International Journal on Semantic Web and Information Systems
Volume 18 • Issue 1

Hybrid Firefly-Ontology-Based Clustering Algorithm for Analyzing Tweets to Extract Causal Factors

Akilandeswari J., Department of IT, Sona College of Technology, India

Jothi G., Department of Computer Applications, Sona College of Arts and Science, India

Dhanasekaran K., Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Kattankulathur, India

Kousalya K., Department of CSE, Kongu Engineering College, India

Sathiyamoorthi V., Department of CSE, Sona College of Technology, India

ABSTRACT

Social media, especially Twitter, has become ubiquitous among people where they express their opinions on various domains. This paper presents a hybrid firefly ontology-based clustering (FF-OC) algorithm that attempts to extract factors impacting a major public issue that is trending. In this research work, the issue of food price rise and disease, which was trending during the time of the investigation, is considered. The novelty of the algorithm lies in the fact that it clusters the association rules without any prior knowledge. The findings from the experimentation suggest different factors impacting the rise of price in food items and diseases such as diabetes, flu, zika virus. The empirical results show the significant improvement when compared with artificial bees colony, cuckoo search algorithm, particle swarm optimization, and ant colony optimization-based clustering algorithms. The proposed method gives an improvement of 81% in terms of DB index, 79% in terms of silhouette index, 85% in terms of C index when compared to other algorithms.

KEYWORDS

Association Rules, Firefly, Social Media, Swarm Intelligence, Twitter, Wordnet

1. INTRODUCTION

In the recent past, social media analytics has gained tremendous momentum due to the vast amount of user-generated data. This data explosion is a gold mine for researchers to perform various analytics to extract insights on different aspects. People vent out their opinions and feelings about diverse topics through social networking sites such as Twitter and Facebook. Globally, Twitter has 330 million monthly active users and 145 million daily active users (Murthy, 2018). There are 22.2 million active members in India. Through Twitter, roughly 500 million tweets are generated per day all over the world and 63% of Twitter users range from 35 to 65 years old. There has been a wide recognition

DOI: 10.4018/IJSWIS.295550

for Twitter which has become a powerful gauge of public sentiments for a spectrum of issues such as social, medical, socio-economic factors on happiness, climate policies, politics, understanding public perception of crime, public health, movie sales, stock market and much more. By generating association rules, taxonomies are built. The main aim of this research work is to extract causal factors from the text-based association rules which are useful in forming a collective opinion on a topic.

Many researchers have used the social media platform to analyze the text posted by the users to obtain the outcomes or insights related to public issues. Association rule mining techniques are used to mine social media data (Chen et al, 2020). Researchers have applied evolutionary computing techniques or nature-inspired optimization techniques to generate meaningful association rules. However, only a few have experimented with the clustering of text-based association rules. A hybrid Firefly – Ontology-based clustering algorithm (FF-OC) is proposed to cluster the association rules. The clusters thus formed provide an insight into the causal factors affecting a topic discussed widely in social media. The methodology uses a keyword-based ontology to compute the similarity among features (Zamazal, 2020). Firefly based algorithmic technique is one of the most popular meta-heuristic based algorithms developed by Xin-She Yang (Yang, 2008) based on the behavior of real fireflies. The algorithm is efficiently applied in many domains such as text clustering, image processing, data analytics, and classification. The advantage of the firefly algorithm over the existing swarm intelligence-based algorithms is that it converges quickly (Xie et al., 2019). Since the algorithm automatically split its population into subgroups, each subgroup can find the best global solution.

The following enumerates the contributions of the proposed work:

1. Generating association rules from the text corpus collected from Twitter is a challenging task. Generally ARs reflect an association or context between items in the left hand side and right hand side of the rule. In this work, WordNet is used to incorporate such context for each of the association rules generated from the text features extracted from the twitter data.
2. The algorithm generates a huge number of ARs. Many of them do not contribute to the effectiveness of the accuracy of the results. Therefore an algorithm is designed to prune the repetitive ARs.
3. Though many ARs are pruned, since the collected text corpus is large, the resultant set is still huge. An evolutionary, nature-inspired algorithm is designed for clustering the ARs by performing an exhaustive exploration of all the ARs.
4. The main contribution and novelty lie in the fact that no prior knowledge is applied during the clustering process. From the clusters thus generated, causal factors are identified.

The algorithm that is proposed in this paper differs from the existing works in literature in the way that the association rules are clustered without any prior knowledge. Further insights are provided as impacts or consequences. The theoretical implication of this work is that a new model for clustering association rules is formulated. The clusters that are obtained from the model give insights into various factors that affect the domain of interest. For instance, the government may take action on implementing certain regulations based on the factors impacting which are the main outputs from the research. The main research objectives are listed as follows:

1. Identify the causal relationships among the concepts/features in the user-generated content by clustering the generated association rules.
2. Determine the factors that impact a public related issue. Food price rise and disease (diabetes, flu, and zika virus) related tweets are used to test the proposed methodology.
3. Analyze the tweets' text corpus to determine the association among the features.
4. Prune the association rules to extract only meaningful rules.
5. Cluster the association rules based on the semantic distance of the features. The clusters thus formed convey the topical polarity or the factors that are associated with the issues.

The remaining sections are organized as follows: Section 2 discusses the related methodologies and algorithms in the literature. Section 3 introduces the data collection and preprocessing of the proposed methodology. The details of the association rule extraction algorithm, the pruning algorithm is explained in Section 4. The proposed method is explained in Section 5. Section 6 discusses the results obtained by the implementation of the algorithm. Section 7 concludes the paper with the insights gathered from the proposed methodology.

2. RELATED LITERATURE

The main purpose of the clustering method is to combine data points that are similar in context. Evolutionary computing-based clustering has recently gained the attention of researchers.

Recently, the researchers investigated the Twitter data using various hybrid optimization algorithms namely, Hybrid Spider Monkey optimization with k-means clustering (Shekhawat et al. 2020), hybrid cuckoo search algorithm (Pandey et al. 2017), and Two-step Artificial Bees Colony (ABC) algorithm (Sahoo et al. 2017). These algorithms incorporate the intelligence of nature inspired algorithms and clustering algorithm. The efficiency of the proposed methods are compared with the other nature inspired algorithms using the classification evaluation measures such as precision, recall, accuracy and t-test.

Twitter data has been used to detect the crime-related activities Sandagiri et al (2021), and eyewitness messages (Zahra et al. 2020). In these applications machine learning approaches are applied to classify the tweets. Hasan et al. (2019) developed an event detection system called TwitterNews+ and Nolasco et al. (2019) detects subevents in a certain event from the Twitter data. Transaction-based Rule Change Mining algorithm (Adedoyin-Olowe et al., 2016) is proposed to extract interesting topics or events presented in the Twitter conversation. Feedforward neural network based approach is applied to detect event causality from tweets (Kayesh et al 2019). Gencoglu et al. (2020) and Doan et al. (2019) has developed a causal inference approach to determine the causal relationships between covid-19 pandemic characteristics and health related topics. A text abstract method is proposed by Hu et al. (2017) to assess the top- k most insightful sentences from online hotel reviews. The authors Ahmed et al. (2016) investigated the Twitter conversation as a campaign tool during the 2014 Indian general elections. The tweets are analyzed to determine the relevance of food and fuel price increase amongst Indonesians during the specific period (Pulse, 2014).

Though few works are found in the literature of applying evolutionary algorithms and nature-inspired algorithms for clustering, there are very few works related to clustering the text-based association rules. The proposed methodology attempts to select the text features from the tweets which best describes the context of the tweets and generated association rules. Those association rules are then clustered the proposed FF-OC algorithm and studied the characteristics of the cluster which implied the factors impacting the topic of interest.

Unlike every other approach, the uniqueness of the proposed methodology lies in the clustering of the dataset without known classes. In the literature, the clustering algorithm clusters the data points with known and defined classes. Those clusters are validated against known clusters. For this study, the number and properties of clusters or their characteristics are not known in advance. Depending on the cluster similarity, clusters are formed and the common features of every cluster are identified to arrive at the factors impacting a trending event. This makes the work more challenging and novel. Table 1 summarizes the related work.

3. DATA COLLECTION AND PREPROCESSING

3.1 Twitter Data Collection

In this research, data from Twitter is collected as the API has been provided to extract tweets with respect to different domains and regions. The tweets are collected from Twitter's streaming API and

Table 1. Summary of the related work

Author and year	Objective	Algorithm utilized	Twitter Dataset & Keywords	Results of the evaluation metrics

Sandagiri et al (2021)	Machine learning approach is employed to predict crimes using twitter data.	Long short-term memory (LSTM)	Twitter Dataset
			● Crime
				Accuracy=82.5%,
				Precision=86.4%,
				Recall=80.4%

Gencoglu et al. (2020)	Developed a causal inference approach helps to determine the causal relationships between Covid-19 pandemic characteristics	Bayesian Network	Twitter corpus
			● covid-19
			● coronavirus
				Average AUROC = 0.833
				Precision = 0.8223
				Recall = 0.7248

Zahra et al. (2020)	The eyewitness messages are categorized into direct, indirect, and vulnerable eyewitnesses	Random Forest	Twitter Streaming API Disaster-related tweets like, ● Floods ● Earthquakes ● Hurricanes ● Wildfire	Avg. Precision =0.7472 Avg. Recall=0.6709 Avg. F-Score=0.6911 AUC = 0.938

Shekhawat et al. (2020)	A hybrid natural inspired algorithm is used to classify tweets.	Hybrid Spider Monkey optimization with k-means clustering	Twitter-sanders2 dataset	Avg. accuracy = 94.45%
				Precision=87.04%,
				Recall=83.47%

Doan et al. (2019)	Extract the causalities from the health related tweets using NLP techniques	Natural Language Processing (NLP)	Twitter corpus ● Stress ● insomnia ● headache	Precision=83.43%

Kayesh et al (2019)	Developed feedforward neural network based approach to detect event causality from tweets.	Feed-forward neural network	Twitter Stemming API
			● Common wealth Games 2018
				Accuracy=69.94%,
				Precision=67.46%,
				Recall=61.96%

Hasan et al. (2019)	Developed an event detection system namely, TwitterNews+, that incorporates specialized inverted indices and an incremental clustering	Nearest Neighbor Algorithm	Twitter streaming data
			● Events2012 corpus
				Recall = 0.96
				Precision = 0.89

Nolasco et al. (2019)	Developed a method that automatically detects subevents in a certain event.	Latent Dirichlet Allocation (LDA	Twitter API
			● Brazil's political protests
			● Zika Virus
				Precision = 0.736
				Accuracy =75.23

Hu et al. (2017)	Evaluate the top- k most insightful sentences from online hotel reviews.	K-Medoids algorithm	Hotel Reviews	Precision = 0.7286
				Recall = 0.7023

Pandey et al. (2017)	Developed a hybrid cuckoo search algorithm (combined with cuckoo search and k-means clustering) for Twitter analysis.	Cuckoo Search algorithm	Twitter-sanders-apple, Twitter Dataset
			● sports
			● saints
				Precision = 0.74
				Accuracy =77.99

Adedoyin-Olowe et al. (2016)	Analyzed the Twitter conversation using Transaction-based Rule Change Mining algorithm to extract interesting topics or events	Transaction-based Rule Change Mining algorithm	Tweets
			● FA cup final 2012
			● US elections 2012
				Precision = 0.72
				Recall = 0.70

Ahmed et al. (2016)	Analyzed the 2014 general election-related tweets and observed the insight of first-time voters and internet accessibility in election	Latent Dirichlet Allocation (LDA)	Twitter Dataset ● 2014 Indian general elections	Accuracy = 71%

related to the keywords such as diabetes, flu, zika virus, and food price. The tweets are collected during the period from February 2017 to January 2018 and stored for analysis. In this research work, to experiment with the methodology, around 1,11,000 tweets are collected which are related to “Diabetes”, 31,258 and 14,158 tweets are collected related to “Flu” and “Zika virus” respectively. These tweets are based on healthcare that helps to find the cause of diseases. Around 21,096 tweets are collected related to “Food Price”. The tweets are collected in the English language. The sample tweets are shown in Figure 1.

