Fake News Detection Based on Multi-view Fuzzy Clustering Algorithm

Hoang Thi Canh[1,2], Pham Huy Thong[3(✉)], Le Truong Giang[3], and Phan Dang Hung[3]

[1] Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
[2] Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
htcanh@ictu.edu.vn
[3] Hanoi University of Industry, Hanoi, Vietnam
{thongph,letruonggiang,phanhung}@haui.edu.vn

Abstract. The rapid development of technology and the internet has enabled users to access and share a large amount of information from various sources. This brings many benefits, but also the emergence of false and inaccurate information, also known as fake news. Fake news can lead to misunderstandings and significant impacts on the economy and society. Therefore, detecting and minimizing fake news is necessary. Machine learning algorithms and artificial intelligence technology can be used to detect and eliminate fake news. In this paper, we propose a new method for detecting fake news using multi-view fuzzy clustering on multi-view data collected from multiple sources. Our proposed method first extracts features from multi-view data, such as the title, content, and social media engagement of news articles. It then uses multi-view fuzzy clustering to group the news articles into clusters. Finally, it uses a semi-supervised learning algorithm to classify the clusters as either real or fake news. Additionally, the paper provides experimental results to evaluate the effectiveness and accuracy of the proposed algorithm.

Keywords: Multi-view data · multi-view clustering · fuzzy clustering

1 Introduction

In the contemporary era, the rapid evolution of technology and the internet has ushered in a plethora of opportunities for individuals to access and disseminate information from a myriad of sources. This accessibility has ushered in a multitude of advantages, including the ability to acquire knowledge, facilitate work, and provide entertainment. Nonetheless, this ease of access has also precipitated

N. Thi Dieu Linh et al. (Eds.): ADHOCNETS 2023, LNICST 558, pp. 150–166, 2024.
https://doi.org/10.1007/978-3-031-55993-8_12

certain challenges, most notably the proliferation of false and inaccurate information, leading to misunderstandings and exerting a detrimental influence on the decisions of individuals and communities. Consequently, in recent years, the detection and mitigation of false information have emerged as focal points of interest among researchers. A pivotal solution to tackle this issue involves the employment of machine learning algorithms and artificial intelligence technology to identify and eradicate counterfeit information, ensuring that users can harness information accurately and efficiently.

The term "multi-view data" refers to information collected from diverse sources, methodologies, or viewpoints regarding a particular entity [1]. This data is characterized by multiple perspectives, with each viewpoint furnishing distinct attributes for the purposes of knowledge discovery, offering varied insights about the same subject with varying degrees of precision and reliability. However, these different views often encompass complementary information that can be harnessed. Through the amalgamation of insights from multiple viewpoints, a more comprehensive and accurate representation of entities can be obtained, thus enriching the processes of data analysis and decision-making.

Data clustering stands as a pivotal challenge in the realm of data mining, with the objective of identifying and unveiling significant data clusters within extensive datasets, thereby furnishing information to support the decision-making process [2]. This process entails segregating an initial dataset into clusters, where the elements within each cluster exhibit similarity to one another, while those in different clusters demonstrate dissimilarity [3]. This clustering approach facilitates data mining, especially within the context of substantial datasets, by effectively organizing data based on their inherent characteristics [4,5]. However, the majority of existing clustering algorithms are tailored for single-view data, while contemporary practical challenges often involve multi-view data. Consequently, there is an imperative need to devise advanced clustering methodologies that can adeptly unearth knowledge from these multi-view datasets, propelling multi-view clustering into the forefront of research interest in recent times.

Multi-view Clustering (MvC) represents a data clustering methodology that harnesses numerous independent viewpoints to uncover groups of similar data within these perspectives [6]. The amalgamation of information from diverse viewpoints and the revelation of shared implicit knowledge across these perspectives deliver substantial benefits to data clustering. Applying multi-view clustering to identify fake news clusters through datasets collected from multiple sources emerges as an innovative and efficacious approach.

In contrast to single-view clustering methods, multi-view clustering boasts a plethora of advantages, including enhanced clustering quality, diminished reliance on individual viewpoints, and more efficient processing of intricate data [7]. Nevertheless, multi-view clustering is not devoid of challenges, encompassing the necessity for independent viewpoints, the complexity of merging distinct clustering methodologies, and the management of a multitude of viewpoints, entailing costs associated with data collection, processing, and storage [7].

The issue of fake news detection is besieged by its own set of complexities and difficulties. Fake news continues to evolve and proliferate, employing sophisticated forgery techniques and artificial intelligence to generate highly convincing counterfeit content. The dearth of annotated data compounds the problem, allowing fake news to disseminate rapidly across social media and other digital platforms, thereby rendering detection and control arduous tasks. To surmount these challenges, a multifaceted approach, drawing upon advanced methodologies and techniques, including machine learning, natural language processing, social network analysis, and community contributions, is indispensable to combat and identify fake news.

