D-FEND: A Diffusion-Based Fake News Detection Framework for News Articles Related to COVID-19

Soeun Han
Hanyang University
Seoul, Republic of Korea
sosilver@hanyang.ac.kr

Yunyong Ko
Hanyang University
Seoul, Republic of Korea
koyunyong@hanyang.ac.kr

Yushim Kim
Arizona State University
Arizona, United States
ykim@asu.edu

Seong Soo Oh
Hanyang University
Seoul, Republic of Korea
ohseongsoo@hanyang.ac.kr

Heejin Park
Hanyang University
Seoul, Republic of Korea
hjpark@hanyang.ac.kr

Sang-Wook Kim*
Hanyang University
Seoul, Republic of Korea
wook@hanyang.ac.kr

ABSTRACT

The social confusion caused by the recent pandemic of COVID-19 has been further facilitated by fake news diffused via social media on the Internet. For this reason, many studies have been proposed to detect fake news as early as possible. The content-based detection methods consider the difference between the contents of true and fake news articles. However, they suffer from the two serious limitations: (1) the publisher can manipulate the content of a news article easily, and (2) the content depends upon the language, with which the article is written. To overcome these limitations, the *diffusion-based* fake news detection methods have been proposed. The diffusion-based methods consider the difference among the diffusion patterns of true and fake news articles on social media. Despite its success, however, the lack of the diffusion information regarding to the COVID-19 related fake news prevents from studying the diffusion-based fake news detection methods. Therefore, for overcoming the limitation, we propose a diffusion-based fake news detection framework (**D-FEND**), which consists of four components: (C1) diffusion data collection, (C2) analysis of the data and feature extraction, (C3) model training, and (C4) inference. Our work contributes to the effort to mitigate the risk of infodemics during a pandemic by (1) building a new diffusion dataset, named CoAID+, (2) identifying and addressing the class imbalance problem of CoAID+, and (3) demonstrating that D-FEND successfully detects fake news articles with 88.89% model accuracy on average.

KEYWORDS

fake news detection, diffusion-based detection, COVID-19 dataset

*Corresponding author

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507134

Figure 1: Social confusion caused by fake news articles.

ACM Reference Format:
Soeun Han, Yunyong Ko, Yushim Kim, Seong Soo Oh, Heejin Park, and Sang-Wook Kim. 2022. D-FEND: A Diffusion-Based Fake News Detection Framework, for News Articles Related to COVID-19. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event,* . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3477314.3507134

1 INTRODUCTION

During COVID-19, countries around the globe have experienced the serious problem of infodemics, which refers to "too much information including false or misleading information in digital and physical environments during a disease outbreak". The expansion of social media and Internet use have facilitated the influence of false or misinformation, and caused confusion about the disease and the public health response, as well as mistrust in health authorities and governments [34]. Mitigating the risk of such fake news has become a priority to health authorities and researchers. Among different types of false or misinformation, fake news – a news article that is created *intentionally* with false information that can be verified, and distributed for a malicious purpose – has gained the particular interest of researchers and practitioners because of its devastating consequences [32]. For example, in Iran, the fake news that methanol has a positive effect in treating COVID-19 has led to more than 800 casualties (see Figure 1) [1].

Previous studies have proposed methods to detect fake news efficiently, particularly by analyzing the contents of news articles [2, 3, 25, 26] (i.e., *content-based* detection method). However, fake news has also become more sophisticated and nuanced by imitating true news. As a result, it has become more difficult to detect fake news by using only the content information. Two limitations of detecting fake news using the content are: (1) the publisher can manipulate it

[1] https://www.bbc.com/news/world-53755067

easily [23, 31], and (2) it depends upon the language with which the article is written [23]. To overcome these limitations, the *diffusion-based* fake news detection methods have been widely studied [17, 21, 27, 31, 33, 34, 38]. This approach analyzes the difference between true news and fake news with respect to their diffusion patterns on social media (e.g., Twitter), and uses such a difference to detect fake news [17, 31, 38]. Unlike the content information, the diffusion of a news article is (1) not easy to be manipulated by publishers and (2) independent of the language with which the news is written. Thus, the diffusion information of a news article has the potential for effectively detecting fake news articles.

Unfortunately, the deficiency of the diffusion information about COVID-19 related fake news articles becomes a hurdle to activate the study on the diffusion-based fake news detection. For addressing the problem, in this work, we propose a comprehensive framework, named as **D**iffusion-based **F**ak**E** **N**ews **D**etection (**D-FEND**), based on the diffusion information of news articles on social media. D-FEND includes the entire process required to detect fake news: (C1) diffusion data collection; (C2) analysis of the data and feature extraction; (C3) model training; and (C4) inference. In (C1), D-FEND collects the information on the way each news article is diffuses on social media. Note that we collect the diffusion information based upon the existing well-recognized dataset, CoAID [9], and build a new diffusion dataset, CoAID+. In (C2), D-FEND analyzes the collected data and extracts 18 useful features (9 structural and 9 temporal features). In (C3), D-FEND trains the four popular machine learning models -– decision tree (DT), random forest (FR), support vector machine (SVM), and deep neural network (DNN) – based on the extracted features. Finally, in (C4), when an unknown news article is given as an input, D-FEND predicts whether it is true or fake by using the models trained in (C3).

