MCWDST
A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media
Truica, Ciprian Octavian; Apostol, Elena Simona; Nicolescu, Radu Catalin; Karras, Panagiotis

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2023.3331220

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media

CIPRIAN-OCTAVIAN TRUICĂ, ELENA-SIMONA APOSTOL, RADU-CĂTĂLIN NICOLESCU, AND PANAGIOTIS KARRAS

1Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, 060042 Bucharest, Romania
2Department of Computer Science, University of Copenhagen, 2100 Copenhagen, Denmark

Corresponding author: Elena-Simona Apostol (elena.apostol@upb.ro)

This work was supported by the National University of Science and Technology Politehnica Bucharest through the PubArt Program.

ABSTRACT The widespread availability of internet access and handheld devices confers to social media a power similar to the one newspapers used to have. People seek affordable information on social media and can reach it within seconds. Yet this convenience comes with dangers; any user may freely post whatever they please and the content can stay online for a long period, regardless of its truthfulness. A need arises to detect untruthful information, also known as fake news. In this paper, we present an end-to-end solution that accurately detects fake news and immunizes network nodes that spread them in real-time. To detect fake news, we propose two new stack deep learning architectures that utilize convolutional and bidirectional LSTM layers. To mitigate the spread of fake news, we propose a real-time network-aware strategy that (1) constructs a minimum-cost weighted directed spanning tree for a detected node, and (2) immunizes nodes in that tree by scoring their harmfulness using a novel ranking function. We demonstrate the effectiveness of our solution on five real-world datasets.

INDEX TERMS Fake news detection, fake news propagation, network-aware fake news mitigation, real-time network immunization.

I. INTRODUCTION

With the accelerated technology adoption by a growing number of users, social media have become the main medium for the dissemination of information on current news and events [1]. While these new media bring several benefits (e.g., a large number of consumers reached, instant and continuous updates on one’s topics of interest), they also enable the spread of harmful information in the form of fake news, and may thus polarize public discourse regarding critical topics (e.g., elections [2], vaccination [3], health hazards [4]) and threaten democratic values [5]. Because of its detrimental effects on society at large [6], [7], [8], the fake news phenomenon has been studied by scientists and practitioners alike; fake news is defined as news articles that intentionally contain verifiably false misleading information inconsistent with factual reality [9], [10], [11], [12], [13], [14], [15]. To mitigate the threat of fake news, journalists have started to manually classify news and offer websites with fact-checking mechanisms that provide a verdict regarding its veracity, such as PolitiFact and Snopes. However, such solutions may fail in high-velocity information spreading social media, as news appears and spreads much faster than any manual verification; by the time it is checked, the news may have been already shared with many sources and its negative effect may have taken hold. In this paper, we propose new models and strategies for misinformation detection and mitigation to address the current real-world challenges posed by fake news.

In particular, we aim to:

(\textbf{O}_1) Propose new deep learning architectures to accurately detect fake news in social media; and
(O₂) Propose new real-time strategies to mitigate the spread of detected fake news in a social network.

To reach these objectives, we answer the following questions:

(Q₁) Can we improve the accuracy of fake news detection using new deep learning architectures?

(Q₂) Can we immunize nodes that spread harmful content using network information in real-time?

To answer (Q₁) and achieve objective (O₁), we propose two novel stack deep learning architectures that utilize convolutional and bidirectional LSTM layers. We show the effectiveness of these architectures with ample benchmarking on four real-world datasets. To answer (Q₂) and reach objective (O₂), we propose a real-time algorithm for network immunization, which builds a minimum-cost weighted directed spanning tree for a detected source node and chooses nodes in that tree to immunize by scoring their potential harmfulness using a ranking function. The proposed ranking functions consider the following network information for each node:

1. How long a chain of followers is (i.e., the length of the diffusion path).
2. How many nodes it reaches (i.e., the spread of information), and
3. How fast it spreads information (i.e., the information diffusion speed).

To show the effectiveness of network immunization, we evaluate our method on one real-world Twitter dataset. In summary, the main contributions of this paper are:

1. New deep learning architectures for fake news detection;
2. New real-time fake news mitigation strategy consisting of an algorithm for building the minimum-cost weighted directed spanning tree for a given node and a network-aware node harmfulness ranking function;
3. Benchmarking on multiple real-world datasets to evaluate the efficiency of our deep learning architectures for fake news detection;
4. Evaluation of our real-time mitigation algorithm on a real-world Twitter dataset.