Researchers worldwide have diligently sought solutions to the aforementioned challenges. Seminal research on multi-view clustering methods by Bickel and Scheffer [8] in 2004 laid the foundation for subsequent developments. Their work extended K-means and Expectation Maximization (EM) clustering methods to accommodate multi-view environments, specifically handling text data featuring two conditionally independent perspectives. Subsequently, numerous multi-view clustering methodologies have been proposed [9–11].

The efficacy of the MvC algorithm hinges on two pivotal principles: the complementary principle and the consensus principle [7]. The complementary principle underscores the importance of leveraging multiple distinct viewpoints to comprehensively and accurately depict data entities. Although individual viewpoints provide adequate information for specific knowledge discovery tasks, disparate perspectives often encompass supplementary data that can be leveraged. In contrast, the consensus principle aims to maximize consistency across multiple viewpoints.

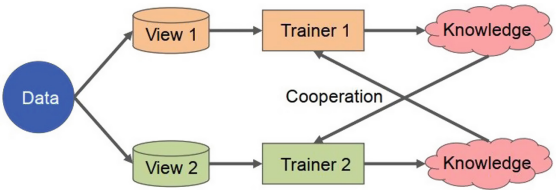

Fig. 1. General process of the co-training algorithm [7]

The co-training algorithm, introduced by Blum and Mitchell in 1998 [12], has emerged as a pioneering technology, firmly establishing itself as one of the most prominent methodologies within the realm of multi-view learning. The primary objective of this algorithm is to maximize consensus across viewpoints, thereby achieving the broadest consensus and enhancing clustering performance. The general procedure of the conventional co-training algorithms is shown in Fig. 1. According to the procedure, the algorithm is trained alternately in order to maximize the consistency of the two distinct views by using prior information or by learning knowledge from each other.

The prevailing body of research in fake news detection predominantly relies on supervised methods, which necessitate the construction of a classification model using diverse feature sets encompassing news content [13], message propagation [14], and social contexts [15]. While these techniques exhibit promising results, they are burdened by a significant drawback: they require a meticulously annotated dataset for training the classification model. The accumulation of a substantial number of annotations is a labor-intensive and time-consuming endeavor, involving in-depth analysis of news content and the inclusion of supplementary corroborative evidence, such as authoritative reports. While crowdsourcing for annotation collection may alleviate the burden of expert review, it carries the potential risk of compromising annotation quality [16].

Despite the multitude of previous research endeavors aimed at addressing the challenge of fake news detection, none have employed multi-view fuzzy clustering as a means to analyze and detect counterfeit news. Consequently, the primary objective of this paper is to bridge this research gap by introducing a fuzzy clustering methodology leveraging the co-training algorithm on a multi-view dataset for the identification of fake news clusters characterized by distinct attributes. This dataset is sourced from diverse origins, and this methodology is denominated as MCFC (Multi-View Co-trained Fuzzy Clustering). The efficacy of this proposed approach is rigorously evaluated through experiments conducted on datasets obtained from a range of sources, with the overarching goal of fake news detection.

The remaining sections of this paper are structured as follows: Sect. 2 delineates the approach to hard and fuzzy clustering on single-view data. Section 3 provides an overview of the proposed methodology. Section 4 presents experimental results and analysis, alongside comparative assessments. Finally, the conclusion section furnishes key takeaways and outlines prospective avenues for future research in subsequent publications.

2 Related Work

The detection of fake news has become a prominent focus in recent research within the realm of social media studies. Current strategies for identifying fake news can be broadly categorized into two principal groups: those centered on the analysis of news content and those oriented towards scrutinizing social contexts [17].

Approaches grounded in news content typically involve the extraction of linguistic and visual attributes. Linguistic elements, encompassing lexical and syntactic features, are harnessed to capture the distinctive writing styles and sensational headlines frequently associated with counterfeit news items [18]. Conversely, visual attributes are employed to unveil manipulated images or unique characteristics of visuals used in fake news dissemination. Models reliant on news content-based features can be further classified into two categories: (1) knowledge-based methods, which rely on external sources to validate the accuracy of claims within news content, and (2) style-based techniques, which detect stylistic manipulations like deceit and subjectivity [19].

In contrast, social context-driven approaches incorporate features derived from user profiles, post content, and social networks. User profiles are utilized to assess user characteristics and credibility, while features extracted from user-generated posts reflect social responses such as stances [20]. Network features are acquired by constructing specific social networks, including diffusion networks or co-occurrence networks [21]. Social context models can be categorized as either stance-based, which leverage user opinions to infer news veracity, or propagation-based, which employ propagation techniques to model unique information dissemination patterns [20–22]. It is worth noting that these approaches primarily fall under the category of supervised methods, focusing on feature extraction and their utilization within supervised learning frameworks [19].