In addition, we identify that CoAID+ has a serious *class imbalance* problem, which may result in significantly degrading the model accuracy. To tackle this problem, we adopt a state-of-the-art over-sampling method, the synthetic minority over-sampling technique (SMOTE) [7], and empirically verify the effectiveness of SMOTE on fake news detection. Via extensive experiments using the four machine learning models on the CoAID+ dataset, we demonstrate that D-FEND successfully detects fake news articles with 88.89 % model accuracy on average.

The primary contributions of this work are as follows:

- Constructing a new diffusion dataset, named CoAID+, and providing CoAID+ publicly to vitalize the study on diffusion-based fake news detection.
- Proposing a comprehensive framework for effectively detecting fake news related to COVID-19, named D-FEND based on the diffusion information of news articles.
- Extensive evaluation validating the effectiveness of D-FEND in fake news detection, successfully detecting fake news articles with 88.89% accuracy on average.

2 RELATED WORK

2.1 Fake News Detection

Here, we present the definition of fake news and well-known fake news detection methods. Let a refer to a news article, which is expressed as an embedding vector $a \in \mathbb{R}^d$, where d is the number of features of a news article. Then, formally, the problem of fake news detection is defined as follows.

Problem definition. *Given a news article $a \in \mathbb{R}^d$, the goal of fake news detection is to predict whether the news article a is true or fake, i.e., $F(a) \rightarrow \{0, 1\}$ such that,*

$$F(a) = \begin{cases} 1, & \text{if } a \text{ is a fake news article,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Here, $F(\cdot)$ is the prediction function that we aim to train. The existing fake news detection methods are classified into *content-based* and *social context-based* methods according to whether the content information or the social context information of news articles is used [23, 31, 32, 34, 36].

Content-based detection. Content-based detection is a method to detect fake news by using the difference in linguistic characteristics that appear in the content of true news and fake news. For example, in [3], the syntactic and semantic differences between true news and fake news content are used for fake news detection, and in [2, 25, 26] the writing style and frequency of words written in news articles are considered. However, because these linguistic features are relatively easy to manipulate, fake news has been created recently by imitating true news very closely, which makes it increasingly difficult to detect fake news using only content information.

Social context-based detection. Social context-based detection identifies fake news using information from users who consume news on social media and various user engagement information. On social media, user profile information, user relationship network information, and interaction information between users, such as 'like', 'retweet', or 'share' can be used. Such social context information is difficult to manipulate artificially, so it detects fake news more effectively [23]. In [4, 37], the authors utilize the differences in the individual characteristics of users who consume true news and fake news on social media, and the users' relationship network information. On the other hand, in [15, 17, 31, 37, 38] user behavior information on social media is used for fake news detection. Jin et al. [15] and Tacchini et al. [37] use the users' opinions and stances that appear in social media posts. Kucharski [17], Shu et al. [31], and Vosoughi et al. [38] detect fake news articles using the information that each news item spreads through users, based upon the intuition that true news and fake news have different diffusion patterns on social media.

Table 1: Statistics of existing fake news datasets.

Dataset	# of News Articles	Social context	
		# of Tweets	# of Actions
FakeCovid [29]	5,182	-	-
ReCOVery [41]	2,029	140K	-
CoAID [9]	3,921	150K	135K
BuzzFeedNews [1]	1,627	-	-
LIAR [40]	12,836	-	-
CREDBANK [22]	-	6,000K	-
FakeNewsNet [30]	23,196	694K	1,821K

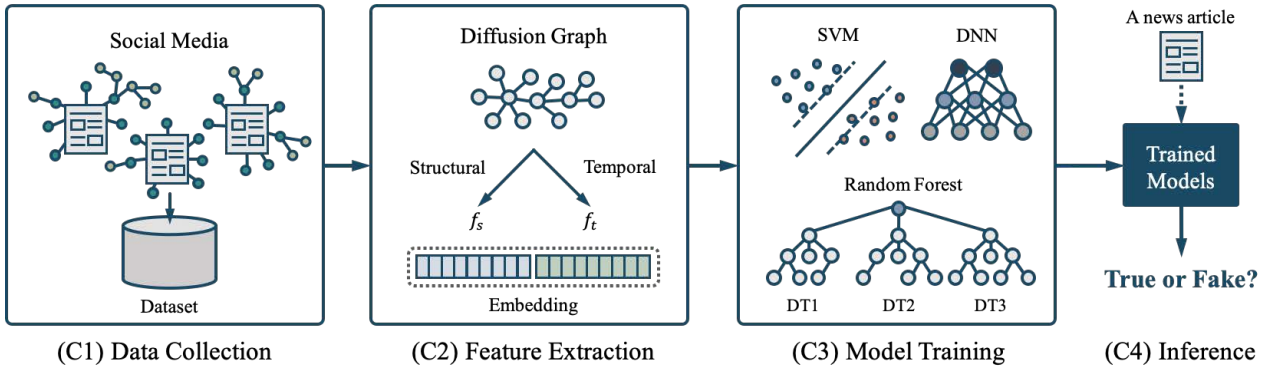


Figure 2: The overview of our proposed framework (D-FEND).

