International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Ant Colony Decision Tree Method to Detect the Suspicious News

Asha Kumari, Balkishan

Abstract*: The ease of availability and low cost of social media platforms has made the human vulnerable to devour the information from social media. Social media serves the unverified information which easily gets disseminated among people through different groups, applications, and other online social platforms. These fake news can lead to any suspicious activities which urges to derive the term suspicious news. The escalation of these suspicious news has open a lot of research ventures to detect and attenuate the impact on human. These suspicious activities have become a nuisance for legitimate users. Despite the presence of existing methods for the suspicious news detection exists, but the continuous growth of such activities is difficult to manage with an individual approach. In this research work, an ensemble approach is considered to detect suspicious news content. Here, Ant Colony Decision Tree method is ensembled for the detection of suspicious news (ACDTDSN). This approach uses the heuristic function and pheromone trail to obtain the optimal solution. The overall functionality of system is based on the content based approach for the detection of suspicious news with steps of dataset consideration, pre-processing, feature selection, and classification. The experimentation is performed using the FakenewNet dataset which consist of BuzzFeed and PolitiFact categories of news content. The results of the proposed ACSTDSN framework are accessed with the performance evaluation measures.

***Keywords*: Fake News, Suspicious News, Ant Colony Optimization, Decision Tree, Artificial Intelligence, Text Mining.**

I. INTRODUCTION

In the recent years, the growth and ease of social media to disseminate information has been raised. This makes to exacerbate the human intentions and reviews towards any product, entity, any event or towards other human beings. It also deludes the people towards the unauthentic news. These kinds of activities also impact on the live share markets, political opinions, and other kind of financial bodies. The available instance related to share market price drop is related to the United Airlines. In the year 2008, somebody posted the fake news of bankrupt status of the organization which badly impacts their share market and fallen their share prices by 76% in less than an hour of time. After the detection of news as fake news, much of the share price has been retained by the organization but the stock price drop was there for the entire week [1]. The other instances also came in light during the period of natural disasters like hurricane [2] and earthquake in the Japan [3]. At that time of natural disaster, the dissemination of fake stories increases the panic of people.

These kinds of life threatening fake stories can only be considered in the category of suspicious news.

The detection of suspicious news is the procedure to access and detect the authenticity about some of the pieces of news. There is the role of four major components behind the fake news. These components are social scenario, news content, target users, and fake news creator. In the recent years, the researchers have started focusing on these research topics. This can be noticed by available statistics of research publications indexed in Scopus database from year 2015 to 2018 concerning the fake news [4]. This statistics is illustrated in fig. 1. The detection of fake news statistics from year 2015 to 2018 is illustrated in fig. 2 [4]. The change in graph (fig. 1 and fig. 2) from year 2015 to 2018 can clearly indicate the shifting focus of researchers towards fake news detection.

The detection of fake news is a challenging research issue that should be addressed completely and quickly. There is the availability of a lot of online websites to check the authenticity of fake news along with the lot of articles that guides to analyze the suspicious news. Even after the availabilities of these facilities, it is hard to completely control the false news entities. The reason behind this uncontrolled situation is the increasing social media platforms. Moreover, the major factor is that fake news also disseminated by bots and cyborgs. These entities can spread and create the content automatically in a quick manner as compared to human. In this manner, fake news are disseminated by both the human and robotic entities. The target of these entities depends on the agenda of disseminators. The fake advertisements or fake reviews can target the online customers, fake health news can target the old age people, fake educational news can target the students and parents, etc. In this way, the fake news directly or indirectly can affect the human life and can lead to suspicious activities.

These suspicious activities can be reduced by improving the existing or designing the new research methodologies. Some of the existing research studies are discussed in next section II of literature review. These methods works well in some of circumstances but fails due to the presence of single research methodology or may be the other reason.

Manuscript published on November 30, 2019.
* Correspondence Author
Asha Kumari*, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India, Email: asha198829@gmail.com
Balkishan, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India, Email: balkishan248@gmail.com

In this research work, machine learning based decision tree and artificial swarm intelligence based ant colony optimization are ensembled to detect the suspicious news. The proposed framework is termed as ACDTDSN which consists of four phases of dataset consideration, pre-processing, feature selection, and classification.