In addition, this paper introduces a novel fuzzy clustering methodology applied to multi-view datasets for the purpose of detecting fake news clusters. This approach leverages established research platforms such as the K-means and FCM algorithms to achieve its objectives.

2.1 K-Means

The K-means algorithm stands as a prominent clustering technique within the realms of data analysis and machine learning. Its primary objective revolves around dividing a given dataset into K separate clusters. Within this partitioning process, each cluster finds its representation through a centroid. This methodological approach is instrumental in unearthing concealed patterns within the data while effectively grouping data points that share similarities together [23]. This algorithm's fundamental principle hinges on an objective function, which plays a pivotal role in driving the clustering process. The mathematical formulation for this objective function is depicted as follows:

$$J_m(u, v) = \sum_{k=1}^{N} \sum_{j=1}^{c} u_{kj}.||x_k - v_j||^2 \rightarrow Min \tag{1}$$

With constraint conditions:

$$\begin{cases} \sum_{j=1}^{c} u_{kj} = 1 \\ u_{kj} \in \{0, 1\} \end{cases}$$

In which: X is the source dataset $X = \{x_1, x_2, ..., x_N\}$ with N data points, c is the number of clusters, v is the set of cluster centers, u_{kj} is the membership degree of element k in cluster j.

By solving the objective function (1), we obtain the formulas for cluster centroids (2) as follows:

$$v = \frac{\sum_{k=1}^{N} u_{kj} x_k}{\sum_{k=1}^{N} u_{kj}} \tag{2}$$

The K-means algorithm is a robust clustering technique renowned for its proficiency in data segmentation through centroid-based partitioning. It boasts several merits, including simplicity, speed, and scalability. However, it is not without limitations. Notably, K-means exhibits sensitivity to the initial positions of cluster centers and demands the predefined number of clusters. These constraints have the potential to impact its overall performance.

2.2 Fuzzy C-Means

Fuzzy C-Means (FCM), proposed by Bezdek and colleagues in 1984 [24], is a widely recognized fuzzy clustering algorithm rooted in Zadeh's theory of fuzzy sets [25]. FCM serves the purpose of data partitioning into clusters by considering the similarity among data points. In the FCM approach, each data point is associated with all clusters, represented by values between 0 and 1, denoting the degree of similarity between the data point and each cluster. The central focus of the FCM fuzzy clustering algorithm revolves around optimizing the distances between data points and cluster centroids [26]. This optimization is achieved through the objective function defined by formula (3).

$$J_m(u,v) = \sum_{k=1}^{N}\sum_{j=1}^{c}(u_{kj})^m.||x_k - v_j||^2 \rightarrow Min \tag{3}$$

With constraint conditions:

$$\begin{cases} \sum_{j=1}^{c} u_{kj} = 1 \\ u_{kj} \in [0,1] \end{cases}$$

In which: X is the source dataset $X = \{x_1, x_2, ..., x_N\}$ with N data points, c is the number of clusters, v is the set of cluster centers, u_{kj} is the membership degree of element k in cluster i, m is the fuzzy parameter.

By solving the objective function (3), we obtain the formulas for cluster centroids (4) and membership degrees (5) as follows:

$$v = \frac{\sum_{k=1}^{N}(u_{kj})^m x_k}{\sum_{k=1}^{N}(u_{kj})^m} \tag{4}$$

$$u_{kj} = \left(\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}\right)^{-1} \tag{5}$$

The FCM (Fuzzy C-Means) method allows a data point to have membership degrees in multiple clusters, leading to enhanced clustering performance when compared to traditional methods like K-means. This flexibility in membership assignment enables FCM to capture complex patterns and relationships in the data, making it a valuable technique for various clustering tasks.

3 The Proposed Method

In this study, we introduce a novel approach to fuzzy multi-view clustering by employing a co-training algorithm. Our method is applied to a diverse multi-view dataset gathered from multiple sources, featuring distinctive characteristics, including variations in the number of records within each view, discrepancies in the number of attributes across different views, and a complex many-to-many relationship between the views.

3.1 Algorithm Idea

The MCFC algorithm consists of three steps as follows:

Step 1: Fuzzy clustering for unlabeled data in each view. In this step, the Fuzzy C-Means (FCM) algorithm is independently applied to each view (viewA and viewB) to partition data points into clusters. The result of this step includes cluster centroids and corresponding membership degrees for each view. The membership degrees are denoted as: $\overline{u}^A$ (membership degree of a data point on viewA with respect to the centroids) and $\overline{u}^B$ (membership degree of a data point on viewB with respect to the centroids).

Step 2: Calculate the dependency degrees of data points in viewA on the cluster center in viewB (and vice versa).