The rest of this paper is structured as follows: In Section II, we present the state of the art. In Section III, we present the proposed architecture, deep learning models for fake news detection, minimum-cost weighted directed spanning tree algorithm for a given source node and harmful node ranking function. In Section IV, we detail the implementation of our architecture. In Section V, we describe the datasets and analyze in detail the experimental validation of our solution. Section VI provides an in-depth discussion of results, and Section VII concludes the paper and hints at future research.

II. RELATED WORK

In this section, we present the approaches of previous research on the two tasks of interest to our work: detecting fake news and mitigating its spread.

A. FAKE NEWS DETECTION

Nakamura et al. [16] introduced a dataset comprising Reddit posts along with an architecture designed to identify posts that contain fake pieces of information using image data, title data, or both combined. The advantages of working with such a dataset are its diversity, large size, and multidimensionality, which allows researchers to treat fake news detection as a 2-way, 3-way, or 6-way classification problem. Further, the authors combine text and image features to classify whether a post spreads fake pieces of information or not. Text is embedded using the BERT [17] and InferSent [18] models, while image data is extracted using the VGG16 [19], ResNet50 [20], and EfficientNet [21] models.

Kumar et al. [22] tackle the fake news detection challenge via sentiment analysis, implementing seven deep learning architectures, such as long short-term memory (LSTM), bidirectional LSTM, convolutional neural networks (CNN), and ensemble models that combine the aforementioned. After performing fake news classification using these models, the authors evaluate the models’ performance against classic machine learning algorithms, such as logistic regression and support vector machines.

Khan et al. [23] address the fake news detection task via classical machine learning algorithms and more advanced deep-learning and neural network models, and compare their performance on two well-known datasets: Liar and Fake or Real News, along with a self-built dataset, which the authors claim to be more variate and denser than the other two.

Granik and Mesyura [24] tackle the fake news detection problem by using the Naive Bayes classification algorithm, which, despite its simplicity, yields good results on the task. Moreover, the authors suggest further adjustments that can be made to improve the results of the aforementioned method.

Pérez-Rosas et al. [25] collect two original fake news detection datasets and use a linear kernel SVM algorithm to perform classification. The first dataset is built by collecting real news from reliable US news websites and adding pieces of fake information to them to turn them into fake news. For the second dataset, the authors target celebrity news and search the web for pairs of fake and real articles regarding celebrity gossip. The classification model takes as input a set of predefined linguistic features or a mix of them, such as unigrams or bigrams, punctuation marks, psycholinguistic features, readability, and syntax. The authors evaluate the model’s performances using these features on the two created datasets.

Nguyen et al. [26] suggest that implicit correspondences between articles can help improve the performance of a fake news detection algorithm. The authors convert the task of detecting whether an article contains untruthful pieces of information into an inference problem in a Markov random field and transform the algorithm that solves this kind of problem into a deep neural network.
Truică et al. [27] propose the use of transformer-based sentence embeddings and transfer learning for fake news detection for news articles in both English and German.

In their study, Truică and Apostol [28] show empirically that the embedding used for encoding the data is one of the key factors in accurately determining the veracity of news articles.

Mayank et al. [29] propose DEAP-FAKED, a new model that uses Natural Language Processing techniques, Graph Neural Networks, and Knowledge graphs to identify Fake News. The experimental results on the Kaggle dataset show this approach improves the performance of misinformation detection. We use the DEAP-FAKED model in our comparison.

Truică and Apostol [15] designed MisRoBÆRTa, a transformed-based ensemble model used for multi-class fake news detection. The ensemble uses RoBERTa [30] and BART [31] to encode the textual content of the news. The experimental results on the Kaggle and The Fake News Corpus (FNC) datasets show how this approach improves the overall performance on the task of misinformation detection. We use MisRoBÆRTa to compare the results obtained by our proposed models on Kaggles.

Raza and Ding [32] propose FND-NS, a framework that learns useful representations for predicting fake news. FND-NS uses both the textual content and the social context to determine the veracity of news articles and social media posts. For the Fakeedit corpus, we compare the results of our models with the ones obtained by FND-NS.