These are discussed in section III of research methodology. The results of the system are analyzed with the evaluation parameters of recall, precision, and f-measure by experimentation on FakenewsNet dataset. The result along with the discussion and comparison are illustrated in section IV. The discussions in the form of conclusion and future direction that can be attained are discussed in section V.

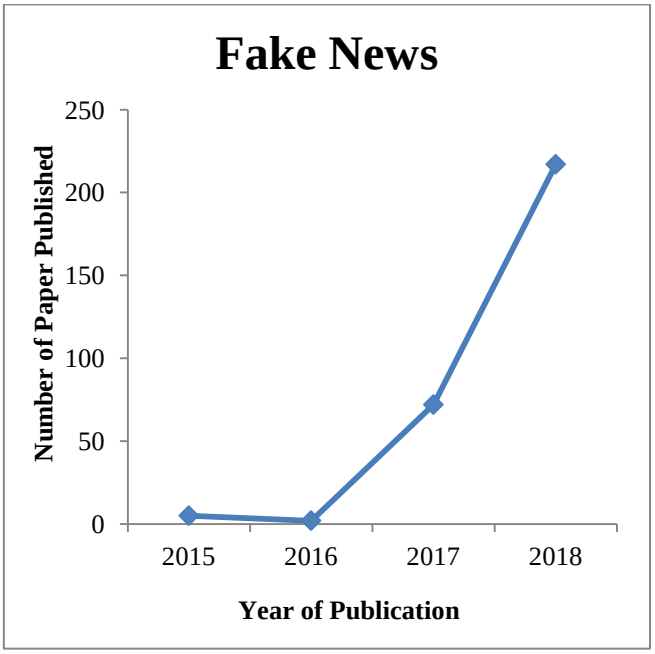

Fig. 1: Research Publications concerning the Fake News from the year 2015 to 2018

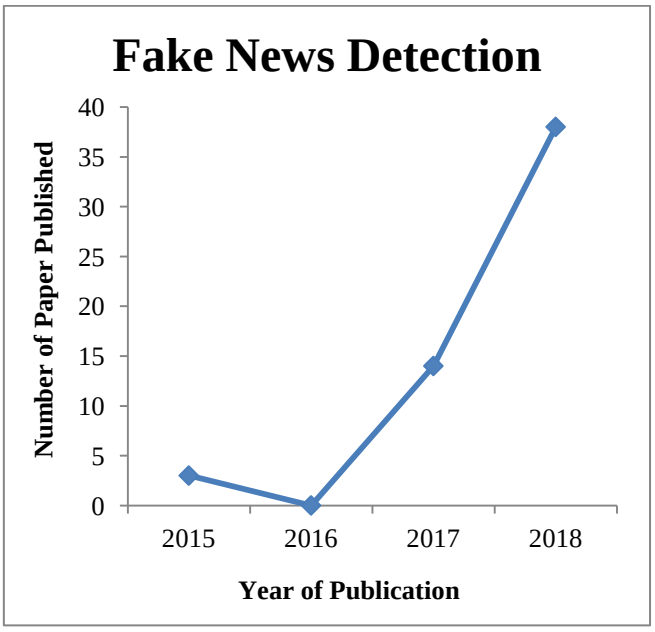

Fig. 2: Research Publications on Fake News Detection from the year 2015 to 2018

II. LITERATURE REVIEW

The online social media platforms have the ability to influence the lives of human in various positive and negative aspects. The fake news greatly affect the public opinion and able the change the overall final resulting scenario. At the time of sensitive situations, the dissemination of fake news stories can leave the harmful impact on people. There are various existing concepts and methods for the suspicious news detection. Some of the quality contributions in the field of fake and suspicious news are discussed here.

Ahmed et al. [5] have analyzed the fake news detection using the manually considered research dataset. The research experimentation was conducted based on the n-gram feature sets with the use of term frequency and TF-IDF approaches for the feature extraction. The classification is conducted and compared using the machine learning classifiers of decision tree, k-nearest neighbor, linear support vector machine, support vector machine, and stochastic gradient descent. The results values in terms of accuracy indicate the superiority of results for the linear support vector machine along with feature extractor of TF-IDF.

