Binary Bat Algorithm for text feature selection in news events detection model using Markov clustering

Wafa Zubair Al-Dyani, Farzana Kabir Ahmad and Siti Sakira Kamaruddin

Cogent Engineering (2022), 9: 2010923
COMPUTER SCIENCE | RESEARCH ARTICLE

Binary Bat Algorithm for text feature selection in news events detection model using Markov clustering

Wafa Zubair Al-Dyani1,2*, Farzana Kabir Ahmad1 and Siti Sakira Kamaruddin1

Abstract: Feature Selection (FS) phase is crucial in the Event Detection (ED) model. Several studies have captured the most informative features using various filter and wrapper FS methods. Recently, FS methods based on swarm intelligence algorithms have been employed to determine the relevant features. Nevertheless, ED from sparse and high-dimensional feature space resulting from a massive number of news documents with different text lengths is a challenging task. Such feature space consists of redundant, irrelevant, and noisy data, which misguide the detection process and substantially affect the reliability of the ED model. Hence, this study proposes a novel Binary Bat Algorithm (BBA) and Markov Clustering Algorithm (MCL) to improve the performance of the ED model. To the best of our knowledge, BBA is employed for the first time in this study in the context of the ED field. The proposed method is tested on 10 benchmark datasets and 2 primary Facebook news datasets using the average of several evaluation metrics such as F-measure (F), Precision (PR), Recall (R), and Selected Feature Ratio (SFR). Comparative experiments against the basic MCL, Binary versions of the Genetic Algorithm and Particle

ABOUT THE AUTHOR

Farzana Kabir Ahmad is a senior lecturer at School of Computing, Universiti Utara Malaysia, MALAYSIA. She holds a Bachelor degree of Computer Science (with Honours) from Universiti Sains Malaysia in 2003 and a Master degree in Computer Science from the same university later in 2005. She pursued her Ph.D. in Computer Science (Bioinformatics) from Universiti Teknologi Malaysia in 2012 and her doctoral work involves the development of synergy network for breast cancer progression. Her main research interests are in machine learning and data mining projects that seek hidden information from huge, complex data set and finally generate/built models to ease human decision-making process. With the emergence of big data era, the data complexity, and the challenge to understand them has doubled that makes this work more interesting. At the moment, most of her researches are related to events detection, fake news detection and predictive modelling main in social tension environment studies. She is engaging with big data analytics, bio-inspired algorithms, and text mining-based research.

PUBLIC INTEREST STATEMENT

A lot of real-world events are happening around the world. The newswires sources are the most essential sources for obtaining information about such events. News documents are published on various platforms, including news official websites and news official pages on Facebook and Twitter. However, identifying the most significant real-world events from the massive number of published news documents on such platforms becomes very challenging and difficult task. Given the importance of identifying real-world events for many decision-makers in various fields to achieve their different goals, many researchers over the years have built different Event Detection (ED) models. Despite the existence of numerous ED models, changes in the contents of news documents, as well as the pace with which these documents are disseminated and increased in quantity, require the improvement of such models to cope with these changes. As a result, this study attempts to improve the underline methods utilised in improving the model’s performance from news documents with varying text lengths and contents.
Swarm Optimization are implemented in this study. The empirical results proved that BBA-MCL outperforms other methods on most datasets based on F and PR metrics. Furthermore, the statistical results confirmed that the BBA-MCL FS method has significantly enhanced MCL performance with p-value = 0.003, by generating the most informative features. Ultimately, this work concludes that BBA-MCL obtains significant features and effectively detects real-world events from heterogeneous news text documents.

**Subjects:** Algorithms & Complexity; Artificial Intelligence; Evolutionary Computing; Machine Learning - Design; Computer Science (General); Data Preparation & Mining; Surface Engineering-Materials Science

**Keywords:** Binary Bat Algorithm (BBA); Markov Clustering (MCL); Feature Selection (FS); Event Detection (ED); wrapper methods; heterogeneous news; text clustering

**1. Introduction**
Event Detection (ED) is defined as “the process of automatically identifying real-world events from different data streams” (Fu et al., 2014). ED falls under the research umbrella of Topic Detection and Tracking (TDT) and has been aggressively studied over the past decade (Panagiotou et al., 2016). TDT defines a topic as “a collection of events/stories that inform about the same subject”, meanwhile an event is defined as “a specific thing that happened in specific time and location”, which requires research to answer questions such as what is the event, when and where it has happened as well as who was involved (Goswami & Kumar, 2016). In general, there are two types of ED models namely; (a) New Event Detection (NED) models (i.e., online ED models), and (b) Retrospective Event Detection (RED) models (i.e., offline ED models). NED focuses on the detection of recently occurred events from online data streams, while RED identifies hidden events from the historic repository in an off mode style. Recently, ED on digital news documents has received a lot of attention as it is considered a good source to exchange numerous information about real-world events (Gashi & Ahmeti, 2021; Mele et al., 2019; Wada, 2021). Examples of such news documents are those published by various newswire organizations on different Internet platforms such as the official news websites and the news pages on various Social Media (SM) sites, e.g., Facebook, Twitter, and Instagram. In addition, it is reported that ED from multiple news sources is more effective in comparison to a single news source (Leban et al., 2014). Therefore, many researchers in the literature have proposed models to detect events from multiple heterogeneous news sources, whereby news documents are varied in structure, written styles, language, or length (Mhamdi et al., 2018; Moutidis & Williams, 2019; Prasad et al., 2018; Rasouli et al., 2020; Wei et al., 2018; Yu & Wu, 2018). Based on our previous survey (Al-Dyani et al., 2020), ED models have been extensively studied by many researchers due to their advantages to the community such as:

(i) In the news analysis area, several researchers have utilized ED models to investigate events mostly reported by news channels. In these studies, they are interested to determine e.g., what are the most-discussed events by each news channel (Salloum, Al-Emran et al., 2017), which news channel most frequently publishing news articles (Salloum, Mhamdi et al., 2017), or which kind of events that people are highly attracted to and interested in sharing about it (Salloum, Al-Emran, Shaalan et al., 2017). Such ED is helpful in providing early warning and faster responses to events caused by natural or man-made disasters.

(ii) ED model could help news channels’ managers in recognizing the most popular real-world events among news readers through analysing the meta-data information associated with the discovered events from Facebook news posts (e.g., number of likes, comments, sharing,
engagement, etc). Consequently, this can assist them in improving their strategies for selecting the type of news to be published in the future.

(i) ED models can help in organizing the published news documents into various events that could be beneficial for readers in finding their desired documents easily and effectively. ED can help policymakers in different disciplines to make the right decisions as they could obtain valuable knowledge about the past hidden real-world events (Ramadan & Mohd, 2011).