2.2 Datasets

In this section, we introduce various public datasets for the fake news detection. Table 1 shows brief statistics of existing datasets.

COVID-19 related datasets. Novel COVID-19 related datasets have been built to tackle the infodemic problem by researchers and practitioners [9, 29, 41]. FakeCovid [29] is a new COVID-19 related dataset that consists of 5,182 news articles that are verified as true or fake via 92 fact-checking sites. ReCOVery [41] includes the various content information of each news article such as textual, visual (e.g., images on the article), and temporal information, and the social context information for the news articles. CoAID [9], built by the PIKE research group in Pennsylvania State University in 2020, consists of 3,921 labeled news articles (i.e., verified as true or fake) and 150,002 tweets posting one of the news articles on Twitter. The news articles and tweets have been collected from 9 famous media outlets from December 2019 to July 2020. Unlike most other existing COVID-19 related datasets, CoAID includes not only the content information of news articles but also the social context information (e.g., the information of the tweets posting news articles on Twitter). By exploiting the strength of this dataset, it is possible to further extend the dataset to include the *diffusion* of each news article (i.e., how each news article is diffused on social media such as Twitter). Therefore, the CoAID dataset has been widely adopted in various fake news detection studies [10, 12, 35]. We also use the CoAID dataset as a seed dataset to additionally collect the diffusion information of news articles (see in Section 3.1).

Other fake news datasets. BuzzFeedNews [1] consists of the content information of 1,627 news articles published in 9 famous media outlets during the 2016 United States presidential election. LIAR [40] collected 12.8K labeled short statements (i.e., identified as fake or real news) from PolitiFact [2], a fact-checking website, and classified the collected articles into 6 classes (pants-fire, false, barely true, half-true, mostly-true, and true), depending on how much misinformation is included in the article. CREDBANK [22] collected news articles for a thousand cases that happened for 96 days from October 2015, and their related 60 million tweets. FakeNewsNet [30] includes both the content and social context information of news articles collected from two famous fact-checking sites, GossipCop [3] and PolitiFact.

[2] https://www.politifact.com
[3] http://www.gossipcop.com/

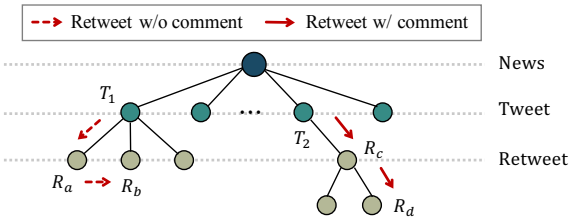


Figure 3: Difference between retweets with/without comment in the perspective of diffusion.

3 THE PROPOSED FRAMEWORK

In this section, we describe an unified framework, named as **D**iffusion-based **FakE** **N**ews **D**etection (**D-FEND**), based on the diffusion information of news articles on social media. As illustrated in Figure 2, D-FEND consists of four components: (C1) Diffusion data collection, (C2) feature extraction, (C3) model training, and (C4) inference.

3.1 Diffusion data collection

As explained in Section 2, we collect the diffusion information of news articles based on the CoAID dataset [9]. Among the 3,921 news articles in the CoAID dataset, only 2,763 articles were diffused on social media. We collect only the news articles diffused through social media users (i.e., news articles having diffusion information), thus excluding the news articles without 'retweet' from our data collection. Consequently, we use only 1,096 news articles as seed nodes for diffusion, where there are 1,017 true news and 79 fake news articles.

In addition, we consider only 'retweet with comment' for our data collection since 'retweet without comment' does not reflects the diffusion process of each news article. A retweet on Twitter can be classified into the 'retweet without comment' or 'retweet with comment'. The 'retweet w/o comment' represents a tweet that simply retweets another tweet without comment. While, the 'retweet with comment' represents a tweet that retweets another tweet with its own opinion together. Figure 3 shows the different diffusion pattern of the two types of retweets. First, let us show the example for 'retweet w/o comment' (see the left side of Figure 3). Assume that user a retweets tweet T_1 without comment (i.e., R_a), and then user b retweets user a's (re)tweet R_a without comment (i.e., R_b) as well (i.e., $T_1 \rightarrow R_a \rightarrow R_b$ in order). Although user b