Shu et al. [6] have proposed the semi-supervised approach by integrating the attributes of news instance, publisher bias, and user engagement. The proposed approach was termed as TriFN which represents the framework to detect the fake news with tri-relationship. The authors have also proposed two datasets of PolitiFact and BuzzFeed News for the detection of fake news. The research experimentation conducted on the mentioned dataset using the proposed TriFN approach indicates the remarkable result values for the mentioned approach.

Fernández-Reyes and Shinde [7] have considered the recurrent neural network and convolutional neural network approaches for the detection of fake news. The research experimentation was performed based on the publically available Liar dataset. The evaluated results indicate the improved performance of concepts based on deep neural network as compared to considered machine learning and other classifiers. But the combinational approach of convolutional and recurrent neural network has not attained much improvement in result values.

Tagami et al. [8] have initially created the research dataset for the fake news detection by collecting the Japanese language posts from twitter. The authors have worked for the prediction of suspicious articles and casting posts. The classification of the dataset entities was performed using the Lachine learning classifiers of long short term memory, random forest, decision tree, support vector machine, and logistic regression. The experimental evaluation indicates the lackness of decision tree and random forest algorithms as compared to other mentioned concepts. The authors have promised to reduce the system cost by 50% as compared to manual analysis of suspicious news detection.

Atodiresei et al. [9] have presented a real time application to detect the fake news. The authors have used the dataset entities from Twitter and detected the fake news & their respective suspicious fake users.

Retrieval Number: D6794118419/2019©BEIESP
DOI:10.35940/ijrte.D6794.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The research work was conducted with the consideration of natural language processing modules, naïve bayes classifier, support vector machine, and maximum entropy approach. The authors have attained the considerable detection reports but with the limitation of method to use in English language and sometimes system lacks in case of novel fake news from trustworthy channel. Aldwairi and Alwahedi [10] have used the machine learning classifiers to identify the fake news posts from the online social media platforms.

The authors have used the WEKA tool to train and classify the test data. The research database was collected from the online social media platforms along with their URL information. The classifiers of naïve bayes, random forest, logistic, and bayes net have been used by authors for the detection and classification of fake news. The authors have noted the outstanding performance evaluation results for the logistic classifier as compared to other mentioned classifiers.

Vishwakarma et al. [11] have proposed rule based system to test the veracity of news and information available on the social media platforms. The authors conducted the research work by extracting the textual data from the images and the veracity of news was analyzed by comparing the results with the top 15 web links available on the Google. The authors explained the research methodology in the four modules of extraction of textual data from image, extraction of entity, web based processing, and process unit. The research experimentation also conducted by considering the dataset of PHEME, FakeNewsNet, and BuzzFeed Election. The authors have noted the remarkable results but the system lacks in case of local news due to availability of lesser information on web for local news.

Reis et al. [12] have worked on the dataset related to US election of year 2016 for the detection of fake news and posts on social media. The authors have tried to do the dataset based experimentation along with the real time practice experimentation. The classification of the system is conducted using the machine learning classifiers of XGBoost, support vector machine with RBF kernel, random forests, naïve bayes, and k-nearest neighbor. The research experimentation indicates the superiority of result values in case of XGBoost and random forest algorithms. The authors have noted the considerable classification with dataset and around 40% of misclassification of true news as fake news.

III. RESEARCH METHODOLOGY

In this section, the step by step procedure of proposed ACDTDSN approach is elaborated. The modules of ACDTDSN approach are dataset consideration, pre-processing, feature selection, and classification. Initially the experimentation based dataset is considered. Then, pre-processing of the raw data is performed. The pre-processed data is used to select the feature components. The final classification is performed using the ant colony decision tree based integrated classifier. These modules are illustrated as the work flow in the fig. 3.

A. Dataset Consideration

The research experimentation of proposed ACDT approach is conducted using the FakenewsNet dataset. The dataset of FakenewsNet consists of two sub-categories of news: BuzzFeed and PolitiFact news based on the ground truth verification from buzzfeed.com and politifact.com respectively. This Buzzfeed dataset category consists of 182 news and PolitiFact contains 240 news. These dataset categories carry both the fake news and true news. In this work, fake news are termed as suspicious news as the fake or untrue activities can lead to any threatening incident in both the physical or mental senses. The division of news in the Buzzfeed and PolitiFact dataset categories is described in fig. 4. and fig. 5 respectively.