B. FAKE NEWS MITIGATION

Sharma et al. [33] start out with graphs constructed from two Twitter datasets and design two independent diffusion paths for real and fake posts, each with a separate set of learned parameters. Using these paths and the obtained parameters, they determine the user characteristics that matter most in news spread and propose two methods to mitigate the spread of fake news: blocking a node (i.e., the user sending the news) or blocking the edge (i.e., the news transmission path). Their experiments indicate that an account that spreads fake news has a low follower count and no account description available, while an account that posts real news has a large number of followers, and is associated with well-known sources.

Saxena et al. [34] consider a social network as a graph in which there are three types of users: positive — which spread true news, negative — which spread false news, and neutral — which are influenced by the other two types. The authors model these influences probabilistically, where neutral nodes influenced by one of the two types of news are less likely to change stance, the longer the diffusion process lasts. To mitigate false information circulating through the graph, the authors design an algorithm that assigns each true news broadcaster a metric called “truth score”, and selects the top-k sources with the best scores.

Shu et al. [35] contribute to the field of false news detection and mitigation by building a comprehensive dataset. They collect news from PolitiFact, GossipCop, and E! Online. The first two are platforms where the level of truth of news is labeled following a fact-checking process, while the latter is considered a credible news source. The dataset includes user features, i.e., number of followers, number of followees, user location (if mentioned in the profile description), user comments or retweets, as well as post features, i.e., post content (written or visual). These features are obtained by searching the news collected from the 3 sources using the Twitter search tool. By virtue of its diverse features, the dataset can be used in both fake news detection and mitigation tasks. Their best-performing model on this dataset is CNN [35], which we will also use in our comparison.

Sayyadiharikandeh et al. [36] address the problem of fake news mitigation assuming that non-human entities in social networks are set up to spread false information in an automated way. The authors implement a mechanism to detect such entities, which they call Botometer. They build several models to detect people and different types of bots in a social network. The models are aggregated into an ensemble learning algorithm, which produces the final classification results.

Nevertheless, none of the above approaches assembles a combined fake news detection and mitigation pipeline, as we do, further discussed in the following section. Firstly, we identify the harmful nodes that spread fake news in the network using a novel deep learning architecture. Secondly, we detect and rank harmful nodes using a novel algorithm that uses the direct weighted graph structure. Lastly, we immunize the network using a blocking strategy based on the ranked list of harmful nodes.

III. METHODOLOGY

Figure 1 presents our pipeline for fake news detection and mitigation. We mine a Social Media platform in real-time to detect nodes that spread harmful content; for each such node, we immunize the network with a real-time mitigation strategy. To be labeled as true or fake, a post passes through a preprocessing stage, where its content is edited and converted to a real-number matrix, which is passed to the detection stage. The detection stage is represented by one of our two proposed architectures, which outputs a post label. In case the post is false, we start the mitigation process; we construct a propagation path, i.e., a minimum-cost weighted directed spanning tree starting from each node that spreads fake news, compute a score for and rank the harmfulness of each node, and eventually extract the top-k most harmful nodes.

A. PREPROCESSING

The preprocessing module cleans the text and minimizes the vocabulary while preserving meaning; it comprises three steps: (1) punctuation marks and stop words removal, (2) lemmatization, and (3) word encoding using word embeddings.

Due to their high frequency of occurrence and specificity, stop words and symbols do not add relevant information when fed into the detection model. Thus, they may better be removed...
We use two pre-trained word embedding models and a model we trained on the chosen datasets. We transform words with similar forms into one unique form. The purpose of classification is to predict previously unseen vector representation; since we generate word embeddings in distinct filter, and adding a bias term. Every filter has an convolution operation between the text window and every passed to the network, creating $k$ size $125864$ embeddings or map every word of a post to a $N$ the Convolutional layer with $k$ size $L$ features by means of a convolutional layer, from which it extracts the best-generated features using a max-pooling layer, while $3BiLSTM$ feeds the initial features directly to the $BiLSTM$ and dropout sequences.

In the stacked CNN-3BiLSTM architecture (Figure 2), we use an embedding layer where we pass word embeddings as weight parameters, followed by a convolutional layer with one group of 128 filters and kernel size 3 with ReLU as activation function and a max-pooling layer with pool size 2 and the number of strides set to 2. We add a dense layer with 256 units, followed by three BiLSTM layers having dropout layers in-between. We set the number of units for the BiLSTM layers to 64 and the dropout rate to 0.2. The architecture ends with two dense layers: one with 128 linear activation units and one with 1 sigmoid activation unit.