Algorithm 1 Collecting diffusion information in D-FEND

Require: News article dataset $D = \{a_1, a_2, ..., a_n\}$, $G \in \mathbb{R}^n$
1: Initialize $G \leftarrow \emptyset$
2: **for each** news article $a_i \in D$ **do**
3: $T_i \leftarrow$ get_tweets(a_i)
4: **for** each tweet $t \in T_i$ **do**
5: $RT \leftarrow$ get_retweets_with_comment(t)
6: $G[i]$.append(RT)
7: **end for**
8: **end for**
9: **Return** G

retweets user a's tweet R_a , it appears as if user b retweets the original tweet T_1, which implies that we cannot obtain the diffusion process of the news articles. Next, in the case of a retweet with comment (see the right side of Figure 3), when user c retweets tweet T_2 with a comment (i.e., R_c), and then user d retweets user c's (re)tweet R_c with a comment (i.e., $T_2 \rightarrow R_c \rightarrow R_d$ in order), we can identify the order of the diffusion process across users. As such, we collect only the retweets having comments to exploit the diffusion information of each news article on social media. The process of collecting diffusion information is described in Algorithm 1.

In this way, we construct a new dataset, named as CoAID+ [4] that includes the diffusion information of each news article on social media. In CoAID+, the diffusion process of each news article is represented by a graph, which consists of nodes and edges. In the diffusion graph, there are three types of nodes (news, tweet, and retweet) and each edge means whether the relationship between the two nodes. For instance, the edge <T_1, R_a> means that user a retweets tweet T_1 (i.e., generating retweet R_a). The diffusion graphs for news articles in CoAID+ are used to extract useful features for fake news detection in the next component. Table 2 shows the descriptive statistics of CoAID+.

Table 2: Descriptive statistics of CoAID+

Feature Name	Fake	True
# of news	157	2,606
# of tweets	9,745	140,257
# of retweets	3,528	45,287
# of nodes	85.51	70.69
Max. depth	1.80	1.57

3.2 Feature extraction

We extract useful features from the collected dataset and analyze the difference between true news and fake news. We extract two types of diffusion features of news articles: structural and temporal features. A *structural feature* indicates a static property of a news article such as the shape and size of the diffusion graph. A *temporal feature* represents a dynamic property of a news article such as how quickly the news article is diffused in a given time. Among the various features of existing excellent works [31, 38], we select 9 structural features and 9 temporal features in total.

Structural features. First, let us explain the structural features of the diffusion graph (a graph per news article).

- (S1) Maximum depth: it means that how far the corresponding news article is diffused on social media.
- (S2) Number of nodes: it represents the total number of nodes (i.e., a news article, tweets, and retweets)
- (S3) Maximum width at a certain hop: it means that how widely the corresponding news article is diffused on social media.
- (S4) Average distance of all node pairs: it means that how close a pair of nodes are to each other.
- (S5) Maximum out-degree: it indicates the number of out-neighbor nodes of the most influential node.
- (S6) Number of tweets that first posted the news article: it represents the number of the "tweet" nodes, thus excluding retweet nodes.
- (S7) Depth from the news article to the influential posting: it represents the distance – i.e., the number of hops – from the news node (i.e., seed) to the most influential node.
- (S8) Number of tweets with retweets: it represents the number of "tweet" nodes that have the out-neighbors (i.e., retweet nodes).
- (S9) Fraction of tweets with retweets: it represents the ratio of the tweet nodes having their out-neighbors (i.e., retweet nodes) to the tweet nodes without their out-neighbors.

Temporal features. Next, let us explain the temporal features.

- (T1) Average time difference between the adjacent retweet nodes: it indicates how quickly the (mis)information is diffused through users on social media.
- (T2) Time difference between the first tweet and the last retweets: it indicates the total amount of time for the diffusion of the corresponding news article (i.e., the lifespan of the diffusion).
- (T3) Time difference between the first tweet and the tweet with maximum out-degree: it represents the required time to reach the most influential user that is highly likely to spread the news article further.
- (T4) Time difference between the tweet and its last retweet: it indicates the lifespan of the diffusion from the corresponding tweet node.
- (T5) Average time difference between the adjacent retweets in the deepest path: it indicates how quickly the (mis)information is exchanged among end users.
- (T6) Time difference between the first and last "tweets" posting the news article: it indicates how long period of time users on social media newly start to spread the news article (i.e., generate a new "tweet" – branch of the diffusion graph).
- (T7) Average time among tweets posting the news article: it indicates how frequently the news article is tweeted by users on social media.
- (T8) Time difference between the first tweet and its first retweet: it represents how quickly the news article spreads in the early stage of the diffusion after its first tweet is generated.
- (T9) Average time difference between tweets and their first retweet: it represents how quickly the news article spreads in the early stage of the diffusion after tweets posting the corresponding news article are generated.

Prior to training models with the extracted features, we compare the extracted features of true news and fake news statistically. Table 2 shows the statistics of 18 features for true news and fake news. The values for structural features of fake news articles are larger

[4] https://anonymous.4open.science/r/CoAID-plus/

Table 3: The extracted structural and temporal features of news articles in CoAID+.