3.2 Preprocessing

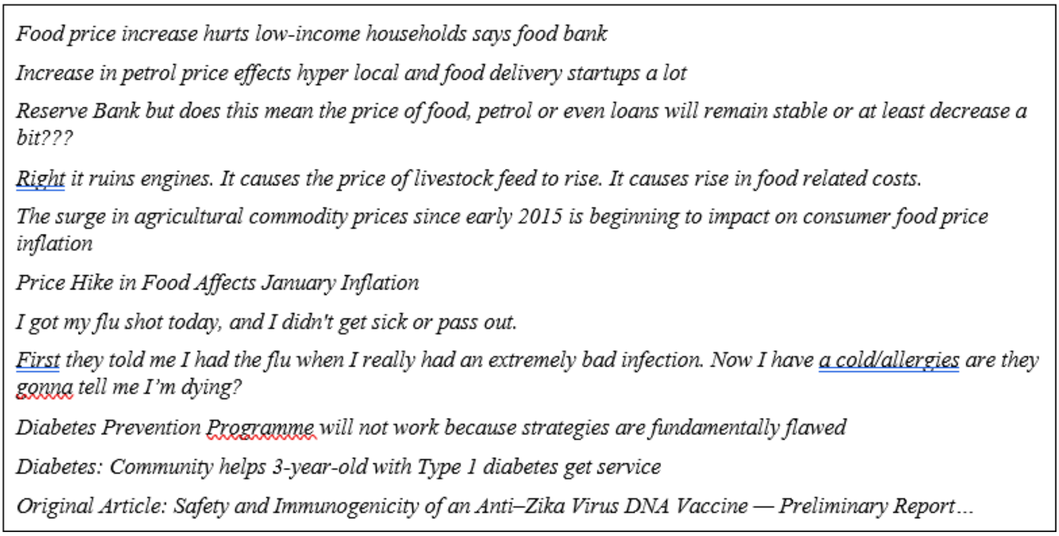
All URLs, hashtag symbols (e.g., # in #fruit), and special characters are removed during preprocessing. Stop words are commonly used terms in a language that reduce the processing speed and increase the memory space. These stop words are therefore removed from tweets. In many cases, the inflective words have similar meanings. The main goal of the stemming algorithm is to reduce the word to its stem or root form. It plays a vital role in text mining and also reduces storage and processing time. To perform stemming, Porter’s stemming algorithm is used. Using the Part of Speech (POS) tagger in the Natural Language Processor Parser (Murthy et al. 2019), the nouns (NN), adjectives (JJ), verbs (VBP), and adverbs (RB) are tagged. The sample output of POS tagging is shown in Figure 2.

3.3 Feature Extraction

Generally, nouns, adjectives, verbs, and adverbs carry opinion on the context. These words from the feature vector for the algorithm to identify the causal factors. There are many tools available to extract the feature word from the text. In this work, Parts of speech (POS) tagger is used to extract those kinds of words from the collected tweets. POS tags are responsible for reading the language of the text and assigning parts of the speech ie. to each word for some specific token. The extracted features are then fed into Bag of Words (BOW) vectors for discovering association rules. For example, the extracted keywords as follows:

BOW= [(‘ dislike ‘, VB),(‘ increase ‘, NN), (‘hike’, NN), (‘ mainly’, RB), (‘Confirm’,VB)]

Figure 1. Sample Tweets



Food price increase hurts low-income households says food bank

Increase in petrol price effects hyper local and food delivery startups a lot

Reserve Bank but does this mean the price of food, petrol or even loans will remain stable or at least decrease a bit???

Right it ruins engines. It causes the price of livestock feed to rise. It causes rise in food related costs.

The surge in agricultural commodity prices since early 2015 is beginning to impact on consumer food price inflation

Price Hike in Food Affects January Inflation

I got my flu shot today, and I didn't get sick or pass out.

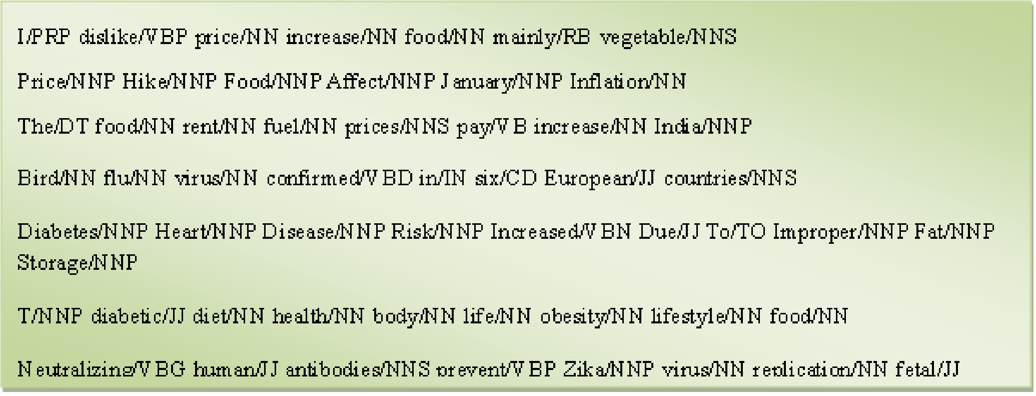
First they told me I had the flu when I really had an extremely bad infection. Now I have a cold/allergies are they gonna tell me I’m dying?

Diabetes Prevention Programme will not work because strategies are fundamentally flawed

Diabetes: Community helps 3-year-old with Type 1 diabetes get service

Original Article: Safety and Immunogenicity of an Anti–Zika Virus DNA Vaccine — Preliminary Report…

Figure 2. Sample output of POS Tagging



```
I/PRP dislike/VBP price/NN increase/NN food/NN mainly/RB vegetable/NNS
Price/NNP Hike/NNP Food/NNP Affect/NNP January/NNP Inflation/NN
The/DT food/NN rent/NN fuel/NN prices/NNS pay/VB increase/NN India/NNP
Bird/NN flu/NN virus/NN confirmed/VBD in/IN six/CD European/JJ countries/NNS
Diabetes/NNP Heart/NNP Disease/NNP Risk/NNP Increased/VBN Due/JJ To/TO Improper/NNP Fat/NNP
Storage/NNP
T/NNP diabetic/JJ diet/NN health/NN body/NN life/NN obesity/NN lifestyle/NN food/NN
Neutralizing/VBG human/JJ antibodies/NNS prevent/VBP Zika/NNP virus/NN replication/NN fetal/JJ
```

4. GENERATING ASSOCIATION RULES

The BOW vector is represented as a transaction in which each WordNet stem represents an item. The Apriori algorithm is used for finding the hidden connections among the content of the transactional tweets. For discovering strong association rules, the algorithm for finding them is constrained by minimum support and confidence thresholds (Dehkharghani et al., 2014; Hamed et al., 2014).

Definition 1 (Twitter Association Rule)

Let $F = \left\{fk_1, fk_2, \ldots, fk_n\right\}$ be a set of n attributes called features or keywords. Let T $= \left\{t_1, t_2, \ldots, t_m\right\}$ be a set of tweets from the Twitter dataset. Each tweet in T has a unique ID and encompasses a subset of the features in F. A tweet is represented as an n-tuple $t_i = \left\{fk_1, fk_2, \ldots, fk_n\right\}$ where n is the number of features or keywords and the value of fk_1 is 1 if the keyword is presented in the tweet t_i or 0 otherwise. A twitter association rule is defined as follows $fk_i \rightarrow fk_j$ where $fk_i, fk_j \subseteq F$ and $fk_i \cap fk_j = \varnothing$ (Cagliero, 2013).

As an example, a rule from the Twitter dataset could be $\left\{food\ price\right\} \rightarrow \left\{fast\ food\right\}$ which relates the features/keywords food price and fast food.

Definition 2 (Support and Confidence)

The Support of a rule A Þ B is the probability of the features or keywords {A, B} available in the dataset. This gives an idea of how often the rule is relevant (Cagliero, 2013):

Support (A Þ B) = P({A,B}) (1)

The Confidence of a rule A Þ B is the conditional probability of B given A. That is the occurrence of feature A in a tweet along with the feature B. The accuracy of the rule is measured as follows:

Confidence (A Þ B) = P(B|A) = support({A,B}) / support(A) (2)

The following are the steps involved in the generation of Association rules (Apriori algorithm):

- Initialize the minimum support and confidence threshold.
- Generate feature sets that are frequently occurred in the BOW vector.
- Compute the confidence value to find strong associations between the features.
- Generate Association Rules based on the confidence.

A sample set of rules generated by the association rule mining algorithm is presented in Table 2.

4.1 Elimination of Redundant Association Rules

All the subsets of frequent itemsets are considered in most of the traditional association rule mining algorithms. Therefore, these algorithms generate thousands or even millions of rules. However, many of these rules have the same meaning or are redundant rules. In the majority of the cases, the number of redundant rules is significantly larger than that of essential rules. The redundant rules affect the importance of information. Thus, it is necessary to eliminate the redundant rules to improve the superiority of the information. The linear algorithm of the proposed methodology removes the redundancy of association rules thereby improves the quality of rules and decreases the size of the rule list (Akilandeswari and Jothi 2016).