Step 3: Multi-view co-training fuzzy semi-supervised clustering. In this step, the results from viewA are used as training data for viewB, and vice versa. The goal is to maximize cross-consensus across all views and achieve the broadest consensus.

3.2 Algorithm Details

Building on the ideas presented in the previous section, this part will elaborate on the modeling of the proposed approach. The objective function of the method is represented by three components, as follows:

$$
\begin{aligned}
J(u^A, u^B, v^A, v^B) = & \sum_{k_A=1}^{N_A} \sum_{j=1}^{c} {u_{k_Aj}^A}^2 ||x_{k_A}^A - v_j^A||^2 \\
& + \sum_{k_B=1}^{N_B} \sum_{j=1}^{c} {u_{k_Bj}^B}^2 ||x_{k_B}^B - v_j^B||^2 \\
& + \sum_{k_A=1}^{N_A} \sum_{j=1}^{c} (u_{k_Aj}^A - u_{k_Aj}^{AB})^2 ||x_{k_A}^A - v_j^A||^2 \\
& + \sum_{k_B=1}^{N_B} \sum_{j=1}^{c} (u_{k_Bj}^B - u_{k_Bj}^{BA})^2 ||x_{k_B}^B - v_j^B||^2 \\
& + \sum_{k_A=1}^{N_A} \sum_{j=1}^{c} (u_{k_Aj}^A - \overline{u}_{k_Aj}^A)^2 ||x_{k_A}^A - v_j^A||^2 \\
& + \sum_{k_B=1}^{N_B} \sum_{j=1}^{c} (u_{k_Bj}^B - \overline{u}_{k_Bj}^B)^2 ||x_{k_B}^B - v_j^B||^2 \rightarrow Min
\end{aligned}
\tag{6}
$$

With constraint conditions:

$$
\begin{cases}
\sum_{j=1}^{c} u_{kj}^A = \sum_{j=1}^{c} u_{kj}^B = 1 \\
u_{kj} \in [0, 1]
\end{cases}
$$

Explanation: $x_{k_A}^A, x_{k_B}^B$ are the data points on viewA and viewB respectively. v_j^A, v_j^B are the cluster centers in viewA and viewB.$u_{k_Aj}^A, u_{k_Bj}^B$ represent the dependency degree of data point k to cluster j in viewA and viewB respectively. $\overline{u}_{k_Aj}^A, \overline{u}_{k_Bj}^B$ represent the dependency degree of data point k_A, k_B to cluster j after using the FCM algorithm on viewA and viewB respectively.

The objective function (6) is represented by three components, as follows:

– The component represents fuzzy clustering:

$$
\sum_{k_A=1}^{N_A} \sum_{j=1}^{c} {u_{k_Aj}^A}^2 ||x_{k_A}^A - v_j^A||^2
$$

and

$$
\sum_{k_B=1}^{N_B} \sum_{j=1}^{c} {u_{k_Bj}^B}^2 ||x_{k_B}^B - v_j^B||^2
$$

– The component represents co-training:

$$
\sum_{k_A=1}^{N_A} \sum_{j=1}^{c} (u_{k_Aj}^A - u_{k_Aj}^{AB})^2 ||x_{k_A}^A - v_j^A||^2
$$

and

$$\sum_{k_B=1}^{N_B}\sum_{j=1}^{c}\left(u_{k_Bj}^{B}-u_{k_Bj}^{BA}\right)^2||x_{k_B}^{B}-v_j^{B}||^2$$

– The component represents semi-supervised:

$$\sum_{k_A=1}^{N_A}\sum_{j=1}^{c}\left(u_{k_Aj}^{A}-\overline{\mathrm{u}}_{k_Aj}^{A}\right)^2||x_{k_A}^{A}-v_j^{A}||^2$$

and

$$\sum_{k_B=1}^{N_B}\sum_{j=1}^{c}\left(u_{k_Bj}^{B}-\overline{\mathrm{u}}_{k_Bj}^{B}\right)^2||x_{k_B}^{B}-v_j^{B}||^2$$

The cluster centers are calculated according to the following formula:

$$v_j^A=\frac{\sum_{k_A=1}^{N_A}\left[{u_{k_Aj}^{A}}^2+\left(u_{k_Aj}^{A}-u_{k_Aj}^{AB}\right)^2+\left(u_{k_Aj}^{A}-\overline{\mathrm{u}^{A}}_{k_Aj}\right)^2\right].x_{k_A}^{A}}{\sum_{k_A=1}^{N_A}\left[{u_{k_Aj}^{A}}^2+\left(u_{k_Aj}^{A}-u_{k_Aj}^{AB}\right)^2+\left(u_{k_Aj}^{A}-\overline{\mathrm{u}^{A}}_{k_Aj}\right)^2\right]} \tag{7}$$

$$v_j^B=\frac{\sum_{k_B=1}^{N_B}\left[{u_{k_Bj}^{B}}^2+\left(u_{k_Bj}^{B}-u_{k_Bj}^{BA}\right)^2+\left(u_{k_Bj}^{B}-\overline{\mathrm{u}^{B}}_{k_Bj}\right)^2\right].x_{k_B}^{B}}{\sum_{k_B=1}^{N_B}\left[{u_{k_Bj}^{B}}^2+\left(u_{k_Bj}^{B}-u_{k_Bj}^{BA}\right)^2+\left(u_{k_Bj}^{B}-\overline{\mathrm{u}^{B}}_{k_Bj}\right)^2\right]} \tag{8}$$

