Improving Fake News Detection Using K-means and Support Vector Machine Approaches

Kasra Majbouri Yazdi, Adel Majbouri Yazdi, Saeid Khodayi, Jingyu Hou, Wanlei Zhou, Saeed Saedy

***Abstract*—**Fake news and false information are big challenges of all types of media, especially social media. There is a lot of false information, fake likes, views and duplicated accounts as big social networks such as Facebook and Twitter admitted. Most information appearing on social media is doubtful and in some cases misleading. They need to be detected as soon as possible to avoid a negative impact on society. The dimensions of the fake news datasets are growing rapidly, so to obtain a better result of detecting false information with less computation time and complexity, the dimensions need to be reduced. One of the best techniques of reducing data size is using feature selection method. The aim of this technique is to choose a feature subset from the original set to improve the classification performance. In this paper, a feature selection method is proposed with the integration of K-means clustering and Support Vector Machine (SVM) approaches which work in four steps. First, the similarities between all features are calculated. Then, features are divided into several clusters. Next, the final feature set is selected from all clusters, and finally, fake news is classified based on the final feature subset using the SVM method. The proposed method was evaluated by comparing its performance with other state-of-the-art methods on several specific benchmark datasets and the outcome showed a better classification of false information for our work. The detection performance was improved in two aspects. On the one hand, the detection runtime process decreased, and on the other hand, the classification accuracy increased because of the elimination of redundant features and the reduction of datasets dimensions.

***Keywords*—**Fake news detection, feature selection, support vector machine, K-means clustering, machine learning, social media.

I. Introduction

DETECTING fake news has become a new research topic in recent years as the continuous spread of false information has raised the need for assessing the authenticity of digital content. Fake news is mostly created to influence people's perceptions in order to distort consciousness and decision-making [1], [2]. Although the dissemination of false information on the Internet is not a new phenomenon, the extensive usage of social media increases its negative impact on society and also more creation of fake news medium. These days, by the growth of technologies, information is distributed very quickly and its impact on social networks is incredible as it can be reinforced and affect millions of users remarkably in a few minutes [3]. Fact-checking, information validation, and verification is a long-term issue that influences all types of media.

For validating and authenticity of the information, classification and prediction are needed based on the previous training, so a classifier is usually used for that purpose [4], [5]. Designing an efficient classifier with less computational complexity and high precision is the goal of this paper.

One of the main issues of the previous works is that they usually involve all detection features, which causes high computational complexity. That also results in a low classification precision due to the consideration of redundant unrelated features in the detection algorithm. High-dimensional datasets decrease the functionality of the classifier in two aspects; on one hand, the volume of computation is increased, and on the other hand, the models created on the high dimensional data have less generalization so it increases the overfitting. Therefore, reducing the dimensions of the datasets can decrease the computational complexity and improve the classification algorithm performance [6]- [8].

News related data are usually described with many features and it is possible that most of them are unrelated and redundant for the desired data mining. The large number of these unrelated features makes a negative impact on fake news detection algorithm performance whilst the computational complexity is very high too. Besides, minimizing the dimensions of the dataset by removing unrelated redundant features is a challenging task in data mining and machine learning.

This paper is organized as follows. The second section reviews the previous works on fake news detection approaches. The third section describes the proposed method. Evaluation and analysis discussion of the proposed method is described in section four and finally, the last section gives the conclusion of this paper.

II. Review of Literature

There are two categories of important researches in automatic classification of real and fake news up to now:

- In the first category, approaches are at conceptual level, distinction among fake news is done for three types: serious lies (which means news is about wrong and unreal events or information like famous rumors), tricks (e.g. providing wrong information) and comics (e.g. funny news which is an imitation of real news but

Kasra Majbouri Yazdi and Jingyu Hou are with the School of Information Technology, Deakin University, 3125, Australia (e-mail: kmajbour@deakin.edu.au, jingyu.hou@deakin.edu.au).

Adel Majbouri Yazdi is with the Department of Computing, Kharazmi University, Tehran, Iran (e-mail: majbourii.adel@gmail.com).

Saeid Khodayi is with the Faculty of Computer & Electrical Engineering, Qazvin Azad University, Qazvin, Iran (e-mail: s.khodayi20@gmail.com).