Definition 3 (Redundant Rules)

Let A→B and A' →B' be two association rules with confidences con and con' respectively. A → B is said to be a redundant rule if A' ⊆ A, B' ⊆ B and con ≤ con' (Batbarai et al. 2014). For example, consider a rule set R having three rules such as:

$$\left\{inflation\right\} \rightarrow \left\{oil, petrol\right\}, \left\{inflation, oil\right\} \rightarrow \left\{petrol\right\}$$

Table 2. Rules Generated by Association Rule Mining Algorithm

Rules Generated by Association Rule Mining Algorithm	
R1	expense -> hike, nature
R2	hike -> expense, nature
R3	nature -> expense, hike
R4	expense, hike -> nature
R5	expense, nature -> hike
R6	hike, nature -> expense
R7	hike -> market
R8	market -> hike
R9	expense -> hike, money
R10	hike -> expense, money
R11	money -> expense, hike
R12	expense, hike -> money
R13	expense, money -> hike
R14	hike, money -> expense

$$\left\{oil, petrol\right\} \rightarrow \left\{inflation\right\} and \left\{inflation, petrol\right\} \rightarrow \left\{oil\right\}$$

Except for the first rule all the remaining rules are redundant and they do not convey any extra knowledge or information.

In this algorithm, a non-redundant rule set is initialized with an empty set. The entire item set of the first rule is stored in intial_node and the next rule is stored in the present_node. If the present_node is not null then the two nodes are compared. If the present_node item set points to the intial_node item set then that rule is eliminated from the ruleset. Otherwise, the rule remains in the ruleset. In Table 2, rules 1, 7, and 9 are valid rules. Other rules are redundant as they are a simple combination of the valid rules.

The linear elimination algorithm removes the redundant rules efficiently and improves the quality of mining without any loss of information or knowledge. The Non-redundant association rules that are generated by the linear elimination algorithm are presented in Table 3.

5. THE PROPOSED FIREFLY – ONTOLOGY-BASED CLUSTERING ALGORITHM (FF-OC)

Xin-She Yang proposed the Firefly Algorithm (FA) which is a population-based metaheuristic optimization algorithm. It simulates the flashing behavior of fireflies. The following assumptions are made in the algorithm: (1) all the fireflies are unisex and therefore a firefly will be attracted to other fireflies; (2) The attraction is proportional to the brightness and decreases as the distance increases. For any two fireflies, the less luminous firefly will travel towards the brighter one. The firefly algorithm needs to address two issues: 1. variation in the light intensity and 2. representation of attractiveness. These two issues can be customized to fit different kinds of problems (Yang & Deb, 2009; Iztok et al., 2013).

Yang established the fact that the fireflies can automatically subgroup themselves since near attraction is more impactful than the long-distance attraction. When evaluated for the performance with respect to efficiency and success rate, FA is more superior to other evolutionary algorithms (Yang & He, 2013). The global solution can be found among all the local optima.

The firefly algorithm is applied in the proposed methodology to explore a large set of rules. As the number and characteristics of the clusters are not defined in advance, the nature-inspired algorithm has to go over every rule and characterize the rule into a cluster.

5.1 The Light Intensity and Attractiveness of the Firefly

The light intensity of the firefly is considered as an objective function. In this research work, the similarity of the two fireflies is computed which is a maximization problem. In other words, a firefly with a higher intensity of light would be attracted to another firefly with the same higher intensity. Let n be the total number of data in the population. Each firefly is represented as x_i where i=1,2,….n. The intensity of light is determined as follows:

Table 3. List of Non-redundant rules

Non-Redundant Rules	
R1	expense -> hike, nature
R7	hike -> market
R9	expense -> hike, money

$$I_i = f\left(x_i\right), 1 \le i \le n \tag{3}$$

The attraction is proportionate to the light intensity between the two neighboring fireflies (Yang, 2008). Each firefly has its unique attractiveness β which implies how strongly it attracts other fireflies in the swarm. The distance r_{ij} between two fireflies x_i and x_j is given as:

$$r_{ij} = x_i - x_j \tag{4}$$

The attractiveness function $\beta\left(r\right)$ is determined as follows:

$$\beta\left(r\right) = \beta_0 e^{-\gamma r^2} \tag{5}$$

where β_0 is the attractiveness of the firefly at $r = 0$ and γ is the light absorption coefficient. The movement of the firefly is given as:

$$x_i\left(t+1\right) = x_i\left(t\right) + \beta_0 e^{-\gamma r^2}\left(x_i - x_j\right) \tag{6}$$

If β_0 value is 0, the algorithm simply becomes a random walk.

The pseudo-code of the standard firefly algorithm is given in Algorithm 1 (Yang, 2008).

5.2 Identifying Similarity Measures Using WordNet Ontology

Ontology is a hierarchical structure that is formed by a set of nodes out of which one is a root node (r). Wu and Palmer's (Sim_{wu}) similarity measure is used to compute the similarity of two ontology elements w_1 and w_2. The principle of similarity computation is based on the path length ($len\left(w_1\right), len\left(w_2\right)$) of the nodes w_1 and w_2 from the root node r and the depth $dep\left(r\right)$

Algorithm 1. Standard Firefly algorithm

```
Initialize firefly algorithm
Define light intensity $I_i$ by fitness function $f\left(x_i\right)$
Repeat
    For $i = 1$ to number of fireflies
        For $j = 1$ to number of fireflies
            If $(I_i < I_j)$ move firefly $i$ toward $j$ by eq. (4)
              Vary attractiveness by eq. (5)
              Evaluate new solution and update light intensity
            End if
        end for $j$
    end for $i$
Rank new solutions and find the current best
Next-generation
Until (convergence)
```

of the Least Common Subsumer (LCS) from the root node r. The node is an LCS if it connects the senses (contents) of both the elements. The similarity measure (Sim_{wu}) is defined by the following equation (Shet et al. 2012, Alharbi et al. 2020):

$$Sim_{wu}\left(w_1, w_2\right) = \frac{2^{*} dep\left(r\right)}{len\left(w_1\right) + len\left(w_2\right) + 2^{*} dep\left(r\right)} \tag{7}$$

$$dep\left(r\right) = the\ \ depth\ \ of\ \ the\ \ LCS\ \ from\ \ the\ \ ontology\ \ root$$

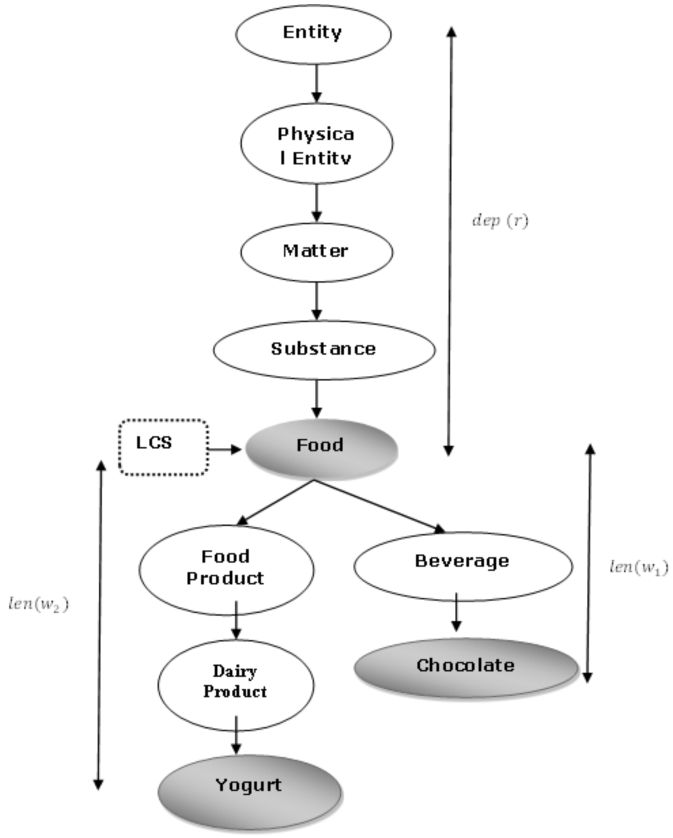
$$len\left(w_1\right) = No.\ \ of\ \ edges\ \ between\ \ w_1\ \ and\ \ LCS$$

$$len\left(w_2\right) = the\ \ path\ \ length\ \ between\ \ w_2\ \ and\ \ LCS$$

The advantage of the Wu and Palmer method is that it is simple to calculate and gives more accurate similarity when compared to other similarity measures. This measure is therefore embraced as a basis for calculating the similarity between the features in the association rules.

As an illustration (Figure 3), let us consider the computation of the similarity between Chocolate and Yogurt:

Figure 3. A sample Ontology Tree



$$dep\left(r\right) = 4$$

$$len\left(Chocolate\right) = 2$$

$$len\left(Yogurt\right) = 3$$

$$Sim_{wu}\left(Chocolate, Yogurt\right) = \frac{2*4}{2 + 3 + \left(2*4\right)} = 0.6153$$

The similarity of chocolate and yogurt is 0.6153 which implies that both the elements are similar.

5.3 The Proposed FF-OC Algorithm

The similarity measure described above is used as the objective function of the firefly to cluster the association rules. The FF-OC algorithm uses the following objective function to determine the similarity between two association rules using ontology:

$$Objective\ function\ F = \left\{AR : CR, Avg\left(Sim_{wu}\left(x_i, x_j\right)\right) > \vartheta\right\} where\ i \neq j \tag{8}$$

In this algorithm, each firefly is initially placed in a rule for exploration. The initial light intensity I_i of each firefly can be related to the depth of each synset and is determined using (3) with f(x) being the similarity between the features. In the proposed FF-OC algorithm, the depth between each synset and the immediate root is considered as the initial intensity of light for the firefly. The distance r_{ij} between two fireflies x_i and x_j is computed using the ontology-based similarity measure which is given below:

$$r_{ij} = Sim_{wu}\left(a_i, a_j\right) \tag{9}$$

The proposed algorithm FF-OC is given in Algorithm 2.

Each firefly is placed with an association rule $\left(x_i = x_1, x_2, \ldots, x_n\right)$. Each of the fireflies is initialized with the light intensity. The light intensity of each firefly is considered as a fitness function and computed using an ontology-based similarity measure. If the light intensity, $I_i < I_j$ then x_i moves towards x_j i.e. maximum similarity value. The solutions are evaluated using the objective function.

The attractiveness between the two fireflies is computed using equation (8). In this work, an ontology-based similarity measure is used to compute the distance r_{ij} between two fireflies x_i and x_j. If the intensity of light of the firefly is greater than the threshold then put it into one cluster. This process is continued until all the association rules are clustered. The time complexity of the proposed FF-OC algorithm is $O(n^2 t)$ where n is the population size and t is the number of iterations.

5.4 Illustration

The sample input is presented in Table 4. Let R1, R2 … R9 represents association rules and set the threshold value ϑ as 0.5.

The light intensity is initialized of each synset and presented in Table 5.