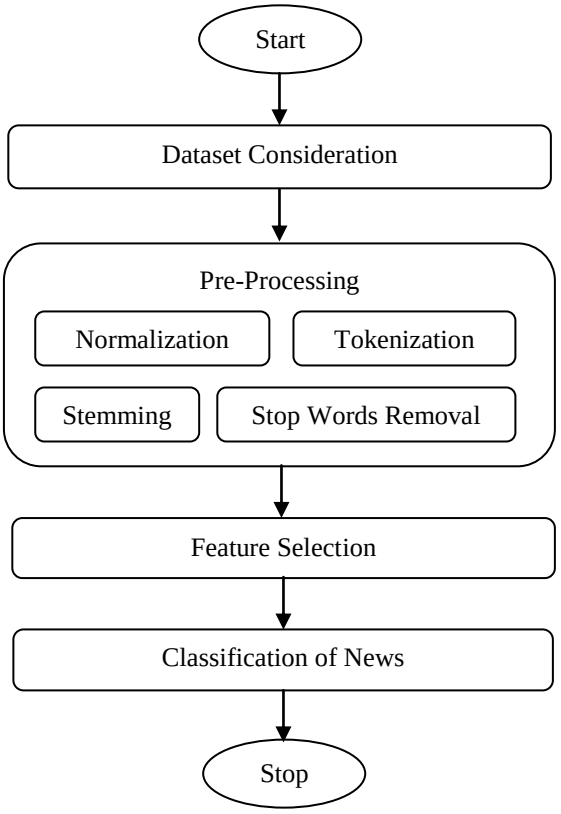

Fig. 3: The Proposed ACDTDSN Framework Workflow

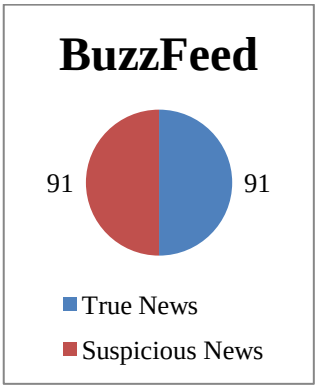

Fig. 4: BuzzFeed news Distribution

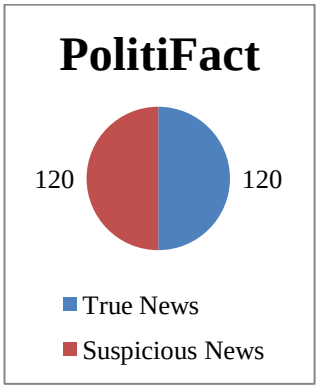

Fig. 5: PolitiFact news Distribution

The statistics illustrated in fig. 4 and fig. 5 indicates that there is the equal distribution of true news and suspicious news in the dataset.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Moreover, the publisher information is also available for the dataset categories. There are the 91 and 9 publishers for the buzzfeed and politifact dataset categories respectively.

B. Pre-processing

The raw data considered in the previous step is needed to pre-process for the extraction of feature set. The pre-processing involves the steps of normalization, tokenization, porter stemming, and stop words removal.

Step 1: The first step of pre-processing is normalization. Normalization is considered to detect and normalize the data available in the other than English language. Here, Google translator is exploited to translate the other language based news into English language.

Step 2: The normalized data attained in the previous step is processed for tokenization. The unigram tokenization is applied to tokenize the sentences. Here, each word in the dataset is separated based on the 'space' separator. The separated words are considered as tokens. These token are further processed for the stemming process.

Step 3: Stemming process is performed using the porter stemmer approach that removes the adjoining affix & suffix of the word and converts the work into the basic root word. For instance, there are various feasible words for the word 'teach' such as 'teaching', 'teaches', 'teacher', 'teachers', 'teachable'. The stemming processing removes the suffix of 'ing', 'es', 'er', 'ers', 'able' respectively from the mentioned words and converts these words into root word 'teach'. These stemmed tokens are accessed to check and remove the stops words.

Step 4: Stop words removal process removes the insignificant words such as 'in', 'an', 'are', 'you', and 'who', 'that', etc. The removal of stop words process reduces the search space and improves the system performance.

