. At the first level, facts and myths being circulated on the net about the women's mensural health during the pandemic are discovered. Facts must be separated from the myths circulated on Twitter, in order to understand the authentic opinions and problems of people. Myths themselves prove as an important resource to understand the reasons behind people's responses and reactions to the government's policies. Identification of myths is also important so that public awareness plans can be launched and myth busters issued by the concerned government agencies. At the second level of abstraction, association rules are discovered from the selected categories of tweets, classified at abstraction level one. To the best of our knowledge, this is one of the first works leveraging the use of association rule mining for deriving meaningful knowledge about menstrual health from tweets. Upon unearthing the associations between the frequently used words in the tweets, they are subjected to four-stage postprocessing rule filters, that enable the bi-directional analysis of these linkages, in order to aid consolidated inferencing about selected classes of tweets.

Results of the framework presented in the paper show that the facts discovered by the system correctly identified the menstrual problems faced by women during the COVID pandemic, and also pointed to COVID-related stress as a probable cause for the same. Some very popular rumors were discovered by the presented framework that could represent some of the valid reasons behind vaccine hesitancy. The framework was also able to discover important associations from facts and myths about menstrual health and the pandemic being circulated on social media, in the form the tweets. The results are also able to clearly establish the efficacy of discovering association rules from tweets, especially when compared to the works that are devoted to just mining the top frequent words in tweets, say, via techniques such as clustering.

Keywords: COVID-19, Knowledge Differentiation, Menstrual Health, Classification, Association Rule Mining.

E-mail address: skaur@ip.du.ac.in, rumjotkaur219@gmail.com

1. INTRODUCTION

The COVID-19 pandemic not only took the world by a frenzy and grimly impacted the health, education, research, and world economy but also led to loneliness, fear, anxiety, and depression in society at large. Recent psychological studies have argued that in fact the stress caused due to COVID-19, initially recognized by some studies as monophobia (simple excessive fear), is actually a form of non-trivial adjustment disorders, severe forms of which have been recognized as Post Traumatic Stress Disorder and COVID Stress Disorder [1], [2], [3]. Many studies have reported that women are emotionally more disturbed by stressful events than men [4], [5], [6], [7]. Fear, mood disorders such as anxiety and depression, and acute life stressors affect hypothalamic–pituitary–gonadal (HPG) and hypothalamic–adrenal axes and lead to menstrual cycle irregularities [8], [9], [10]. Studies even confirm that the stress experienced by women during the COVID-19 pandemic could parallel the changes in menstrual cyclicity caused by other acute life stressors such as war, natural disasters, displacement, and famine [8]. It is pertinent therefore to study the ramifications of the COVID-19 pandemic on women's menstrual health.

Though substantial resources and research have been dedicated globally to the study of the COVID-19 pandemic, its origins, ramifications, vaccinations, etc., it is disappointing that almost no research studies, especially from the machine learning community, have focused on this very important yet oft-neglected issue of women's health [11], [12], [13]. Only a handful of medical (largely survey-based) studies on the aforementioned topic have been conducted, some of which are reviewed in Section 2. Interestingly, even these works stress the need for further research [11], [12], [13].

During the COVID-19 pandemic, medical practitioners in most nations were swamped, the medical facilities were locked down and the medical supplies and supply chains were broken or choked. In these unprecedented times, it was the micro-blogging platforms such as Twitter, that served as platforms where people could voice their unhindered opinions and share their experiences, about the diverse facets of their lives [14], [15]. The tweets from this period, therefore, deserve to be recognized not just as mere text messages, but as an important resource, that not just reflect people's viewpoints, thought processes, and attitudes, but also sketch a picture of what they underwent. Technologies such as machine learning, artificial learning, and data mining can be applied to such tweets to know the urgent, actionable needs of people in times of crisis [14]. Such knowledge proves to be a crucial tool to guide the policymakers to understand the severity of situations and take steps accordingly, especially during the time of a crisis or just after one [16]. However, misinformation spread on social media platforms blemishes the efficacy of such platforms. Unverified rumors and myths mislead people about treatment plans, medical advisories, and precautions, fueling anxiety, fear, and stress [17], [18]. It is vital therefore to build automated competencies that can flag off misleading information on social media platforms.

This paper proposes an automated, integrated framework to perform the following tasks:

- Identify the facts as well as myths being circulated on the social media platforms such as Twitter by classification (Section 4). Three classifiers were selected for this task, implemented, and evaluated using the metrics such as precision, recall, and F1-score. The most suitable among the three was selected. The task of identifying misinformation has also been taken care of by some other studies, some of which are reviewed in Section 2. However, most of these works stop at the identification of misinformation.

- An important characteristic to gauge the interestingness of discovered knowledge is its understandability and actionability [19], [20], [21]. It is important to design systems that produce results that are easy to comprehend, deploy, and possibly analyzable at different levels of abstraction. For instance, counter-strategies such as myth busters can only be issued by nodal medical agencies such as the world health organization, if the discovered misleading information is comprehensible. Similarly, strategies can be designed to plan the course of action for future emergencies if the critical information retrieved from social media platforms facts about exigencies faced by people, and their viewpoints on the important topics are in a form that is understandable and ready to be leveraged. Realization of such a task can be achieved by application of the principle of knowledge differentiation, wherein knowledge at a higher level of abstraction is derived from the knowledge at a lower level of abstraction, over a period of time [22]. Higher level consolidated knowledge provides a way of comprehending and leveraging the discovered knowledge at a lower level of abstraction from different perspectives. Accordingly, association rules, representing knowledge at abstraction level two are generated from the classified tweets, representing knowledge discovered at abstraction level one. This allows us to draw consolidated inferences about the selected class of tweets by unearthing previously unknown and interesting linkages between different words used in tweets of the classes interesting and/or useful to us.

- It is important to note here that most of the works for discovering critical information from the tweets so far have been centered around clustering and classification [23], [24], [25], [26], [16]. To the best of our knowledge, this is one of the select few works leveraging the use of association rule mining for deriving meaningful knowledge. The choice of using association rules in this work is premeditated. The study of tweets has generally been used to yield a list of most frequently encountered words or topics, around which conversations are centered. In comparison, associations don't just stop at the discovery of frequent data items. They are derived from the frequent itemsets. They represent more structured information, in the form of rules, X →Y, where X and Y represent frequent itemsets. These rules give us an opportunity for in-depth, bidirectional analysis of the linkages i.e., X →Y and Y →X. So not only are the hitherto hidden relationships between the data items X and Y revealed, but it is also possible to inspect these rules bidirectionally with the help of metrics such as confidence and lift, etc. to establish whether the dependence of X on Y is more meaningful or vice versa.

Results of the presented framework not only show the successful discovery of tweets representing facts and myths about menses-related symptoms during COVID-19 but also successfully identify neutral and not relatable tweets and filter them out before the task of association rule mining. It was found that the neutral and non-relatable tweets formed almost half of the total tweets. Filtering them out led to computation time and space benefits and also reduced the number of relevant rules produced, post-processed, and analyzed. The fact and myth associations discovered by the framework unearthed important and interesting linkages between menses-related symptoms, COVID-19, some COVID-19 vaccinations, and stress.

The organization of the paper is as follows: Section 2 reviews the related work. Section 3 presents the general template of the framework proposed in this paper. It also lays down the implementation details of the framework and the data used. Section 4 presents the classification framework to discern authentic information from misleading information on Twitter. The results and inferences drawn from the classified tweets are also discussed. Section 5 details the generation, postprocessing, and usage of association rules drawn from the classified tweets. The interesting results and their inferences are also discussed in this section. Section 6 concludes the paper.

2. RELATED WORK

In this section, we present an overview of the works closer to our work. As mentioned in the introduction, to the best of our knowledge, this paper is one of the novel machine learning studies, that attempts to rattle the taboo surrounding women's menstrual health and the impact on it due to the COVID-19 pandemic. Also, most of the work dedicated to information retrieval revolves around other data mining techniques such as clustering and classification. This work plans to leverage the rich semantic knowledge provided by association rules to consolidate and derive understandable, actionable knowledge. Also, the principle of knowledge differentiation doesn't find application in any machine learning works so far.

