An Effective Prediction of Events in Social Networks Using Influence Score of Communities

B. S. A. S. Rajita[a], Yaganti Bhanu Vikas, Pritish Prashant Moharir, Deepa Kumari[b] and Subhrakanta Panda[c]

CSIS, BITS-Pilani Hyderabad Campus, Hyderabad, India

Keywords: Social Networks, Influence Score, Community, Derived Feature, Event Prediction, ML Models.

Abstract: In real-life social networks (SN), dynamic community evolution changes the structure of that network. Hence, a comprehensive framework is imperative for predicting community evolution, which this research refers to as an 'event'. This research studies how the influence of peer nodes in a social network often triggers community evolution. Therefore, this paper proposes calculating the communities' new derived feature called *Influence Score (IS)*, to predict their events. Thus, it is imperative to compute the communities' influence score (as a derived feature) and study its suitability for accurately predicting events using Machine Learning (ML) models. The experimental results show that *derived features* together with *community features* are more effective in predicting community events. The implementation and significance of the presented approach on the dataset show that IS, as an added feature, improved the accuracy of the ML models by approximately 6.6%. Additionally, it considerably improved other parameters, including F-measure, recall, and precision. This paper also presents a comparative analysis with other derived features. It shows an improvement in the accuracy by approximately 1.5% and 0.8%. The results also indicate that the IS score improved the accuracy of the logistic regression by 2.53% compared to an existing similar approach. Thus, this paper infers that IS as a derived feature is considerably effective in improving the accuracy of ML models in predicting events in SN communities.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Social networks (SN) define the social structures of entities and their interactions (Citron and Way, 2018). SNs are represented graphically as nodes and edges, where nodes stand for distinct entities and edges for interactions between them. SNs reveal the connection structure of entities, providing a comprehensive and detailed view of their behavior. An essential aspect of SN research is detecting communities corresponding to dense subgraphs (Sumith et al., 2017) (Altmami and Menai, 2020).

Over time, communities interact, leading to significant changes in community structures. These changes can trigger various events (*Born, Same, Merge, Split,* and *Dead*) within a community (Palla et al., 2007). The prediction of such events relies on a combination of community and derived features. Community features are based on inherent community properties, while derived features incorporate concealed factors such as collaborative distances between nodes, anomalous behaviors, node influences, and more. The research focussed on derived community features addresses compelling and intricate challenges, as highlighted in (Bhattacharjee et al., 2019).

Existing research (Kong et al., 2019), (Kong et al., 2019), (Duan et al., 2017) and (Duan et al., 2017) on event prediction has traditionally focused solely on community properties, not on the hidden properties. This paper integrates the influence of community members as a derived feature (hidden properties) in conjunction with community properties for predicting community events. This approach recognizes that influential nodes within communities can significantly impact the interactions among other community nodes (Rajita et al., 2020b) (Chakraborty et al., 2016).

Numerous social network analysis approaches (Kubiszewski et al., 2023) (Schäfermeier et al., 2023)

[a] https://orcid.org/0000-0003-0319-9601
[b] https://orcid.org/0000-0002-0696-9790
[c] https://orcid.org/0000-0003-4768-772X

Rajita, B., Vikas, Y., Moharir, P., Kumari, D. and Panda, S.
An Effective Prediction of Events in Social Networks Using Influence Score of Communities.
DOI: 10.5220/0012708300003756
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 13th International Conference on Data Science, Technology and Applications (DATA 2024)*, pages 229-236
ISBN: 978-989-758-707-8; ISSN: 2184-285X
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

suggest that influential nodes provides valuable insights into community event prediction. Understanding these dynamics is fundamental for assessing structural changes and identifying evolving communities. Such analysis substantially impact researchers, affecting citation patterns and collaboration dynamics (Zhang et al., 2021) (Saxena et al., 2021). For instance, consider co-author networks, where authors are represented as nodes, and their collaborative paper publications form the edges. When an influential node in the co-author network changes its research focus from topic A to topic B, this shift can trigger similar transitions in other nodes within the community, thus impacting the entire research landscape.