Regardless of the numerous benefits that ED offered, developing an ED model for multiple heterogeneous news text documents that vary in length could result in high dimensional feature space (Panagiotou et al., 2016). Such space consists of redundant, irrelevant, and noisy features, which misguide the detection methods and substantially, affect the reliability of the ED model (Xue et al., 2016). To overcome this problem, most ED studies have applied different traditional FS techniques such as Term Frequency (TF) (Beigh et al., 2016; Dai & Sun, 2010; Rasouli et al., 2020), and Term Frequency Inverse Document Frequency (TFIDF) (Mhamdi et al., 2018; Prasad et al., 2018; Salloum, Al-Emran et al., 2017; Salloum, Al-Emran, Shaalan et al., 2017; Salloum, Mhamdi et al., 2017) to select relevant words for the subsequent phases of ED model. However, such techniques suffer from the drawback of specifying a threshold to select the top features. In contrast, some researchers have utilized different methods as an alternative to FS techniques, for instance, word embeddings (Hu et al., 2017), LDA (Mele et al., 2019), Part of Speech (POS) (Nanba et al., 2013), and Named Entity Recognition (NER) (Moutidis & Williams, 2019). Nevertheless, the word embedding method requires a large volume of data for training as they are essentially supervised methods and LDA needs a predefined number of topics while POS and NER are heavily dependent on the existing lexicons or dictionaries for extracting different parts of words or named entities. In addition, NER is very difficult to understand, particularly when the same named entities are used to describe different events (Édouard et al., 2017). Furthermore, the NER process requires removing a large portion of news articles as not all documents hold named entities (Moutidis & Williams, 2019), which eventually could reduce the overall performance of the ED model. Apart from all the above studies, several researchers have proceeded directly to the ED phase, and totally ignored the FS phase that caused poor performance of the ED model (Mhamdi et al., 2018).

Given the limitations in the existing ED studies in solving the curse of high dimensionality for detecting events from heterogeneous news documents, this study has motivated to propose a novel wrapper FS method based on Binary Bat Algorithm (BBA) and Markov Clustering (MCL) method namely, BBA-MCL. BBA (Nakamura et al., 2012), is a Meta-Heuristic Algorithm (MHA), which has shown better results in solving FS problems for various data mining applications, particularly in intrusion detection (A. Enache & Science, 2015), community detection (Sharma & Annappa, 2016), e-fraud detection (Akinleye & Adewumi, 2018), anomaly detection (A.-C. Enache & Sgarciu, 2014), etc. Despite the brilliant success of BBA in these research fields, BBA has never been used in ED (Al-Dyani et al., 2020; Al-Dyani et al., 2018). To effectively fill this gap, this paper has introduced the wrapper FS BBA-MCL method. The key idea is to wrap the BBA with a well-known graph clustering method MCL. MCL recently has been used successfully in the ED domain (Chen et al., 2017). Based on our best knowledge, this study is the first to introduce such a wrapper FS method to select the optimal feature subset in order to identify the events from multiple heterogeneous news documents in which news documents vary in length. The rest of the paper is organized as follows. Sections II describes the materials and methods used in this study. Experimental results and discussion are presented in Section III and Section IV, respectively. Finally, conclusion is stated in Section V.
2. The materials and method

2.1. Bat algorithm
Bat Algorithm (BA) was developed by Yang (Yang, 2010) who was inspired by the echolocation property of microbats. The basic BA consists of four main steps as follows:

Initialize bat population: the population of bats is set using randomly selected values from a collection of real numbers i.e., from lower to higher values. Then, the solutions are assessed using Equation (1)

\[ x_{ij} = x_{\text{min}} + \varphi (x_{\text{max}} - x_{\text{min}}) \]  

(1)

where \( i = 1, 2, \ldots N \), \( j = 1, 2, \ldots d \), \( x_{\text{min}} \) and \( x_{\text{max}} \) are lower and higher borders for dimension \( j \), respectively. \( \varphi \) is a randomly selected value from [0,1].

- Updating frequency, velocity, and new solutions: within this step, the position \( x_i \) and velocity \( v_i \) of every single bat are updated during the iterations. Hence, the new solution \( x_i^t \) and velocity \( v_i^t \) at time step \( t \) can be computed using Equations (2)-(4)

\[ f_i = f_{\text{min}} + (f_{\text{max}} - f_{\text{min}}) \theta. \]  

(2)

\[ v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*) f_i. \]  

(3)

\[ x_i^t = x_i^{t-1} + v_i^t. \]  

(4)

where \( \theta \) is uniformly chosen from [0,1], \( x_* \) is the current best global solution (i.e., location) recognized from all \( N \) bats. In addition, each bat has a frequency rate uniformly selected from \([f_{\text{min}}, f_{\text{max}}]\). For the local search, every single bat walk around the best solution found so far, and hence, a new solution for every single bat is generated locally by means of Equation (5)
\[ x_{\text{new}} = x_{\text{old}} + \varepsilon A^T \] (5)

where \( \varepsilon \) is randomly chosen from \([-1,1]\), and \( A_t = \langle A^T \rangle \) represents the average loudness of all bats at \( t \)th time step. While \( x_{\text{old}} \) is the best solution identified by a specific mechanism.

- Updating pulse rates \( (r) \) and loudness \( (A) \): using a random technique, different initial values are assigned to the \( A \) and \( r \) parameters of BA. Then, \( A \) and \( r \) are updated utilizing Equations (6) and (7); if the new solutions are improved.

\[ A^{t+1}_i = \alpha A^{t}_i \] (6)

\[ r^{t+1}_i = r^{t}_i \left( 1 - \exp(-\gamma t) \right) \] (7)

where \( \gamma \) and \( \alpha \) are constants; for any \( 0 < \alpha, \gamma < 1 \).

- Evaluation, Saving, and Ranking Best Solutions: If the accomplished solutions satisfy the given condition, then they will be stored conditionally as best solutions. Finally, a ranking step will be performed on all bats to identify the current best solution \( (x_*) \).

The fundamental steps of an ordinary BA can be summarized in Algorithm 1 (see Figure 1).

### 2.2. Binary Bat Algorithm

BBA was introduced by Nakamura et al. (Nakamura et al., 2012). BBA has a similar structure as the basic BA but with a slight difference in the update position equation of the BA, where it is replaced with binary vectors through applying one of the transfer functions (e.g., sigmoid function) using Equation (8)

\[ S(v^0_i) = \frac{1}{1 + e^{-v^0_i}} \] (8)

hence Equation (4) of generating a new BA’s location is replaced with Equation (9)

\[ x^t_i = \begin{cases} 1 & \text{if } S(v^0_i) > \sigma \\ 0 & \text{if otherwise} \end{cases} \] (9)

in which (1) Means the feature is selected and (0) indicates that the feature is not selected, where \( \sigma \in (0,1) \).
Table 1. Characteristics of the text news datasets

| Dataset name       | Short name | #Documents | #Features | #Events |
|--------------------|------------|------------|-----------|---------|
| 20newsgroup        | DS1        | 100        | 2894      | 5       |
| 20newsgroup        | DS2        | 100        | 2529      | 10      |
| 20newsgroup        | DS3        | 100        | 2823      | 20      |
| 20newsgroup        | DS4        | 200        | 4637      | 10      |
| 20newsgroup        | DS5        | 200        | 5383      | 20      |
| 20newsgroup        | DS6        | 300        | 6556      | 20      |
| News aggregator    | DS7        | 800        | 1085      | 4       |
| News aggregator    | DS8        | 2000       | 2106      | 4       |
| News articles      | DS9        | 1467       | 14,770    | 56      |
| RSS news posts     | DS10       | 2095       | 3818      | 56      |
| Facebook news posts| DS11       | 1139       | 3742      | 33      |
| Facebook news posts| DS12       | 1074       | 3420      | 16      |

2.3. Markov clustering algorithm

MCL was developed by Dongen (Van Dongen, 2000). It has become a popular clustering algorithm in many fields such as the bioinformatics domain for identifying protein interaction, ED (Manaskasemsak et al., 2016), topic detection (Chen et al., 2017), document clustering (Altuncu et al., 2019), spam detection (El-mawass et al., 2018), analysis network interaction (Szilágyi & Szilágyi, 2014), video and image processing (Bustamam et al., 2018). This success is due to MCL’s unique features that differ from the existing graph-based techniques (Bustamam et al., 2018). One of the most significant characteristics of MCL is that no predefined number of clusters is needed. This uniqueness of MCL is very vital, especially when dealing with events as the number of events cannot be predicted in real life and thus, such a technique is appropriate to be used in ED models. Additionally, MCL could handle noisy data well and is very fast compared to other graphical clustering techniques (Chen et al., 2017).