So, we review the literature related to all aspects of our work, classified according to their relevance as follows. The first set of works is devoted to the identification of misleading information. These are summarized in subsection A. Some works dedicated to retrieving useful information from tweets are reviewed next, in subsection B. Some medical papers that did attempt to study the ramifications of the COVID-19 pandemic on women's menstrual health are summarized next (subsection C).

A. Misinformation Detection

Prediction of whether a tweet story is fake or real, with the help of a neural network-based GCAN (Graph-aware Co-Attention Network) model has been taken up in [27]. The GCAN model is typical in its approach involving the short text content, retweet sequence of users along with user profile i.e., extraction of textual features, and application of supervised learning methods. The proposed GCAN model creates features that quantify the participation of a user in social networks and generates the representation of words present in the source tweet. It also models how the source tweets propagate and captures the correlation between the source tweet and the user propagation to detect whether a tweet is likely to be true or false.

To detect fake information spreaders, bot-based fake information dissemination has been studied in [15]. User-based features and content-based features from Twitter have been employed to compute statistics such as TF-IDF, Bag of Words, and average mean time to tweet. Sentiment analysis using the VADER lexicon is used to compute the average sentiment score of a user's tweets. The work claims to be more accurate than the other machine learning classifiers like MLP, decision tree, and random forest.

An advanced framework to identify tweets with fake news content has been proposed in [28]. The framework statistically analyses the Twitter user accounts, uses reverse image searching, and performs cross-verification of fake news sources for the same. A study to identify misinformation spread about vaccines from Russian trolls on Twitter has been taken up in [29]. Descriptive, bivariate, multivariable negative binomial regression has been applied to infer that detect misinformation about personal dangers, civil liberty violations, and vaccine conspiracies. A study to detect fake news about COVID-19 from twitter for multiple Indic Languages has been reported in [30]. BERT (Bidirectional Encoder Representations from Transformers) model has been applied to an annotated dataset of Hindi and Bengali tweets to identify fake news. A dataset of English and Chinese tweets, related to the protests staged in Hong Kong, in 2020, published by Twitter is processed for the extraction of the top 10 most significant features in [31]. The work employs four different algorithms are, namely the

Naïve Bayes, SVM, C4.5, and Random Forests of C4.5 for used for the training and evaluation of classification models. The extracted linguistic patterns and semantic polarity are used to distinguish tweets spreading fake news.

An adaptation of the BERT model, with three blocks of 1d-CNN, and different kernel-sized convolutional layers has been deployed for fake news detection on the U.S. General Presidential Elections dataset [32]. Recurrent Neural Network (RNN) methods such as Gated Recurrent Unit (GRU) and Long-Short Term Memory (LSTM) methods have been used for textual fake news detection in [33]. The paper compares word embeddings, the numerical representations of the text being investigated, from various sources such as 'glove', an open-source project at Stanford University, 'Twitter' word embeddings, 'news' word embeddings, and 'crawl' word embeddings. Most of the above works stop at the identification of real and fake tweets, unlike the framework in this work that uses the principle of knowledge differentiation to leverage the classified tweets.

A framework for Identification of real and fake news using association rule mining has been presented in [34]. The framework treats 2016 US presidential election campaign tweets as text transactions and converts the tweet corpus into a binary transactional database to mine association rules, using the Apriori algorithm. It is interesting to note that this work attempts to identify fake tweets based on the association rules. The work presented in this paper, however, draws out classified association rules only from the category of tweets, from which the higher order knowledge needs to be drawn. For instance, in this paper, rules are drawn only from the FACT class and the MYTH class. The neutral and not relatable category of tweets don't contribute directly to the actionable knowledge required and are therefore not subjected to association rule mining at all. This proves as an automatic filter, leading to optimization of computing time, and space and limiting the number of association rules generated.

B. *Critical Information Retrieval*

In one of the earliest works (2010), an application called "Hotstream", to mine breaking news from the Twitter timeline was developed by [35]. Proper nouns were used to form groups, which were ranked on popularity and reliability. Another topic detection algorithm proposed by Cataldi et al. [36], in the year 2010, models the life cycle of terms extracted from tweets and identifies emerging terms. A navigable topic graph between the emerging terms with other semantically related keywords is plotted for the detection of the emerging topics.

To analyze the authority of users and their content, the social relationships in the network are ranked with the Page Rank algorithm. Two approaches, namely the Bag-of-Words approach and the network-based approach were adopted for the classification of Twitter trending topics by Lee et. Al. in 2011 [25]. An accuracy of 65% was reported with the Naive Bayes Multinomial classifier, applied on the word vectors constructed using trending topic definition and tweets, and tf-idf weights, in the bag of words approach. The top 5 similar topics for a given topic are based on the number of common influential users. are used to classify the topics using a. In the network-based classification method, C5.0 decision tree learner was applied to the categories of similar topics and the number of common influential users to predict a trending topic with a claimed efficacy of 70%. A business case study for retrieval of real-time disaster information was taken up by Zheng et. Al. [31]. Two systems were designed - an information network system for web-based systems and a browser system for mobile devices. The systems support report summarization and use a probabilistic model to dynamically generate query forms and information dashboards, to gain insights about the disaster situation and for making decisions. TRCM (Transaction-based Rule Change Mining), a system to study the evolution of linkages mined from the hashtags of tweets, and the evolution of real-life events, has been presented by Olowe et. Al. [37]. Some works from the year 2010 to the year 2013, devoted to topic detection, have also been reviewed in the paper.

In more recent studies, two matrix decomposition algorithms, namely, Rolling-ONMF and Sliding-ONMF, are proposed and compared, for monitoring the weekly evolution of the most discussed topics regarding COVID-19 [38]. Discovery and comparison of the topics, sentiments, and emotions, from the tweets of two news agencies from Iran and Turkey, have been undertaken in [39]. Differences between the topics of conversation among men, women, and other gender minorities, on Twitter, have been studied in [23].

C. *Menstrual Health and COVID*

A review of works related to the COVID-19 pandemic and menstrual health has been taken up in [13]. The study establishes that women have indeed experienced menstrual changes (altered menstrual duration, frequency, and volume) because of COVID-19 and the vaccinations, which also led to an increase in vaccine hesitancy. They claim that menstruation-related symptoms can become a source of economic burden through decreased productivity and therefore determining the scale of menstrual problems, their cause, and the impact on those who menstruate and wider society will allow the identification of new preventative and therapeutic strategies.

Menstrual irregularities and abnormal uterine bleeding after the first and second doses of the COVID-19 vaccine and the time of these disturbances have been investigated by [40]. 205 out of 369 women with gynecological or non-gynecological diseases, who had undergone hormonal or non–hormonal treatments, were in premenopausal or menopausal period, or who had irregular menstrual cycles in the last 12 months before the vaccine were excluded from the study. A questionnaire was designed which consisted of 26 multiple–choice questions. The results claimed that women had menstrual cycle irregularities regardless of the type of vaccine and phase of the menstrual cycle.

Menstrual cycle data of US residents aged 18-45 was monitored using an application named "Natural Cycles" by [11], to study the relationship between the length of the menstrual cycle and COVID-19 vaccination. The results claimed that the participants had normal cycle length for three consecutive cycles before the vaccination. However, after vaccination, a small change in cycle length was noted. The link between COVID-19 vaccination and menstrual disturbances among women aged 18-30 in Norway was studied by [41]. Mobile-phone questionnaires were used to collect reports. They found an increased risk of menstrual disturbances, for instance, heavier and longer bleeding duration than usual after vaccination. An online survey was conducted in May 2021 in United Kingdoms by [42], to evaluate the effect of COVID-19 vaccination on the menstrual cycle, deduce the factors responsible for the disturbances, and identify patterns of symptoms in patients. For analysis, 4989 participants who were pre-menopausal and vaccinated were selected. According to the findings, 80% of pre-menopausal vaccinated individuals did not report any menstrual changes up to 4 months after their vaccination. Participants who had used oral contraceptives were smokers and had positive COVID status reported menstrual changes.