Algorithm 2. The Proposed FF-OC Algorithm

```
$Algorithm : FF - OC(AR, \vartheta, n)$
$Input : AR, the\ set\ of\ Association\ Rules;$
        $n, Total\ no.\ of\ fireflies(i.e., Association\ Rules);$
$Output : Clustered\ Rules(CR);$
$Initialize\ the\ population\ randomly$
$Generate\ initial\ fireflies\ x_i\ (i = 1, 2, \ldots n)\ corresponding\ to\ each\ association\ rule$
NS= $No.\ of\ synsets\ in\ the\ AR;$
$Objective\ function\ F = \{AR : CR, Avg(Sim_{wu}(x_i, x_j)) > \vartheta\}\ where\ i \neq j$
                    $Avg(Sim_{wu}(x_i, x_j)) = Sim_{wu}(a) / NS$
$Repeat$
$CR = \{\}$
$for\ each\ firefly,\ initialize\ light\ intensity\ I_i\ by\ fitness\ function\ f(x_i)$
  $for\ i = 1 : n\ No.\ of\ fireflies$
    $If(i \neq j)$
        $If\ (I_i < I_j)\ move\ firefly\ x_i\ toward\ x_j\ by\ eq.(4)$
                $Compute\ Attractiveness\ i.e.,\ Distance\ r_{ij}\ by\ eq.(9)$
            $Rank\ the\ fireflies$
            $If\ the\ intensity\ of\ two\ fireflies\ x_{ij}(I_{x_{ij}} > \vartheta)$
            $CR = CR \cup x_{ij}$
            $end\ if$
        $end\ if$
          $Evaluate\ new\ solution\ and\ update\ light\ intensity$
      $end\ if$
      $end\ for\ j$
$end\ for\ i$
$Until(convergence)$
```

The attractiveness of each of the fireflies placed in each rule is computed using the Wordnet ontology:

Rule 1(firefly 1) & Rule 2 (firefly 2): 0.3541
Rule 1(firefly 1) & Rule 3 (firefly 3): 1.0390
Rule 1(firefly 1) & Rule 4 (firefly 4): 0.8201
Rule 1(firefly 1) & Rule 5 (firefly 5): 0.3011

Table 4. Sample Input

Input: Association Rules	
Threshold $\vartheta = 0.5$	
R1	food -> chocolate
R2	pen -> pencil
R3	milk -> butter
R4	bread -> jam
R5	egg -> price
R6	exam -> time
R7	curd -> cheese
R8	car -> journey
R9	car -> drive

Table 5. Light intensity value

Synset	Depth
Food	0.3
Egg	0.5
Curd	0.5
Milk	0.5
Pen	1.0
Car	1.0
bread	0.5
Exam	0.8

Rule 1(firefly 1) & Rule 6 (firefly 6): 0.2678
Rule 1(firefly 1) & Rule 7 (firefly 7): 1.0078
Rule 1(firefly 1) & Rule 8 (firefly 8): 0.2618
Rule 1(firefly 1) & Rule 9 (firefly 9): 0.2843

The intensity of firefly 1 is compared to all the other fireflies. The intensity of the firefly 3, 4, and 7 are greater than the threshold similarity of the other fireflies. Therefore, these fireflies are grouped in one cluster C1 = [1, 3, 4, 7]. Similarly, firefly 2 is compared with the remaining flies and grouped in the second cluster and the process is repeated until all fireflies (rules) are clustered. In this process the clusters C2 = [2, 5, 6] and C3 = [8, 9] are formed. The clustered rules are presented in Table 6.

Table 6. Clustered Rules

Cluster 1	Cluster 2	Cluster 3
food -> chocolate milk -> butter bread -> jam curd -> cheese	pen -> pencil egg -> price exam -> time	car -> journey car -> drive

6. EXPERIMENTAL ANALYSIS

6.1 Experimental Setup

All the experiments are performed on an Intel Corei5 processor with 4 GB memory. The clustering algorithms are executed in MATLAB. The efficiency of the proposed approach is compared with the existing nature inspired based clustering algorithms namely, Artificial Bees Colony (Karaboga, 2005), Cuckoo Search Algorithm (Yang & Deb, 2009), Particle Swarm Optimization (Kennedy, 1995), Improved Particle Swarm Optimization (Mehdi, 2021) and Ant Colony Optimization (Dorigo 1996; 1997). To compare the proposed methodology in the similar terms of the other algorithms in the literature, tuning parameter values which are used in the existing algorithms are considered. In this research work, default parameter values available in the MATLAB are utilized for implementing the algorithms. During the implementation, the parameter values are assigned in trial and error manner. The values assigned by different authors (Yarpiz, 2021; 2021; Seyedali, 2021; Mehdi, 2021; Xin-She Yang, 2021) are tried. It is found that the default values available in MATLAB itself gave good optimization. In table 7, the parameter values used during the implementation after trying with different values are given. The number of clusters in the existing clustering algorithms is specified to 5 based on the elbow method.

6.2 Results and Discussion

During the first phase of the implementation, stop words removal, stemming and POS tagging are performed. Features/keywords extracted are stored in the Bag of Words vector. A representative sample of features selected from the tweets related to diabetes, flu, zika virus, food price rise is shown in Figure 4.

Each tweet is considered a transaction. The Apriori algorithm generates association rules with the support and confidence as 0.5. The algorithm generates a large number of rules and out of those rules, many are redundant. A linear algorithm as discussed in section 3.5 is applied to eliminate the redundant rules. Figure 5 demonstrates the total number of rules generated by the association rule mining algorithm based on the support and confidence value and the number of redundant rules eliminated by the linear algorithm.

It is observed that the linear elimination algorithm efficiently removes the redundant rules. On average 25% of rules are eliminated. This reduction decreases the memory utilization and time consumption for processing. Figure 6 shows the elapsed time of the elimination algorithm.

The proposed Firefly – Ontology-based clustering algorithm clusters the association rules with the objective function deciding the intracluster similarity. In this experiment, the algorithm parameter Absorption coefficient γ is initialized as 1, the threshold value ϑ as 0.5, and the δ as 0.97. Table 10 shows the representative clusters formed with a sample of 250 tweets in the domain of food price rise.

The proposed algorithm generates 12 clusters for food price rise related data. It can be observed that cluster 1 and cluster 2 represent the food price rise being impacted due to the rise in the price of household things, market price, oil, fuel (petrol), fast food, power, organic products, noodles, and due to natural disaster. Cluster 3 represents the impact of the country’s economy, health-related issues,

Table 7. Tuning parameter

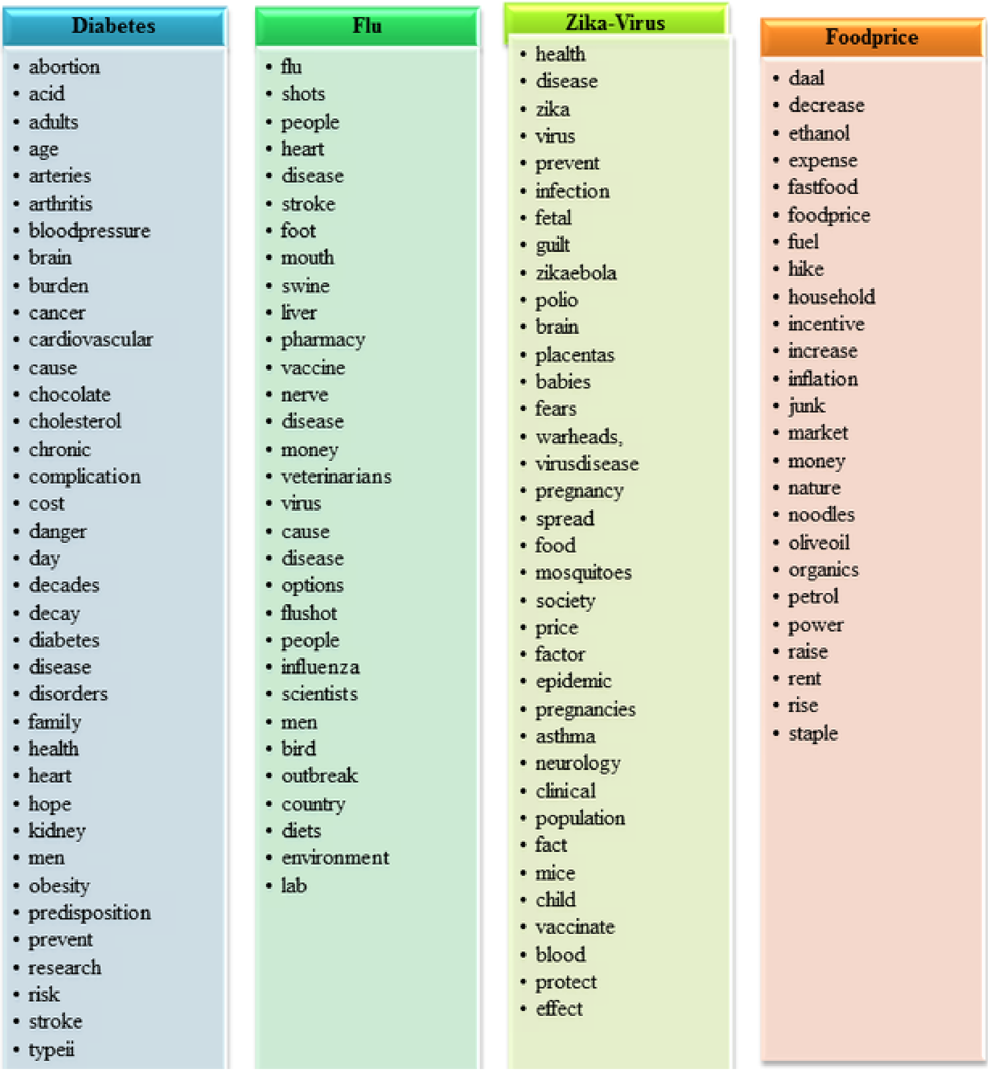
Clustering algorithm	Tuning Parameters	Parameter values
PSO	Population size (particles)	Total No. of association rules
	Maximum Iteration	150
	No. of Clusters	5
	Initial weights w	0.62
	$C1$ (Personal Learning Coefficient)	1.5
	$C2$ (Global Learning Coefficient)	2.0
ACO	Population size (ants)	Total No. of association rules
	Maximum Iteration	150
	No. of Clusters	5
	tau0 (Initial Pheromone)	10
	$Alpha$ (Pheromone Exponential Weight)	1
	Rho (Evaporation Rate)	0.5
ABC	Population Size (Colony Size)	Total No. of association rules
	Maximum Iteration	150
	a (Acceleration Coefficient)	1
	$nVar$ (Number of Decision Variables)	5
	No. of Clusters	5
Cuckoo Search	N (Number of nests)	Total No. of association rules
	Maximum Iteration	150
	No. of Clusters	5
	Pa (Discovery rate of alien eggs/ solutions)	0.25
	Nd (bounds of the search domain)	15
Improved PSO	Population size (particles)	Total No. of association rules
	Maximum Iteration	150
	No. of Clusters	5
	Initial weights w	0.62

continued on following page

Table 7. Continued

Clustering algorithm	Tuning Parameters	Parameter values
Firefly	Population (No. of fireflies)	Total No. of association rules
	Maximum Iteration	150
	Alpha (ϑ)	0.2
	gamma (γ)	1.0
	delta (δ)	0.97