Initially, MCL was inspired by the idea of randomly walking through a graph G, whereby MCL intends not to leave the cluster until it traverses as many nodes within the cluster. MCL algorithm has three main operations, namely expansion, inflation, and pruning. The expansion process opens new flows and increases the flow among existing nodes in the transition probability matrix M. In other words, expansion tends to introduce new non-zero values in M through generating new edges and eliminating the old ones that are not needed in the graph. In contrast, the inflation process aims to reinforce the strong edges and deteriorate the weak edges, hence helping in the elimination of weak edges and removing zero values in the M matrix using the pruning process. In the inflation process, every single element of the matrix M is raised by the inflate parameter r. This operation is known as the Hadamard operation, and it is calculated using Equation (10)

\[ Minf = Inflate(M, r) = \frac{M(i,j)^r}{\sum_{k=1}^{n} M(k,j)} \]  

(10)

The pruning process carries out after the inflation operation within each iteration in order to save memory. In the nutshell, expansion, inflation, and pruning processes are iteratively implemented until the transition probability matrix or also known as a stochastic matrix (M) is converged. The
matrix $M$ is converged when there is no more change in its elements or there are slight changes from the elements in the previous matrix of MCL. Consequently, the whole given graph by this matrix is partitioned into many clusters without any overlapping. The pseudo code of the MCL algorithm is presented in Algorithm 2 as shown in Figure 2.

2.4. The proposed methodology
The ED model consists of five phases that include (1) data preparation phase, (2) text pre-processing phase, (3) FS phase, (4) ED phase, and (5) evaluation phase. The experiment has conducted using Python (version 3.7) on a machine with 8 GB RAM in Windows 10 environment. The following subsections explain each phase in the ED model.

(1) Datasets Preparation Phase: Several datasets have been used in this study, which includes 10 benchmark datasets (secondary datasets) and 2 primary Facebook datasets (see Table 1). These datasets have been segmented into different groups as being done by studies in (Prasad et al., 2018), (Abualigah & Khader, 2017; Abualigah, Khader, Al-Betar et al., 2016; Abualigah, Khader, AlBetar et al., 2016; Huang et al., 2013). The main idea behind this segmentation is to create textual datasets that are sparse and contain high dimensional feature space. In addition, the datasets are used to test and validate the performance of the proposed methods with different quantities of documents, diverse attributes, and various numbers of events. For instance, 20 new group was divided into six different datasets (DS1-DS6), and the news aggregator was broken into two datasets (DS7 and DS8), and Facebook news posts were divided into two datasets (DS11 and DS12). The number of documents and events are determined by using stratified random sampling, while the number of features is obtained using TFIDF method.

To demonstrate, DS1 contains 100 random documents that belong to five categories. DS2 includes 100 random documents that belong to 10 categories. DS3 consists of 100 documents that belong to 20 categories. DS4 contains 200 random documents that belong to 10 categories. DS5 includes 200 documents that belong to 20 categories. Finally, DS6 consists of 300 documents that belong to 20 categories. D7 and D8 were created from news aggregator datasets by randomly
selecting 800 and 2000 documents, respectively, which were assigned in 4 categories. DS9 contains 1467 documents and DS10 consists of 2095 documents that were distributed over 56 events. D11 and D12 were constructed from Facebook news posts. D11 includes 1139 posts that belong to 33 events, meanwhile, D12 contains 1074 posts that belong to 16 events.

(1) Text Pre-Processing Phase: Text preprocessing is a vital phase in the ED model to enhance the detection of events from the huge volume of news text documents. The present study has applied the standard preprocessing techniques that have been widely used in the area of ED. Specifically, five pre-processing steps were employed: stop words removal, URLs removal, tokenization, stemming, and text document representation with term weighting scheme TFIDF.

(2) Feature Selection Phase (Proposed BBA-MCL): Let \( n \) be several news documents that contain a set of features \( F = \{ f_{i,1}, f_{i,2}, \ldots, f_{i,j}, \ldots, f_{i,m} \} \), where \( m \) is the total of all exclusive features for the \( n \) documents, \( i \) is the number of news document, and \( j \) is the number of features. Let \( SF = \{ s_{f_1,i}, s_{f_2,i}, \ldots, s_{f_k,i} \} \) is a set of informative features selected by the proposed BBA-MCL method with a new dimension of features (i.e., feature space), \( t \) is the new number of unique features, and \( s_{f_i,j} \in \{0, 1\} \), if \( s_{f_i,j} = 1 \), then \( j \) feature is selected, while 0 indicates...
that $j$ feature is discarded. Indeed, the FS phase works in parallel with the subsequent ED phase as the BBA FS method is wrapped with the MCL ED method as it is depicted in Figure 3.

1) Event Detection Phase: In this phase, the MCL ED works simultaneously with BBA from the FS phase whereby the swarm of bats in the BBA-MCL method starts with randomly initialized solutions and enhances its population to find the global optimum solution. The collection of bats is represented as vectors (i.e., rows) and each bat has several positions. The $j$th location in the bat indicates the state of the $j$th feature i.e., selected ($j = 1$) or not selected ($j = 0$). Subsequently, each solution generated by the BBA is used to construct an undirected-weighted graph that is given to the MCL to partition it into distinct clusters. The granularity of the clustering process for MCL is measured by the $Q$ criterion. The large value of $Q$ shows a

| Algorithm | Parameter | Value |
|-----------|-----------|-------|
| BBA       | Number of Bats | 20    |
|           | Max iterations | 100   |
|           | Loudness A    | 0.5   |
|           | Emission Rate $r$ | 0.5  |
|           | Alpha         | 0.9   |
|           | Gamma         | 0.9   |
|           | Fmin          | 0     |
|           | Fmax          | 2     |
|           | Stopping criterion | Max iteration |
| GA        | Number of Individuals | 20   |
|           | Max iterations | 100   |
|           | Selection ratio | Roulette wheel |
|           | Crossover ratio | 0.9   |
|           | Mutation      | 0.005 |
|           | Number of iterations | 100 |
|           | Stopping criterion | Max iteration |
| BPSO      | Number of Particles | 20   |
|           | Max iterations | 100   |
|           | $C1$, $C2$    | 2,2   |
|           | Inertia Weight ($w$) | Is decrease linearly from 0.9 to 0.4 |
|           | Max velocity  | 6     |
|           | Stopping criterion | Max iteration |
| MCL       | Expansion     | 2     |
|           | Inflation (vary depending on preliminary experiments) | 1.5 (DS3, DS12), 2 (DS5, DS6, DS7), 3 (DS8, DS10), 5 (DS2), 6 (DS6, DS11), 7 (DS1), 9 (DS4) |
|           | Loop value    | 1     |
|           | Pruning_threshold (vary depending on preliminary experiments) | 0.1 (DS1-DS6, DS12), 0.01(DS8-DS10), 0.001(DS7) |
|           | Pruning_frequency | 1    |
|           | Convergence_check_frequency | 1    |

Table 2. Initial parameters setting for BBA, GA, BPSO, and MCL algorithms
good quality of the cluster’s structure and its value is always less than one and may be negative.