C. Feature Selection

The news can be predicted as the suspicious or true news based on the features available. The features are the essential components of for the final classification. In this research work, the features related to textual content, URL information, and user behavior are exploited. The considered features are shown in fig. 6.

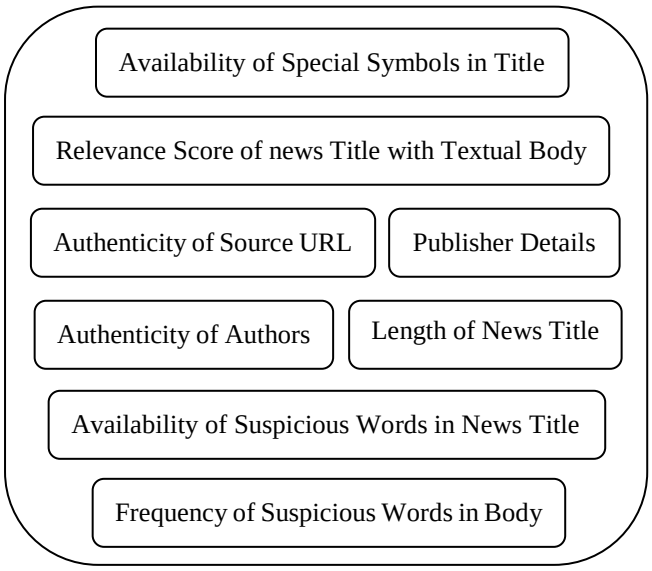

Fig. 6: Features Selected for Classification

The feature components illustrated in fig. 6 are selected for the classification of news as the suspicious or true. The next module is the classification of news.

D. Classification

The classification of the news is the last module of proposed ACDTDSN framework. The classification of the news is conducted using the integrated approach of ant colony optimization and decision tree classifier. This integrated approach is termed as Ant Colony Decision Tree (ACDT) approach [13]. In this amalgamated approach, the ants construct the decision trees based on the heuristic functions and pheromone trail. The decision trees are constructed using the CART algorithm to lower down the possible mean squared errors.

The process of ACDT based classification begins with the dissemination of ants among the available nodes and node splitting is conducted using twoing splitting rule. The initial pheromone is noted along with the pheromone after the movement of ants. After the movement of ants from beginning point, the possible route combinations of decision tree are determined and stored in the matrix format. The process continues till the best possible prediction classification of news attained. The classified predictions of news are stored and results are evaluated by comparing the evaluated predictions with ground truth values.

IV. RESULTS AND COMPARISON

The results of the proposed ACDTDSN framework are evaluated by experimentation on window based computer machine with 8GB of RAM, and Intel I5-Core processor. The simulation software of MATLAB is used for the experimentation. The performance of the system is accessed using the evaluation measures of recall, precision, and f-measure. The formulations of these measures are based on the false negative (FN), false positive (FP), true negative (TN), and true positive (TP). These are evaluated based on the predicted results and ground truth values. The confusion matrix for these TP, TN, FP, and FN is mentioned in table I.

Table I: Confusion Matrix for Result Evaluation

		Ground Truth	
		Suspicious News	**True News**
Predicted Results using ACDTDSN	**Suspicious News**	True Positive	False Positive
	True News	False Negative	True Negative

Based on the confusion matrix definitions, the formulations of recall, precision, and f-measure are illustrated in (1) to (3).

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F - Measure = 2 * \quad (3)$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$\frac{Recall*Precision}{Recall+Precision} \quad (3)$$

Further, these formulations are used to evaluate the results of proposed ACDTDSN framework for both the BuzzFeed and PolitiFact dataset categories of FakenewsNet dataset. For testing, the overall dataset is divided in the ratio of 60:40.

The 60% data is utilized for the learning of proposed framework and 40% is further used for the testing. The testing results for the BuzzFeed category and PolitiFact category are evaluated as illustrated in table II

Table II: Evaluated Results using ACDTDSN Approach

	BuzzFeed Category	PolitiFact Category
TP	34	46
FN	02	02
FP	03	04
TN	33	44
Recall (%)	94.44	95.83
Precision (%)	91.89	92.00
F-Measure (%)	93.15	93.88

The evaluated results illustrated in table II are further utilized for the comparison with existing methods used by Shu et al. [6] and Vishwakarma et al. [11]. These authors have also evaluated the results in terms of recall, precision, and f-measure. The comparison of proposed ACDTDSN approach with existing concepts is illustrated in fig. 7 and fig. 8 respectively for the BuzzFeed and PolitiFact Categories.