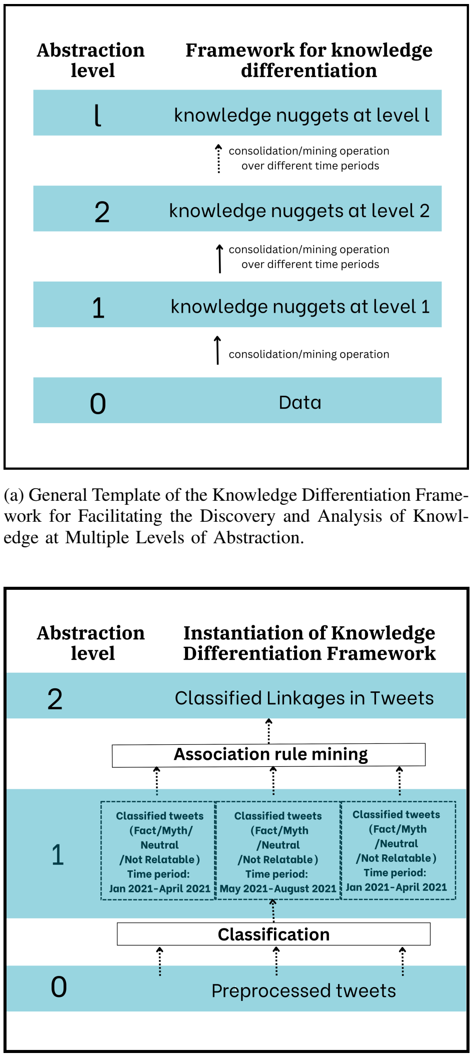
Experiences of 1273 people in the UK who had kept the record of their menstrual cycles and vaccination dates, were studied by [12]. The participants were asked to anonymously fill out a web form with details such as their age, height, menstrual cycle length, use of hormonal contraception, and gynecological condition if any. The findings reported that there is no association between the type of COVID-19 vaccine and changes in periods. They claimed that people who had a diagnosis of a menstrual or gynecological condition were not more likely to report a change in flow than those who did not have such a diagnosis. They found that people with endometriosis reported earlier periods than usual and people with polycystic ovaries reported late periods than usual.

3. RESEARCH FRAMEWORK DEPLOYING KNOWLEDGE DIFFERENTIATION

In this section, we lay out the general template of the knowledge differentiation framework, used in this work, to extract, classify, and mine association rules, to develop an overall understanding and leverage the tweets related to the covid-19 pandemic and menstrual health, posted by people worldwide. The schematic layout of the framework proposed in this paper, based on the knowledge differentiation template presented earlier, is detailed next, in Section 3A. The details of the data extraction and pre-processing are outlined, and the details of the programming environment and the underlying implementation are also laid down in Section 3B.

A. *Research Framework*

The useful information and/or the misinformation mined from the tweets can be viewed as knowledge nuggets at the lowest level of abstraction, say abstraction level one. The knowledge nuggets at abstraction level one can be used to derive knowledge nuggets at abstraction level two, for providing multiple perspectives of the knowledge discovered, better comprehension, and hence better actionability. In general, this process of deriving knowledge at higher levels of abstraction from the lower levels of abstraction, consolidated over multiple time windows, via the application of some consolidation/mining functions, etc., is called knowledge differentiation [22]and is summarized in figure 1. Figure 1a shows how the general framework of knowledge differentiation, presented in Figure 1b, is instantiated in our work presented in this paper.


Abstraction level
Framework for knowledge differentiation
l
knowledge nuggets at level l
consolidation/mining operation over different time periods
2
knowledge nuggets at level 2
consolidation/mining operation over different time periods
1
knowledge nuggets at level 1
consolidation/mining operation
0
Data
Abstraction level
Instantiation of Knowledge Differentiation Framework
2
Classified Linkages in Tweets
Association rule mining
1
Classified tweets (Fact/Myth/ Neutral /Not Relatable) Time period: Jan 2021-April 2021
Classified tweets (Fact/Myth /Neutral /Not Relatable) Time period: May 2021-August 2021
Classified tweets (Fact/Myth /Neutral /Not Relatable) Time period: Jan 2021-April 2021
Classification
0
Preprocessed tweets

(a) General Template of the Knowledge Differentiation Framework for Facilitating the Discovery and Analysis of Knowledge at Multiple Levels of Abstraction.

(b) Instantiation of the Knowledge Differentiation Framework in the Study.

Figure 1

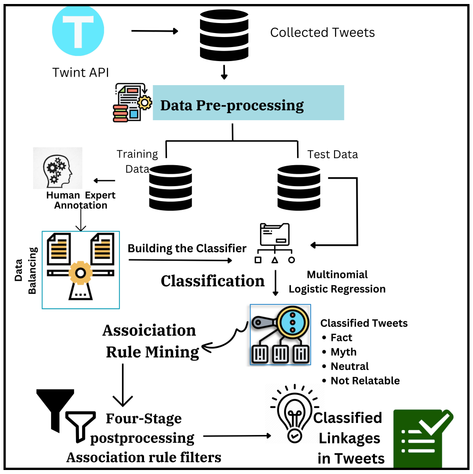


Figure 2. Detailed schematic diagram of the implemented framework based on the principle of Knowledge Differentiation.

Figure 2 shows the detailed schematic diagram of the framework implemented in this paper based on knowledge differentiation. Twitter, a microblogging site, with a very large number of subscribers of different age groups and nationalities, has proven to be a critical resource for researchers, to study the opinions, sentiments, and emotions of people. More importantly, it has also been used to study what people underwent and to draw conclusions about the state and structure of amenities available to people, especially in times of a health crisis such as the COVID-19 pandemic. It is important to note that most of the discourses concerning the state of people are either constrained or controlled. In contrast, Twitter provides a platform for people to express their views without any restrictions. Talking about the menstrual health of women has often been considered a taboo. In times of pandemic, when general healthcare was compromised in most nations, women's health took a definite back seat. With constrained means of approaching medical professionals, a large number of women took to social media platforms to express concerns about their health. It is for this reason that the framework presented in this paper chose to harness the power of these unconstrained communications about women's health from people of all ages, and all nationalities.

The framework uses a blend of supervised predictive techniques such as classification as well as unsupervised, descriptive techniques such as association rule mining. The pre-processed tweets were classified using the best classifier out of the pool of three classifiers, namely, Multinomial Naïve Bayes, Multinomial Logistic Regression, and Passive Aggressive, evaluated with the help of three classifier evaluation metrics (details in Section 4). The MYTH and FACT classes of tweets were further subjected to association rule mining and post-processing rule filters (details in Section 5) to draw out novel, useful and interesting MYTH and FACT classified linkages.

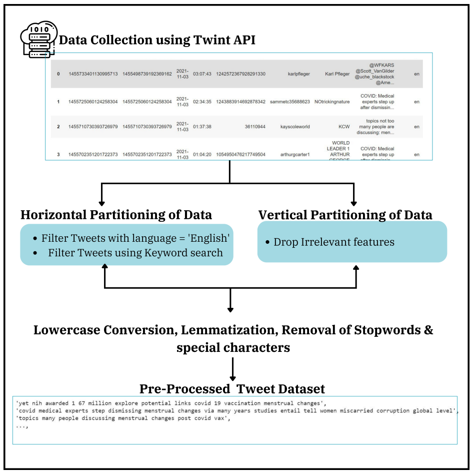
B. Data Details

An advanced open-source python library, called TWINT was chosen for data extraction from Twitter. Twint was preferred over the Twitter API, for its ease of setup and speed. It was used to scrape Twitter data without the restriction of any rate limits and API keys. Pre-processing of raw data puts it in a form that has been treated for missing data, potentially free of noise, subjected to feature selection, etc., and amenable for the application of data mining/ machine learning algorithms to draw out knowledge. Figure 3 depicts the steps taken to pre-process the tweets used in this study, to make them amenable for further study. The process of tweet pre-processing is enumerated as follows. 1. The data was first horizontally partitioned by extracting only the English language tweets, using the following key phrases to extract the tweets related to menses and COVID-19: "menses OR menstrual OR periods AND vaccination OR covid", where OR and AND denote the logical operators. A total of 3399 tweets were chosen for study over three-time widows chosen from a time span of 1st Jan 2021 to 3rd Nov 2021. 2. The data was vertically partitioned i.e., only the relevant columns were retained. Remaining columns such as the irrelevant or redundant columns were dropped for the study. 3. The tweets were converted to lower case. 4. All the punctuations, emoticons, and special characters (@, URLs, #) were removed using the re library. 5. The tweets were tokenized i.e., split into the smaller units called tokens using the function word_tokenize(), to estimate frequencies of different words using data mining models. 6. Lemmatizing aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*. It is different and more powerful than stemming because it uses lexical knowledge bases to find the correct base form of the word. The NLTK class WordNetLemmatizer() was used to perform lemmatization. 7. With the help of the stopword library of NLTK, the stop words, the common words that don't add much value to the text, were removed from the tweets to get the data in the final shape desired for further study.