In the stacked 3BiLSTM architecture (Figure 3), we use an embedding layer with weights set as the word embeddings, then stack three 128 units BiLSTM layers with dropout layers in-between, with a dropout rate of 0.2. As in the previous model, the architecture ends with two Dense layers.

Concretely, we propose two deep learning models, namely CNN-3BiLSTM and 3BiLSTM. The main difference between them is that CNN-3BiLSTM creates new features through a convolutional layer, from which it extracts the best-generated features using a max-pooling layer, while $3BiLSTM$ feeds the initial features directly to the $BiLSTM$ and dropout sequences.

B. FAKE NEWS DETECTION

The purpose of classification is to predict previously unseen items based on inferences derived by training on a set comprising news articles/posts and their labels. We first describe the basic elements of the two proposed architectures and motivate why we chose them.

We use an Embedding layer at the start of our models to store word embeddings that we obtain through the continuous bag-of-words model or pre-trained models. Thereby, our architectures have access to a word embedding dictionary. In general, an embedding layer would either generate embeddings or map every word of a post to a real-number vector representation; since we generate word embeddings in the preprocessing stage, we use the embedding layer with the latter function.

We employ a Bidirectional Long Short-Term Memory (BiLSTM) layer because the news we predict represents word sequences, and this type of layer is specialized in learning long-term dependencies in text. A further reason to employ such a layer is that it uses two simple LSTM architectures that look both forward and backward in the sequence, such that given a sequence element, both previous and future elements are available to the network.

We use a Convolutional layer with $N$ filters and kernel size $k$ to extract patterns from a $k$-size window in the data passed to the network, creating $N$ features by means of a convolution operation between the text window and every distinct filter, and adding a bias term. Every filter has an associated channel where it stores features. The final result of the Convolutional layer with $N$ filters and size $k$ applied on a post of length $L$ consists of $N$ channels of length $L - k + 1$, each containing the new features obtained by its corresponding filter. We equalize post lengths in the preprocessing stage by zero-padding or truncation, such that the post length is constant, leading to constant channel size for all posts. Following the Convolutional layer, we use a max-pooling layer with pooling size $p$ to decrease the size of the feature channels by grouping elements into groups of length $p$ and choosing only the feature with the maximum value. We use dropout layers to deactivate a percentage of the outputs coming from the previous layer; thus, we reduce overfitting by creating artificial noise and improve the generalization capacity of the network when new, unknown data is fed for prediction. We use dense layers with linear activation as a connection between the network layers, and a final dense layer with a softmax activation function to produce the classification result, i.e., the probability of the post to be true.

Concretely, we propose two deep learning models, namely CNN-3BiLSTM and 3BiLSTM. The main difference between them is that CNN-3BiLSTM creates new features through a convolutional layer, from which it extracts the best-generated features using a max-pooling layer, while $3BiLSTM$ feeds the initial features directly to the $BiLSTM$ and dropout sequences.

In the stacked CNN-3BiLSTM architecture (Figure 2), we use an embedding layer where we pass word embeddings as weight parameters, followed by a convolutional layer with one group of 128 filters and kernel size 3 with ReLU as activation function and a max-pooling layer with pool size 2 and the number of strides set to 2. We add a dense layer with 256 units, followed by three BiLSTM layers having dropout layers in-between. We set the number of units for the BiLSTM layers to 64 and the dropout rate to 0.2. The architecture ends with two dense layers: one with 128 linear activation units and one with 1 sigmoid activation unit.

C. FAKE NEWS MITIGATION

Having detected a source of fake news, we mitigate its spread by rating each user with a network-aware ranking score that represents how influential that user is. This ranking score weights in the following network-based goals: firstly, we aim to immunize nodes that are followed extensively, by followers.
ourselves having many followers; thus, we take into account the height of the subtree spanning out from them, i.e., the length of diffusion paths. Secondly, we wish to immunize nodes that spread information to multiple nodes; therefore, we take into account the size of the spanning subtree, i.e., the number of nodes. Lastly, we intend to immunize nodes that spread information faster than others; ergo, we immunize nodes having high information diffusion speed. A high ranking score by these criteria implies that a user has a high impact on news spread. After calculating this score for each user, we select the top-\(k\) most harmful users to be proposed for removal or added to a blacklist.