Structural Features					Temporal Features (sec.)				
Feature #	Fake		True		Feature #	Fake		True	
	Mean	Median	Mean	Median		Mean	Median	Mean	Median
S1	2.59	2.00	2.47	2.00	T1	25,427	0	63,877	0
S2	164.53	63.00	159.23	53.00	T2	3,274,014	888,347	3,271,199	1,026,472
S3	127.77	58.00	126.21	45.00	T3	450,698	102,811	907,908	133,686
S4	2.25	2.18	2.20	2.12	T4	2,714,204	1,728,996	2,729,019	2,050,785
S5	18.33	3.00	19.44	2.00	T5	534,308	23,223	825,426	22,469
S6	118.94	48.00	117.57	43.00	T6	36,442	0	51,541	0
S7	1.04	1.00	1.04	1.00	T7	88,948	38,605	106,606	43,454
S8	10.28	3.00	8.35	2.00	T8	258,692	31,962	560,844	66,466
S9	0.17	0.10	0.10	0.06	T9	187,146	9,334	119,822	9,550

than those of true news articles only except for S5, indicating that *fake news articles tend to spread more widely* than true news articles. While, the values for temporal features of fake news articles are smaller than those of true news articles, which implies that *fake news articles are diffused across users faster and have shorter lifespan* than than true news articles. Thus, via the statistical analysis, we observed that there is a meaningful (statistical) difference among fake and true news articles in terms of the diffusion process. This result implies that *these diffusion features are highly likely to be effective on detecting fake news.* We will verify the effectiveness of the structural and temporal features in fake news detection in Section 4.3.

3.3 Model training and Inference

Now, we train four popular machine learning models using the CoAID+ dataset. First, we represent each news article as an embedding vector based on the extracted features (i.e., 9 structural and 9 temporal features). Formally, an embedding vector for a news article a, $\lambda_a \in \mathbb{R}^{m+n}$ is represented by

$$\lambda_a = \{s_1, s_2, ..., s_m, t_1, t_2, ..., t_n\}, \quad (2)$$

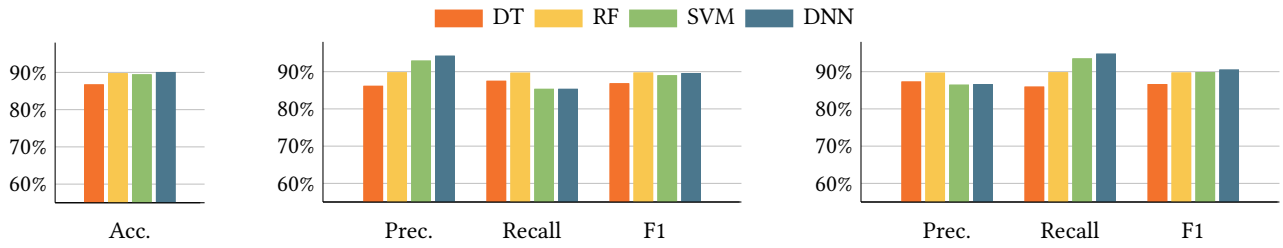
where, s_i is the i^{th} structural feature, t_i is the i^{th} temporal feature, m is the number of structural features, and n is the numbers of temporal features. (i.e., $m = 9$ and $n = 9$ in our setting).

Model training. We use the following four machine learning models in our framework. Let us describe each of the machine learning models briefly.

- Decision tree (DT): a non-parametric model, where each of the extracted features is used to divide a branch of the decision tree. Note that we apply the branch pruning technique to alleviate the over-fitting problem.
- Random forest (RF): an ensemble model, a collection of multiple decision trees, where the classification result is decided by a majority class from the multiple decision trees. This model is more robust to the over-fitting problem than a single decision tree model.
- Support vector machine (SVM): a non-parametric model that aims to find the optimal decision boundary to classify data samples with two classes. We use the radial basis function (rbf or Gaussian) kernel with varying the hyperparameter γ and C, where γ controls the range of each data sample and C controls how much the model allows outliers. We will verify the impact of each hyperparameter on the model accuracy and provide the best values for the hyperparameters in Section 4.4.
- Deep neural network (DNN): a parametric model that consists of multiple layers of perceptrons, where each layer has a non-linear activation function. The non-linearity by the activation function of each layer helps to classify more complex data samples [19]. We will also verify the impact of the number of layers and the size of a model on the model accuracy, and provide the best combination that maximizes the model accuracy in Section 4.4.

Model evaluation and inference. To evaluate the trained models, we use the widely used cross-validation method, Leave-One-Out Cross-Validation (LOOCV) [5, 16]. The process of LOOCV is as follows. It selects one of the n data samples as the test sample for validating models, and uses the remaining n-1 samples for training models. Then, the selected test sample will be used to evaluate the model trained with the remaining n-1 samples. This process is repeated by n times for all n data samples, and the n results are averaged. Finally, when an unknown news article is given as an input, the inference part of D-FEND predicts whether the new article is true or fake by using the trained model.