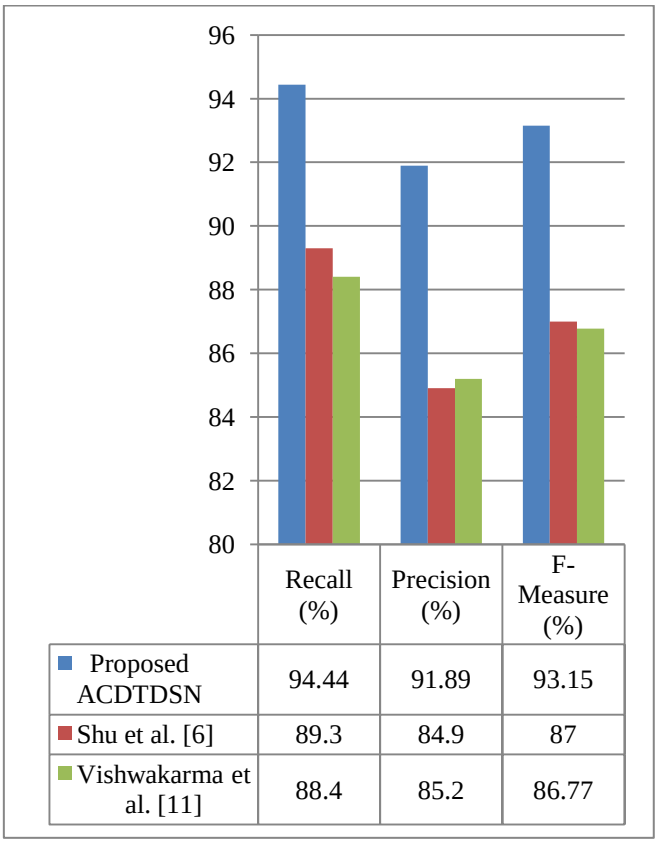

Fig. 7: Comparison of Results for BuzzFeed Category

The results and comparison illustrated in fig. 7 indicate that proposed ACDTDSN approach have achieved the better performance results for the BuzzFeed dataset category in terms of all the evaluation measures. The proposed ACDTDSN approach have f-measure score of 93.15% which is higher as compared to f-measure score evaluated by Shu et al. (87%) and Vishwakrama et al. (86.77%). As f-measure is the evaluated from the recall and precision, the higher the f-measure score indicate the overall higher efficacy.

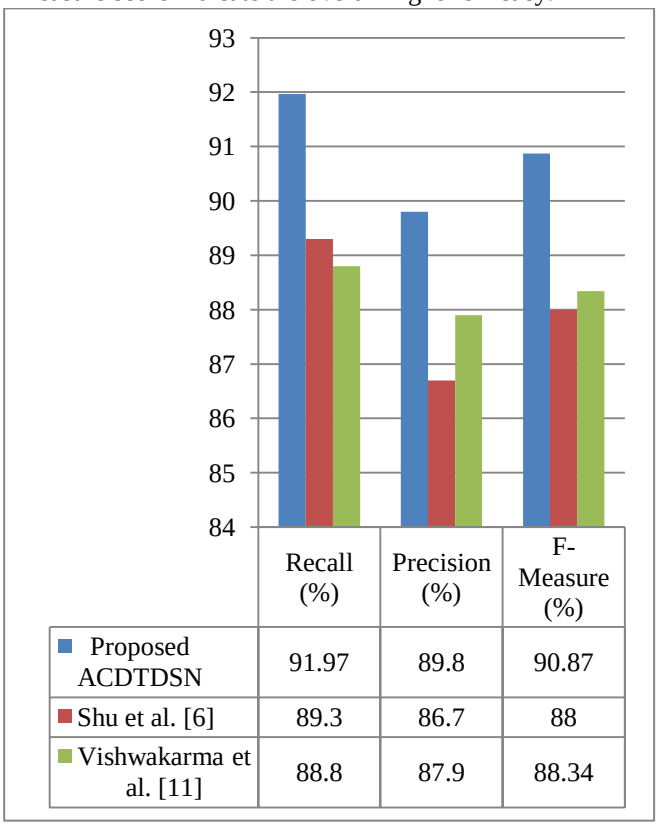

Fig. 8: Comparison of Results for PolitiFact Category

Alike the results illustrated in fig. 7 for the BuzzFeed Category, the fig. 8 illustrate the result comparison for the PolitiFact dataset category. The comparison result indicate the higher efficacy of proposed ACDTDSN approach as compared to other considered concepts used by Shu et al. [6] and Vishwakarma et al. [11].