At each iteration, the fitness value is computed for every single BBA-MCL solution to determine if an improvement is identified to accept and save or reject it. Then, the solutions for all BBA-MCL are ranked and the solution with the maximum fitness value is recognized as the optimal feature subset at that iteration. This process is repeated during iterations until the stopping criterion for BBA is met i.e., iteration number reached 100. Finally, the output event clusters based on the optimal feature subset with the highest $Q$ value are recognized as the final output event clusters. The pseudocode of the proposed wrapper BBA-MCL FS method is shown in Algorithm 3 (see Figure 4).

### Table 3. Performance of methods based on average $F (F_{avg})$

| Datasets | MCL | GA-MCL | BPSO-MCL | BBA-MCL |
|----------|-----|--------|----------|---------|
| DS1      | 0.147 | 0.356 | 0.332 | 0.360 |
| DS2      | 0.288 | 0.353 | 0.338 | 0.357 |
| DS3      | 0.303 | 0.309 | 0.301 | 0.322 |
| DS4      | 0.195 | 0.348 | 0.347 | 0.349 |
| DS5      | 0.263 | 0.281 | 0.303 | 0.311 |
| DS6      | 0.243 | 0.258 | 0.280 | 0.283 |
| DS7      | 0.515 | 0.552 | 0.550 | 0.555 |
| DS8      | 0.303 | 0.310 | 0.302 | 0.310 |
| DS9      | 0.578 | 0.617 | 0.668 | 0.672 |
| DS10     | 0.568 | 0.618 | 0.597 | 0.573 |
| DS11     | 0.814 | 0.840 | 0.829 | 0.843 |
| DS12     | 0.559 | 0.596 | 0.564 | 0.577 |

### Table 4. Performance of methods based on average $PR (PR_{avg})$

| Datasets | MCL | GA-MCL | BPSO-MCL | BBA-MCL |
|----------|-----|--------|----------|---------|
| DS1      | 0.428 | 0.488 | 0.494 | 0.519 |
| DS2      | 0.344 | 0.363 | 0.377 | 0.387 |
| DS3      | 0.239 | 0.244 | 0.249 | 0.265 |
| DS4      | 0.258 | 0.377 | 0.390 | 0.393 |
| DS5      | 0.216 | 0.221 | 0.255 | 0.274 |
| DS6      | 0.200 | 0.196 | 0.233 | 0.242 |
| DS7      | 0.897 | 0.989 | 0.986 | 0.992 |
| DS8      | 0.960 | 0.976 | 0.974 | 0.975 |
| DS9      | 0.445 | 0.467 | 0.535 | 0.548 |
| DS10     | 0.645 | 0.662 | 0.666 | 0.671 |
| DS11     | 0.801 | 0.835 | 0.828 | 0.853 |
| DS12     | 0.616 | 0.720 | 0.728 | 0.739 |
### Table 5. Performance of methods based on average $R (R_{\text{avg}})$

| Datasets | MCL | GA-MCL | BPSO-MCL | BBA-MCL |
|----------|-----|--------|----------|---------|
| DS1      | 0.080 | 0.282  | 0.251    | 0.278   |
| DS2      | 0.170 | 0.349  | 0.309    | 0.333   |
| DS3      | 0.200 | 0.431  | 0.382    | 0.414   |
| DS4      | 0.110 | 0.325  | 0.314    | 0.315   |
| DS5      | 0.175 | 0.389  | 0.376    | 0.366   |
| DS6      | 0.163 | 0.383  | 0.355    | 0.345   |
| DS7      | 0.385 | 0.383  | 0.381    | 0.386   |
| DS8      | 0.169 | 0.184  | 0.179    | 0.184   |
| DS9      | 0.919 | 0.911  | 0.889    | 0.868   |
| DS10     | 0.679 | 0.580  | 0.555    | 0.501   |
| DS11     | 0.849 | 0.844  | 0.830    | 0.834   |
| DS12     | 0.555 | 0.510  | 0.461    | 0.473   |

### Table 6. Performance of methods based on SFR

| Datasets | GA-MCL | BPSO-MCL | BBA-MCL |
|----------|--------|----------|---------|
| DS1      | 0.59   | 0.49     | 0.49    |
| DS2      | 0.63   | 0.62     | 0.54    |
| DS3      | 0.65   | 0.56     | 0.55    |
| DS4      | 0.81   | 0.64     | 0.69    |
| DS5      | 0.72   | 0.59     | 0.61    |
| DS6      | 0.68   | 0.63     | 0.63    |
| DS7      | 0.81   | 0.68     | 0.54    |
| DS8      | 0.78   | 0.70     | 0.61    |
| DS9      | 0.76   | 0.75     | 0.44    |
| DS10     | 0.82   | 0.73     | 0.64    |
| DS11     | 0.81   | 0.72     | 0.82    |
| DS12     | 0.84   | 0.70     | 0.67    |

### Table 7. Average computational time (in seconds) on different fs methods

| Datasets | GA-MCL | BPSO-MCL | BBA-MCL |
|----------|--------|----------|---------|
| DS1      | 122.99 | 315.62   | 9.30    |
| DS2      | 51.92  | 111.30   | 4.00    |
| DS3      | 25.85  | 150.67   | 7.10    |
| DS4      | 159.59 | 449.96   | 26.50   |
| DS5      | 134.05 | 520.20   | 56.60   |
| DS6      | 356.58 | 886.11   | 160.00  |
| DS7      | 1920.37| 5108.23  | 2025.60 |
| DS8      | 10,710.10| 27,830.11| 17,925.10|
| DS9      | 7779.60| 16,443.46| 10,426.96|
| DS10     | 19,179.30| 50,421.10| 14,026.02|
| DS11     | 2558.40| 11,421.85| 5402.60 |
| DS12     | 1156.60| 4349.02  | 1766.70 |
Figure 5. Convergence graphs of comparative algorithms for all datasets DS1-DS12.
(1) Evaluation Phase: In this study, Q was used in the proposed BBA-MCL FS method as a fitness function in order to evaluate its performance in terms of selecting the optimal feature subset based on the high quality of the cluster’s structure. Newman (2004) introduced Q to assess the quality of the generated clusters. Q is built on the idea that the given graph is a random graph that has no clear structure for clusters. It is calculated by computing the density of intra-cluster edges and comparing it with the density of inter-cluster edges. Q has various formulas based on the type of graph under the research e.g., undirected-weighted graph or directed-weighted graph. Since an undirected weighted graph was built and fed into MCL for the experiment of this study, therefore the formula that is represented by Equation (11) was utilized:

\[ Q = \frac{1}{2m} \sum_{ij} \left[ w_{ij} - \frac{k_i k_j}{2m} \right] \delta (C_i, C_j) \] (11)

where \( m \) is the number of edges, \( k_i \) is the degree of node \( i \), while \( k_j \) is the degree of node \( j \), \( C_i \) is the community that node \( i \) belongs, \( C_j \) is the community that node \( j \) belongs to, and \( \delta (C_i, C_j) = 1 \), if \( i \) and \( j \) belong to the same community, otherwise it equals to 0.

On the other hand, the performance of the proposed wrapper BBA-MCL FS method has been evaluated utilizing three evaluation measurements: F-measure (F), Precision (PR), and Recall (R). F, PR, and R are common measures used to evaluate different FS methods in the text clustering domain (Abualigah, Khader, Hanandeh et al., 2018a). In addition, the same metrics have been employed to estimate the performance of various detection methods in the ED domain (Rasouli et al., 2020; Wei et al., 2018). F is a famous measure used extensively in the field of text clustering (Hong et al., 2015). F is based on two metrics: PR and R, where PR and R are calculated using Equations 12 and 13, respectively.

\[ PR(i,j) = \frac{n_{ij}}{n_j} \] (12)

\[ R(i,j) = \frac{n_{ij}}{n_i} \] (13)

where \( n_{ij} \) indicates the number of documents of class \( i \) in cluster \( j \), \( n_j \) represents the number of documents of cluster \( j \) and \( n_i \) indicates the number of documents of class \( i \).