The Lagrange multiplier method is used to determine the degree of u_{kj}^{A} and u_{kj}^{B} :

$$u_{k_Aj}^{A}=\frac{1-\sum_{i_A=1}^{N_A}\Delta_{i_Aj}}{\sum_{i_A=1}^{N_A}\frac{{d_{k_Aj}^{A}}^2}{{d_{i_Aj}^{A}}^2}}+\Delta_{k_Aj} \tag{9}$$

Explanation:

$$\Delta_{k_Aj}=\frac{u_{k_Aj}^{AB}+\overline{\mathrm{u}}_{k_Aj}^{A}}{3}$$

$$u_{k_Bj}^{B}=\frac{1-\sum_{i_B=1}^{N_B}\Delta_{i_Bj}}{\sum_{i_B=1}^{N_B}\frac{{d_{k_Bj}^{A}}^2}{{d_{i_Bj}^{B}}^2}}+\Delta_{k_Bj} \tag{10}$$

Explanation:

$$\Delta_{k_Bj}=\frac{u_{k_Bj}^{BA}+\overline{\mathrm{u}}_{k_Bj}^{B}}{3}$$

The algorithm diagram is depicted in Fig. 2.

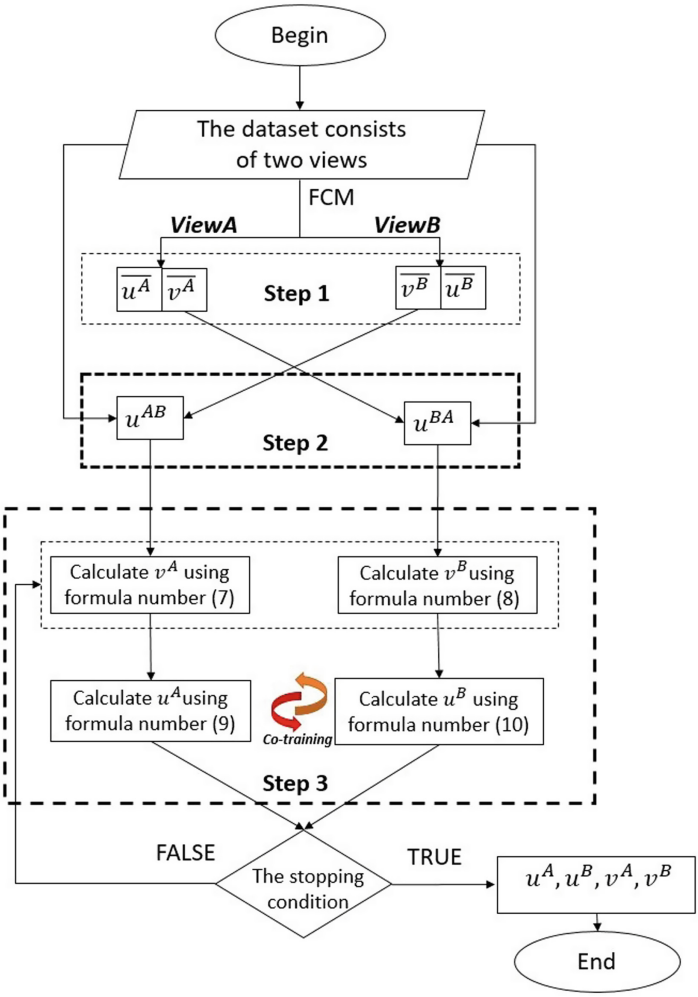

Fig. 2. Algorithm diagram

3.3 MCFC Algorithm Table

Algorithm 1: MCFC

Input
- The dataset X consists of two views:
viewA: $X^A = \{x_1^A, x_2^A, x_3^A, \ldots, x_n^A\}$
viewB: $X^A = \{x_1^A, x_2^A, x_3^A, \ldots, x_n^A\}$
The number of records in viewA and viewB is equal.
The number of attributes in viewA and viewB may differ.
- The number of clusters is c
- The number of iterations is Maxstep
- The allowable error ϵ