Figure 4. Sample Features



Diabetes

- abortion
- acid
- adults
- age
- arteries
- arthritis
- bloodpressure
- brain
- burden
- cancer
- cardiovascular
- cause
- chocolate
- cholesterol
- chronic
- complication
- cost
- danger
- day
- decades
- decay
- diabetes
- disease
- disorders
- family
- health
- heart
- hope
- kidney
- men
- obesity
- predisposition
- prevent
- research
- risk
- stroke
- typeii

Flu

- flu
- shots
- people
- heart
- disease
- stroke
- foot
- mouth
- swine
- liver
- pharmacy
- vaccine
- nerve
- disease
- money
- veterinarians
- virus
- cause
- disease
- options
- flushot
- people
- influenza
- scientists
- men
- bird
- outbreak
- country
- diets
- environment
- lab

Zika-Virus

- health
- disease
- zika
- virus
- prevent
- infection
- fetal
- guilt
- zikaebola
- polio
- brain
- placentas
- babies
- fears
- warheads,
- virusdisease
- pregnancy
- spread
- food
- mosquitoes
- society
- price
- factor
- epidemic
- pregnancies
- asthma
- neurology
- clinical
- population
- fact
- mice
- child
- vaccinate
- blood
- protect
- effect

Foodprice

- daal
- decrease
- ethanol
- expense
- fastfood
- foodprice
- fuel
- hike
- household
- incentive
- increase
- inflation
- junk
- market
- money
- nature
- noodles
- oliveoil
- organics
- petrol
- power
- raise
- rent
- rise
- staple

Figure 5. Number of Rules Generated vs No. of Rules Without Redundancy

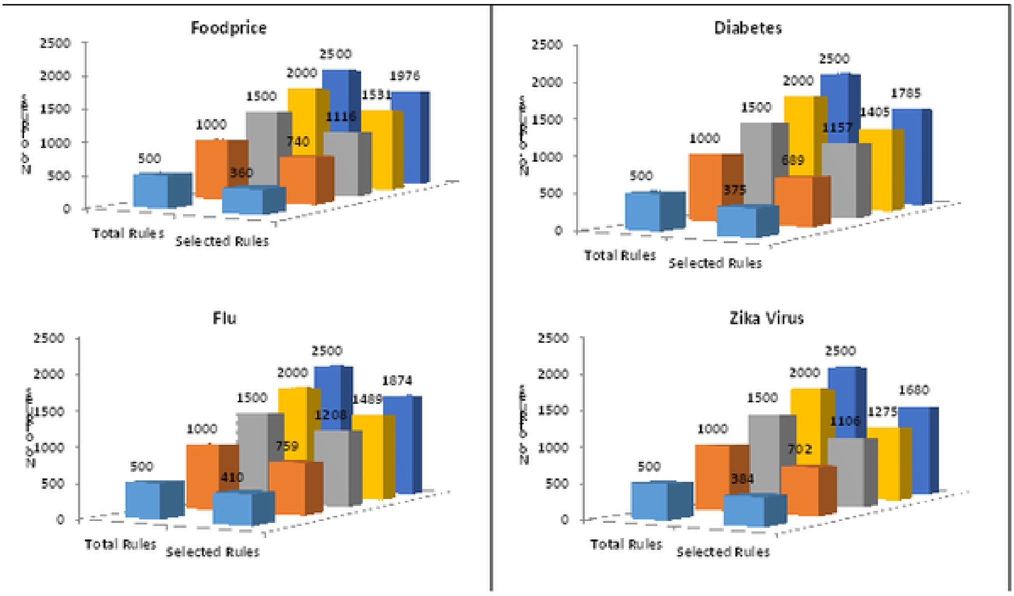
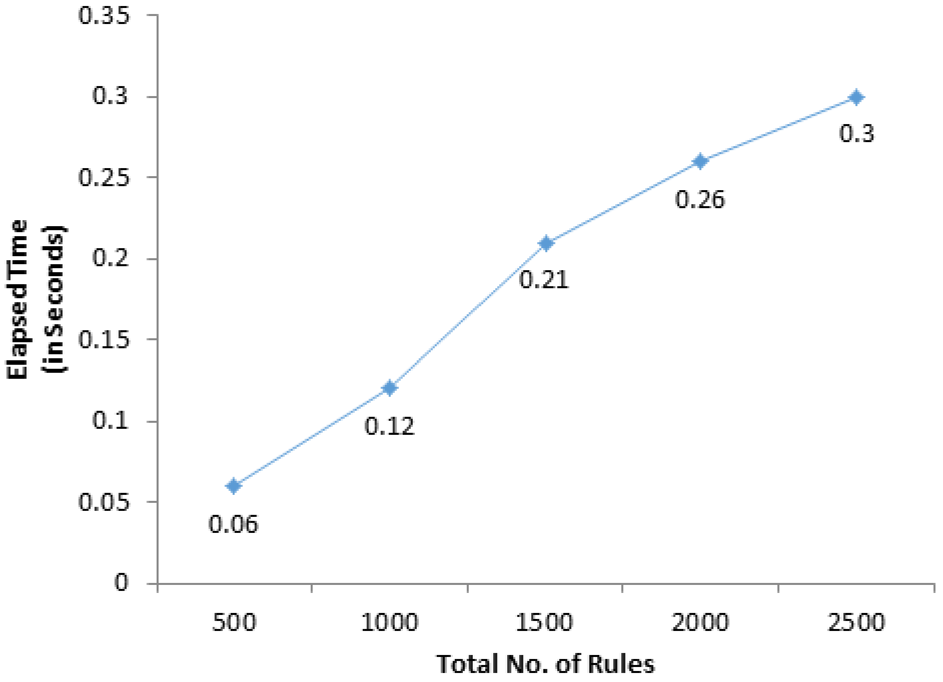


Figure 6. Elapsed Time of Linear Elimination Algorithm



and a hike in household things. Cluster 4 predicts the factors as the rise in the share market and dollar value. The impression from cluster 5 is that the factors are related to natural causes. Cluster 6 infers the factor as inflation. Similarly, the factors inferred from cluster 8 are a rise in fuel, food, and junk food. Cluster 9 infers the impact of a rise in fast-food prices. Cluster 10 gives the impression of the factor as the demand or scarcity in natural products. Cluster 11 infers the factor as consumer behavior and Cluster 12 as the rise in daal price. From the above findings, the factors causing the food price hike can easily be inferred. Table 8 shows clusters of diabetes-related tweets.

Similarly, the proposed algorithm clusters the diabetes-related tweets into seven groups. From Table 9 above, the risk factors can be found as heart problems, lifestyle adaptions, blood pressure, sleep-related problem, cholesterol, obesity, Parkinson stroke, and so on. Table 10 shows clusters of zika virus related tweets.

6.3 Performance Metrics and Assessment

Cluster validation is a significant problem in clustering analysis because the result of clustering needs to be validated in most applications. Three different validation measures are applied to verify the results of the proposed clustering algorithm as these measures will detect the correct number of clusters in many experiments.

6.3.1 Davies-Bouldin (DB) Index

In this method, the ratio of the sum of intra-cluster distance and the inter-cluster distance is computed. It is defined as follows (Davies & Donald, 1979):

$$DB = \frac{1}{n}\sum_{i=1}^{n} \begin{matrix} max \\ i \neq j \end{matrix} \left\{ \frac{S_n\left(Q_i\right) + S_n\left(Q_j\right)}{S\left(Q_i, Q_j\right)} \right\} \tag{4}$$

where n is denoted as the total number of clusters, S_n is the average of all intra-cluster distance and $S\left(Q_i, Q_j\right)$ is the inter-cluster distance. The performance of the clustering algorithm is good if the computed value is low.

6.3.2 Silhouette Index

The silhouette index value is measured based on the silhouette width. The index is computed as follows (Rousseeuw, 1987):

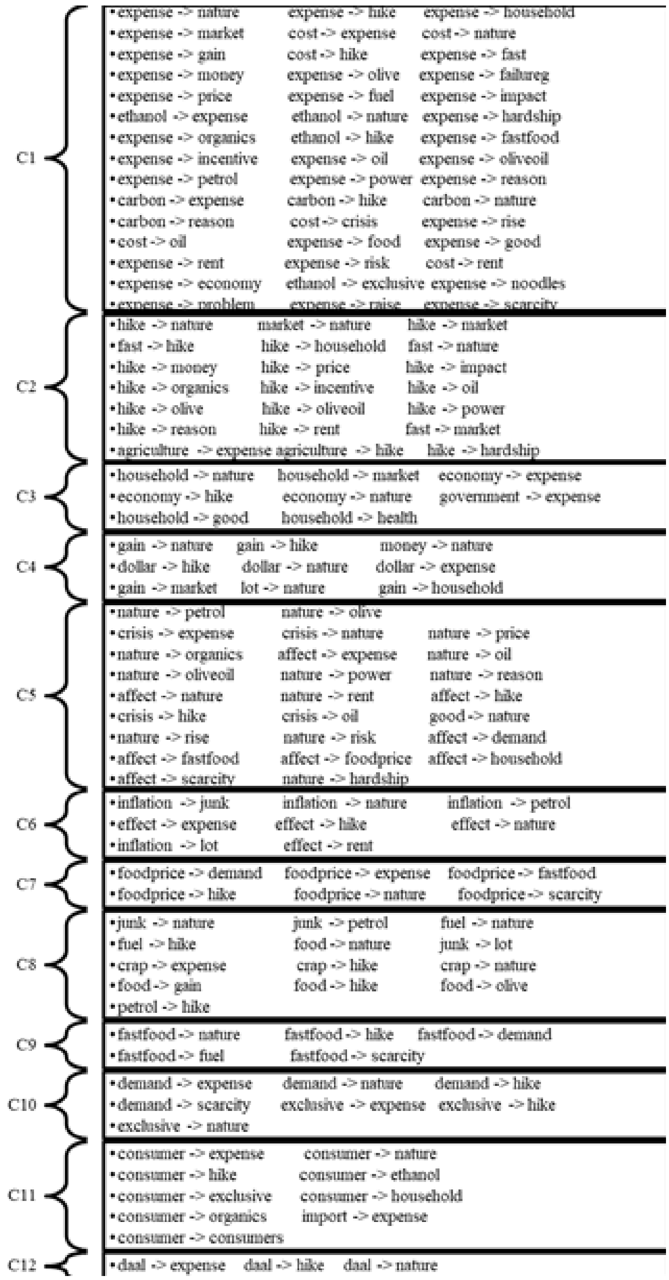
$$S\left(i\right) = \frac{\left(b\left(i\right) - a\left(i\right)\right)}{\max\left\{a\left(i\right), b\left(i\right)\right\}} \tag{5}$$

In this equation, $a\left(i\right)$ is the average difference between the i^{th} sample and all the other samples in the same cluster; $b\left(i\right)$ is the minimum of the average difference between i^{th} sample and all samples in a different cluster. If the index value is high, then it can be inferred that the clustering algorithm performs well.