The remaining sections of this paper are as follows: Section 2 provides the background of the research. Section 3 delves into the details of our proposed methodology. Section 4 discusses the implementation and comparative study of the empirical results obtained from the proposed approach. Finally, in Section 5, we wrap up the work by presenting our findings and giving insights regarding future research directions.

2 BACKGROUND

This section covers the following topics: social networks (SNs), social network analysis (SNA), community detection, and community mining techniques applied to understand the proposed work. A SN is a collection of graphs over a range of timestamps represented as $\{1,2,\ldots,T\}$.

The term *SNA* refers to the structural change analysis in these SN graphs. These structural changes can be effectively captured by the communities within the SN.

To gain a clearer understanding of Community Detection, let's examine the social network (SN) depicted in Figure 1. This network comprises 10 nodes and 15 edges, and it serves as an illustrative example for showcasing the functionality of the Louvain algorithm, which is employed in this investigation to identify communities within the provided SN. Firstly, the Louvain technique treats individual nodes of the input graph as a separate community (single node as a set). Louvain randomly selects a node and its neighboring nodes in each iteration. Subsequently, the algorithm calculates the modularity associated with the randomly chosen node with each adjacent node. The combination that yields the highest modularity leads to combining the corresponding nodes, forming a new community. This process iterates until no additional progress in the modularity of the community is achievable.

Modularity is mathematically defined as $A_{i,j} - \frac{k_i k_j}{E}$, where $A_{i,j}$ represents the adjacency value (the count of the edges between nodes j and i), k_i is the degree of node-i, k_j is the degree of node-j. E is the total number of edges in the social network (SN). The process of detecting the communities of a toy example SN is illustrated in Figure 1.

Influence Score

This section delves into the Influence Score (IS) concept for a community. IS relies on various parameters, including Influencing Power (IP), the number of active or positive nodes (A), Similarity between neighboring nodes denoted as S_{ij}, and the Degree of each node, D_i. To ensure a comprehensive understanding, we provide mathematical definitions of these parameters and then illustrate the process of IS calculation is represented in Figure 2.

The IS of a community is defined as the average value of positive Influence-Power (IP) of active nodes, represented as $\frac{1}{A}\sum_{i=1}^{i=A} IP_i$, where A corresponds to the count of active nodes, which are nodes with positive Influence-Power values.

The Influence-Power (IP_i) of each node (node-i) measures its ability to influence other nodes within the community. Specifically, the Influence-Power (IP) of a node is computed as $IP_i = D_i - \sum_{j\in N_i} \frac{S_{ij}}{\sum_{j\in N_i} S_{ij}} * D_j$. This calculation involves using the Degree D_i of each node (N_i) and the Similarity between each neighboring node S_{ij}.

Where nodes of the community is defined as:
$N_i = \{i \varepsilon V\{i,j\}\varepsilon E\}$.

The Degree of each node is defined as:
$D_i = \sum_{i=1}^{N}(d^{+}_{C_G}(V_i) + d^{-}_{C_G}(V_i))$. Here, *Similarity* ($S_{ij}$) quantifies the Jaccard Similarity between neighboring nodes, which is calculated as $S_{ij} = \frac{|JS_i \cap JS_j|}{|JS_i \cup JS_j|}$.

The Jaccard Similarity of each node is measured as follows: $JS_i = \{j \varepsilon V\{i,j\}\varepsilon E\} \cup i$. The computation of the IS involves the following steps, including determining the *Degree of each node* (which can be obtained from the community detection approach), assessing the *Similarity (S) of each node*, calculating the *influencing power (IP) of each node*, and then considering only positive IP nodes. Finally, the IS of the community is computed as the average value of positive IP values. Figure 2 displays the process of calculating the Influence Score for *Community-1* of the SN shown in Figure 1.

Community mining to identify the evolution of communities detects structural changes in a sequence of communities over a period of time. Supposing C_{G1}

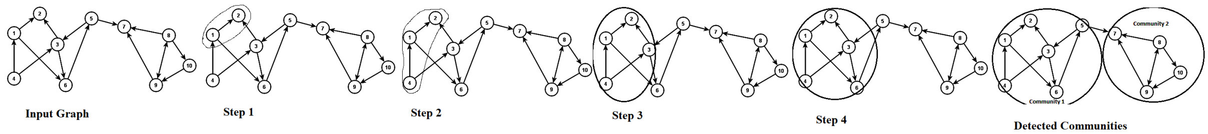


Figure 1: Example of Community Detection using Louvain Technique.