Wanlei Zhou is with the School of Software, The University of Sydney, 2006, Australia (e-mail: wanlei.zhou@uts.edu.au).

Saeed Saedy is with the Faculty of Electrical Engineering, Shahid Beheshti University Iran (e-mail: s.saedy@hotmail.com).

contain bizarre contents) [9].

- In the second category, linguistic approaches and reality considerations techniques are used at a practical level to compare the real and fake contents [10].

Linguistic approaches try to detect text features like writing styles and contents that can help in distinguishing fake news. The main idea behind this technique is that linguistic behaviors like using marks, choosing various types of words or adding labels for parts of a lecture are rather unintentional, so they are beyond the author's attention. Therefore, an appropriate intuition and evaluation of using linguistic techniques can reveal hoping results in detecting fake news.

Rubin et al. [11] studied the distinction between the contents of real and comic news via multilingual features, based on a part of comparative news (The Onion, and The Beaverton) and real news (The Toronto Star and The New York Times) in four areas of civil, science, trade and ordinary news. She obtained the best performance of detecting fake news with a set of features including unrelated, marking and grammar.

Balmas et al. [12] believe that the cooperation of information technology specialists in reducing fake news is very important. In order to deal with fake news, using data mining as one of the techniques has attracted many researchers. In data mining based approaches, data integration is used in detecting fake news [13]. In the current business world, data are an ever-increasing valuable asset and it is necessary to protect sensitive information from unauthorized people. However, the prevalence of content publishers who are willing to use fake news leads to ignoring such endeavors. Organizations have invested a lot of resources to find effective solutions for dealing with clickbait effects. However, the employees who continue visiting such websites will endanger the companies with cyber-attacks [14].

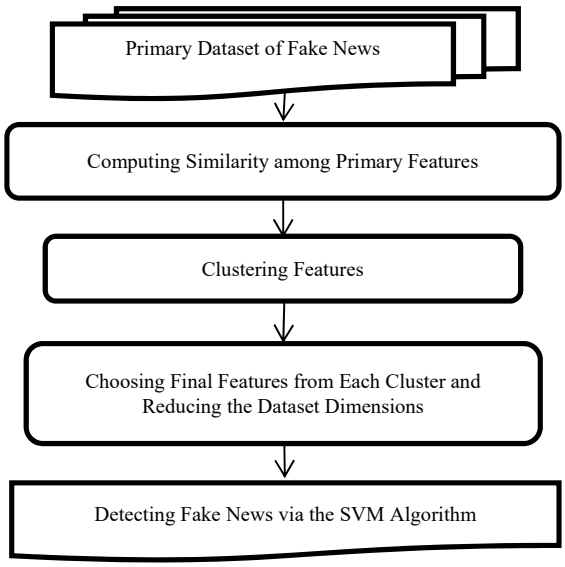

Fig. 1 Flowchart of the Proposed Method

III. Proposed Method

Feature selection is also known as attribute selection method searches among the available subsets of primary features and selects the appropriate ones to form the final selective subset. In this technique, the primary features are transferred into a new space with fewer dimensions. No new features are made but only several features are chosen and the irrelevant and redundant features are removed.

Our proposed method of choosing features and detecting fake news has four main steps. The first step is computing similarity between primary features in the fake news dataset. Then, features are clustered based on their similarities. Next, the final attributes of all clusters are selected to reduce the dataset dimensions. Finally, fake news is detected using the SVM classifier. Fig. 1 shows the flowchart of our method.

A. Computing Similarity among Features

As mentioned earlier, the similarity between attributes[1] needs to be calculated for clustering primary features. In that regard, we assume a weighted undirected graph $G = (F, E, w_F)$ where, $F = \{F_1, F_2, \ldots, F_n\}$ shows a set of n features each of which is represented as a node in the graph and $E = \{(F_i, F_j): F_i, F_j \in F\}$ shows the edges of the graph, $w_F: (F_i, F_j) \rightarrow \mathbb{R}$ is a function that shows the similarity (represented as weight) between two features of F_i and F_j. An appropriate criterion for determining the similarity between features can make a great impact on the algorithm's performance. There are various methods of computing similarity between features with different results, so choosing a good criterion is very important. In general, the most commonly used criteria in measuring similarity between features are Euclidean distance, Cosine similarity and also Pearson's correlation coefficient. In this paper, the absolute value of Pearson's correlation coefficient is used to compute the similarity between attributes. Pearson's correlation coefficient between two features F_i and F_j is calculated as follows:

$$W_{ij} = \frac{\sum_p (x_i - \overline{x_i})(x_j - \overline{x_j})}{\sqrt{\sum_p (x_i - \overline{x_i})^2} \sqrt{\sum_p (x_j - \overline{x_j})^2}} \quad (1)$$

where x_i and x_j are the vector elements of F_i and F_j features. Also, $\overline{x_i}$ and $\overline{x_j}$ are the mean of values for x_i and x_j vector elements respectively for p instances. According to (1), the similarity between two fully similar features is 1, but the similarity between two non-similar features is 0.

B. Clustering Features

The clustering features approach is about dividing attributes into several clusters based on their similarities. Therefore, features within a cluster have a higher similarity with each other and the features in different clusters have a lower similarity with each other. In this paper, we use the K-means algorithm for feature clustering. In this algorithm, the

[1] In text mining, typically each position in the input feature vector corresponds to a given word. This representation often called bag of words model [15].

data are classified into K different clusters after several iterations. However, its performance depends on primary conditions and convergence to optimal local points (centers). Also, data vectors that are in a D-dimension space are classified into a pre-specified number of clusters.

K-means start with K randomly selected points in the dataset (i.e. features) as the initial cluster centers. Then, other data entities join the nearest cluster centers to form new clusters with new centers. This process continues until each data entity (feature) is allocated to its closest cluster center. In each iteration, the centers of clusters are updated with their new entities and this continues until no more improvement happens.

In the initial set of k means $m_1^{(1)},...,m_k^{(1)}$, the algorithm proceeds by alternating between the two following steps:

Assignment step: Assigns each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean [16].

$$\mathrm{Si(t)} = \{\mathrm{xp}: ||\ \mathrm{xp} - \mathrm{mi(t)}\ || <: ||\ xp - mj(t)\ ||\ \sqrt{}\ j = 1\ \ldots\ k\} \quad (2)$$

where each xp (feature) is assigned to exactly one, even if it could be assigned to two or more of them.

Update step: Calculates the new means (centroids) of the observations in the new clusters

$$m_i(\mathrm{t}+1) = \frac{1}{|S_i(\mathrm{t})|} \sum_{x_j \varepsilon S_i(\mathrm{t})} x_j \quad (3)$$

The algorithm has converged when the assignments no longer change. It does not guarantee to find the optimum.

The algorithm is often presented as assigning objects to the nearest cluster by distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging [16].

C. Feature Selection

After clustering, the most suitable attributes of each cluster are selected to form the final subset. For this purpose, Fisher score (FS), [17] which is a supervised feature selection method, is used to rank attributes and create the final subset of features. In this technique, the distance between patterns of the same class is as minimum as possible and the distance between patterns[2] of different classes is as maximum as possible. In other words, this specifies the ratio between distributions of patterns among different classes and distributions of patterns within each class. Therefore, higher scores go to features that have a better splitter capability. FS is determined via (4):

$$FS(S,A) = \frac{\sum_{v \in Values\,(S)}\ n_v(\overline{A_v} - \bar{A})^2}{\sum_{v \in Values\,(S)}\ n_v(\sigma_v(A))^2} \quad (4)$$

where $\overline{A}$ is the mean value of the whole set of patterns corresponding to feature A. n_v is the number of patterns of classes with a label. v. $\sigma_v(A)$ and $\overline{A}_v$ are respectively the standard deviation and the mean value of patterns within class v according to feature A. After computing the FS for all features, the features with higher scores are selected to form the final subset.

Once the FS is computed for all features, m final feature with the highest scores is selected from each cluster. Then, after selecting the final features, the dataset dimensions are reduced to $k \times m$ (k is the number of clusters and m is the number of selected features from each cluster).