Data preprocessing for the class imbalance problem. As shown in Table 2, the CoAID+ dataset has the problem of a significant class imbalance, where the ratio of true and fake news articles is 93:7. The class imbalance can cause a serious problem that the trained model is over-fitted to the majority class (i.e., true news articles), which negatively affects the prediction of the model on new data samples [8, 14]. In other words, the model trained on the class-imbalanced dataset is highly likely to predict the class of a given new sample as only the majority class (e.g., true news), indicating that the model is not able to detect fake news articles successfully. For addressing this problem, over-sampling methods have been widely adopted [11, 14, 28, 39]. Over-sampling methods sample the training data with the minority class more than the data with the majority class for balancing the ratio of the numbers of samples that belong to the majority class and the minority class. Among existing over-sampling methods, we selected Synthetic Minority Oversampling TEchnique (SMOTE) [7] as the solution to the class-imbalance problem. SMOTE generates new virtual data points highly likely to follow the data distribution of the minority



(a) Overall accuracy **(b) Three accuracy metrics for true news articles** **(c) Three accuracy metrics for fake news articles**

Figure 4: Comparison of the model quality for four different accuracy metrics.

class and samples the generated data points. We will show the effect of SMOTE on the model accuracy of D-FEND in Section 4.2.

4 EXPERIMENTAL VALIDATION

In this section, we evaluate D-FEND by answering the following evaluation questions:

- EQ1. How accurately does D-FEND detect fake news articles?
- EQ2. Which type of features (structural/temporal) is more effective in fake news detection?
- EQ3. How sensitive are the accuracies of SVM and DNN models in D-FEND to their hyperparameters?

4.1 Set-Up

Datasets and models. As explained in Section 3.3, we use the following four well-recognized decision tree (DT), random forest (RF), support vector machine (SVM), and deep neural network (DNN) models. For DT, we use the GINI index for impurity and set the max depth of the tree as 3 as recommended in *scikit-learn* [24]. For RF, we set the number of decision trees as 100. For SVM, we use the 'rbf' kernel, which is the most widely used kernel [13, 18, 20]. We empirically found the best values for the hyperparameters γ and C, and set γ as 1.0 and C as 10. For DNN, we use the rectified linear unit (RELU) as the activation function, cross-entropy as the loss function, and momentum SGD as the optimizer. We set the learning rate as 0.1 and set the momentum as 0.9. As the training dataset, we use CoAID+, which consists of 1,096 news articles with two labels (i.e., true/false), and their 150,002 tweets. For validating the models, we use the leave-one-out cross-validation (LOOVC) method [6].

Metrics. The goal of this work is to accurately detect fake news articles. Thus, we evaluate the trained models in D-FEND with four accuracy metrics: Accuracy, precision, recall, and F1-score.

- *Accuracy*: the ratio of correctly predicted samples to the total samples, indicating the overall prediction accuracy of the model.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}. \tag{3}$$

- *Precision*: the ratio of correctly predicted positive (negative) samples to the total predicted positive (negative) samples, indicating how many predicted samples are correct.

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

- *Recall*: the ratio of correctly predicted positive (negative) samples to the all samples in the positive (negative) class, indicating how many positive (negative) samples are predicted correctly.

$$Recall = \frac{TP}{TP + FN}. \tag{5}$$

- *F1-score*: the harmonic mean of *Precision* and *Recall*.

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{6}$$

Here, TP and TN denote the number of true positive and negative samples (thus, correctly classified), while FN and FP denote the number of false positive and negative samples (thus, in-correctly classified).

System configuration. We use *Scikit-learn* library 0.24.2 to implement D-FEND on Windows 10 OS. We run our experiments on the machine, which has an Intel i7-9700k CPU with 32 GB memory.

4.2 EQ1. Model accuracy

In this experiment, we evaluate the model accuracies of all models in D-FEND in terms of detecting fake news articles. We train the four models using CoAID+, and measure their model accuracies with the four metrics. Note that we use the LOOCV method to validate the trained models, as we explained in Section 3.3.

Table 4 and Figure 4 show the results, where the x-axis represents the accuracy metrics and the y-axis represents the value of the metric. D-FEND achieves 88.89% of model accuracy on average in fake news detection, indicating that the diffusion information of news articles is quite effective in detecting fake news articles. In particular the DNN model achieves the highest accuracy, compared to the other machine learning models. This is because the non-linearity, added by passing through the activation functions in the DNN model, helps to classify more complex data samples. The RF model *outperforms* the DT model about 3.0%, where this improvement comes from effectively addressing the over-fitting problem by the *ensemble* method. Overall, D-FEND achieves good results in all metrics including precision and recall not only for the majority class (i.e., true news) but also for the minority class (i.e., fake news). This result indicates that our selected solution to the class imbalance problem (i.e., SMOTE) successfully addresses the problem, thereby preventing the problem of model over-fitting.

4.3 EQ2. Effectiveness of features

As explained in Section 3.2, we use structural and temporal diffusion features for detecting fake news articles. For more in-depth analysis of the diffusion features, in this experiment, we verify the effectiveness of each type of features (i.e., structural and temporal features) on fake news detection. We compare the three versions of

Table 4: Comparison of the model quality for various metrics.