\( F \) for cluster \( j \) is calculated using Equation (14)

\[ F(j) = \frac{2 \times PR(i,j) \times R(i,j)}{PR(i,j) + R(i,j)} \] (14)

| Algorithms   | Mean Rank | Ranking |
|--------------|-----------|---------|
| BBA-MCL      | 2.71      | 1       |
| GA-MCL       | 1.96      | 2       |
| BPSO-MCL     | 1.33      | 3       |
| p-value      | 0.003     |         |
Table 9. Results of Wilcoxon signed-rank test based on $F_{AVG}$

| Algorithms         | Total Datasets | +  | -  | =  | p-value |
|--------------------|----------------|----|----|----|---------|
| BBA-MCL—GA-MCL    | 12             | 9  | 2  | 1  | 0.154   |
| BBA-MCL—BPSO-MCL  | 12             | 11 | 1  | 0  | 0.028   |

$P_R(i,j)$ denotes the precision of documents of class $i$ in cluster $j$, while $R(i,j)$ is the recall of documents of class $i$ in cluster $j$, and hence the $F$ for all clusters is computed using Equation (15):

$$F = \sum_{j}^{n} \frac{m_i}{m_j} \max(n(i,j)). \tag{15}$$

Average values of $F$, $PR$, and $R$ are computed using Equation (16):

$$M_{AVG} = \frac{1}{w} \sum_{i=1}^{w} M_i^j \tag{16}$$

where $M$ in $M_i^j$ represents either $F$, $PR$, or $R$ value in $i$th run. Besides the three evaluation metrics mentioned earlier, the Selected Feature Ratio ($SFR$) metric is also used and computed using Equation (17):

$$SFR = \frac{1}{w} \sum_{i=1}^{w} \frac{\text{length}(x)_i^j}{|D|} \tag{17}$$

where $w$ is the total number of runs, $|D|$ is the whole number of features, and $\text{length}(x)_i^j$ is the length of the selected feature subset in the $i$th run.

The proposed method is evaluated using the average (AVG) of $F$, $PR$, and $R$ on 10 benchmark datasets and 2 primary Facebook news datasets. Furthermore, the $SFR$ values are also recorded for these datasets. Moreover, comparative experiments against the basic MCL, Binary versions of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) are also implemented to observe the proposed method’s performance.

2.5. Parameter settings

The parameters of the proposed BBA-MCL are set to the following values: $A = 0.5$, $r = 0.5$, $a = 0.9$, and $\gamma = 0.9$. These values are adopted as several studies have confirmed that these values have achieved promising results (Akinyelu & Adewumi, 2018; Alomari et al., 2017). In the case of MCL, different values for the inflation parameter and the pruning threshold are assigned to the MCL according to the dataset used (see Table 2). Such values are determined based on several preliminary experiments. For the rest of the MCL’s parameters, the default values used by the original study are employed (Van Dongen, 2000). The population size for all comparative algorithms is set to 20 and they were terminated after 100 iterations. The results from the proposed method BBA-MCL are registered for 10 independent runs and compared to the results obtained by the standard MCL without applying any FS technique, GA-MCL, and BPSO-MCL. The parameters for GA and BPSO are referred to the values set in (Mirjalili et al., 2014).
3. Experimental results

This section presents the experimental results for the proposed wrapper BBA-MCL FS method performance over the 12 datasets from three different perspectives as follows:

3.1. Evaluation metrics

The performance of the proposed BBA-MCL FS method and other comparative methods based on the total average value of the evaluation metrics $F$, $PR$, $R$, $SFR$ (i.e., $F_{AVG}$, $PR_{AVG}$, $R_{AVG}$, and $SFR$), and computational time for ten independent runs are given in Table 3, IV, V, VI, and VII respectively. The best results are shown in bold text.

Table 3, shows that the proposed BBA-MCL FS method has recorded the highest $F$ rates in 10 datasets (i.e., DS1-DS9, and DS11) in comparison to MCL, GA-MCL, and BPSO-MCL. Such results indicate that the selected feature subset by the proposed BBA-MCL FS method is the most optimal and informative feature subset that lead to obtaining high-quality event clusters. In addition, these results are consistent with what has been found in the literature, that the BBA has a better ability in terms of exploring the feature space compared to GA and PSO (Emary et al., 2014; Ye et al., 2018). On the other hand, GA-MCL has rated the second where it obtained the second-highest $F$ scores in two datasets (i.e., DS10 and DS12) and exhibited better performance than BPSO-MCL in most datasets.

Table 4 outlines the scores of PR for MCL and the applied wrapper FS methods. It reveals that BBA-MCL has achieved the best PR score, except for the DS8 dataset, where GA-MCL is slightly better than BBA-MCL. BPSO-MCL came in second place where it has shown better PR values compared to GA-MCL in nine datasets i.e., DS1-DS6, DS8-DS10, and DS12. Table 5 demonstrates that GA-MCL has accomplished the highest $R$ values as compared to MCL and other FS methods in seven datasets i.e., DS1-DS6, and DS8. The basic MCL has placed second where it exhibited the best $R$ scores in four datasets i.e., DS9-DS12 while BBA-MCL has shown best $R$ values in only two datasets i.e., DS7 and DS8.

Table 6 outlines the SFR for the applied FS methods on the 12 datasets. It is clear that BBA-MCL was able to select the lowest number of features in 9 datasets (i.e., DS1-DS3, DS6-DS10, and DS12) and at the same time, it has recorded the best $F$ (see Table 3). To highlight, the smaller value of SFR is referring to informative features. Table 6 shows BBA-MCL offers consistently better results than other based line methods though, in DS4 and DS5, PSO has obtained slightly better results with differences of 0.05 and 0.02, respectively.

Table 7 presents the obtained computational time for the different FS methods, namely GA-MCL, BPSO-MCL, and BBA-MCL. The computational execution time analysis illustrates that the proposed BBA-MCL has relatively outperformed other comparative FS methods with the shortest time recorded in 7 datasets (DS1-DS6 and DS10). These results prove that the selected feature subsets by BBA-MCL have preserved the significant features that lead to the clustering improvement (see Table 3) while reducing computational time (see Table 7) in comparison with other FS methods. GA-MCL has been observed in the second rank with the best short execution time in 5 datasets (DS7-DS9, DS11, and DS12). In contrast, BPSO-MCL has performed poorly with a longer execution time in all of the datasets.

3.2. Convergence rate

In this section, the convergence behaviour of all applied FS methods based on the best fitness value (i.e., $Q$ score) for overall datasets (i.e., DS1-DS12) are drawn in Figure 5. The figure illustrates that the proposed wrapper BBA-MCL FS method has converged to the best solution in most datasets. In contrast, other comparative methods (i.e., GA-MCL and BPSO-MCL) have converged faster than BBA-MCL which indicates that GA and BPSO algorithms have an early convergence rate.
that in turn make them fall into local optimum solutions. On contrary, BBA has a much better convergence rate than comparative methods due to its unique feature, which is represented in its possession of two important parameters: $A$ and $r$ parameters (Gandomi & Yang, 2014; Gupta et al., 2019). Such parameters assist BBA to reduce the convergence rate through balancing between the exploration and exploitation processes and hence, to avoid falling into local optimal solutions (Fister et al., 2014; Yadav & Phogat, 2017).