6.3.3 C Index

C index (Hubert & Schultz, 1976) is defined as follows:

Table 8. Association rule clustering for food price rise dataset

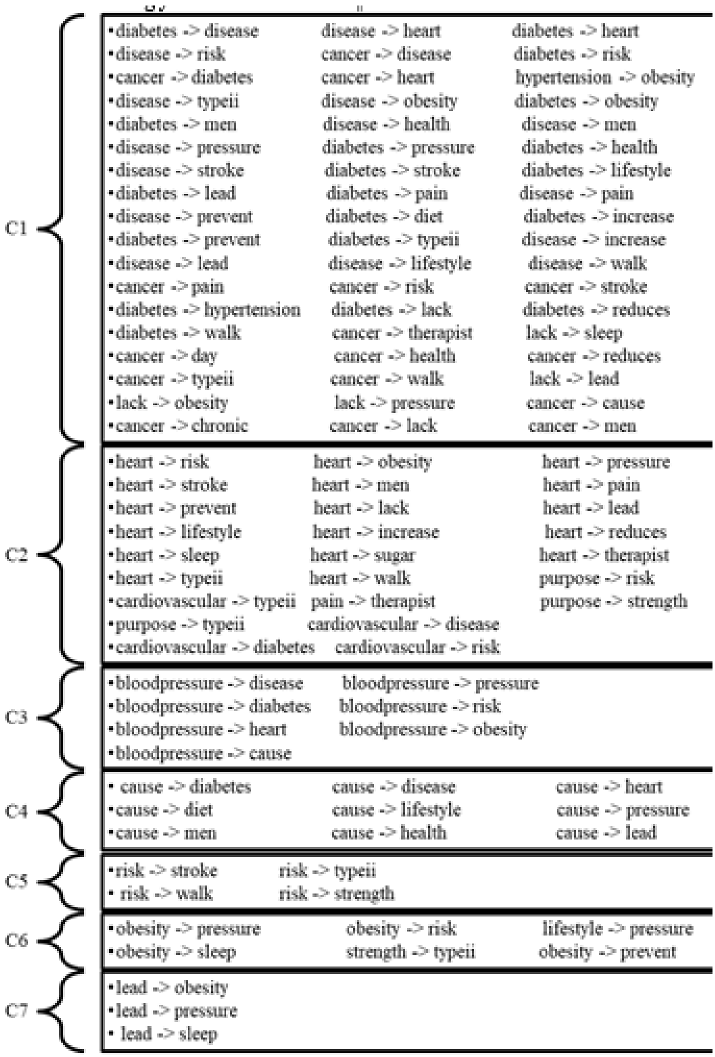


C1	•expense -> nature	expense -> hike	expense -> household
	•expense -> market	cost -> expense	cost -> nature
	•expense -> gain	cost -> hike	expense -> fast
	•expense -> money	expense -> olive	expense -> failureg
	•expense -> price	expense -> fuel	expense -> impact
	•ethanol -> expense	ethanol -> nature	expense -> hardship
	•expense -> organics	ethanol -> hike	expense -> fastfood
	•expense -> incentive	expense -> oil	expense -> oliveoil
	•expense -> petrol	expense -> power	expense -> reason
	•carbon -> expense	carbon -> hike	carbon -> nature
	•carbon -> reason	cost -> crisis	expense -> rise
	•cost -> oil	expense -> food	expense -> good
	•expense -> rent	expense -> risk	cost -> rent
	•expense -> economy	ethanol -> exclusive	expense -> noodles
	•expense -> problem	expense -> raise	expense -> scarcity
C2	•hike -> nature	market -> nature	hike -> market
	•fast -> hike	hike -> household	fast -> nature
	•hike -> money	hike -> price	hike -> impact
	•hike -> organics	hike -> incentive	hike -> oil
	•hike -> olive	hike -> oliveoil	hike -> power
	•hike -> reason	hike -> rent	fast -> market
	•agriculture -> expense	agriculture -> hike	hike -> hardship
C3	•household -> nature	household -> market	economy -> expense
	•economy -> hike	economy -> nature	government -> expense
	•household -> good	household -> health	
C4	•gain -> nature	gain -> hike	money -> nature
	•dollar -> hike	dollar -> nature	dollar -> expense
	•gain -> market	lot -> nature	gain -> household
C5	•nature -> petrol	nature -> olive	
	•crisis -> expense	crisis -> nature	nature -> price
	•nature -> organics	affect -> expense	nature -> oil
	•nature -> oliveoil	nature -> power	nature -> reason
	•affect -> nature	nature -> rent	affect -> hike
	•crisis -> hike	crisis -> oil	good -> nature
	•nature -> rise	nature -> risk	affect -> demand
	•affect -> fastfood	affect -> foodprice	affect -> household
	•affect -> scarcity	nature -> hardship	
C6	•inflation -> junk	inflation -> nature	inflation -> petrol
	•effect -> expense	effect -> hike	effect -> nature
	•inflation -> lot	effect -> rent	
C7	•foodprice -> demand	foodprice -> expense	foodprice -> fastfood
	•foodprice -> hike	foodprice -> nature	foodprice -> scarcity
C8	•junk -> nature	junk -> petrol	fuel -> nature
	•fuel -> hike	food -> nature	junk -> lot
	•crap -> expense	crap -> hike	crap -> nature
	•food -> gain	food -> hike	food -> olive
	•petrol -> hike		
C9	•fastfood -> nature	fastfood -> hike	fastfood -> demand
	•fastfood -> fuel	fastfood -> scarcity	
C10	•demand -> expense	demand -> nature	demand -> hike
	•demand -> scarcity	exclusive -> expense	exclusive -> hike
	•exclusive -> nature		
C11	•consumer -> expense	consumer -> nature	
	•consumer -> hike	consumer -> ethanol	
	•consumer -> exclusive	consumer -> household	
	•consumer -> organics	import -> expense	
	•consumer -> consumers		
C12	•daal -> expense	daal -> hike	daal -> nature

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \qquad (6)$$

where S is the total of distances of all pairs of patterns from a similar cluster. The number of pairs is denoted as l. In this equation, S_{min} is the sum of the total of l smallest distances of all pairs and

Table 9. Association Rule Clustering for Diabetes Dataset



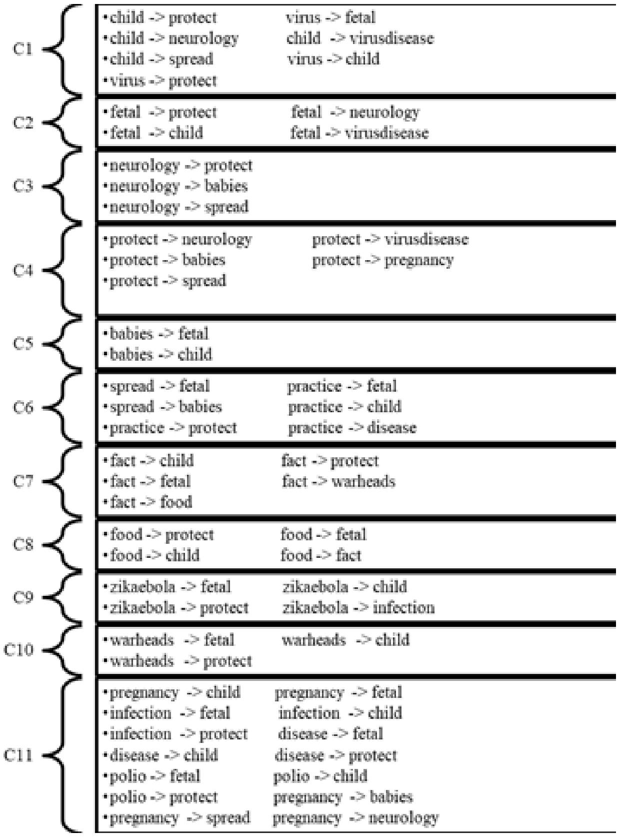
C1	•diabetes -> disease	disease -> heart	diabetes -> heart
	•disease -> risk	cancer -> disease	diabetes -> risk
	•cancer -> diabetes	cancer -> heart	hypertension -> obesity
	•disease -> typeii	disease -> obesity	diabetes -> obesity
	•diabetes -> men	disease -> health	disease -> men
	•disease -> pressure	diabetes -> pressure	diabetes -> health
	•disease -> stroke	diabetes -> stroke	diabetes -> lifestyle
	•diabetes -> lead	diabetes -> pain	disease -> pain
	•disease -> prevent	diabetes -> diet	diabetes -> increase
	•diabetes -> prevent	diabetes -> typeii	disease -> increase
	•disease -> lead	disease -> lifestyle	disease -> walk
	•cancer -> pain	cancer -> risk	cancer -> stroke
	•diabetes -> hypertension	diabetes -> lack	diabetes -> reduces
	•diabetes -> walk	cancer -> therapist	lack -> sleep
	•cancer -> day	cancer -> health	cancer -> reduces
	•cancer -> typeii	cancer -> walk	lack -> lead
	•lack -> obesity	lack -> pressure	cancer -> cause
	•cancer -> chronic	cancer -> lack	cancer -> men
C2	•heart -> risk	heart -> obesity	heart -> pressure
	•heart -> stroke	heart -> men	heart -> pain
	•heart -> prevent	heart -> lack	heart -> lead
	•heart -> lifestyle	heart -> increase	heart -> reduces
	•heart -> sleep	heart -> sugar	heart -> therapist
	•heart -> typeii	heart -> walk	purpose -> risk
	•cardiovascular -> typeii	pain -> therapist	purpose -> strength
	•purpose -> typeii	cardiovascular -> disease	
	•cardiovascular -> diabetes	cardiovascular -> risk	
C3	•bloodpressure -> disease	bloodpressure -> pressure	
	•bloodpressure -> diabetes	bloodpressure -> risk	
	•bloodpressure -> heart	bloodpressure -> obesity	
	•bloodpressure -> cause		
C4	• cause -> diabetes	cause -> disease	cause -> heart
	•cause -> diet	cause -> lifestyle	cause -> pressure
	•cause -> men	cause -> health	cause -> lead
C5	•risk -> stroke	risk -> typeii	
	• risk -> walk	risk -> strength	
C6	•obesity -> pressure	obesity -> risk	lifestyle -> pressure
	•obesity -> sleep	strength -> typeii	obesity -> prevent
C7	•lead -> obesity		
	•lead -> pressure		
	• lead -> sleep		

S_{max} is the total of l largest distances out of all pairs. A minimum value of the *C* index indicates a good clustering.