D. Detection of Fake News

After creating the final feature set and reducing the dataset dimensions, fake news can be detected by using a classifier. In this paper, we use the SVM classifier which is one of the supervised learning methods used for classification and regression. The goal in SVM is to separate fake news data with hyperplane and extend it to non-linear boundaries. The following equations are used in SVM to detect fake news:

$$If\ Y_i = +1\ ,\ wx_i + b \geq 1 \quad (5)$$

$$If\ Y_i = -1\ ,\ wx_i - b \leq 1 \quad (6)$$

$$For\ all\ i; \quad y_i\ (w_i\ +\ b)\ \geq\ 1 \quad (7)$$

In the above equations, x is the vector of fake news data, y is the class label of the news which can be either 1 or -1, and w is the weight vector. If the training data are suitable[3], then each vector of the test data is located in radius r of the training data vector. Now if the selected hyperplane is at the farthest possible distance from the data, then it maximizes the margin between points of classes.

The distance of the closest point to the main point on hyperplane can be found by maximizing the x on the hyperplane. Similarly, the same strategy is applied to all points on the other side. Therefore by subtracting the two distances (i.e. (5) and (6)), we obtain the distance from the hyperplane to the nearest point. So the maximum margin is $M\ =\ 2\ /\ ||w||$. At this stage, we have a quadratic optimization problem that needs to be solved for w and b. To resolve this, the quadratic function needs to be optimized with linear constraints. The solution includes creating a dual problem where a Langlier's multiplier of α_i is associated. We need to find w and b so that $\Phi\ (w)\ =\ ½\ |w'||w|$ is minimized.

According to (5) to (7), we have [18] $for\ all\ \{(x_i, y_i)\}: y_i\ (w\ *\ x_i\ +\ b)\ \geq\ 1$, we have:

$$w\ = \Sigma \alpha_i\ *\ x_i;\ b =\ y_k -\ w\ *x_k \text{ for any } x_k \text{ like } \alpha_k \neq 0 \quad (8)$$

where α_i is a Langlier's multiplier. Finally, the classifier function is as:

$$f(x)\ =\ \Sigma \alpha i\ y_i\ x_i *\ x + b \quad (9)$$

[2] The pattern is data. In this case, the pattern is news. A pattern has 2 parts; class and features, so news contents are the features and news type (e.g. real or fake) is the class.

[3] A suitable training data mean that it is not very different from the test (educational) data.

IV. Experimental Results

This section presents the evaluations of the proposed method on different datasets and discusses the comparison results with a feature extraction-based method [19]. At first, the used datasets and their features are introduced. Then, the used classifier approach is described and finally, the evaluation results are discussed.

A. Datasets

We used several datasets with various features to evaluate our proposed method:

Buzz Feed News: This dataset has a full sample of published news on Facebook from 9 well-known news agencies for one week close to the American Election in 2016, from 19 to 23 September and also 26 and 27 September. It includes 1627 papers 826 of which are related to the main political wing, 356 papers are for left-wing and 545 papers are for the right wing.

BS Detector: This dataset was gathered by a browser extension called BS which was made for studying the authenticity of the created news.

LIAR: This dataset was gathered by a website PolitiFact reality using its API. It includes 12836 brief statements with labels that were collected from different sources such as published news, TV and radio interviews, election speeches and, etc. These samples are classified as real, mostly real, semi-real and wrong classes.

B. Classifiers

We used three classifiers, SVM, Decision Tree (DT) and Naïve Bayes (NB), to evaluate the performance of the proposed method through applying different classifiers on the experimental datasets.

DT: It is a popular tool for classification and prediction. It is created based on the training data and each of its paths (from root to leaf) presents a rule for classification. Each node in this tree corresponds to a feature and each edge corresponds to an offspring and shows a possible value for that feature.

NB: is a learning approach for classifying data according to their occurrence possibility. This classifier is based on a simplified assumption so that features are considered conditionally and independent from each other based on the target class.

V. Results and Discussion

We did several simulations and experiments using various classifiers to evaluate the performance of the proposed method on different datasets. The dataset was divided into two parts of training and test data randomly so that 66% of the dataset is considered as the training data and the rest as the test data. Also, in all experiments, after specifying the training and test dataset, each method of feature selection was executed 10 times and the average of 10 executions was used to compare different methods. The precision of the classification was used as the criteria to compare the performance of different methods.