Figure 4 presents an example that evaluates every node that may spread a harmful article for a given source node. Firstly, given a weighted directed graph \(G = (V, E)\) with positive costs \(t_{u,v} \in \tau\), where \(\tau = \{t_{u,v}|(u, v, t_{u,v}) \in E\}\), with each \(t_{u,v}\) measuring the time required to propagate the harmful content from \(u\) to \(v\), we build a minimum-cost weighted directed spanning tree (MCWDST) \(T\) rooted at the detected source node \(r\), using Algorithm 1.

Algorithm 1 works as follows. We initialize (1) the tree’s set of vertices \(V_t\) to \(\{r\}\) and \(E_t\) to the empty set, and (2) the sets of untested vertices \(V_c\) to \(\{r\}\) and the set of untested edges \(E_c\) to all edges except the ones that point to \(r\) (Lines 1-4). While there are still untested vertices (Line 5), we initialize, in each iteration, a set for vertices \(V_n\) and one for edges \(E_n\) (Lines 6-7) which are needed to update \(V_c\) and \(E_c\) at the end of the iteration (Lines 20-21). We pass over all vertices \(n \in V_c\) and find the edges in \(E_c\) where \(n\) is the parent of \(v\) (Lines 8-9). If the node \(v\) is not present in \(V_t\) then we add \(v\) to \(V_t\) and \(E_n\) and the edge \((n, v)\) to \(E_t\) and \(E_n\) (Lines 10-14), otherwise, it means that there is an edge pointing to \(v\) in \(E_t\) (Lines 15-17). In this case, we verify if it is cheaper to pass through this edge than to use the existing one (Line 16). If it is cheaper, we update \(E_t\) by removing the existing edge (Line 17) and adding the new edge to \(E_n\) (Line 18). We update \(V_c\) with the newly added nodes in \(V_n\), and \(E_c\) by removing the nodes in \(E_n\) (Lines 20-21). When the algorithm exits the while block, it returns the propagation tree \(T\). The complexity is \(O(|V||E|)\). Having obtained this tree, we evaluate the potential harmfulness of each node using a ranking function and sort the nodes in descending order of their scores to obtain a leaderboard of the most harmful nodes.

Our ranking function assesses each node \(n\) (Equation (1)) in real-time in terms of three components that meet the criteria enumerated previously:

\[
\text{score}(n) = \text{height}(n) \times \text{size}(n) \times \text{speed}(n)
\]
we normalize the value of a timestamp into the ranking function the hole propagation paths, i.e., how many levels of context to predict words instead of using a distributed representation of the words to predict the context. For processing, we use: (1) SciKit-Learn [38] LabelEncoder to transform labels from categorical to numerical, (2) regular expressions to clean the text, (3) NLTK [39] WordNetLemmatizer to extract the lemmas, and (4) Keras [40] Tokenizer to vectorize the clean text corpus.

To obtain word embeddings, we consider two approaches: (1) training a Word2Vec CBOW model on each dataset using Gensim [41], and (2) using the pre-trained models GloVe pre-trained and Word2Vec pre-trained.

We train 100-dimensional Word2Vec CBOW [42] embeddings using a learning rate of 0.025 and a word window size of 5 for 5 epochs. We chose CBOW over Skip-gram because it is faster to train and it uses distributed representations of context to predict words instead of using a distributed representation of the words to predict the context. For the pre-trained embeddings we chose: (1) GloVe pre-trained: a 100-dimensional GloVe [43] embedding trained on Wikipedia 2014 and Gigaword 5, and (2) Word2Vec pre-trained: 300-dimensional Word2Vec embeddings trained on Wikipedia data.
B. FAKE NEWS DETECTION
In the Detection module, we use the Keras interface of the TensorFlow library to build our networks; we compile the model using the Adam optimizer with binary cross-entropy as the loss function, train the models for 200 epochs using a validation split of 0.2 and a batch size of 128, and add an early-stopping callback that monitors the validation loss.

C. FAKE NEWS MITIGATION
We construct propagation trees by reading the graph’s adjacency matrix, along with the timestamps associated with each node. The propagation tree is the minimum-cost weighted directed spanning tree starting from a source node that spreads fake news. We iterate through all nodes in the tree and assign a ranking score to each. To calculate the function of descendant timestamps, we use the built-in Python library statistics. We sort the scores dictionary by descending values and take the first $k$ nodes as the most harmful. We visualize the tree by Graphviz.