V. CONCLUSION

Misinformation and fake news always lead to threatening and suspicious outcomes. The increasing social media platforms and growing users may lead to hit the target users with their agenda. The research on fake news is still growing and expanding for the better prediction of such kind of fake news. The most of the existing work is based on individually machine learning and other classification methods which lacks in some cases. In this research work, the concepts of decision tree and ant colony optimization are ensembled and used for the detection of suspicious news (ACDTDSN). The proposed ACDTDSN framework is tested on the FakenewsNet dataset in which dataset is verified from buzzfeed.com and politifact.com. This makes the nomenclature of categories as BuzzFeed and PolitiFact.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

These categories consist of both the fake/suspicious news and true news. The results are determined using the evaluation measures of recall, precision, and f-measure. The proposed ACDTDSN approach has achieved the f-measure score of 93.15% and 90.87% for the BuzzFeed and PolitiFact news respectively.

The evaluated experimental results are also compared with existing methods. The comparison results indicate that the proposed ACDTDSN framework outperformed as compared to existing concepts.

REFERENCES

1. C. Carvalho, N. Klagge, and E. Moench, "The persistent effects of a false news shock", Journal of Empirical Finance, Vol. 18, no. 4, 2011, pp. 597-615.
2. A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy", In Proceedings of the 22nd international conference on World Wide Web companion, 2013, pp. 729-736.
3. M. Takayasu, K. Sato, Y. Sano, K. Yamada, W. Miura, and H. Takayasu, "Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study", PLoS one, Vol. 10, no. 4, 2015, pp. 1-18.
4. A. Bondielli, and F. Marcelloni, "A survey on fake news and rumour detection techniques." Information Sciences, Vol. 497, 2019, pp. 38-55.
5. H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using N-gram analysis and machine learning techniques", In International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, Cham, 2017, pp. 127-138.
6. K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection." Project: Fake News Detection and Mitigation on Social Media, 2017.
7. F. C. Fernández-Reyes, and S. Shinde, "Evaluating Deep Neural Networks for Automatic Fake News Detection in Political Domain", In Ibero-American Conference on Artificial Intelligence, Springer, Cham, 2018, pp. 206-216.
8. T. Tagami, H. Ouchi, H. Asano, K. Hanawa, K. Uchiyama, K. Suzuki, K. Inui "Suspicious News Detection Using Micro Blog Text." In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong, 2018, pp. 648-657.
9. C.S. Atodiresei, A. Tănăselea, and A. Iftene, "Identifying Fake News and Fake Users on Twitter." Procedia Computer Science, Vol. 126, 2018, pp. 451-461.
10. M. Aldwairi, and A. Alwahedi, "Detecting fake news in social media networks." Procedia Computer Science, Vol. 141, 2018, pp. 215-222.
11. D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search." Cognitive Systems Research, Vol. 58, 2019, pp. 217-229.
12. J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised Learning for Fake News Detection." IEEE Intelligent Systems, Vol. 34, no. 2, 2019, pp. 76-81.
13. U. Boryczka and J. Kozak, "Ant colony decision trees–a new method for constructing decision trees based on ant colony optimization." In International Conference on Computational Collective Intelligence Conference, Springer, Berlin, Heidelberg, 2010, pp. 373-382.

AUTHORS PROFILE

Asha Kumari received her M.tech degree in Computer Science from Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. She is currently working as an Assistant Professor in Department of Computer Science, Bhaskaracharya College of Applied Sciences, New Delhi, India. Her research interests include Data mining, machine learning, Software Engineering.

Balkishan received his PhD degree in Computer Science from Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. He is currently working as an Assistant Professor in Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, India. His research interests include Soft computing, Artificial Intelligence, Data Mining, Software Engineering.