Tables I-III show the results of classification for SVM, DT, and NB classifiers. The values in the tables are the mean value of classification precision in 10 independent executions for the proposed method and feature extraction-based method [19].

TABLE I
Classification Results Using SVM Classifier

	Proposed method	Feature Extraction-based method [19]
BuzzFeedNews	95.34	89.76
BS Detector	93.89	90.78
LIAR	94.19	91.76

TABLE II
Classification Results Using DT Classifier

	Proposed method	Feature Extraction-based method [19]
Buzz Feed News	93.16	90.23
BS Detector	93.19	91.19
LIAR	92.58	91.43

TABLE II
Classification Results Using NB Classifier

	Proposed method	Feature Extraction-based method [19]
Buzz Feed News	92.28	91.52
BS Detector	91.57	90.06
LIAR	91.64	91.87

As results show, in almost all classifiers and datasets, the proposed method has better outcomes. For example, in SVM and DT, for all three datasets and also for NB classifier, Buzz Feed News and BS Detector datasets, the proposed method has better performance, and just in NB and on LIAR dataset, the performance is 0.23% less than the other method. Moreover, the performance results show that SVM Classifier achieved higher precisions compared with DT and NB classifiers.

VI. Conclusion

Over the last few years, the issue of fake news and its effects on society has attracted more and more attention. In the fake news detection issue, the problem of predicting and classifying data needs to be validated using training data. Since the majority of fake news datasets have many features that most of them are irrelevant and redundant, so reducing the number of those features could improve the precision of fake news detection algorithm. Therefore, a method of fake news detection via feature selection is proposed in this paper. In the feature selection phase, the primary features are divided into several clusters using the k-means clustering method based on the similarity between features. Then the final feature set is chosen from each cluster, based on the appropriateness of the features. Finally, after specifying the final set of features, the dimension-reduced dataset is created using the final set and in the next phase, the SVM classifier is used to predict the fake news. After implementing the proposed method, we evaluated the performance of the proposed method on different datasets. The simulation results

showed that the proposed method achieved better outcomes than the comparison method which used a feature extraction approach for detecting fake news.

REFERENCES

[1] Gravanis, G., et al., Behind the cues: A benchmarking study for fake news detection. Expert Systems with Applications, 2019. 128: p. 201-213.
[2] Zhang, C., et al., Detecting fake news for reducing misinformation risks using analytics approaches. European Journal of Operational Research, 2019.
[3] Bondielli, A. and F. Marcelloni, A survey on fake news and rumour detection techniques. Information Sciences, 2019. 497: p. 38-55.
[4] Ko, H., et al., Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system. Cognitive Systems Research, 2019. 55: p. 77-81.
[5] Zhang, X. and A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 2019.
[6] Robbins, K.R., W. Zhang, and J.K. Bertrand, The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. Journal of Mathematical Medicine and Biology, 2008: p. 1-14.
[7] Alirezaei, M., S.T.A. Niaki, and S.A.A. Niaki, A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Systems with Applications, 2019. 127: p. 47-57.
[8] Zakeri, A. and A. Hokmabadi, Efficient feature selection method using real-valued grasshopper optimization algorithm. Expert Systems with Applications, 2019. 119: p. 61-72.
[9] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. News in an online world: The need for an “automatic crap detector”. Proceedings of the Association for Information Science and Technology, 52(1):1–4.
[10] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4.
[11] Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of NAACL-HLT, pages 7–17.
[12] Balmas, M., 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. Communication Research 41, 430–454.
[13] Pogue, D., 2017. How to stamp out fake news. Scientific American 316, 24–24.
[14] Aldwairi, M. and A. Alwahedi, Detecting Fake News in Social Media Networks. Procedia Computer Science, 2018. 141: p. 215-222.
[15] Mehdi H.A, Nasser G.A, Mohammad B, Text feature selection using ant colony optimization, Expert Systems with Applications, 2009
[16] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.
[17] Quanquan Gu, Zhenhui Li, and J. Han, Generalized Fisher Score for Feature Selection. In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2011
[18] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX
[19] Reis, J.C., Correia, A., Murai, F., Veloso, A., Benevenuto, F. and Cambria, E., 2019. Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), pp.76-81.