Output u^A, u^B, v^A, v^B

BEGIN

Step1: Fuzzy clustering for unlabeled data in each view

1.1. Applying the FCM algorithm to viewA, we obtain $\overline{u}^A$ and $\overline{v}^A$.

1.2. Applying the FCM algorithm to viewB, we obtain $\overline{u}^B$ and $\overline{v}^B$.

Step2:

2.1. Calculate the dependency degree u^{AB} of viewA on the cluster center v^B using the formula number (5).

2.2. Calculate the dependency degree u^{BA} of viewB on the cluster center v^A using the formula number (5).

Step3: Semi-supervised Fuzzy co-training clustering

3.1. Init t = 0

Repeat:

3.2. $t = t + 1$

3.3. Update v^A using formula (7)

3.4. Update v^B using formula (8)

3.5. Update u^A using formula (9)

3.6. Update u^B using formula (10)

Until:

$Max\{||u^{A(t+1)} - u^A(t||, ||u^B(t+1) - u^B(t||\} \leq \epsilon$ or $t \geq$ Maxstep

END

4 Experimental Results

4.1 Environmental Configuration

To validate the performance of the proposed method, the research team conducted simulations using datasets collected from two sources. The datasets include: **Fake and real news dataset** https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset (It's called viewA), **Fake News** https://www.kaggle.com/competitions/fake-news/data (It's called viewB).

The detailed information about the dataset presented in Table 1.

Table 1. Summary of the multi-view data sets

	No. of clusters	No. of objects	No. of variables
ViewA	2	4	44898
ViewB	2	4	5070

The experimental setup was performed on an Apple MacBook Air M1 2020 with a configuration of 8GB/256GB/7-core GPU, using Python programming language version 3.10. To assess performance, we use the following criteria: ACC (accuracy score), DB (Davies-Bouldin index), NMI(Normalized Mutual Information) and ARI (Adjusted Rand Index).

i) Clustering accuracy [27]: Cluster accuracy measures the level of accuracy in labeling data points within each cluster. A higher ACC value indicates better clustering performance.

$$\text{ACC} = \frac{\text{Number of correctly classified points}}{\text{Total number of points}}$$

in which: Number of correctly classified points refers to the total number of points in the clustering that are assigned the correct label. Total number of points represents the total number of points in the clustering.

ACC is used to evaluate the quality of a clustering. It measures the ratio of correctly classified instances by the model. A higher ACC indicates a higher proportion of correct classifications, suggesting a better clustering model. In the case of binary clustering, ACC can also be understood as the accuracy of correctly predicting classifications.

ii) Clustering quality: We use the DB measure (Davies-Bouldin index) [28]. Cluster quality measures the separation and cohesion among clusters. A smaller DB value indicates better cluster quality. To compute the DB index, several quantities are involved. Let us denote by δ_k the mean distance of the points belonging to cluster C_k to their barycenter $G^{\{k\}}$:

$$\delta_k = \frac{1}{n_k} \sum_{i \in I_k} \|M_i^{\{k\}} - G^{\{k\}}\| \tag{11}$$

Let us also denote by

$$\Delta_{kk'} = d(G^{\{k\}}, G^{\{k'\}}) = \|G^{\{k'\}} - G^{\{k\}}\|$$

the distance between the barycenters $G^{\{k\}}$ and $G^{\{k'\}}$ of clusters C_k and C'_k.

One computes, for each cluster k, the maximum M_k of the quotients $\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}$ for all indices $k' \neq k$. The Davies-Bouldin index is the mean value, among all the clusters, of the quantities M_k:

$$C = \frac{1}{K} \sum_{k=1}^{K} M_k = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}} \right) \tag{12}$$

The DB index serves as a criterion to compare different clustering algorithms or evaluate the performance of a single algorithm. A lower DB index indicates better clustering quality, where clusters are more compact and well-separated. Conversely, a higher DB index suggests poorer clustering performance, with clusters being less cohesive and more overlapping.

iii) ARI(Adjusted Rand Index) [27]: The Adjusted Rand Index is a measure used to assess the similarity between two clusters or between a cluster and the ground truth, for the purpose of evaluating the quality of data clustering.

The formula to calculate ARI for a single view can be expressed as follows:

$$\text{ARI} = \frac{\text{RI} - \text{Expected\_RI}}{\max(\text{RI}_{\max} - \text{Expected\_RI}, 0)} \tag{13}$$

in which: ARI represents the Adjusted Rand Index. RI denotes the Rand Index, which measures the agreement between the predicted clusters and the true labels. Expected_RI represents the expected Rand Index under the assumption of random label assignments. RI_max is the maximum possible Rand Index, calculated as the expected Rand Index when the predicted cluster assignments perfectly match the true labels.