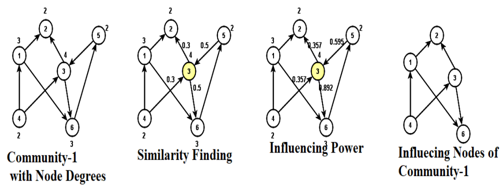


Figure 2: Example of Influence Score Finding of a Community.

to C_{Gt} are sequence of communities over the years, y_1 to y_t, $1 \leq t \leq T$. Community mining calculates similarity scores between sequences of communities and finds the structural changes. These structural changes are categorized as five *events*, namely, *Born*, *Dead*, textitSame, *Merge*, and *Split*. Next, one needs to develop a method to specify the reason for these changes in the nodes of the community (Kumari et al., 2023) (Rajita et al., 2020a).

3 PROPOSED METHODOLOGY

This section elucidates the envisaged framework, illustrated in Figure 3, designed for the identification of communities, feature extraction from these communities, computation of Influence Scores (IS), and event prediction leveraging Machine Learning (ML) models. The framework is implemented on a dataset sourced from DBLP, comprising 1.8 million publications penned by over 1 million authors across numerous journals or conference proceedings series.

The methodology involves several key steps, such as community detection, feature extraction, IS computation, and ML-based event prediction. The algorithmic intricacies of each step within the framework, as represented in Figure 3, are critical for achieving the intended outcomes. Further details on the specific algorithms and techniques employed in each phase are necessary for a more comprehensive understanding.

3.1 Approach

The proposed methodology follows the steps as given below: Step 1: The first step is to process the data, convert XML into a vector structure, and present it graphically for each year.

Step 2: This step uses the *Louvain method* to identify communities of graphically represented data of each year. Table 1 presents the experimental findings from the comparative study of nine well-known community detection methods. Compared to the other eight approaches, **Louvain** finds the communities in less *Time* with greater *Clustering Coefficient(CC)* and *Modularity(M)*.
Step 3: In this step, the community events are detected according to the algorithm suggested in (Bommakanti and Panda, 2018). According to Palla et al. (2007), these identified events are designated as *Dead*, *Born*, *Merge*, *Same*, and *Split*.
Step 4: This step is to identify the community's direct features (Section 3.1.1) and the IS score (Section 3.1.2) of the communities.
Step 5: This step gives IS score and community features as input to the ML model to predict events.

3.1.1 Community Features

Determining the evolutionary patterns of the communities requires computing the community features. Community features refer to characteristics and properties exhibited by subsets of nodes within a social network. These properties provide information on the underlying structure of the community. They are derived through computational methods to determine each community's properties, thereby getting the underlying information of the community structure. This work identified **13** direct community features for better internal connectivity. The detected **13** community features are represented in Table 2.

According to Saganowski et al. (Saganowski et al., 2019), the *Filter method technique* validates the significance of 13 community features on the communities' events. This method justifies the relevance through a correlation matrix using **Pearson correlation**. A relationship between the events (the output variable) and the community features (the independent variables) is depicted by the *Pearson correlation heatmap* in Figure 4. Typically, the correlation coefficient, which runs from -1 to 1, is used to quantify the relationship between the independent and dependent variables. The range of correlation values for the features of the community, which fall between -0.8 and 0.8, is shown in Figure 4. Consequently, it can be inferred from the experimental findings in Fig-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