Model	Class	Acc.	Prec.	Recall	F1-score
DT	True	0.8663	0.8606	0.8741	0.8673
	Fake		0.8721	0.8584	0.8652
RF	True	0.8963	0.8967	**0.8958**	**0.8962**
	Fake		**0.8959**	0.8968	0.8963
SVM	True	0.8933	0.9283	0.8525	0.8888
	Fake		0.8636	0.9341	0.8975
DNN	True	**0.8997**	**0.9414**	0.8525	0.8947
	Fake		0.8652	**0.9469**	**0.9042**

D-FEND: in D-FEND^{+S}, the models are trained with only 9 structural features, in D-FEND^{+T}, the models are trained with only 9 temporal features, and in D-FEND, the models are trained with all 18 structural and temporal features.

Figure 5 shows results, where the x-axis represents the accuracy metrics and the y-axis represents the value of the metric. First, both types of features (i.e., structural and temporal) are quite effective on fake news detection. D-FEND^{+S} and D-FEND^{+T} achieve 84.22% and 82.26% of model accuracies on average, respectively. These results imply that the diffusion of true and fake news on social media differs in terms of structure and temporal aspects. In addition, D-FEND, trained with both types of features, achieves the accuracy comparable to or higher than the other two versions, trained with one of the types of features (up to 11.45% higher accuracy). This result indicates that (1) the structural and temporal features do not interfere with each other and (2) for effectively detecting fake news articles, it is critical to consider both static and dynamic properties of the diffusion of news articles.

4.4 EQ3. Hyperparameter sensitivity

In this experiment, we evaluate the hyperparameter sensitivity of the SVM and DNN models in D-FEND and provide the best hyperparameter values, effectively detecting fake news articles. First, as explained in Section 3.3, for the SVM model, we evaluate γ, which the range of each data sample and C, which controls how much the model allows outliers. In other words, as γ gets smaller, the wider range of each data sample is considered, and as C gets larger, the more outliers are allowed. We measure the SVM model accuracy with varying $\gamma = 10, 1, 0.1$ and $C = 10, 1, 0.1$. Also, in the case of the DNN model, as the numbers of layers and parameters increase, the capacity of the model becomes larger. Thus, we measure the accuracy of the DNN model with varying the number of layers (3 and 5) and the size (small, medium, and large).

Table 5 shows the results. The accuracy of the SVM model tends to vary, depending on the hyperparemters γ and C, which indicates that the accuracy of the SVM model is sensitive to γ and C and it is important to set the hyperparameters carefully. Based on the results, we recommend to set γ as 1 and C as 10. While, the DNN model always achieves higher than 89% of accuracy for all combinations, and the accuracy tends to be consistent across different sizes of the models. As a result, the DNN model in D-FEND not only effectively detects fake news articles, but also rarely requires much trial-and-error tuning to find the best value for the hyperparameter.

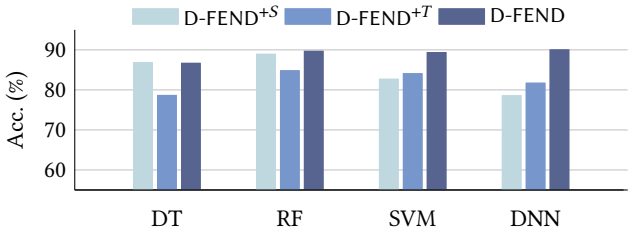


Figure 5: Effectiveness of the structural and temporal features on the accuracy of D-FEND.

5 CONCLUSIONS

In this work, we proposed a comprehensive diffusion-based fake news detection framework, named as D-FEND that consists of four components: (C1) diffusion data collection, (C2) analysis of the data and feature extraction, (C3) model training, and (C4) inference. With D-FEND, we built a new diffusion dataset CoAID+ and provide it publicly available to contribute to vitalizing the studies on diffusion-based fake news detection. In addition, we identified that our CoAID+ has a serious class imbalance problem, which may result in significantly degrading the model accuracy. To tackle this problem, we adopted a state-of-the-art over-sampling method, the synthetic minority over-sampling technique (SMOTE), and empirically verified the effectiveness of SMOTE on fake news detection. Via extensive experiments using four machine learning models on the CoAID+ dataset, we demonstrated that D-FEND successfully detects fake news articles with 88.89% model accuracy on average. In future work, we plan to extend D-FEND by adding more state-of-the-art fake news detection models and collecting larger size of diffusion datasets than CoAID+.

ACKNOWLEDGMENTS

The work was supported by the National Research Foundation of Korea (NRF) under Project Number 2020R1A2B5B03001960 and 2018R1A5A7059549, and Institute of Information Communications Technology Planning Evaluation (IITP) under Project Number 2020-0-01373.