The ARI is computed by subtracting the expected Rand Index from the actual Rand Index and dividing it by the maximum possible difference between the Rand Index and the expected Rand Index. The ARI ranges between -1 and 1:

- A value close to 1 indicates a high agreement between the predicted clusters and the true labels, beyond what would be expected by chance. It suggests that the clustering algorithm has accurately captured the underlying structure of the data.
- A value close to 0 suggests random agreement, meaning that the clustering results are not significantly better than random chance.
- A negative value implies a disagreement that is worse than random, indicating that the clustering results are worse than random chance.

Therefore, a higher ARI value is desirable, indicating a better agreement between the predicted clusters and the true labels. A value of 1 indicates a perfect clustering solution, while values close to 0 or negative values indicate poor or random clustering results.

iv) NMI(Normalized Mutual Information) [27]: The Normalized Mutual Information is a clustering metric that quantifies the degree of information shared between two clusterings, providing a normalized measure of their similarity. The formula to calculate NMI for a single view can be expressed as follows:

$$\text{NMI} = \frac{2 \times \text{MI}(V, L)}{\text{H}(V) + \text{H}(L)} \tag{14}$$

in which: NMI represents the Normalized Mutual Information. $\text{MI}(V, L)$ denotes the Mutual Information between the predicted clusters in view V and the true labels L. It measures the shared information or agreement between the two sets. $\text{H}(V)$ is the entropy of the predicted clusters in view V. It quantifies

the uncertainty or disorder within the predicted clusters. H(L) represents the entropy of the true labels. It measures the uncertainty or disorder within the true labels.

The NMI is calculated by dividing the twice the Mutual Information by the sum of the entropies of the predicted clusters and the true labels. This normalization accounts for the differences in the sizes of the clusters and the number of unique labels. By using the NMI for a single view, we can evaluate how well the clustering algorithm captures the structure and patterns within that particular view. It provides insights into the quality of clustering within that view and helps in assessing the performance of different algorithms or parameter settings for that specific view.

4.2 Results and Discussion

The MCFC method is compared with two other methods, K-means [23] and FCM [24]. We installed the K-means and FCM algorithms on the viewA and viewB datasets. Afterward, we merged the two datasets from two views: viewA (title, text) and viewB (title, text) to validate the proposed algorithm. The comparison results are shown in Table 2.

In Table 2 , MCFC achieved better values according to the criteria of ACC, ARI, and NMI. Therefore, the MCFC method outperformed the K-means and FCM methods in achieving cluster accuracy.

In Fig. 3, a chart presents the results of three algorithms (K-means, FCM, and MCFC) based on metrics such as ACC, DB, ARI, and NMI. According to the chart, the MCFC algorithm demonstrates identical clustering results across both views. Thus, the proposed algorithm can distinguish between real and fake news equally on both views.

Table 2. Table of experimental results comparing three algorithms (Bold values indicate the best results)

Algorithms	**Kmeans**			**FCM**			**MCFC**		
Data	ViewA	ViewB	Average	ViewA	ViewB	Average	ViewA	ViewB	Average
ACC	0.6588	0.4724	0.5656	0.6588	0.4724	0.5656	**0.8722**	**0.8722**	**0.8722**
DB	0.5102	0.4997	0.5050	**0.5100**	**0.4994**	**0.5047**	0.8702	0.8702	0.8702
ARI	0.1009	0.0028	0.0518	0.1016	0.0027	0.0521	**0.5541**	**0.5541**	**0.5541**
NMI	0.0742	0.0022	0.0382	0.0748	0.0021	0.0384	**0.4529**	**0.4529**	**0.4529**