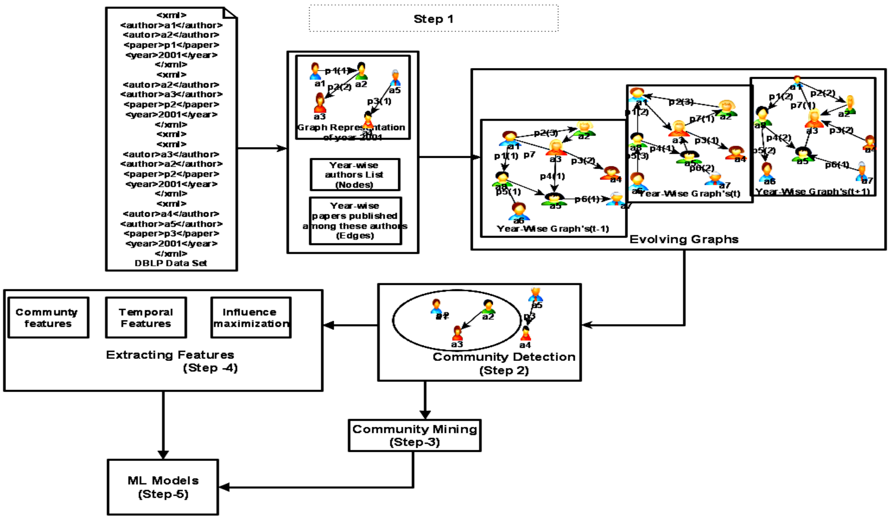


Figure 3: A Proposed Framework for Event Prediction in a Social Network.

Table 1: Comparative Analysis of **Nine** Well-Known Community Detection Algorithms (Rajita et al., 2020b) (Rajita et al., 2021c).

Community Detection Algorithm	Clusterin Co-efficient (CC)	Modularity (M)	Time (in sec)	No.of Communities
Louvain Algorithm	**0.829**	**0.959**	**96.42**	**634402**
Multilevel Algorithm	0.729	0.959	140.23	471504
Fastgreedy Algorithm	0.729	0.949	1756.43	482463
Label propagation Algorithm	0.879	0.879	8321.54	473423
Infomap Algorithm	0.669	0.869	34331.34	595354
Walktrap Algorithm	0.649	0.839	59962.16	548659
Eigenvector Algorithm	0.609	0.769	5332.34	469372
Spinglass Algorithm	0.549	0.779	42083.45	450
Edgebetweeness Algorithm	0.449	0.359	445921.25	12654

Table 2: Direct Features of the Community along with their ID's.

S.No	Feature ID	Feature Name
1	f1	Number of Nodes
2	f2	Number of edges
3	f3	Degree
4	f4	Intra connection
5	f5	Interconnection
6	f6	Conductance
7	f7	Density
8	f8	Clustering Coefficient
9	f9	Connected components
10	f10	Closeness
11	f11	Neighborhood
12	f12	Activeness
13	f13	Aging.

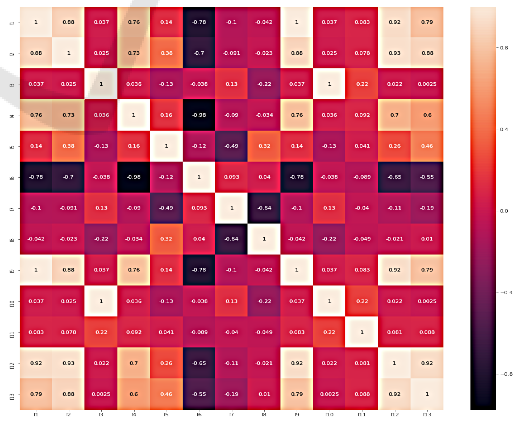
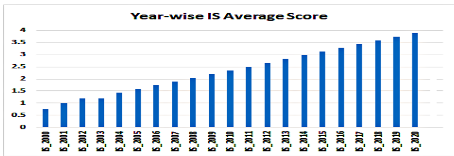
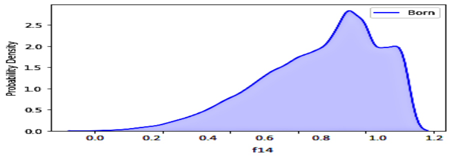


Figure 4: Correlation between Community Features and Events.

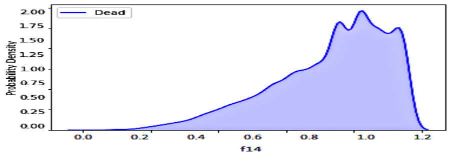
ure 4 that the suggested experiment found that community features aid in predicting community events. The existing ML models for event prediction represent only known entities (node or edge features) and time-varying relationships. This proposed work aims to develop a method that uses multiple and related granularity levels to calculate the IS of the communities in a series of temporal graphs (Papanikolaou et al., 2017) (Sakiyama et al., 2020) (Leskovec et al., 2008).