3.3. Statistical analysis results
To statistically illustrate the comparison, Table 8 presents the results obtained from the Friedman test based on the $F_{AVG}$ measure. Table 8 outlines the ranks of the comparative FS methods and the $p$-value was calculated at $(0.003)$, which is less than $(0.05)$ that was assumed as a significant level for all datasets. This demonstrates that the proposed wrapper BBA-MCL FS method outperforms other comparative FS methods significantly. The top rank is allocated to BBA-MCL according to its mean rank followed by GA-MCL and BPSO-MCL.

To evaluate the performance of the BBA-MCL FS method, the multi-problem Wilcoxon signed-rank test is implemented, and the achieved results were given in Table 9. BBA-MCL performed better in 9 datasets compared to GA-MCL, while it performed worse in 2 datasets and has one tied result. Similarly, BBA-MCL outperforms BPSO-MCL in 11 datasets and loses in one dataset. In the case of BBA-MCL and GA-MCL, although BBA-MCL is better than GA-MCL, the difference between averages of $F$ scores is not statistically significant i.e., the $p$-value is $0.154$. On the other hand, results of BBA-MCL and BPSO-MCL revealed that the difference between them is statically significant i.e., the $p$-value is $0.028$.

4. Discussion
To sum up, Table 3, IV and V have shown that the wrapper FS method based on any optimization algorithm either BBA, GA, or BPSO has a superior clustering (detection) performance in terms of $F$, $PR$, and $R$ values compared to the basic MCL (without any FS method) on the majority of the datasets. The reason behind such poor performance of MCL is that it does not work well on large-scale datasets with sparse and high-dimensional feature space. In addition, MCL suffers from the early convergence rate that results in many meaningless clusters (Setiawan et al., 2016). This causes news documents to be placed in the wrong clusters thus, obtaining low $F$, $PR$, and $R$ values.

The main objective of the ED model is being able to measure and detect how many news documents are correctly assigned to their correct event clusters. The most important metric used to measure this in the ED field is the $F$ metric, for which our proposed BBA-MCL FS method has shown significant results compared to other baseline methods (see Table 3). BBA differs from GA and PSO in terms of having an automatic zooming technique that is controlled by its $A$ and $r$ parameters (Gupta et al., 2019). Parameter $A$ is responsible for controlling the performance and continuity of the global and random search whereas parameter $r$ controls the implementation of the local search (Gandomi & Yang, 2014; Gupta et al., 2019). Hence, BBA offers an advantage to select the optimal feature subset which contributes to obtaining high $F$ values which indicate that high-quality clusters (events) were produced. This results are aligned with findings from other studies as well (Abasi et al., 2020; Liu et al., 2019).

To point out, BBA_MCL has achieved the best $PR$ scores for almost all datasets (see Table 4), for which BBA_MCL has also attained the highest $F$ scores (see Table 3). This happens due to the exploration ability of BBA, which guides it to the regions in feature space where most informative feature subsets about events are found. As result, BBA-MCL has succeeded in placing news documents in the correct clusters (events) in almost all datasets and increasing the $F$ and $PR$ scores.
On the other hand, slight low R values are observed for BBA-MCL in comparison to GA-MCL (refer to Table 5). This indicates that GA-MCL is able to select features that lead to generating more disparate clusters with documents belonging to different events (high R values). Meanwhile, BBA-MCL has selected features that lead to producing more compact clusters with documents belonging to the same events (high PR values). Despite the low R scores obtained by BBA-MCL, yet it achieved the best F scores in the majority of datasets, which is reported in the literature to be the most important evaluation metric in ED and FS domains (Abualigah, Khader, Hanandeh et al., 2018b; Rasouli et al., 2020; Wei et al., 2018). The results from Table 6 confirms the ability of the proposed wrapper BBA-MCL FS method to select the minimum number of features that can improve the performance of the MCL in detecting events.

Table 7, shows the ability of BBA-MCL in achieving the best clustering performance (see Table 3) in a short time for several datasets (DS1-DS6, and DS10). However, it has recorded the second shortest execution time for the other datasets (DS7-DS9, DS11, and DS12). This happens may be due to the high-dimensional feature spaces for such datasets, for which BBA consumes more time to explore these spaces searching for the optimal feature subsets. In addition, the BBA algorithm has more steps to be implemented than GA and PSO, which could affect the execution time criterion. Although BBA-MCL has reserved a longer execution time for some datasets, it has achieved a better clustering performance based on the F score, in which the F measure is reported to be a vital metric according to many ED studies. Such studies have focused on reporting F results rather than execution time for evaluating ED model performance over the historical heterogeneous news text documents (Huang et al., 2013; Mele et al., 2019; Nanba et al., 2013; Prasad et al., 2018).

The statistical analysis is also conducted in this study using the Friedman rank test to obtain a significant statistical difference between the methods. Table 8 confirms the superiority of BBA-MCL in terms of discovering high accurate real-world events from multiple heterogenous news text documents, whereby a lower p-value = 0.003 has been achieved. This evidence reveals significant detection performance of BBA-MCL compared to other baseline FS methods. The findings were proved by Wilcoxon signed-rank test (see Table 9), which illustrated the effectiveness of the BBA-MCL through the number of victories it achieved over other baseline FS methods in relation to different datasets.

One point worth noting is the slight poor performance of the BBA-MCL, which was observed on some datasets, for example, DS10 and DS12 for F and DS8 for PR. The reason behind such performance is might be due to the early convergence rate of BBA, which makes it unable to explore the whole feature space effectively. This convergence behaviour of BBA is mainly based on the setting up values for the A and r parameters of BBA (Dhal & Das, 2018). Since in this study, the predefined fixed values have been assigned to such parameters, it might be not optimal values for some datasets (i.e., DS8, DS10, and DS12). As a matter of fact, the values of these parameters are mainly affected by the application domain, scope, and the size of the given datasets (Sheng et al., 2020). However, the best values for the different parameters of BBA are still uncertain and very challenging to be determined (Bangyal et al., 2018; Barbosa & Vasconcelos, 2018). To address this problem, our future study is focused on improving the BBA and MCL in order to enhance the performance of the proposed wrapper BBA-MCL FS method.

5. Conclusion
ED on multiple heterogeneous news documents suffers from the problem of high-dimensional feature space, which affects the overall performance of the ED model. Earlier ED works have either neglected the FS phase or have applied traditional FS methods, which failed in capturing the most informative features. To overcome this problem, many researchers from the text mining field have
introduced various wrapper FS methods based on MHAs, including BBA. The wrapper FS method based on BBA has achieved better results in different data mining applications. However, it has not been applied in the context of ED. To fill this gap, this work presents a novel wrapper FS method based on the BBA and MCL. A total of 12 news text datasets are used to evaluate the performance of BBA-MCL against the standard MCL (without FS method) and two wrapper FS methods, namely, GA-MCL and BPSO-MCL. The evaluation process based on $F_{AVG}$, $PR_{AVG}$, $R_{AVG}$, SFR, and computational time metrics, has shown superior performance of BBA-MCL in terms of choosing a minimal optimal feature subset and obtaining the highest $F$ and $PR$ scores. Additionally, BBA-MCL has shown a better convergence rate compared to other wrapper FS methods. The statistical test results confirm that BBA-MCL has outperformed other FS methods significantly at 0.003. Despite the outstanding performance of the BBA-MCL, several drawbacks are identified such as the fast convergence behaviour and exploration capability of BBA that lead to poor results in some datasets. For future work, BBA-MCL can be further improved by enhancing the convergence behaviour and exploration capability of BBA. In addition, more techniques can be developed for tuning or controlling the $A$ and $r$ parameters of BBA to operate effectively on different data sizes. Finally, additional datasets and other comparative MHAs can be included and tested.