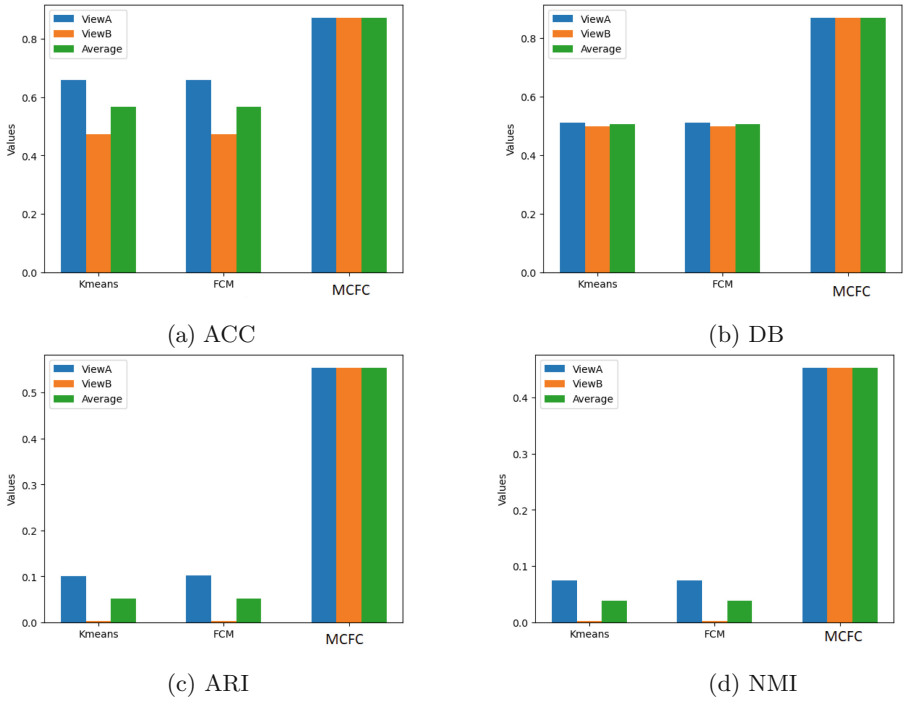

Fig. 3. The chart shows the metrics ACC, DB, ARI, NMI

5 Conclusions and Future Works

This study presents MCFC, a novel co-training fuzzy multi-view clustering model. MCFC is specifically designed to analyze multi-view data originating from diverse sources, effectively capturing the inherent characteristics of such data. Empirical investigations carried out as part of this study reveal that the MCFC method surpasses both K-means and FCM methods in terms of clustering accuracy. Furthermore, this method excels in accurately classifying genuine and fake news, offering users a reliable means of accessing and utilizing information with precision and efficiency. These promising findings serve as a catalyst for future research endeavors within the realm of multi-view clustering.

Despite its evident effectiveness when dealing with multi-source data, the MCFC model exhibits certain limitations. Notably, it involves a substantial number of parameters, and the repetitive co-training process contributes to prolonged computational time and inefficiency. These challenges become particularly pronounced when handling multi-source data with more than two distinct perspectives. Addressing the intricacies of multi-view data, especially when derived from divergent sources with varying perspectives, necessitates the continuous development of innovative algorithms in forthcoming research initiatives.

Acknowledgments. This work has been supported by Vietnam National University, Hanoi under the Project, code: QG.23.66.

References

1. Li, X., et al.: A multi-view model for visual tracking via correlation filters. Knowl. Based Syst. **113**, 88–99 (2016)
2. Tuan, T.M., et al.: A new approach for semi-supervised fuzzy clustering with multiple fuzzifiers. Int. J. Fuzzy Syst. **24**(8), 3688–3701 (2022)
3. Al-Amri, S.S., Kalyankar, N.V.: Image segmentation by using threshold techniques. arXiv preprint arXiv:1005.4020 (2010)
4. Thong, P.H., et al.: Picture-neutrosophic trusted safe semi-supervised fuzzy clustering for noisy data. Comput. Syst. Sci. Eng. **46**(2) (2023)
5. Huan, P.T., et al.: TS3FCM: trusted safe semi-supervised fuzzy clustering method for data partition with high confidence. Multimed. Tools Appl. **81**(9), 12567–12598 (2022)
6. Zhu, Z., et al.: Shared Subspace Learning for Latent Representation of Multi-View Data. J. Inf. Hiding Multim. Signal Process. **5**(3), 546–554 (2014)
7. Yang, Y., Wang, H.: Multi-view clustering: a survey. Big Data Min. Anal. **1**(2), 83–107 (2018)
8. Bickel, S., Scheffer, T.: Multi-view clustering. ICDM **4**, 2004 (2004)
9. Ye, F., et al.: New approaches in multi-view clustering. Recent Appl. Data Cluster. **195** (2018)
10. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning. arXiv preprint arXiv,1304.5634 (2013)
11. Sun, S.: A survey of multi-view machine learning. Neural Comput. Appl. **23**, 2031–2038 (2013)
12. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory (1998)
13. Wang, W.Y.: "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
14. Wu, L., Huan, L.: Tracing fake-news footprints: characterizing social media messages by how they propagate. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2018)
15. Ma, J., et al.: Detect rumors using time series of social context information on microblogging websites. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (2015)
16. Kim, J., et al.: Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (2018)
17. Shu, K., et al.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
18. Potthast, M., et al.: A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638 (2017)
19. Yang, S., et al.: Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01) (2019)

20. Jin, Z., et al.: News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30(1), pp. 2972–2978 (2016)
21. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (2017)
22. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures. In: 2015 IEEE 31st International Conference on Data Engineering, pp. 651–662. IEEE (2015)
23. Ikotun, A.M., et al.: K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. (2022)
24. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2–3), 191–203 (1984)
25. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
26. Tuan, T.M., Thong, P.H., Ngan, T.T.: An improvement of trusted safe semi-supervised fuzzy clustering method with multiple fuzzifiers. J. Comput. Sci. Cybernet. **38**(1), 47–61 (2022)
27. Wang, J., Liu, Y., Ye, W.: FMvC: fast multi-view clustering. IEEE Access **11**, 12808–12820 (2023)
28. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 224–227 (1979)