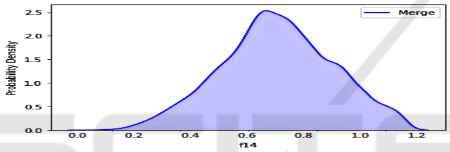
(a) Average value of IS at each Year.



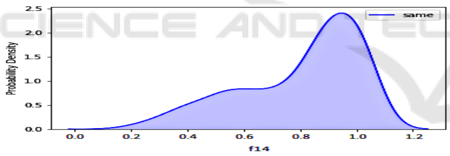
(b) PDF between IS feature (f14) versus Born Event.



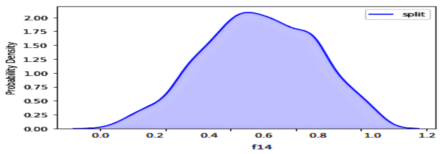
(c) PDF between IS feature (f14) versus Dead Event.



(d) PDF between IS feature (f14) versus Merge Event.



(e) PDF between IS feature (f14) versus Same Event.



(f) PDF between IS feature (f14) versus Split Event.

Figure 5: Average values of IS at each Year and Events and PDF of IS feature versus Five Events.

3.1.2 Influence Score

How to know that IS is a good measure for detecting event changes? The answer to the above question lies in *Probability Distribution Function (PDF)* and *Poisson distribution*.

Probability Distribution Function (PDF): This function is helpful because it tells about the probability of an event occurring in a given interval (In Figure 5, the pdf interval is between 2.0 and 2.5). To demonstrate how well the newly derived feature (IS) works to improve the accuracy of the ML models, it examines whether it impacts the target variable (Events). That is to find the relation between IS (new derived feature) and Events in two steps (Uddin et al., 2012). In Step One, the average of the newly derived feature (IS) at each time period is computed and shown in Figure 5a. According to Figure 5a, the average IS value increased annually, which supports the first step. The second step is to identify the impact of IS on each Event (**Born**, **Dead**, **Merge**, **Split**, and **Same**) by using the *Probability Density Function (PDF)*. It is because the PDF finds the probability across all the possible outcomes concerning events. Figure 5 illustrates the outcome of this step, showing the PDFs of the following events: Born (2.5), Dead (2.0), Merge (2.5), Same (2.5), and Split (2.0). The averages of IS (from Figure 5a) match these PDF values. The inference from Figure 5 is that IS score considerably impacts *Born*, *Merge*, and *Same* events and a lesser impact on *Dead* and *Split* events.

Poisson Distribution: The Poisson Distribution is chosen because it helps to identify which independent variables (features of our data set) impact the dependent variable (event change). The data is grouped year-wise to help organize and derive essential conclusions on how the communities have changed. The beta values are the estimated *Poisson regression coefficients* for the model (Leskovec et al., 2008). The IS Score's greater beta values (0.0289) indicates it is a good fit for identifying event changes. Algorithm 1 gives the proposed algorithm's pseudo-code for computing the communities' IS score.

Data: Community CG(E,V)
Result: IS Score
initialization A = [] ;
for *each node-i in CG* **do**
  Consider D_i for each node (calculated in Community Detection) ;
  $JS_i = \{j\varepsilon V|\{i,j\}\varepsilon E\} \cup i.$;
  $S_{ij} = \frac{JS_i \cap JS_j}{JS_i \cup JS_j}.$;
  $IP_i = D_i - \sum_{j\varepsilon N_i} \frac{S_{ij}}{\sum_{j\varepsilon N_i} S_{ij}} * D_j.$;
  if *($IP_i > 0$)* **then**
    $A = A \cup i$ IS = $\frac{1}{|A|} \sum_{i=1}^{i=T} IP_i$ (A is active or positive IP nodes).;
  end
end
return IS Score;

Algorithm 1: IS Algorithm.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

4 PERFORMANCE AND COMPARATIVE ANALYSIS OF RESULTS

The necessary experimental setup, performance analysis, and comparative analysis are covered in this section.