Figures 7, 8, and 9 give the performance comparison of the proposed Firefly Ontology-based clustering (FF-OC) methodology with the ABC based clustering, CSA based clustering, ACO based clustering, PSO based clustering, and Improved PSO based clustering. The Euclidean distance measure is used to compute the objective function for the ACO based algorithm. From the figures, it can be observed that the proposed methodology's cluster is validated with good scores. In figure 5, it can be inferred that the proposed algorithm gives minimum value except for Zika dataset. With respect to Zika dataset, ABC and CSA perform well. From Figures 6 and 7, it is noted that the FF-OC algorithm gets appropriate values for the respective indexes when compared to the existing algorithms.

Table 11 summarizes the results of the proposed method and the existing methods which are discussed in Section 2 (related work).

Table 10. Association Rule Clustering for Zika Virus Dataset



C1	•child -> protect •child -> neurology •child -> spread •virus -> protect	virus -> fetal child -> virusdisease virus -> child
C2	•fetal -> protect •fetal -> child	fetal -> neurology fetal -> virusdisease
C3	•neurology -> protect •neurology -> babies •neurology -> spread	
C4	•protect -> neurology •protect -> babies •protect -> spread	protect -> virusdisease protect -> pregnancy
C5	•babies -> fetal •babies -> child	
C6	•spread -> fetal •spread -> babies •practice -> protect	practice -> fetal practice -> child practice -> disease
C7	•fact -> child •fact -> fetal •fact -> food	fact -> protect fact -> warheads
C8	•food -> protect •food -> child	food -> fetal food -> fact
C9	•zikaebola -> fetal •zikaebola -> protect	zikaebola -> child zikaebola -> infection
C10	•warheads -> fetal •warheads -> protect	warheads -> child
C11	•pregnancy -> child •infection -> fetal •infection -> protect •disease -> child •polio -> fetal •polio -> protect •pregnancy -> spread	pregnancy -> fetal infection -> child disease -> fetal disease -> protect polio -> child pregnancy -> babies pregnancy -> neurology

Figure 7. Cluster Validation – DB Index

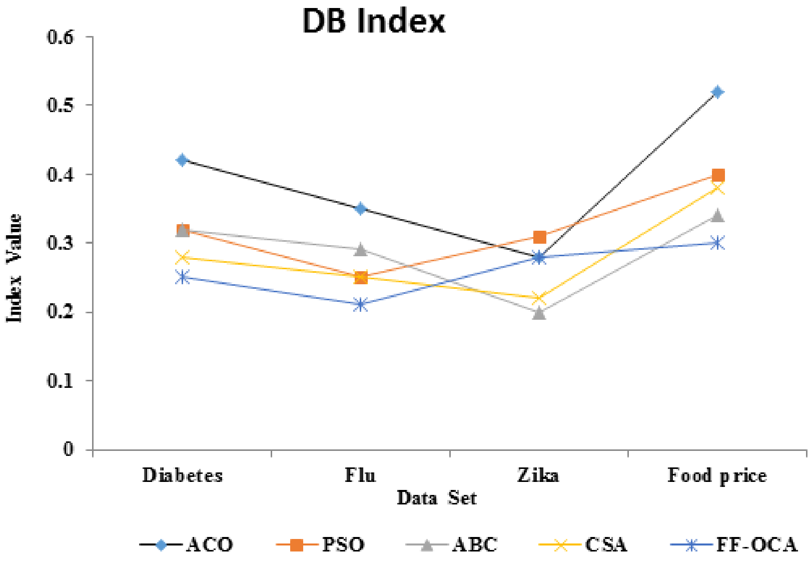


Figure 8. Cluster Validation – Silhouette Index

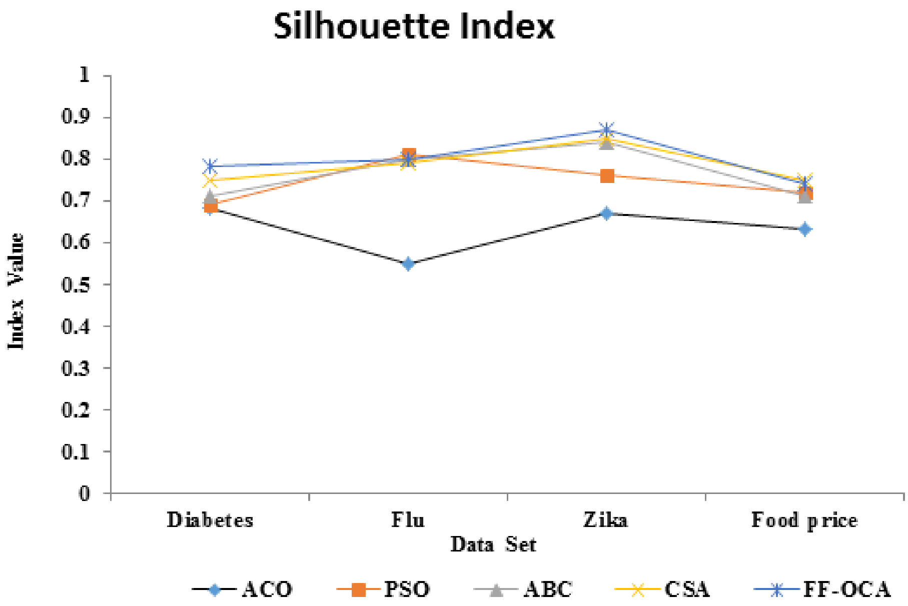


Figure 9. Cluster Validation – C Index

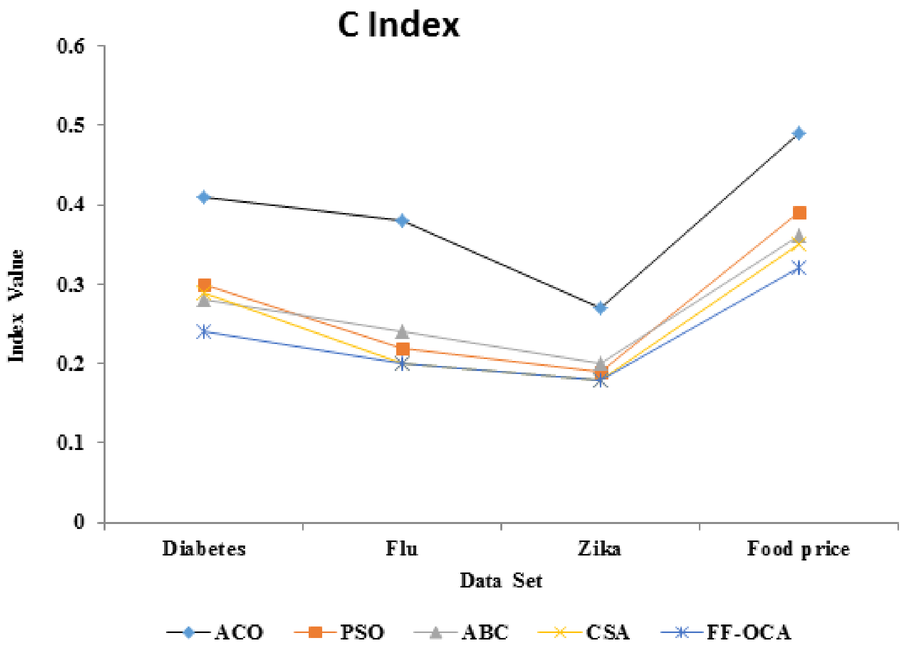


Table 11. Evaluation summary

Author and year	Algorithm	Objective	Datasets	Keywords	No. of Tweets	Precision	Recall	Accuracy
Sandagiri et al (2021)	Long short-term memory (LSTM)	To predict crimes using twitter data.	Twitter Dataset	• Crime	432	0.8250	0.8040	86.40%
Gencoglu et al. (2020)	Bayesian Network	Developed a causal inference approach helps to determine the causal relationships between Covid-19 pandemic characteristics	Twitter corpus	• Covid-19 • Coronavirus	954,902	0.8223	0.7248	74.17%
Zahra et al. (2020)	Random Forest	The eyewitness messages are categorized into direct, indirect, and vulnerable eyewitnesses	Twitter Streaming API Disaster-related tweets	• Floods • Earthquakes • Hurricanes	6000	0.7472	0.6709	68.00%
Shekhawat et al. (2020)	Hybrid Spider Monkey optimization with k-means clustering	A hybrid natural inspired algorithm is used to classify tweets.	Twitter-sanders2	Twitter-sanders2 dataset	25000	0.8704	0.8347	94.45%
Doan et al. (2019)	Natural Language Processing	Extract the causalities from the health related tweets	Twitter corpus	• Stress • insomnia • headache	240,000	0.7816	0.8343	82.14%
Kayesh et al (2019)	Feed-forward neural network	To detect event causality from tweets.	Twitter Stemming API	• Common wealth Games 2018	207,705	0.6746	0.6196	69.94%
Hasan et al. (2019)	Nearest Neighbor Algorithm	Developed an event detection system namely, TwitterNews+, that incorporates specialized inverted indices and an incremental clustering	Twitter Stemming API	• Events2012 corpus	6000	0.8600	0.8200	85.00%
Hu et al. (2017)	K-Medoids algorithm	Evaluate the top- k most insightful sentences from online hotel reviews.	Hotel Reviews	Hotel Reviews	955	0.7286	0.7023	71.00%
Pandey et al. (2017)	Cuckoo search algorithm	Developed a hybrid cuckoo search algorithm for Twitter analysis.	Twitter-sanders-apple	• Sports • saints	3965	0.7400	0.7000	77.99%
Adedoyin-Olowe et al. (2016)	Transaction-based Rule Change Mining algorithm	Analyzed the Twitter conversation to extract interesting topics or events	Twitter Dataset	• FA cup final 2012 • US elections 2012	20000	0.7200	0.7000	70.00%
FF-OC - The proposed method	Firefly and Ontology based clustering	To analyzing the tweets and extract the casual factors	Twitter Stemming API	• Diabetes • Flu • Zika virus • Food price	1110000	NA	NA	89.00%

As prior class labels are not known for the implementation of the proposed algorithm on the dataset, the evaluation metrics like precision, recall, F1 score cannot be computed. Table 11 shows that the proposed methodology gives better accuracy i.e., 89% when compared to other algorithms except the algorithm proposed by Shekhawat et al. This is due to the fact that, the number of tweets considered by the authors is less than 2500. In this research work, around one million tweets are collected and analyzed.

7. CONCLUSION

Social media analytics has become an increasingly researched domain for different kinds of analysis that encompasses identifying flu trends to customer behavior. Individuals use social media to provide opinions on varied topics. A large amount of data thus generated by the users is mined to extract actionable knowledge with the help of various data mining techniques. A hybrid FF-OC algorithm is proposed to efficiently cluster the association rules that are generated from the text corpus of tweets. The factors inferred from the clusters can be used as a decision support tool by the business organizations to study the impact of flash sales, the health originations to determine the factors affecting epidemics, and the government to perceive the public opinion on any new policies. The proposed algorithm's performance is promising in terms of the quality of the factors identified. The insights gained by applying the proposed approach can be used by any organization to propose actions to mitigate the crisis reported by the people in social media.