D. GRAPHICAL USER INTERFACE
We build a user-friendly graphical user interface consisting of three pages: a landing page, a fake news classification page, and a fake news mitigation page. For detection (Figure 5), the user inputs the content of the article or post in a text box, clicks the “Predict” button, and gets the classification result as a verdict; the results of preprocessing are thereby applied to the text, with stop words and punctuation removed, and lemmatization of the remaining words. Figure 6 presents the fake news mitigation page. The user can opt to visualize a node’s minimum-cost weighted directed spanning tree in two formats: human-friendly format as source — (destination, timestamp), or Twitter15 format. The format is toggled by a check box. Upon pressing “Submit”, the tree image is shown along with a ranking of the most harmful nodes therein according to our evaluation metric.

V. EXPERIMENTAL RESULTS
In this section, we present the datasets we use for evaluation, compare the performance of the two proposed architectures, and analyze how each metric in our ranking function performs, individually and collectively, in mitigating harmful content propagation.

A. DATASETS AND PREPROCESSING ANALYSIS
We use four datasets to evaluate the detection models and one dataset to evaluate the mitigation metrics. We split each dataset for classification in training and testing sets using an 80%-20% ratio. From the training set, we extract 20% for validation.

The Fake News Corpus (FNC) dataset contains over 9.4 million articles labeled using 11 predefined labels. We chose a balanced subset of 20,000 articles tagged as either reliable or fake. As the original dataset is automatically filtered, we manually verified each article’s content to ensure it is correctly labeled.

The Kaggle dataset, downloaded from the Kaggle Fake News Detection challenge, consists of 20,800 articles tagged as reliable or unreliable. After eliminating null rows, the resulting dataset consists of 10,387 reliable and 10,374 unreliable tagged news.

The GossipCop dataset consists of 22,155 news titles, article URLs, tweet IDs of users that retweeted the article, and tags that label articles as true or fake. As the dataset is initially imbalanced, we select 5,323 articles representing both tags.

The Fakeddit is a multimodal dataset that allows 2-way, 3-way, or 6-way classification using images, text, or both. We use the title of the post and the 2-way label to perform our task. We edit the dataset to exclude entries that contain images and select 107,742 posts with a 1:1 true-to-fake ratio.

The Twitter15 contains news articles and their labels, along with information on the directed network structure among users, followers, and followees. We use this dataset to...
analyze the performances of our mitigation module. It consists of 1490 news articles labeled as true, unverified, non-rumor, or false. We select 361 articles identified as false, construct their propagation traces, and compute scores for the nodes.

Table 1 presents the features of the datasets with cleaned content, where every article contains words that are lemmatized. In terms of average and maximum text length, we can group the datasets into two categories: large ones (i.e., FNC, Kaggle) and small ones (i.e., Fakeddit, GossipCop). The articles in FNC and Kaggle consist of sequences of sentences or phrases, while the posts in Fakeddit and GossipCop consist of a summary sentence. In terms of size, the largest dataset is Fakeddit and the smallest is GossipCop.

TABLE 1. Post-processing stage datasets details.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Article</th>
<th>Average Token</th>
<th>Maximum Token</th>
<th>True Articles</th>
<th>Fake Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNC</td>
<td>20000</td>
<td>412</td>
<td>10736</td>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td>Kaggle</td>
<td>20761</td>
<td>419</td>
<td>12059</td>
<td>10387</td>
<td>10374</td>
</tr>
<tr>
<td>GossipCop</td>
<td>10546</td>
<td>8</td>
<td>26</td>
<td>5323</td>
<td>5323</td>
</tr>
<tr>
<td>Fakeddit</td>
<td>107447</td>
<td>6</td>
<td>41</td>
<td>53975</td>
<td>53767</td>
</tr>
</tbody>
</table>