4.1 Experimental Setup

The co-authorship SN is modeled in this paper using a presented framework. All experiments are conducted on an NVIDIA GeForce GTX TITAN X GPU using Python 3.0. Various machine learning models, including ***Naive Bayes***, ***Decision Tree***, ***SVM***, ***Neural Networks***, and ***Logistic Regression***, are evaluated through 10-fold cross-validation.

4.2 Performance Analysis

Community features mentioned in Section 3.1.1 (community features) and Influence Score mentioned in Section 3.1.2 (derived features) are considered for the experent. Results obtained by comparing IS scores as an additional feature in addition to community features (shown in Table 4) are compared with the 13 community features (shown in Table 3).

Table 3 suggests that, in the absence of IS score as a feature, the efficacy of **Neural Network**, **SVM**, and **Logistic Regression** is equivalent.

Accuracy measures the expected occurrences of events by the training events. A more accurate model gives precise, meaningful, and relevant results. A poor accuracy suggests that the specific machine learning model is inappropriate for the given dataset. Higher precision machine learning models are always superior to the current low-accuracy models (Chen et al., 2017). Experimental results also confirm increased accuracy after including the IS score as a feature. It is evident from Tables 3 and 4 that the Logistic Inclusion of IS score improved the Regression model's accuracy from 81.47% to 85.07%. It implies that the predicted events are now approximately 3.6% more accurate.

Precision calculates the positive impact of the expected occurrences based on the training events. A lower rate of incorrect predictions is implied by high precision, and vice versa for low precision. The findings in Tables 3 and 4 demonstrate that the Logistic Regression model's precision increased from 79.68% to 84.69%. Hence, the precision of the predicted events improved approximately by 5.01% after including IS scores as an additional feature.

The ratio of successfully predicted positive events to all expected events is known as the recall. It evaluates how comprehensive the ML models are. A poor recall score suggests that there are a lot of false negative values in the prediction. Table 4 and Table 3 results reveal that the Logistic Regression model's recall value increased from 76.37% to 84.53%. Therefore, adding IS scores as an extra feature increased the recall score of the predicted events by about 8.16%.

The precision and recall weighted average is the F-measure or F1 score. It represents ML models' robustness (it does not miss a significant number of cases) and preciseness (the number of instances it classifies appropriately). A lower F-score suggests that more balance needs to be added to the dataset. Tables 3 and 4 results demonstrate that the Logistic Regression model's F-score increased from 76.83% to 84.57%. Because of this, the F-score of the anticipated events increased by roughly 7.74% after adding IS scores as a feature.

Table 3: Performance of ML models based on only Community Features.

ML Models	Accuracy	Precision	Recall	F-measure
Decision Tree	52.92	51.59	50.89	51.24
Naive Bayes	73.92	72.61	71.72	72.16
Neural Network	81.84	72.14	75.53	74.25
SVM	81.17	79.92	71.82	70.85
Logistic Regression	81.47	79.68	76.37	76.83

Table 4: Performance of ML models based on Community Features + IS score.

ML Models	Accuracy	Precision	Recall	F-measure
Decision Tree	76.82	76.48	76.41	76.79
Naive Bayes	78.91	78.31	78.71	78.89
Neural Network	81.78	81.63	81.27	81.94
SVM	81.97	81.57	81.59	81.58
Logistic Regression	**85.07**	**84.69**	**84.53**	**84.57**

4.3 Comparative Analysis with Existing Work

This section compares this paper's proposed methodology with some existing works. The number of connected neighbors shared by nodes v_i and v_j is used by *Influence maximization using greedy strategy* (Zhang et al., 2021) to identify active nodes. The correlation between v_i and v_j's number of connected neighbors is confirmed via a greedy technique. So, it identified only the association of two nodes. However, it did not include the overall community structural change. Hence, it misses out on those nodes that the nodes in other communities influence. Therefore, the existing *Influence maximization using greedy strategy* (Zhang et al., 2021) is less accurate in predicting certain events. So, it fails to find some critical events such as **Merge**, **Same**, and **Born**. The presented IS technique considers the influence of other nodes and nodes influencing other nodes. So, Similarity, S_i of each node (node-i), measures how much its neighbors influence node-i. And Influence-Power IP_i of each node (node-i) measures how much node-i can influence others. This method helps to find the probability of an author likely to change the research interest based on the IS score of the community. It experimentally proved that using PDF, the proposed IS score helps detect more critical events such as **Merge**, **Same**, and **Born**.