Acknowledgements
The authors of this paper would like to express thanks to the editor and all reviewers for the valuable comments.

Funding
The authors received no direct funding for this research.

Author details
Wafa Zubair Al-Dyani$^{1,2}$
E-mail: wafazb1084@gmail.com
ORCID ID: http://orcid.org/0000-0003-4192-6320
Farzana Kabir Ahmad$^1$
E-mail: farzana85@uum.edu.my
ORCID ID: http://orcid.org/0000-0002-0374-6559
Siti Sakira Kamaruddin$^1$
E-mail: sakira@uum.edu.my
ORCID ID: http://orcid.org/0000-0002-5937-7801

1 Department of Computer Science, School of Computing, College of Arts and Science, Universiti Utara Malaysia, Sintok Kedah, Malaysia.
2 Department of Computer Science, College of Computing and Information Technology, Hadhramout University, Hadhramout, Yemen.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Correction
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Citation information
Cite this article as: Binary Bat Algorithm for text feature selection in news events detection model using Markov clustering, Wafa Zubair Al-Dyani, Farzana Kabir Ahmad & Siti Sakira Kamaruddin, Cogent Engineering (2022), 9: 2010923.

References
Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Makhadmeh, S. N., & Alyasseri, Z. A. A. (2020). Link-based multi-verse optimizer for text documents clustering. Applied Soft Computing, 87, 106002. https://doi.org/10.1016/j.asoc.2019.106002
Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. The Journal of Supercomputing, 73(11), 4773–4795. https://doi.org/10.1007/s11227-017-2046-2
Abualigah, L. M., Khader, A. T., & Al-Betar, M. A. (2016). Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. In 2016 7th International Conference on Computer Science and Information Technology (CSIT) (pp. 1–6).
Abualigah, L. M., Khader, A. T., Al-Betar, M. A., & Hanandeh, E. S. (2016). Unsupervised text feature selection technique based on particle swarm optimization algorithm for improving the text clustering. In 1st EAI International Conference on Computer Science and Engineering (p. 169).
Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018a). Hybrid clustering analysis using improved krill herd algorithm. Applied Intelligence, 48(11), 4047–4071. https://doi.org/10.1007/s10489-018-1190-6
Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018b). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. Journal of Computational Science, 25, 456–466. https://doi.org/10.1016/j.jocs.2017.07.018
Akinjelu, A. A., & Adewumi, A. O. (2018). On the performance of cuckoo search and bat algorithms based instance selection techniques for SVM speed optimization with application to e-fraud detection. KSII Transactions on Internet and Information Systems, 12(3), 1348–1375. https://doi.org/10.3837/ktiis.2018.03.021
AL-Dyani, W. Z., Ahmad, F. K., & Kamaruddin, S. S. (2020, July). A Survey on event detection models for text data streams. Journal of Computer Science, 16(7), 916–935. https://doi.org/10.3844/jcssp.2020.916.935
Al-Dyani, W. Z., Yahya, A. H., & Ahmad, F. K. (2018). Challenges of event detection from social media streams. International Journal of Engineering & Technology, 7(2.15), 72–75. http://dx.doi.org/10.14419/ijet.v7i15.11217
Alomari, O. A., Khader, A. T., Al-Betar, M. A., & Abualigah, L. M. (2017). Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm.
International Journal of Data Mining and Bioinformatics, 19(1), 32–51. https://doi.org/10.1080/15561863.2017.1360295

Altuncu, M. T., Mayer, E., Yaliraki, S. N., & Barahona, M. (2019). From free text to clusters of content in health records: An unsupervised graph partitioning approach. Applied Network Science, 4(1), 2. https://doi.org/10.1007/s11109-018-0190-9

Bangyal, W. H., Ahmad, J., Rauf, H. T., & Pervaiz, S. (2018). An overview of mutation strategies in bat algorithm. International Journal of Advanced Computer Science and Applications, 9(8), 523–534. https://doi.org/10.14569/ijasca.2018.090866

Barbosa, C. E. M., & Vasconcelos, G. C. (2018). Eight bio-inspired algorithms evaluated for solving optimization problems. In International Conference on Artificial Intelligence and Soft Computing (pp. 290–301). https://doi.org/10.1007/978-3-319-91253-0_28

Beigh, T. M., Upadhyaya, S., & Gopal, G. (2016). Event identification in social news streams using keyword analysis. International Research Journal of Engineering and Technology, 3(5). https://doi.org/10.1109/ictsys.2015.5654957

Bustamam, A., Wisnubroto, M. S., & Lestari, D. (2018). Analysis of protein-protein interaction network using Markov clustering with pigeon-inspired optimization algorithm in HIV (human immunodeficiency virus). In Proceedings of the 3rd International Symposium on Current Progress in Mathematics and Sciences 2017 (ICPMMS2017) (Vol. 2023, no. 1, p. 20229) https://doi.org/10.1063/1.5064226.

Chen, Q., Guo, X., & Bao, H. (2017). Semantic-based topic detection using Markov decision processes. Neurocomputing, 242, 40–50. https://doi.org/10.1016/j.neucom.2017.02.020

Dai, X., & Sun, Y. (2019). Event identification within news topics. In 2010 International Conference on Intelligent Computing and Integrated Systems (pp. 498–502).

Dhal, K. G., & Das, S. (2018). A dynamically adapted and weighted bat algorithm in image enhancement domain. Evolving Systems, 1–19. https://doi.org/10.1007/s12530-018-9216-1

Edouard, A., Cabrio, E., Tonelli, S., & Le Thanh, N. (2017). Graph-based event extraction from Twitter. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 222–230). https://doi.org/10.26615/978-954-452-049-6_031.

El-mawass, N.,NONEINE, P., & Vercouter, L. (2018). Supervised classification of social spammers using a similarity-based markov random field approach. https://doi.org/10.1145/3227966.3227712

Emary, E., Yaman, W., & Hassanien, A. E. (2014). New approach for feature selection based on rough set and bat algorithm. In 2014 9th International Conference on Computer Engineering & Systems (ICCES) (pp. 346–353). https://doi.org/10.1109/icces.2014.7030984

Enache, A., & Science, C. (2015). Intelligent feature selection method rooted in Binary Bat Algorithm for intrusion detection. In 2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics (pp. 517–521). https://doi.org/10.1109/isci.2015.7208259.

Enache, A. C., & Sgarciu, V. (2014). Anomaly intrusions detection based on support vector machines with bat algorithm. In 2014 18th International Conference on System Theory, Control and Computing (ICSTCC) (pp. 856–861). https://doi.org/10.1109/icstcc.2014.6982526.

Fister, I., Yang, X.-S., Fong, S., & Zhuang, Y. (2014). Bat algorithm: Recent advances. In 2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI) (pp. 163–167). https://doi.org/10.1109/cinti.2014.7028669.