B. FAKE NEWS DETECTION RESULTS

Table 2 provides the classification results of the proposed architectures using the 3 embedding methods, i.e., Word2Vec trained, GloVe pre-trained, and Word2Vec pre-trained. The results are the average for each score after 10 distinct executions using random seeding. We use a default padding size of 10 for small-length data (Fakeddit and GossipCop) and 1000 for large-length (i.e., Kaggle and FNC) data. CNN-3BiLSTM obtains the overall best results on Fakeddit, Kaggle, and FNC, although the difference between CNN-3BiLSTM and 3BiLSTM is relatively small. On FNC, the best accuracy is obtained when using CNN-3BiLSTM with GloVe pre-trained, i.e., 98.37. The difference between CNN-3BiLSTM and BiLSTM with the same embedding is insignificant, i.e., 0.22. CNN-3BiLSTM with Word2Vec trained registers an accuracy of 96.38 on Kaggle, with only a small increase of 0.74 over BiLSTM with the same embedding. On GossipCop, CNN-3BiLSTM with Word2Vec pre-trained obtains the best accuracy, i.e., 74.97. On Fakeddit, BiLSTM with Word2Vec pre-trained embeddings obtains the best accuracy. We emphasize that, in fake news detection, recall is an important metric as it tells how many fake articles are correctly classified. On datasets having a large average text length (i.e., FNC and Kaggle), recall scores are between 90 and 100. On small-length data, recall is lower than 77.

Table 3 presents a comparison between the results obtained by our models and the current state-of-the-art models. We can observe that the proposed architecture outperforms or obtains similar results as the current models. MisRoBÆRTa [15] and the proposed CNN-3BiLSTM with Word2Vec-trained embeddings obtain very similar results. The small difference between the models is a direct result of the embedding: MisRoBÆRTa uses two context-aware stat-of-the-art transformers, i.e., BART [31] and RoBERTa [30], as embeddings, while we use a static embedding.

FIGURE 7. The influence of padding on the f1-scores by architecture.
We also study the impact of padding size on performance with large data, i.e., Kaggle and FNC. We do not include small-length data in our analysis, since their average text length is 6-8 and their content consists of one sentence; thus, their padding size range is small. We evaluate the two datasets in terms of f1-score for 4 padding sizes, namely 250, 500, 750, and 1 000. Figure 7 shows the f1-scores we obtain after 10 distinct executions using random seeding. On FNC with 3BiLSTM, the best-performing embedding strategy is Word2Vec pre-trained with a padding size of 750. On Kaggle with 3BiLSTM, the Word2Vec-trained embeddings obtain the best results when the padding size is 500. On FNC with CNN-3BiLSTM, GloVe pre-trained with a padding size of 750 achieves the best f1-score. The Word2Vec-trained embeddings with a 1 000 padding size have the best score on Kaggle with CNN-3BiLSTM. We note that a larger text size does not necessarily translate into better results. We conclude that the padding size has a small impact on model performance.

C. FAKE NEWS MITIGATION RESULTS

To analyze the performance of our proposed algorithm in identifying harmful nodes by measuring the distribution of news by users, we use 361 trees with an average of 335 nodes per tree, a minimum number of nodes of 97, a maximum number of nodes of 2 971, an average tree height of 4, a minimum height of 2, and a maximum height of 10. We first analyze the importance of \( f_t(n) \) in ranking nodes. We divide the interval \((1, 3]\) into two intervals and study the influence of the function \( f_t(n) \) on the obtained subintervals. Table 4 shows the percentages of the scores obtained for the evaluation metric in each range, for the three strategies applied in the \( f_t(n) \) function, when collecting the top-\( k \)-most harmful nodes with \( k \in \{1 000, 2 000, 3 000\} \). A higher score represents a greater capability to spread untruthful information faster. Thus, a score in \((2, 3]\) indicates a very harmful node to be immunized first, a score in \([1, 2]\) indicates a mildly harmful node to be immunized in second priority, where nodes with a score close to 1 suggest that a node is weakly to not harmful and should be immunized in last priority. A node with a score of 1 is a lead node. We observe that when using \( \text{median}(t_n) \), our ranking function scores with higher values more harmful nodes (i.e., harmful nodes are ranked in \((2, 3]\)) than when using \( \text{average}(t_n) \) or \( \text{ratio}(t_n) \) (i.e., harmful nodes are ranked in \((1, 2]\)). Furthermore, as we know which nodes are harmful and which are not, \( \text{rank}(n) \) identifies correctly all the harmful nodes regardless of the method used to compute \( f_t(n) \).