Table 5: Performance Analysis of IS over AS, CD, and Zhang et al.

ML Model	**Logistic Regression**			
Types of Features ↓	Accuracy	Precision	Recall	F-measure
Community Features (CF) + Anomaly Score (AS) (Rajita et al., 2021a)	83.57	83.62	83.59	83.26
CF + Collaborative Distance (CD) (Rajita et al., 2021b)	84.27	83.62	83.53	83.57
CF + IS	**85.07**	**84.69**	**84.53**	**84.57**
zhang et al.'s Influence maximization using greedy strategy (Zhang et al., 2021)	82.54	81.21	81.72	81.36

Section 4.2 inferred that the Logistic Regression (LR) model predicts more accurate events than the remaining four ML models (Decision Tree, Naive Bayes, Neural network, and SVM). The effectiveness of the suggested strategy is evaluated in Table 5 with that of other derived features, including Anomaly Score (Rajita et al., 2021a) and Collaborative Distance (Rajita et al., 2021b), as well as with the Zhang et al. (Zhang et al., 2021) method on the logistic regression model. The accuracy of the proposed approach (community features + Influence score) is 85.07%. It is approximately 2.53%, 1.5% and 0.8% more accurate than *Zhang et al. (Zhang et al., 2021)*, *Anomaly Score*, and *Collaborative Distance* approach, respectively. The precision of the proposed approach is 84.69%. It is about 3.48%, 1.07%, and 1.07% more accurate than *Zhang et al.*, *Anomaly Score*, and *Collaborative Distance* approaches, respectively. The recall of the proposed method is 84.53%. It is approximately 2.81%, 0.94%, and 1.00% improvement over *Zhang et al.*, *Anomaly Score*, and *Collaborative Distance* approach, respectively. Similarly, the F-measure of the proposed model is 84.57%. It is approximately 3.21%, 1.31%, and 1.00% better than *Zhang et al.*, *Anomaly Score*, and *Collaborative Distance* approaches, respectively.

5 CONCLUSION AND FUTURE WORK

The Influence Score, a newly developed feature included in this study, enhanced the ability of machine learning models to predict social network community events. The graphical depiction of the social network community serves as the basis for the newly suggested feature. The similarity and influence power of the nodes determines the IS score. The outcomes of the performance and efficacy of the presented method on the dataset demonstrate that including the IS score as an extra feature significantly increased the ML models' accuracy by about 6.6%. The accuracy of the logistic regression model increased from 81.47% to 85.07%. Additionally, it demonstrated a significant improvement in the other parameters, including F-measure, recall, and precision. A comparative analysis presented in this paper shows that the proposed approach gives better results over other derived features, such as anomaly score and collaborative distance, by approximately 1.5% and 0.8%, respectively. It also indicates that the IS score increased the accuracy of the logistic regression model by 2.53% compared to the Zhang et al. approach.

The authors intend to employ stochastic gradient and GAN techniques to measure Influence Score in the future to increase accuracy even further.

REFERENCES

Altmami, N. I. and Menai, M. E. B. (2020). Automatic Summarization of Scientific Articles: A Survey. *Journal of King Saud University-Computer and Information Sciences*, pages 46–57.

Bhattacharjee, U., Srijith, P., and Desarkar, M. S. (2019). Term Specific tf-idf Boosting for Detection of Rumours in Social Networks. In *11th International Conference on Communication Systems & Networks (COMSNETS)*, pages 726–731. IEEE.

Bommakanti, S. R. and Panda, S. (2018). Events Detection in Temporally Evolving Social Networks. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 235–242. IEEE.

Chakraborty, T., Krishna, A., Singh, M., Ganguly, N., Goyal, P., and Mukherjee, A. (2016). Ferosa: A Faceted Recommendation System for Scientific Articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 528–541. Springer.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., and Ma, J. (2017). A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. *Catena*, 151(2):pages 147–160. Elsevier.

Citron, D. T. and Way, S. F. (2018). Network Assembly of Scientific Communities of Varying Size and Specificity. *Journal of Informetrics Elsevier*, 12(1):pages 181–190.