Fu, Z., Sun, X., Shu, J., & Zhou, L. (2014). Plain text zero knowledge watermarking detection based on asymmetric encryption. 48(Cia), 126–134. https://doi.org/10.14257/ijost.2014.48.21

Gandomi, A. H., & Yang, X.-S. (2014). Chaotic bat algorithm. Journal of Computational Science, 5(2), 224–232. https://doi.org/10.1016/j.jocs.2013.10.002

Gashi, R., & Ahmeti, H. G. (2021). Impact of social media on the development of new products, marketing and customer relationship management in Kosovo. Emerging Science Journal, 5(2), 125–138. https://doi.org/10.28991/esj-2021-01263

Goswami, A., & Kumar, A. (2016). A survey of event detection techniques in online social networks. Social Network Analysis and Mining, 6(1), 107. https://doi.org/10.1007/s13278-016-0416-1

Gupta, D., Arora, J., Agrawal, U., Khanna, A., & de Albuquerque, V. H. C. (2019). Optimized Binary Bat Algorithm for classification of white blood cells. Measurement, 143, 180–190. https://doi.org/10.1016/j.measurement.2019.06.002

Hong, S.-S., Lee, W., & Han, M.-M. (2015). The feature selection method based on genetic algorithm for efficient of text clustering and text classification. International Journal of Advances in Soft Computing and Its Applications, 7(1). https://scholarworks.bwise. krigachon/handle/2020.sw.gachon/11004

Hu, L., Zhang, B., Hou, L., & Li, J. (2017). Adaptive online event detection in news streams. Knowledge-Based Systems, 138, 105–112. https://doi.org/10.1016/j.knosys.2017.09.039

Huang, X., Zhang, X., Ye, Y., Deng, S., & Li, X. (2013). A topic detection approach through hierarchical clustering on concept graph. Applied Mathematics & Information Sciences, 7(6), 2285–2295. https://doi.org/10.12785/amis/070619

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry – Learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web (pp. 107–110). https://doi.org/10.1145/2567948.2577024.

Liu, J., Abbass, H. A., & Tan, K. C. (2019). Evolutionary community detection algorithms. In Evolutionary Computation and Complex Networks (pp. 77–115). Springer.

Manaskasemsak, B., Chinthanet, B., & Rungsawang, A. (2018). Graph clustering-based emerging event detection from twitter data stream. In Proceedings of the Fifth International Conference on Network, Communication and Computing (pp. 37–41). https://doi.org/10.1109/3033288.3033312.

Mele, I., Bahrainian, S. A., & Crestani, F. (2019). Event mining and timelines analysis from heterogeneous news streams. Information Processing & Management, 56(3), 969–993. https://doi.org/10.1016/j.ipm.2019.02.003

Mhamdi, C., Al-Emran, M., & Salloum, S. A. (2018). Text mining and analytics: A case study from news channels posts on Facebook. In Intelligent Natural Language Processing: Trends and Applications (pp. 399–415). Springer.
Mirjalili, S., Mirjalili, S. M., & Yang, X.-S. (2014, September). Binary Bat Algorithm. Neural Computing and Applications, 25(3-4), 663–681. https://doi.org/10.1007/s00521-013-1525-5

Moutidis, I., & Williams, H. T. P. (2019). Utilizing complex networks for event detection in heterogeneous high-volume news streams. In International Conference on Complex Networks and Their Applications (pp. 659–672).

Nakamura, R. Y. M., Pereira, L. A. M., Costa, K. A., Rodrigues, D., Papa, J. P., & Yang, X.-S. (2012). BBA: A Binary Bat Algorithm for feature selection. In 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images (pp. 291–297). https://doi.org/10.1109/sibgrapi.2012.47.

Nanba, H., Saito, R., Ishino, A., & Takezawa, T. (2013). Automatic extraction of event information from newspaper articles and web pages. In International Conference on Asian Digital Libraries (pp. 171–175).

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69(6), 66133. https://doi.org/10.1103/PhysRevE.69.066133

Panagiotou, N., Katakis, I., & Gunopulos, D. (2018). Detecting events in online social networks: Definitions, trends and challenges. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9580, pp. 42–84). Springer.

Prasad, R., Bisandu, D., & Liman, M. (2010). Clustering news articles using efficient similarity measure and N-grams. International Journal of Knowledge Engineering and Data Mining, 5(1), 1. https://doi.org/10.1504/ijkedm.2018.0106103

Ramadan, Q. H., & Mohd, M. (2011). A review of retrospective news event detection. In 2011 International Conference on Semantic Technology and Information Retrieval (pp. 209–214). https://doi.org/10.1109/stair.2011.5995790.

Rasouli, E., Zarifzadeh, S., & Rafsanjani, A. J. (2020). WebKey: A graph-based method for event detection in web news. Journal of Intelligent Information Systems, 54(3), 585–604. https://doi.org/10.1007/s10844-019-00576-7

Salloum, S. A., Al-Emran, M., Abdallah, S., & Shaalan, K. (2017). Analyzing the Arab Gulf newspapers using text mining techniques. International Conference on Advanced Intelligent Systems and Informatics (pp. 396–405). https://doi.org/10.1007/978-3-319-64861-3_37.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017). Mining social media text: Extracting knowledge from Facebook. International Journal of Computing and Digital Systems, 6(2), 73–81. https://doi.org/10.12785/ijcds/060203

Salloum, S. A., Mhamdi, C., Al-Emran, M., & Shaalan, K. (2017). Analysis and classification of Arabic newspapers’ Facebook pages using text mining techniques. International Journal of Information Technology, 1(2), 8–17. https://journals.sfu.ca/ijitls/index.php/ijitls/article/view/12

Setiawan, E. I., Susanto, C. P., Santoso, J., Sumpeno, S., & Purnomo, M. H. (2016). Preliminary study of spam profile detection for social media using Markov clustering: Case study on Javanese people. In 2016 International Computer Science and Engineering Conference (ICSEC) (pp. 1–7). https://doi.org/10.1007/978-3-319-60000-0_5.

Sharma, J., & Annapo, B. (2016). Community detection using meta-heuristic approach: Bat algorithm variants. In 2016 Ninth International Conference on Contemporary Computing (IC3) (pp. 1–7). https://doi.org/10.1109/ic3.2016.7880209.

Sheng, J., Liu, C., Chen, L., Wang, B., & Zhang, J. (2020). Research on community detection in complex networks based on internode attraction. Entropy, 22(12), 1383. https://doi.org/10.3390/e22121383

Szilágyi, T. M., & Szilágyi, L. (2019). A fast hierarchical clustering algorithm for large-scale protein sequence data sets. Computers in Biology and Medicine, 48, 94–101. https://doi.org/10.1016/j.compbiomed.2014.02.016

Van Dongen, S. M. (2000). Graph clustering by flow simulation. University of Utrecht.

Wada, H. (2021). Assessing the social media user’s credibility rating of shared content, and its utilization in decision making. Emerging Science Journal, 5 (2), 191–199. https://doi.org/10.28991/esj-2021-01269

Wei, Y., Singh, L., Buttler, D., & Gogalagher, B. (2018). Using semantic graphs to detect overlapping target events and story lines from newspaper articles. International Journal of Data Science and Analytics, 5(1), 41–60. https://doi.org/10.1109/si.2016.01066-x

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(6), 606–626. https://doi.org/10.1109/tevc.2015.2504420

Yadav, S. L., & Phogat, M. (2017). A review on bat algorithm. International Journal of Computational Science and Engineering, 5(7), 39–43. https://doi.org/10.26438/ijcse.v5i7.3943

Yang, X.-S. (2010). A new metaheuristic bat-inspired algorithm. In Nature Inspired Cooperative Strategies for Optimization (NISCO 2010) (pp. 65–74).

Ye, Z., Hou, X., Zhang, X., & Yang, J. (2018). Application of bat algorithm for texture image classification. International Journal of Intelligent Systems and Applications, 10(5), 42–50. https://doi.org/10.5815/ijisa.2018.05.05

Yu, S., & Wu, B. (2018). Exploiting structured news information to improve event detection via dual-level clustering. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC) (pp. 873–880). https://doi.org/10.1109/dsc.2018.00140.