REFERENCES

[1] [n. d.]. BuzzfeedNews Dataset. https://github.com/BuzzFeedNews/everything.
[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Proceedings of the International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer, 127–138.
[3] Sameer Badaskar, Sachin Agarwal, and Shilpa Arora. 2008. Identifying real or fake articles: Towards better language modeling. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.
[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.
[5] Gavin C Cawley. 2006. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the 2006 IEEE international joint conference on neural network proceedings*. IEEE, 1661–1668.
[6] Gavin C Cawley and Nicola LC Talbot. 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition* 36, 11 (2003), 2585–2592.
[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[8] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter* 6, 1 (2004), 1–6.

Table 5: Hyperparameter sensitivity of SVM and DNN models in D-FEND.

SVM		$\gamma = 0.1$				$\gamma = 1$				$\gamma = 10$			
		Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
$C = 0.1$	True	0.7616	0.7486	0.7876	0.7676	0.6514	0.7810	0.4208	0.5470	0.5688	0.9321	0.1485	0.2561
	Fake		0.7759	0.7355	0.7552		0.6036	0.8820	0.7167		0.5374	0.9892	0.6964
$C = 1$	True	0.8063	0.8093	0.8014	0.8053	0.8746	0.9079	0.8338	0.8693	0.8240	0.7873	0.8879	0.8346
	Fake		0.8033	0.8112	0.8072		0.8464	0.9154	0.8795		0.8715	0.7601	0.8120
$C = 10$	True	0.8402	0.8604	0.8122	0.8356	**0.8933**	0.9283	0.8525	0.8888	0.8225	0.7847	0.8889	0.8336
	Fake		0.8222	0.8682	0.8446		0.8636	0.9341	0.8975		0.8719	0.7561	0.8099
DNN		Small				Medium				Large			
		Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
L3	True	0.8948	0.9267	0.8574	0.8907	0.8948	0.9222	0.8623	0.8913	**0.8977**	0.9354	0.8545	0.8931
	Fake		0.8673	0.9322	0.8986		0.8707	0.9272	0.8981		0.8661	0.941	0.902
L5	True	0.8904	0.9214	0.8535	0.8862	**0.8997**	0.9414	0.8525	0.8947	0.8958	0.9333	0.8525	0.8911
	Fake		0.8636	0.9272	0.8943		0.8652	0.9469	0.9042		0.8643	0.939	0.9001

[9] Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885* (2020).

[10] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. An ensemble deep learning technique to detect COVID-19 misleading information. In *Proceedings of the International Conference on Network-Based Information Systems*. Springer, 163–175.

[11] Adel Ghazikhani, Hadi Sadoghi Yazdi, and Reza Monsefi. 2012. Class imbalance handling using wrapper-based random oversampling. In *Proceedings of the 20th Iranian Conference on Electrical Engineering (ICEE2012)*. IEEE, 611–616.

[12] Sunil Gundapu and Radhika Mamidi. 2021. Transformer based Automatic COVID-19 Fake News Detection System. *arXiv preprint arXiv:2101.00180* (2021).

[13] Shunjie Han, Cao Qubo, and Han Meng. 2012. Parameter selection in SVM with RBF kernel function. In *Proceedings of the World Automation Congress 2012*. IEEE, 1–4.

[14] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[16] Michael Kearns and Dana Ron. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation* 11, 6 (1999), 1427–1453.

[17] Adam Kucharski. 2016. Study epidemiology of fake news. *Nature* 540, 7634 (2016), 525–525.

[18] Bor-Chen Kuo, Hsin-Hua Ho, Cheng-Hsuan Li, Chih-Cheng Hung, and Jin-Shiuh Taur. 2013. A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 1 (2013), 317–326.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[20] Yin Liu and Keshab K Parhi. 2016. Computing RBF kernel for SVM classification using stochastic logic. In *Proceedings of the 2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, 327–332.

[21] Yang Liu and Yi-Fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on Artificial Intelligence*.

[22] Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the international AAAI conference on web and social media*.

[23] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673* (2019).

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[25] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[26] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *arXiv preprint arXiv:1707.03264* (2017).

[27] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 797–806.

[28] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2008. RUSBoost: Improving classification performance when training data is skewed. In *Proceedings of the 2008 19th International Conference on Pattern Recognition*. IEEE, 1–4.

[29] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343* (2020).

[30] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[31] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 626–637.

[32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[33] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 430–435.

[34] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 312–320.

[35] Mirela Silva, Fabrício Ceschin, Prakash Shrestha, Christopher Brant, Juliana Fernandes, Catia S Silva, André Grégio, Daniela Oliveira, and Luiz Giovanini. 2020. Predicting misinformation and engagement in covid-19 twitter discourse in the first months of the outbreak. *arXiv preprint arXiv:2012.02164* (2020).

[36] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the 2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 39–47.

[37] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[38] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[39] Juanjuan Wang, Mantao Xu, Hui Wang, and Jiwu Zhang. 2006. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *Proceedings of the 2006 8th international Conference on Signal Processing*, Vol. 3. IEEE.

[40] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[41] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3205–3212.