Duan, L., Ma, S., Aggarwal, C., Ma, T., and Huai, J. (2017). An Ensemble Approach to Link Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):pages 2402–2416 IEEE.

Kong, X., Shi, Y., Yu, S., Liu, J., and Xia, F. (2019). Academic Social Networks: Modeling, Analysis, Mining and Applications. *Journal of Network and Computer Applications*, 132(3):pages.86–103 Elsevier.

Kubiszewski, I., Concollato, L., Costanza, R., and Stern, D. I. (2023). Changes in Authorship, Networks, and Research topics in Ecosystem Services. *Ecosystem Services*, 59:101–111.

Kumari, D., Yannam, P. K. R., Gohel, I. N., Naidu, M. V. S. S., Arora, Y., Rajita, B., Panda, S., and Christopher, J. (2023). Computational model for breast cancer diagnosis using hfse framework. *Biomedical Signal Processing and Control*, 86:105121.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008). Statistical Properties of Community Structure in Large Social and Information Networks. In *Proceedings of the 17th International Conference on World Wide Web*, pages 695–704. ACM.

Palla, G., Barabsi, A.-L., and Vicsek, T. (2007). Quantifying Social Group Evolution. *Nature*, 446(7136):pages.664–685 Nature Publishing Group.

Papanikolaou, Y., Foulds, J. R., Rubin, T. N., and Tsoumakas, G. (2017). Dense Distributions from Sparse Samples: Improved Gibbs Sampling Parameter Estimators for LDA. *The Journal of Machine Learning Research*, 18(1):pages 2058–2115. JMLR.

Rajita, B., Kota, V. R. A., Perika, K., and Panda, S. (2021a). Anomaly score: A new feature for effective prediction of events in social networks. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 189–192. IEEE.

Rajita, B., Kumari, D., and Panda, S. (2020a). A Comparative Analysis of Community Detection Methods in Massive Datasets. *In: Modelling, Simulation and Intelligent Computing. MoSICom 2020.*, pages 174–183. Springer.

Rajita, B., Narwa, B. S., and Panda, S. (2021b). An efficient apagesroach for event prediction using collaborative distance score of communities. In *International Conference on Distributed Computing and Internet Technology*, pages 271–279. Springer.

Rajita, B., Ranjan, Y., Umesh, C. T., and Panda, S. (2020b). Spark-Based Parallel Method for Prediction of Events. *Arabian Journal for Science and Engineering*, 45(4):pages. 1–17 Springer.

Rajita, B., Shukla, M., Kumari, D., and Panda, S. (2021c). An Experimental Analysis of Community Detection Algorithms on a Temporally Evolving Dataset. In *Recommender Systems*, pages 23–70. CRC Press.

Saganowski, S., Bródka, P., Koziarski, M., and Kazienko, P. (2019). Analysis of Group Evolution Prediction in Complex Networks. *PloS one*, 14(10):e0224194 Public Library of Science San Francisco, CA USA.

Sakiyama, M., Fujii, N., Kokuryo, D., and Kaihara, T. (2020). Visualization of Group Discussion using Correspondence Analysis and LDA in Ideathon. *Procedia CIRP*, 88(1):pages. 595–599. Elsevier.

Saxena, A., Kumar, B., Gupta, A., Sahasrabudhe, N., and Moharir, S. (2021). Influencing opinions of heterogeneous populations over finite time horizons. In *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 474–482. IEEE.

Schäfermeier, B., Hirth, J., and Hanika, T. (2023). Research Topic Flows in Co-authorship Networks. *Scientometrics*, 128(9):5051–5078.

Sumith, N., Annappa, B., and Bhattacharya, S. (2017). Social Network Pruning for Building Optimal Social Network: A User Perspective. *Knowledge-Based Systems Elsevier*, 117:pages 101–110.

Uddin, S., Hossain, L., Abbasi, A., and Rasmussen, K. (2012). Trend and Efficiency Analysis of Co-Authorship Network. *Scientometrics*, 90(2):pages. 687–699. Springer.

Zhang, Z., Li, X., and Gan, C. (2021). Identifying influential nodes in social networks via community structure and influence distribution difference. *Digital Communications and Networks*, 7(1):pages 131–139.