The framework was implemented using python 3.6.9, on the Google Colaboratory. To store the data extracted from Twitter, google colab was connected to google drive. In the google colab, 41.87 GB of disk space and 1.18 GB of RAM was used. Some of the Python libraries used in the framework were: pandas to work on data frames, nest_asyncio to handle any runtime errors, and sys to manipulate different parts of the runtime environment.

4. DISCOVERING MYTHS OR FACTS ABOUT COVID – 19 RAMIFICATIONS ON MENSTRUAL HEALTH

As discussed earlier, the application of machine learning techniques on the social media messages posted by people can help us uncover important and interesting knowledge, reflecting the unhindered viewpoints, of the people. However, sometimes the misinformation overshadows the utility of social media platforms, especially during the time of


Data Collection using Twint API
Horizontal Partitioning of Data
• Filter Tweets with language = 'English'
• Filter Tweets using Keyword search
Vertical Partitioning of Data
• Drop Irrelevant features
Lowercase Conversion, Lemmatization, Removal of Stopwords & special characters
Pre-Processed Tweet Dataset

0	1455733401130995713	1455498739192369162	2021-11-03	03:07:43	1242572367928291330	karlpfleger	Karl Pfleger	@WFKARS @Scott_VanGilder @uche_blackstock @Ame...	en
1	1455725060124258304	1455725060124258304	2021-11-03	02:34:35	1243883914692878342	sammetc35688623	NOtrickingnature	COVID: Medical experts step up after dismissin...	en
2	1455710730393726979	1455710730393726979	2021-11-03	01:37:38	36110944	kayscoleworld	KCW	topics not too many people are discussing: men...	en
3	1455702351201722373	1455702351201722373	2021-11-03	01:04:20	1054950476217749504	arthurgcarter1	WORLD LEADER 1 ARTHUR	COVID: Medical experts step up	en

```
'yet nih awarded 1 67 million explore potential links covid 19 vaccination menstrual changes',
'covid medical experts step dismissing menstrual changes via many years studies entail tell women miscarried corruption global level',
'topics many people discussing menstrual changes post covid vax',
...,
```

Figure 3. Data Collection and Pre-processing

TABLE I. Count of labelled tweets

Training Data	305
Myth	86
Fact	82
Neutral	67
Not Relatable	70

(a) Training Data

Test Data	131
Myth	33
Fact	30
Neutral	38
Not Relatable	30

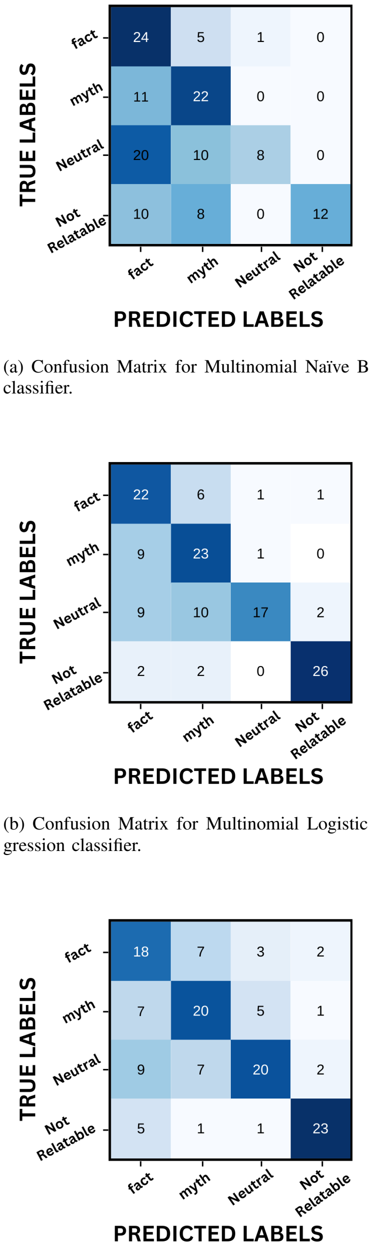
(b) Test Data

a health crisis. For instance, the rumors and fake news circulated about the COVID-19 vaccine ramifications on the menstrual health of women, significantly enhanced vaccine hesitancy and triggered chaos, fueling anxiety and fear. We, therefore, chose to apply machine learning technology, to develop an automated system that can discern false information from real information.

For this purpose, the tweets, pre-processed as detailed in Section 3, were subjected to the task of classification, to study the myths and facts that were being circulated on the social media platforms, about women's mensural health during the COVID-19 pandemic. To facilitate the supervised learning process of classification, out of the 3399 pre-processed tweets, 700 were initially annotated by human experts, chosen anonymously. Interestingly, the criteria for choosing the class labels emerged from the examination of the tweets during the process of annotation. Another advantage of involving manual annotation was that it became clear that many pre-processed tweets contained references to the keyword 'period', which is also a synonym of mensuration, as 'time period'. Hence these were labeled as NOT RELATABLE. At the same time, there were many tweets that merely posed queries on the relationship between the terms 'covid', 'vaccine', and 'menses'. Since these tweets did not state any facts and did not represent the opinion of people on the topic concerned, we chose to label them as NEUTRAL. The tweets that actually represented the opinion of people, were annotated as either MYTHS or FACTS, based on the knowledge of the human annotators, in line with the subject of this paper. Hence though theoretically the problem was to perform bi-classification, it was decided to treat it as a four-class problem, with the class labels: FACT, MYTH, NEUTRAL, NOT RELATABLE. The annotated tweets were later partitioned into the training and the test set. The third advantage that surfaced by deploying human expert annotation was the discovery that the tweets in the training set were not balanced. The number of tweets labeled as MYTH was only 90, which was quite less than the representation of the other classes in the data set. Thus, the tuples for supervised learning (forming the training and test set) were chosen again. This time 436 tweets were annotated and 2963 tweets were left in the untrained dataset. Following class representations were found in the supervised set: class Myth: 119 tweets, class Fact: 112 tweets, class Neutral: 105 tweets, and class Not Relatable: 100 tweets. The annotated data was partitioned into the training and test datasets, for training the classifiers, as shown in Tables Ia and Ib respectively.

A. *Implementation*

Three algorithms, namely, Multinomial Naïve Bayes, Multinomial Logistic Regression, and Passive Aggressive were implemented for the task of classification. A Naive Bayes classifier is a probabilistic classifier, that uses the conditional probability (Bayes Theorem) to classify the tuples, based on the set of features extracted from tweets. The naïve keyword is attributed to the assumption that all the features are considered to be conditionally independent of each other. Logistic Regression, based on Bernoulli's distribution, a subset of the binomial pdf with one as the binomial denominator, is the most common method used to model binary response data. Multinomial Logistic Regression is an extension of the Logistic Regression classifier. The algorithm can be used for multiclass classification problems by splitting them into multiple binary class classification problems. We exploit the One-vs-Rest model in this paper, for performing multi-label classification using Logistic Regression. The Passive-Aggressive algorithm has been known for the classification of large datasets. The classifier, rather than using the dataset as a whole, takes in one piece at a time, adjusting the weights of its model based on each entry's results. For implementation of the classification, the tweets, pre-processed and annotated, as detailed in Section 3B, were subjected to feature extraction with the help of package TfidfVectoizer from sklearn library. TfidfVectorizer (Term Frequency - Inverse Document Frequency) is a class used for feature extraction and to calculate the frequency of each feature, considering their weightage in the whole document. Tweets were converted to the matrix of normalized fractional feature count because all the classifiers are supposed to be initialized on the feature count. The classifiers were initialized on a fractional feature count of the annotated tweets and the function 'TfidVectorizer' was used for feature extraction. train_text_split method was applied on the matrix obtained from feature extraction, to



(a) Confusion Matrix for Multinomial Naïve Bayes classifier.

(b) Confusion Matrix for Multinomial Logistic Regression classifier.

(c) Confusion Matrix for Passive Aggressive classifier .