While the findings of the proposed methodology relate to the food price rise crisis, disease-related tweets, and their impact; the methodology can be applied to other critical applications as well such as factors impacting government policies and natural disasters. The limitation of this work is that manual intervention is required to interpret the results i.e., casual factors.

This research work can be extended further in the following directions:

1. Generate association rules with more than one feature in both antecedent and consequent.
2. Modify the FF-OC algorithm to cluster association rules with more features.
3. Implement the algorithm in a distributed environment to handle a large number of association rules.

REFERENCES

Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., & Gomes, J. B. (2016). A rule dynamics approach to event detection in twitter with its application to sports and politics. *Expert Systems with Applications*, *55*, 351–360. doi:10.1016/j.eswa.2016.02.028

Ahmed, S., Jaidka, K., & Cho, J. (2016). The 2014 Indian elections on Twitter: A comparison of campaign strategies of political parties. *Telematics and Informatics*, *33*(4), 1071–1087. doi:10.1016/j.tele.2016.03.002

Akilandeswari, J., & Jothi, G. (2016). Elimination of Redundant Association Rules—An Efficient Linear Approach. In *Computational Intelligence, Cyber Security and Computational Models* (pp. 171–180). Springer. doi:10.1007/978-981-10-0251-9_18

Alam, S., Dobbie, G., Koh, Y. S., Riddle, P., & Rehman, S. U. (2014). Research on particle swarm optimization based clustering: A systematic review of literature and techniques. *Swarm and Evolutionary Computation*, *17*, 1–13. doi:10.1016/j.swevo.2014.02.001

Alharbi, J. R., & Alhalabi, W. S. (2020). Hybrid Approach for Sentiment Analysis of Twitter Posts Using a Dictionary-based Approach and Fuzzy Logic Methods: Study Case on Cloud Service Providers. *International Journal on Semantic Web and Information Systems*, *16*(1), 116–145. doi:10.4018/IJSWIS.2020010106

Batbarai, A., & Naidu, D. (2014). Approach for rule pruning in association rule mining for removing redundancy. *Int. J. Innov. Res. Comput. Commun. Eng*, *2*(5), 4207–4213.

Cagliero, L. (2011). Discovering temporal change patterns in the presence of taxonomies. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 541–555. doi:10.1109/TKDE.2011.233

Chen, T. Y., Chen, Y. M., & Tsai, M. C. (2020). A Status Property Classifier of Social Media User's Personality for Customer-Oriented Intelligent Marketing Systems: Intelligent-Based Marketing Activities. *International Journal on Semantic Web and Information Systems*, *16*(1), 25–46. doi:10.4018/IJSWIS.2020010102

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. doi:10.1109/TPAMI.1979.4766909 PMID:21868852

Dehkharghani, R., Mercan, H., Javeed, A., & Saygin, Y. (2014). Sentimental causal rule discovery from Twitter. *Expert Systems with Applications*, *41*(10), 4950–4958. doi:10.1016/j.eswa.2014.02.024

Doan, S., Yang, E. W., Tilak, S. S., Li, P. W., Zisook, D. S., & Torii, M. (2019). Extracting health-related causality from twitter messages using natural language processing. *BMC Medical Informatics and Decision Making*, *19*(3), 79. doi:10.1186/s12911-019-0785-0 PMID:30943954

Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, *1*(1), 53–66. doi:10.1109/4235.585892

Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *26*(1), 29–41. doi:10.1109/3477.484436 PMID:18263004

Fister, I., Fister, I. Jr, Yang, X. S., & Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and Evolutionary Computation*, *13*, 34–46. doi:10.1016/j.swevo.2013.06.001

Gencoglu, O., & Gruber, M. (2020). Causal modeling of twitter activity during covid-19. *Computation (Basel, Switzerland)*, *8*(4), 85. doi:10.3390/computation8040085

Hamed, A. A., Wu, X., & Rubin, A. (2014). A twitter recruitment intelligent system: Association rule mining for smoking cessation. *Social Network Analysis and Mining*, *4*(1), 212. doi:10.1007/s13278-014-0212-6

Hasan, M., Orgun, M. A., & Schwitter, R. (2019). Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management*, *56*(3), 1146–1165. doi:10.1016/j.ipm.2018.03.001

Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews–A text summarization approach. *Information Processing & Management*, *53*(2), 436–449. doi:10.1016/j.ipm.2016.12.002

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical & Statistical Psychology*, *29*(2), 190–241. doi:10.1111/j.2044-8317.1976.tb00714.x

Karaboga, D. (2005). An idea based on honey bee swarm for numerical optimization (Vol. 200). Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department.

Kayesh, H., Islam, M., & Wang, J. (2019). *On event causality detection in tweets*. arXiv preprint arXiv:1901.03526.

Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. *Proceedings of ICNN'95-International Conference on Neural Networks*, 4, 1942-1948. doi:10.1109/ICNN.1995.488968

Mehdi, D. (2021). *Improved Particle swarm optimization.*

Mirjalili, S. (2021). *Ant Colony Optimiztion (ACO)*. https://www.mathworks.com/matlabcentral/fileexchange/69028-ant-colony-optimiztion-aco

Mitchell, T. (1996). *Machine learning*. McGraw Hill.

Murthy, D. (2018). *Twitter*. Polity Press.

Murthy, J. S., Siddesh, G. M., & Srinivasa, K. G. (2019). A real-time twitter trend analysis and visualization framework. *International Journal on Semantic Web and Information Systems*, *15*(2), 1–21. doi:10.4018/IJSWIS.2019040101

Nolasco, D., & Oliveira, J. (2019). Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, *93*, 290–303. doi:10.1016/j.future.2018.09.008

Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management*, *53*(4), 764–779. doi:10.1016/j.ipm.2017.02.004

Pulse, U. G. (2014). *Mining Indonesian Tweets to understand food price crises*. UN Global Pulse.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi:10.1016/0377-0427(87)90125-7

Sahoo, G. (2017). A two-step artificial bee colony algorithm for clustering. *Neural Computing & Applications*, *28*(3), 537–551. doi:10.1007/s00521-015-2095-5

Sandagiri, C., Kumara, B. T., & Kuhaneswaran, B. (2021). Deep Neural Network-Based Crime Prediction Using Twitter Data. *International Journal of Systems and Service-Oriented Engineering*, *11*(1), 15–30. doi:10.4018/IJSSOE.2021010102

Shekhawat, S. S., Shringi, S., & Sharma, H. (2020). Twitter sentiment analysis using hybrid Spider Monkey optimization method. *Evolutionary Intelligence*, 1–10.

Xie, H., Zhang, L., Lim, C. P., Yu, Y., Liu, C., Liu, H., & Walters, J. (2019). Improving K-means clustering with enhanced Firefly Algorithms. *Applied Soft Computing*, *84*, 105763. doi:10.1016/j.asoc.2019.105763

Yang, X. S. (2008). *Nature-Inspired Metaheuristic Algorithms*. Luniver Press.

Yang, X.-S. (2021). *The Standard Cuckoo Search (CS)*. https://www.mathworks.com/ matlabcentral/fileexchange/74767-the-standard-cuckoo-search-cs

Yang, X. S., & Deb, S. (2009, December). Cuckoo search via Lévy flights. In *2009 World congress on nature & biologically inspired computing (NaBIC)* (pp. 210-214). IEEE. doi:10.1109/NABIC.2009.5393690

Yang, X. S., & He, X. (2013). *Firefly algorithm: Recent advances and applications*. arXiv preprint arXiv:1308.3898.

Yarpiz. (2021). *Artificial Bee Colony (ABC) in MATLAB*. https://www.mathworks.com/matlabcentral/fileexchange/52966-artificial-bee-colony-abc-in-matlab

Yarpiz. (2021). *Particle Swarm Optimization (PSO)*. https://www.mathworks.com/ matlabcentral/fileexchange/52857-particle-swarm-optimization-pso

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, *57*(1), 102107. doi:10.1016/j.ipm.2019.102107

Zamazal, O. (2020). A Survey of Ontology Benchmarks for Semantic Web Ontology Tools. *International Journal on Semantic Web and Information Systems*, *16*(1), 47–68. doi:10.4018/IJSWIS.2020010103

J. Akilandeswari is Dean Academics and Professor cum Head of Department of IT, Sona College of Technology, Salem. She has experience of over 24 years in Teaching and Research. She is a Fulbright scholar. She had completed M.E and Ph.D. in Computer Science and Engineering from NIT, Tiruchirappalli. She has published papers in various international journals and conferences. Her areas of interest include data mining, web mining, data analytics, and cloud computing.

G. Jothi is currently working as an Assistant Professor in the Department of Computer Applications at Sona College of Arts and Science, Salem. She has completed M. Phil and Ph.D. in Computer Science from Periyar University, Salem. She has published papers in various reputed journals and conferences in the area of data mining, big data analytics, and image processing.

K. Dhanasekaran is currently working as Assistant Professor in the Department of Data Science and Business Systems, School of Computing at SRM Institute of Science and Technology (formerly SRM University), Chengalpattu, Chennai, Tamilnadu, India. He has over 10 years of experience in education industry. His research interest includes machine learning, deep learning, and natural language processing. In his service, he has worked as research coordinator & NBA coordinator. As a project guide, he has received "Best project of the Year" cash award from KSCST-DST. Apart from that, he has received funding from funding Agencies, namely, ISRO (Seminar Grant), CSIR (SYMPOSIA Grant), and has organized various Conferences, Seminars, Workshops etc. for the skill and knowledge development of students and faculty members. He has published 25 research papers, and one Book Chapter in IGI Global, USA. He is also a member of the ISTE. He has been the session chair for two international conferences and has delivered sessions at various events.

K. Kousalya has obtained her Ph.D. degree in Cloud Computing in the year 2010 from Anna University, Chennai. Currently, she is a professor in the Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Tamilnadu. She has published more than 75 research articles in International/National Journals. She has guided more than 15 Ph.D. scholars. Her area of interest includes Cloud Computing, data mining, image processing, and optimization Techniques.

V. Sathiyamoorthi is currently working as a Professor in Computer Science and Engineering Department at Sona College of Technology, Salem, Tamil Nadu, India. He was born on June 21, 1983, at Omalur in Salem District, Tamil Nadu, India. He received his Bachelor of Engineering degree in Information Technology from Periyar University, Salem with First Class. He obtained his Master of Engineering degree in Computer Science and Engineering from Anna University, Chennai with Distinction and secured 30th University Rank.He received his Ph.D degree from Anna University, Chennai in Web Mining. His areas of specialization include Web Usage Mining, Data Structures, Design and Analysis of Algorithm and Operating System. He has published many papers in International Journals and conferences. He has published many books and book chapters in various renowned international publishers. He has also participated in various National level Workshops and Seminars conducted by various reputed institutions.