Figure 4

divide the dataset into training and test datasets. Once the training and test set were ready, the respective libraries, for the chosen classifiers, were imported from the Sci-kit learns package. MultinomialNB, PassiveAggressiveClassifier, LogisticRegression classes were used to implement the Naive Bayes, Passive Aggressive, and Logistic Regression algorithms respectively. For instance, to implement

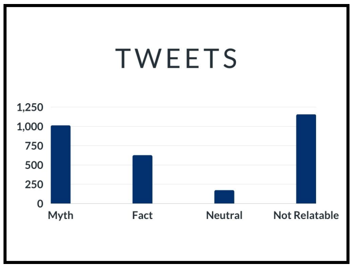
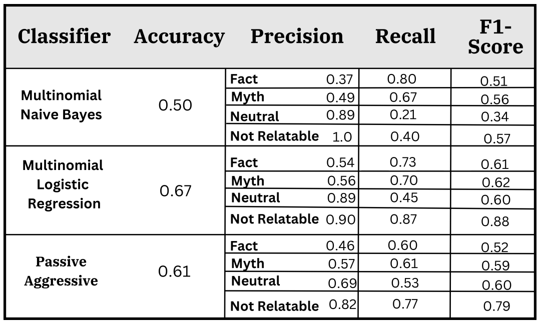


Figure 5. Tweet classification results obtained by implementing Multinomial Logistic Regression Algorithm

TABLE II. Performance evaluation for selecting the classifier



Classifier	Accuracy	Precision		Recall	F1-Score
Multinomial Naive Bayes	0.50	Fact	0.37	0.80	0.51
		Myth	0.49	0.67	0.56
		Neutral	0.89	0.21	0.34
		Not Relatable	1.0	0.40	0.57
Multinomial Logistic Regression	0.67	Fact	0.54	0.73	0.61
		Myth	0.56	0.70	0.62
		Neutral	0.89	0.45	0.60
		Not Relatable	0.90	0.87	0.88
Passive Aggressive	0.61	Fact	0.46	0.60	0.52
		Myth	0.57	0.61	0.59
		Neutral	0.69	0.53	0.60
		Not Relatable	0.82	0.77	0.79

Multinomial Logistic Regression, LogisticRegression function of scikit.learn library was applied with parameters : multi_class ='multinomial', solver='lbfgs'. The sklearn. metric module was used to derive the evaluation metrics: precision, recall, and F1-score.

B. Results

Figures 4a, 4b and 4c depict the confusion matrix of the three classifiers. A comparative analysis of the performance of the chosen classifiers on the test set, evaluated based on the four metrics, namely accuracy, precision, recall and F1-score is presented in Table II. Results show that Multinomial Logistic Regression depicts high accuracy as compared to Multinomial Naïve Bayes and Passive-Aggressive classifier. Multinomial Logistic Regression depicts high precision as well for all the classes except the MYTH class.

Recall metric evaluates the percentage of true positive tweets with respect to the sum of true positives and false negatives and offers a way to measure the percentage of predicted true positives w.r.t. to the actual number of positives. It is applied in cases where the cost of false negatives is high. In the present case, it was chosen as an evaluation metric to safeguard the tweets actually belonging to the MYTH or FACT class being labeled as 'NOT A MYTH' or 'NOT A FACT'. Multinomial Naïve Bayes depicted a better recall value for the class MYTH as compared to the Multinomial Logistic Regression, so we considered the results of the F1 score metric to break the tie between Multinomial Naïve Bayes and Multinomial Logistic Regression classifier. The F1-score is another metric that is more appropriate in situations where it is essential to guard against false negatives and false positives are more crucial than true negatives and true positives. Table II clearly indicates a better F1-score for the Multinomial Logistic Regression classifier than all other classifiers. This confirmed our choice of the Multinomial Logistic classifier. We concluded therefore that, in the current scenario, the Multinomial Logistic Regression classifier would serve our purpose better. Figure 5 plots the results obtained by implementing Multinomial Logistic Regression algorithm over the untrained, pre-processed tweets. 1011 tweets were classified as MYTH, 625 tweets were classified as FACT, 172 tweets were classified as NEUTRAL, and 1154 tweets were classified as NOT RELATABLE.

C. Inferences

The Following inferences could be drawn from the above results:

1) Results show that the number of myths circulating on social media is almost double the number of facts. This may lead to the conclusion that there are a large group of people who were misled or have misconceptions about the effect of the covid vaccine on menstruation. This result also shows how the efficacy of a social media platform that has been used in critical times to disseminate essential facts, has been undermined by people who spread fictitious information.

2) To draw more detailed, accurate inferences, we studied some of the tweets classified under each category randomly. Consider the following tweets classified as MYTH: “scientific evidence covid19 vaccines since caused infertility none planning get pregnant whether pregnant breastfeeding receive vaccine turn”, “gop took another level covid one fave claim non vaccinated woman menstrual cycle could impacted spending time vaccinated person”, “unfortunately entirely correct second point covid vaccines affect fertility appear many instances short term effects menstrual cycles think important menstruators”, and the following tweets classified as FACT: “reported changes menstrual cycle vaccination short-lived people report change find returns normal following cycle studies potential impacts vaccination menstruation underway”, “read studies research amp definitely thing amongst many women enough data say yes covid vaccine causes temporary menstrual irregularities lot women going thru period side effects lol ig see couple years”, “1 3rd 12 300 syrvet respondents said experienced changes menstrual cycle changes menstrual symptoms covid 19 pandemic survey included perceived stress scale assess impact stress asrm2021”. It can be seen that there is a very thin line separating the tweets classified as MYTH and those classified as FACT by the Multinomial Logistic Regression classifier. We discovered that most of the tweets classified as FACT supported the claim that there might be a temporary, small impact on the women’s menstrual cycle due to the COVID-19 vaccines. We could also infer that due to a very thin line separating the claims made by tweets classified as MYTH and those classified as FACT. Apparently, people distorted the FACTs themselves to spread the fake news about the effect of the vaccine on menstruation.

3) It is also noteworthy here that the large number of tweets classified as NOT RELATABLE in the results strengthened our choice of converting a bi-class problem into a multi-class problem. The NEUTRAL and NOT RELATABLE tweets would have otherwise been involuntarily classified as either a MYTH or a FACT. The following tweets should help clarify this stance. Consider the tweets placed in the category NOT RELATABLE: ”friends seeing current covid situation Indonesia seems passed one bad periods lifetime hope everyone family healthy could endure tough time hopefully bright days come”, ”really think covid showed us work collective heal nature even short periods remember posts like month lot people lockdown globally”, and the following tweets placed in the class NEUTRAL: ”going discuss covid vaccines changes menstrual cycle hope ok today x”, ”topic need research covid vaccine affecting women periods nice time atm”.

5. DISCOVERY OF ASSOCIATION RULES

Mining knowledge at higher levels of abstraction provides a concise, consolidated view of looking at the discovered information, that enables better understanding, comprehension, and actionability. We mine association rules, that depict knowledge at abstraction level two, from each category of tweets classified by the Multinomial Logistic Regression classifier, representing knowledge at abstraction level one (Section 4). Association rule mining is a classic descriptive data mining task that is employed to discover hidden linkages between the data items [43]. The linkages mined by the presented framework represent possibly novel and interesting relationships from classified tweets, which allows us to draw consolidated inferences about selected classes of tweets. Since the NEUTRAL and NOT RELATABLE tweets were logically irrelevant for the task at hand, these tweets were not considered for mining associations. This decision was also taken for optimizing processing time, space, and computation efficiency. Also, the number of rules produced by the framework, presented to the domain user for analysis, was reduced by more than half. Since only relevant rules were analyzed, the time of the post-processing phase was substantially reduced and efficiency improved.

A. Implementation

The Apriori algorithm was used to mine the association rules [44]. The algorithm relies on the Apriori property, also known as the anti-monotone property, which states that all non-empty subsets of a frequent itemset must be frequent to prune the search space of all itemsets. Apriori algorithm has been considered a milestone algorithm for association rule mining and has inspired a plethora of algorithms for association discovery. It was chosen for implementation in the current framework because of its inherent simplicity. However, more sophisticated rule discovery algorithms are now available and can be used to improve upon the performance of Apriori in the future work.

Before the application of the mining algorithm, a preprocessing base item filter was set to only include the items related to the synonyms of the word 'periods', 'COVID-19', 'Vaccination', and some major menstrual symptoms, etc. This was done to focus the mining and limit the search space. For mining association rules, the Mlxtend (machine learning extension) library of python was installed. TransactionEncoder, apriori and association_rules classes were imported from Mlxtend library. TransactionEncoder is used to transform the python list of lists into a one-hot encoded Numpy boolean array. The fit method of TransactionEncoder learns the different labels in the list and the transform method converts the list into a NumPy array. This conversion is necessary because apriori only works on one-hot encoded data frames.

For mining, a data frame was initially created for each category of classified tweets i.e., FACT, MYTH, NEUTRAL, and NOT RELATABLE (Section 4). Each tweet in the data frame was tokenized, to convert it into a list of words. These list values were then used to form a one-hot encoded data frame using the TransactionEncoder. The apriori class was next initialized on these data frames with the minimum support of 0.3 or 30% to get the frequent itemsets. Finally, association rules were generated with confidence of 0.7 (70%) and were sorted with the lift value greater than 1.0. The support and confidence of the rules were later diminished over several iterations to find interesting rules with items that belonged to the preprocessing filter. The process of diminishing was halted at the minimum support as 0.04 and minimum confidence as 0.1. In the post-processing phase, the discovered rules were analyzed based on rule metrics such as confidence and lift. The more interesting rules were retained. The less interesting ones were filtered out.

B. Results

A total of 15272 rules were discovered, which were then subjected to the postprocessing phase.

Let DC denote the set of discovered rules for a category c where c = FACT, MYTH, NEUTRAL, NOT RELATABLE. During the post processing phase, the following filters were applied to the sets D_c ∀ c.

1) Template-Based Rule Filters: Multiple rule templates were developed to limit the search space i.e., to only retain the rules interesting enough to be put through further post processing filters and analyzed. The retained rules were added to a set of Interesting Rules I_c, ∀ c and passed on metric filters. The pruned rules were put in a set of rules P_c, ∀ c.Let a discovered rule be of the form: A → C ∈ D_C, where A is the antecedent and C is the consequent.

```
For all (A → C ∈ $D_C$) If (A, C ∈ {synonyms of the word 'periods', 'COVID-19', 'Vaccination' and some major menstrual symptoms}) then
        $I_C = I_C \cup A \rightarrow C$
Else
        $P_C = P_C \cup A \rightarrow C$
Endif
EndFor
$D_c = I_c$ // Assign the set of interesting rules to $D_c$, and pass it to the next stage of postprocessing.
$I_c = \emptyset$.
```

2) Metric Filters: The interesting template-based rules within the desired confidence and lift metric threshold limits were retained. The rest were filtered out and added to the set of pruned rules P_c.

```
For all (A → C ∈ $D_C$)
If (confidence (A → C) ≥ conf_threshold ∧ lift (A → C) ≥ lift_threshold)
then
        $I_C = I_C \cup A \rightarrow C$
Else
        $P_C = P_C \cup A \rightarrow C$
Endif
EndFor
$D_c = I_c$
$I_c = \emptyset$
```

3) Redundant Rule Filter: Some rules, even though they satisfy the minimum metric threshold requirements prove to be logically redundant, if their superset is already present in the ruleset.

```
For all (A → C ∈ $D_C$)
If (∃ A' → C' s.t. A ⊆ A' ∨ C ⊆ C') then
        $P_C = P_C \cup A \rightarrow C$
Else
        $I_C = I_C \cup A \rightarrow C$
Endif
EndFor.
```

4) Human Analytic Filters: The rule set I_C satisfying the above three filters as well as the set of pruned rules P_C were given to domain experts for manual analysis and filtering out of rules which were non interesting. The experts also added some rules ∈ P_C that still seemed interesting but had been pruned initially for failing to meet the metric filters back to I_C. The rationale was that there were some words that were interesting enough for the problem at hand but had dropped low on metric thresholds as their synonyms were used in conversations.

C. Inferences

All the rules $\in I_c$, c = {FACT, MYTH, NEUTRAL, NOT RELATABLE} were analyzed to draw inferences. We present inferences drawn from some interesting rules belonging to FACT and MYTH classes.

1) Association Facts

A total of 10,743 association rules were discovered from the FACT class, which were put through the postprocessing filters described above. We present some rules from the FACT class, with interesting and significant inferences in this subsection.

- Let us consider a set of rules, with positively correlated antecedent and consequence, indicated by a lift greater than 1 (Table III). Rule 592 indicates that the word covid was in 6% of the tweets, associated with the terms menstrual pain. Rule 593 drawn from the same itemset, refines the knowledge to conclude that about 80% of the times the word pain was used in the tweets it was used to refer to the menstrual pain due to or during covid. This is a very significant revelation.

- Consider the rules 597, 598, and 599, drawn from the itemset heavy, menstrual, covid, shown in Table IV. The support of all three rules is the same. The lift of all three rules being greater than one indicates a positive correlation between the rule antecedent and the rule consequent. However, the confidence of rule 598 indicates that 60% of the time word ’heavy’ was used, it was used to refer to heavy menstrual flow during or due to COVID. The confidence of Rule 597 indicates that whenever the word menstrual was cited in the tweets, it was associated with many other words out of which approximately 8% of times it was used with the words heavy and covid. Confidence of Rule 599 suggests around 6% of the time, the word covid was used in tweets, it was used to refer to the heavy menstrual flow, which is understandable since the usage of the word covid is very high and has been used as a topic in many discussions. For drawing conclusions related to menstrual flow, however, it is apparent that Rule 598 seems a more appropriate choice, as compared to Rules 597 and 599.

- Similarly, the rules 283, 284, 285 have same support (Table V). However, rule 283 has higher confidence as compared to the other two rules, implying that 67% of times when a tweet talks about bleeding and covid, the term heavy was present. It is interesting to note that though this is high confidence, positively correlated rule (lift > 1), it is logically redundant. It uncovers the same linkage, discovered from rules 597, 598, and 599, discussed above, between the words covid, heavy and menstrual, implying that due to COVID complications, women experienced heavy bleeding. This is largely because of the use of synonyms menstrual and bleeding. This leads us to an important inference that the support of an item is not truly indicative of its popularity in such cases. We used this rationale to search for even lower support rules than initially envisaged. One solution to this problem caused by synonyms could be to replace all the synonyms of a particular word with the word itself.

- We came across another high confidence rule, with a very significant observation (Table VI). Lift >1 indicates the antecedent and consequent are positively correlated. The rule infers that covid has led to stress. Given, the context of tweets to be the menstrual problems, it would not be wrong to refine the inference to the fact that indeed many menstrual problems were caused by COVID related stress. Even medical studies have established that stress, anxiety and fear caused due to COVID, left some women with severe menstrual disorders akin to those faced by women during the times of war or natural calamities [8].
  Table VII shows that Rule 64 has much higher confidence as compared to rule 65, and indicates that 86% of the times when the word covid appears in the tweets, the word irregular was also present. The inference is that in 86% of tweets talking about covid, people have referred to an irregular period.

- Table VIII shows a low confidence Rule 923, that points out that the impact of COVID-19 vaccines on the women’s periods has also been the center of conversation of a few tweets.

- Rule 10716 with 100% high confidence (Table IX), is another rule that led to the implication of an assertion that COVID experts dismissed the menstrual changes in women. The study revealed that people tweeted the messages supporting the above rule in rage. Even though the changes in the periods due to COVID or after the vaccination have been largely temporary, people wanted to demand an in-depth study of the long-term changes, if any.

The above discussions clearly show the superiority of discovering association rules over just mining the top frequent words say via techniques such as clustering. As discussed earlier, the output of such methods is generally limited to a set of popular words. However, in association rule mining, we have itemsets of various sizes, which can be analyzed and leveraged. The ability to analyze the rules bidirectionally, and vis. a vis. the metric thresholds, as demonstrated by the above results lead to the discovery of more advanced and structured knowledge.

2) Association Myths

813 rules were discovered from MYTH class. In this subsection, we discuss some rules drawn from MYTH class, which led to interesting inferences.

- Rule 228 (Table X) has lift > 1 and confidence of

TABLE III. Confidence and Lift Metrics for Rules 592,593

Rule No	Antecedent	Consequent	Confidence	Lift
592	({‘pain’})	({’menstrual’, ’covid’})	0.861111	1.764572
593	({‘covid’})	({’menstrual’, ’Pain’})	0.062124	1.252505

TABLE IV. Confidence and Lift Metrics for Rules 597-599

Rule No.	Antecedent	Consequent	Confidence	Lift
597	({menstrual’})	({‘heavy’,’covid’})	0.08	1.35
598	({‘heavy’})	({‘menstrual’,’covid’})	0.61	1.25
599	({‘covid’})	({‘menstrual’,’heavy’})	0.06	1.18

TABLE V. Confidence and Lift Metrics for Rules 283-285

Rule No.	Antecedent	Consequent	Confidence	Lift
283	({’bleeding’, ’covid’})	({’heavy’})	0.67	8.61
284	({’heavy’,’covid’})	({’bleeding’})	0.65	7.61
285	({’bleeding’})	({’heavy’, ’covid’})	0.46	7.61

TABLE VI. Confidence and Lift Metrics for Rule 93

Rule No.	Antecedent	Consequent	Confidence	Lift
93	({’covid’})	({’stress’})	0.86	1.25

TABLE VII. Confidence and Lift Metrics for Rules 64,65

Rule No.	Antecedent	Consequent	Confidence	Lift
64	({‘irregular’ })	({‘covid’})	0.86	1.07
65	({‘covid’})	({‘irregular’})	0.05	1.07

TABLE VIII. Confidence and Lift Metrics for Rule 923

Rule No.	Antecedent	Consequent	Confidence	Lift
923	({’vaccine’})	({’periods’,’women’})	0.12	1.06

TABLE IX. Confidence and Lift Metrics for Rules 10685, 10716, 10719

Rule No.	Antecedent	Consequent	Confidence	Lift
10685	({’women’, ’menstrual’, ’medical’})	({’changes’, ’experts’, ’dismissing’, ’step’, ’covid’})	0.87	7.96
10716	({’women’, ’dismissing’}) ({’changes’, ’experts’, ’step’, ’menstrual’, ’medical’, ’covid’})	1	8.92	
10719	({’women’, ’medical’})	({’changes’, ’experts’, ’dismissing’, ’step’, ’menstrual’, ’covid’})	0.87	7.96

TABLE X. Confidence and Lift Metrics for Rule 228

Rule No.	Antecedent	Consequent	Confidence	Lift
228	({’vaccin ated’})	({‘long’, ‘effects’})	0.38	3.10

TABLE XI. Confidence and Lift Metrics for Rule 420

Rule No.	Antecedent	Consequent	Confidence	Lift
420	({’evidence’, vaccine’}) ({’infertility’})	0.9	1.45	

TABLE XII. Confidence and Lift Metrics for Rule 802

Rule No.	Antecedent	Consequent	Confidence	Lift
802	({’may’, ’covid’})	({’infertility’, ’cause’, ’vaccine’})	0.66	4.32

TABLE XIII. Confidence and Lift Metrics for Rule 1380

Rule No.	Antecedent	Consequent	Confidence	Lift
1380	({’covid’, ’ivermectin’}) ({’infertility’})	0.9	1.62	

38%, which shows that 38% of times itemset {long, effects} appeared, it was teamed up with the itemset {vaccinated}. This rule clearly states that people believed vaccinated population will show some long-term effects. Since the context of discussions was menstrual disorders, we can refine the inference to conclude that vaccine will have long-term effects on the menstrual cycle. A study of the medical papers and news articles confirmed that the rule was indeed a MYTH because most of the cases, where vaccine has shown some effects on menstrual cycle, have only been temporary [11], [45].

- A very strong rule 420 (high confidence of 90%), with positively correlated antecedent and consequent (lift > 1), was classified by our framework as a MYTH (Table XI). It could be inferred from the rule that there is evidence that vaccine causes infertility. A further study to investigate the truth of rule 420, led to the discovery that people had confused the spike protein involved in the growth and formation of the placenta with the spike protein on the surface of the covid virus. As the covid vaccine breaks this protein, people concluded that it would also shed or stop the formation of the placenta, thus leading to infertility [44]. It is noteworthy here that it is very important to flag off such strong rules as myths, a feat well accomplished by our framework, in order to curb vaccine hesitancy and to drive away the emotions of anxiety, stress and fear from the society.

- Rule 802, though with confidence less than rule 420 (66%) and (lift > 1), also suggested the role of covid vaccines in causing infertility (Table XII). It was also flagged off as a MYTH.

- Another very strong rule 1380 (high confidence of 90%), with positively correlated antecedent and consequent (lift > 1), was classified by our framework as a MYTH (Table XIII). On inspection, it was found that the rule was similar to rule 420, except that it pointed towards a specific medicine Ivermectin, which wasn’t developed as a covid vaccination, but was used in its treatment. On investigation, we found that some people spread the rumor about infertility caused by ivermectin, on the basis of some study conducted in 2011[36].

6. CONCLUSIONS AND FUTURE DIRECTIONS

The work presented in this paper is an attempt to highlight the need for research on COVID-19 ramifications on women’s mensural health, especially by the machine learning and data mining community. The idea is to appeal to the machine learning community to catch up and complement the efforts of the medical community in this important public health initiative by designing algorithms and frameworks that lend the power of wrangling large amounts of data, pattern detection, and analysis of data, possibly from multiple perspectives. A knowledge differentiation-based framework has been presented in the current work that discovers knowledge and provides insights into the discovered knowledge at different levels of abstraction using the classical data mining tasks of classification and association rule mining. The knowledge so discovered was also subjected to postprocessing filters that aided in the goal of pruning irrelevant, uninteresting, and duplicate rules.

Some very important and interesting conclusions and patterns were drawn from the work. The number of neutral and not relatable tweets also roughly amounted to 80% of the number of tweets containing facts and myths. It was discovered that the number of myths circulating on social media is almost double the number of facts. It could also be analyzed that most of these myths were derived by slightly changing the facts to make them appear more realistic. Associations depicting myths revealed by the implemented system, concerning the long-term effects of covid, its vaccines, and their ramifications on fertility could very well explain the reasons behind slow vaccine acceptance and resistance. The role of covid-related stress in causing menstrual problems was corroborated by high confidence, high conviction rules. Associations depicting facts about women reporting irregular, painful, and heavy periods due to COVID-19 were discovered, which set the platform for the study of long-term ramifications of the pandemic on women’s menstrual health, thus justifying the call for greater research on the slated topic. In fact, a very high confidence, high conviction rule revealed people’s outrage and demand for an in-depth study of the long-term changes in the menstrual health of women.

References

[1] R. Nagarajan, Y. Krishnamoorthy, V. Basavarachar, and R. Dakshinamoorthy, "Prevalence of post-traumatic stress disorder among survivors of severe covid-19 infections: A systematic review and meta-analysis," *Journal of Affective Disorders*, vol. 299, 2022.

[2] L. Trogstad, "Increased occurrence of menstrual disturbances in 18- to 30-year-old women after covid-19 vaccination," 2022.

[3] A. Zervopoulos, A. G. Alvanou, K. Bezas, A. Papamichail, M. Maragoudakis, and K. Kermanidis, "Hong kong protests: Using natural language processing for fake news detection on twitter," *Artifical Intelligence Applications and Innovations*, 2020.

[4] L. L. Carli, "Women, gender equality and covid-19," *Gender in Management*, vol. 35, pp. 647–655, 2020.

[5] T. L. Holbrook, D. B. Hoyt, M. B. Stein, L. Han, and S. W. J, "Gender differences in long-term posttraumatic stress disorder outcomes after major trauma: Women are at higher risk of adverse outcomes than men," *Journal of Trauma Acute Care*, vol. 53, pp. 882 – 888, 2002.

[6] A. V. Mattioli, S. Sciomer, S. Maffei, and S. Gallina, "Lifestyle and stress management in women during covid-19 pandemic: Impact on cardiovascular risk burden," *Am J Lifestyle Med*, vol. 15, 2021.

[7] I. Sandanger, J. F. Nygård, S. T, and M. T, "Is women's mental health moresusceptible than men's to the influence of surroundingstress?" *Soc Psychiatry Psychiatr. Epidemiol*, vol. 39, 2004.

[8] N. Ozimek, K. Velez, H. Anvari, L. Butler, K. N. Goldman, and N. C. Woitowich, "Impact of stress on menstrual cyclicity during the coronavirus disease 2019 pandemic: A survey study," *Journal of Womens Health*, vol. 31, 2022.

[9] E. Young and A. Korszun, "Psychoneuroendocrinology of depression: Hypothalamic-pituitary-gonadal axis," *Psychiatrc Clinics of North America*, vol. 21, 1998.

[10] N. Yunitri, H. Chu, X. L. Kang, H.-J. Jen, L.-C. Pien, H.-T. Tsai, A. R. Kamil, and K.-R. Chou, "Global prevalence and associated risk factors of posttraumatic stress disorder during covid-19 pandemic: A meta-analysis," *International Journal of Nursing Studies*, vol. 126, 2022.

[11] A. Edelman, E. R. Boniface, E. Benhar, L. Han, M. K. A, C. Favaro, J. T. Pearson, and D. B. G, "Association between menstrual cycle length and coronavirus disease 2019 (covid-19) vaccination: A u.s. cohort," *Obstetrics and Gynecology*, vol. 139, pp. 481–489, 2022.

[12] V. Male, "Effect of covid-19 vaccination on menstrual periods in a retrospectively recruited cohort," *medRxiv 2021.11.15.21266317*, 2021.

[13] G. C. Sharp, A. Fraser, G. Sawyer, G. Kountourides, K. E. Easey, G. Ford, Z. Olszewska, L. D. Howe, D. A. Lawlor, A. Alvergne, and J. A. Maybin, "The covid-19 pandemic and the menstrual cycle: research gaps and opportunities," *International journal of epidemiology*, vol. 51, 2021.

[14] S. Madichetty and S. M, "A stacked convolutional neural network for detecting the resource tweets during a disaster," *Multimedia Tools and Applications*, vol. 80, pp. 3927 – 3949, 2021.

[15] C. Monica and N. Nagarathna, "Detection of fake tweets using sentiment analysis," *SN Computer Science*, vol. 1, 2020.

[16] E. L. Warner, J. L. Barbati, K. L. Duncan, k. Yan, and S. A. Rains, "Vaccine misinformation types and properties in russian troll tweets," *Vaccine. 40*, 2022.

[17] T. Balasubramaniam, R. Nayak, K. Luong, and M. A. Bashar, "Identifying covid-19 misinformation tweets and learning their spatio-temporal topic dynamics using nonnegative coupled matrix tensor factorization," *Social Network Analysis and Mining*, vol. 11, p. 57, 2021.

[18] R. K. Kaliyar, A. Goswami, and P. Narang, "Echofaked: improving fake news detection in social media with an efficient deep neural network," *Neural Computing and Applications*, vol. 33, pp. 8597 – 8613, 2021.

[19] L. Cao, "Actionable knowledge discovery and delivery," *WIREs Data Mining Knowledge Discovery*, vol. 2, pp. 149–163, 2012.

[20] L. Cao, D. Luo, and C. Zhang, "Knowledge actionability: Satisfying technical and business interestingness," *International journal of Business Intelligence and Data Mining*, vol. 2, pp. 496–514, 2007.

[21] N. Kalanat, "An overview of actionable knowledge discovery techniques," *Journal of Intelligent Information Systems*, vol. 58, pp. 591 – 611, 2022.

[22] V. Bhatnagar and S. Kochchar, "Modeling support changes in streaming item sets," *International Journal of Systems Science*, vol. 37, pp. 879–891, 2006.

[23] A. Al-Rawi, K. Grepin, X. Li, R. Morgan, C. Wenham, and J. Smith, "Investigating public discourses around gender and covid-19: a social media analysis of twitter data," *Healthcare Informatics Research*, vol. 5, pp. 249–269, 2021.

[24] G. Ifrim, B. Shi, and I. Brigadir, "Event detection in twitter using aggressive filtering and hierarchical tweet clustering," *CEUR Workshop Proceedings*, vol. 1150, pp. 33 – 40, 2014.

[25] K. Lee, D. Palsetia, R. Narayanan, M. M. M. Patwary, A. Agrawal, and C. A, "Twitter trending topic classification," *EEE 11th International Conference on Data Mining Workshops*, pp. 251–258, 2011.

[26] A. Sechelea, T. D. Huu, E. Zimos, and N. Deligiannis, "Twitter data clustering and visualization," *International Conference on Telecommunications (ICT)*, vol. 39, 2016.

[27] Y.-J. Lu and C.-T. Li, "Graph-aware co-attention networks for explainable fake news detection on social media," pp. 505 –514, 2020.

[28] S. Krishnan and A. Chen, "Identifying tweets with fake news," *IEEE International Conference on Information Reuse and Integration (IRI)*, vol. 33, pp. 460 – 464, 2018.

[29] E. Xiao and M. Ferin, "Stress-related disturbances of the menstrual cycle," *Annals of Medicine*, vol. 29(3), 1997.

[30] D. Kar, M. Bhardwaj, S. Samanta, and A. P. Azad, "No rumours please! a multi-indic-lingual approach for covid fake-tweet detection," *Grace Hopper Celebration India (GHCI)*, 2020.

[31] L. Zheng, C. Shen, L. Tang, C. Zeng, T. Li, S. Luis, and S. C. Chen, "Data mining meets the needs of disaster information management," *IEEE Transactions on Human-Machine Systems*, vol. 43, 2013.

[32] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news

detection in social media with a bert-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, pp. 11 765 – 11 788, 2021.

[33] S. Kula, M. Choraś, R. Kozik, P. Ksieniewicz, and M. Woźniak, "Sentiment analysis for fake news detection by means of neural networks," *Computaional Science- ICCS 2020*, vol. 33, pp. 653 – 666, 2020.

[34] J. A. Diaz-Garcia, C. Fernandez-Basso, M. Dolores Ruiz, and M. J. Martin-Bautista, "Mining text patterns over fake and real tweets," *Journal of Healthcare Informatics Research Information Processing and Management of Uncertainty in Knowledge-Based Systems*, vol. 1238, pp. 648–660, 2020.

[35] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 31, 2010.

[36] M. Cataldi, L. D. Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," *MDMKDD: Proceedings of the Tenth International Workshop on Multimedia Data Mining*, 2010.

[37] M. Adedoyin-Olowe, M. M. Gaber, and F. Stahl, "A survey of data mining techniques for social network analysis," *Data Mining Digital Humanities*, p. 5, 2013.

[38] C.-H. Chang, M. Monselise, and C. C. Yang, "What are people concerned about during the pandemic? detecting evolving topics about covid-19 from twitter," *Journal of Healthcare Informatics Research*, vol. 5, pp. 70–97, 2021.

[39] W. Ahmad, B. Wang, H. Xu, M. Xu, and Z. Zeng, "Topics, sentiments, and emotions triggered by covid-19-related tweets from iran and turkey official news agencies," *SN Comput. Sci*, vol. 2, p. 394, 2021.

[40] A. S. Laganà, G. Veronesi, F. Ghezzi, M. M. Ferrario, A. Cromi, M. Bizzarri, S. Garzon, and M. Cosentino, "Evaluation of menstrual irregularities after covid-19 vaccination: Results of the mecovac survey," *Open Meds (Wars)*, vol. 17, pp. 475–484, 2022.

[41] A. Varshney, Y. Kapoor, V. Chawla, and V. Gaur, "A novel framework for assessing the criticality of retrieved information," *Interntional journal of Computing and digital Systems*, vol. 11, 2021.

[42] A. Alvergne, G. Kountourides, M. A. Argentieri, L. Agyen, N. Rogers, D. Knight, G. C. Sharp, J. A. Maybin, and Z. Olszewska, "Covid-19 vaccination and menstrual cycle changes: A united kingdom (uk) retrospective case-control study," 2021.

[43] V. Bhatnagar and S. Kaur, "Association rule mining," 2008.

[44] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules."

[45] S. Taylor, "Clinical and nosological considerations," *Current psychiatry report*, vol. 23, 2021.

Dr. Sarabjeet Kaur Kochhar was awarded her Ph.D. Degree in Computer Science from Department of Computer Science, University of Delhi, Delhi, India. She is an Associate Professor in the Department of Computer Science, Indraprastha College for Women, University of Delhi, Delhi, India, with over twenty years of teaching experience. She has published extensively in international journals and conferences. Her research interests are currently aligned along the fields of Data Mining, Data Analytics, and Natural Language Processing.

Ms. Rumjot Kaur is completed B.Sc. Hons. in Computer Science from Indraprastha College for women, University of Delhi, Delhi, India.Her research interests span the fields of Data Science and Machine learning.