Figure 8 plots the contribution that each component of the proposed ranking function has in a node’s final score. We analyze the top-\( k \) most harmful nodes, where \( k = \{5, 10, 15, 20\} \). The \( f_t(n) \) function influences the score the
most, followed by the $H(n)$ and $A(n)$ functions. Notably, the weight of $f_t(n)$ is more than 50% in all evaluation situations. In all three subgraphs, as we increase $k$, the contribution of $H(n)$ is transferred to $A(n)$. With $k = 20$, $A(n)$ surpasses $H(n)$ (Figure 8(b) and 8(c)). In conclusion, the information diffusion speed has the most contribution to our ranking function. That is a reasonable design, as, for a small $k$, nodes that have long chains of followers (i.e., a longer maximum diffusion path) play a more important role than the nodes that can reach many nodes. As $k$ increases, the impact of the nodes with many followers decreases, and the impact of nodes that reach many nodes increases.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$f_t(n)$</th>
<th>$%$ in (1, 2)</th>
<th>$%$ in (2, 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>average($n$)</td>
<td>68.51</td>
<td>31.49</td>
</tr>
<tr>
<td></td>
<td>median($n$)</td>
<td>60.10</td>
<td>39.90</td>
</tr>
<tr>
<td></td>
<td>ratio($n$)</td>
<td>89.10</td>
<td>10.90</td>
</tr>
<tr>
<td>2000</td>
<td>average($n$)</td>
<td>83.31</td>
<td>16.69</td>
</tr>
<tr>
<td></td>
<td>median($n$)</td>
<td>80.05</td>
<td>19.95</td>
</tr>
<tr>
<td></td>
<td>ratio($n$)</td>
<td>94.55</td>
<td>5.45</td>
</tr>
<tr>
<td>3000</td>
<td>average($n$)</td>
<td>90.20</td>
<td>9.80</td>
</tr>
<tr>
<td></td>
<td>median($n$)</td>
<td>86.70</td>
<td>13.30</td>
</tr>
<tr>
<td></td>
<td>ratio($n$)</td>
<td>96.36</td>
<td>3.64</td>
</tr>
</tbody>
</table>

### D. MITIGATION SCALABILITY RESULTS

We want to test the scalability of our solution and determine the mitigation strategy performance in a real-world setting. We collect from Twitter a graph structure containing over 10,000 nodes and 89,000 edges and split it into 3 subgraphs for scalability testing. Table 5 presents the MCWDST construction and ranking list creation time in seconds for these subgraphs. We can observe that for a small graph, the mitigation is done almost instantly. This is useful when we detect and target the harmful content spreaders within a disinformation network. For a larger network, the entire process is done in near real-time, as it takes less than 5 minutes. Thus, the best strategy is to immunize the network by targeting the topmost harmful nodes. Finally, we compare our method with two preventive network immunization strategies, i.e., NetShield [47] and SparseShield [48] an improved version of NetShield, and one counteractive network immunization in large networks, i.e, DAVA [49].

Although both preventive network immunization algorithms are faster, the main problems with NetShield and SparseShield that we solved with MCWDST are: (1) nodes are blocked in no particular order if they converge, (2) nodes are not ranked, (3) the edge weights are not used in the immunization process, (4) the directed graph structure is not considered, and (5) the number of immunized nodes is dependent on the budget.

Whereas our proposed solution gives an immunization strategy for a weighted directed graph by returning a blocking order by ranking the harmfulness of nodes.

When comparing MCWDST with DAVA, another counteractive network immunization algorithm, we observe that:

1. DAVA blocks all the nodes in no particular order while our MCWDST manages to block nodes in the order given by the information diffusion paths,
2. DAVA does not rank nodes, while MCWDST ranks them from most to least harmful,
3. DAVA is dependent on the budget and immunizes a given number of nodes, while MCWDST immunizes the entire network, and
4. DAVA does not consider the directed graph structure, while MCWDST takes this graph property into account.

### VI. DISCUSSION

This section highlights the results and lessons learned from the implementation and presents an in-depth analysis of the two modules.

### A. FAKE NEWS DETECTION

We observe that the pre-trained models (i.e., Word2Vec pre-trained and GloVe pre-trained) achieve better results on smaller datasets, i.e., GossipCop and Fakeddit, while Word2Vec trained and GloVe pre-trained obtain better results on larger datasets, i.e., FNC and Kaggle. These results of the classification models directly reflect the typical text length in each dataset. For small-length texts, it is better to use pre-trained embeddings as they better encode
semantic, syntactic, and contextual information that cannot be found in short statements. This is mitigated in large datasets by the texts’ length, which offers sufficient semantic, syntactic, and contextual information to train accurate and data-specific word embeddings. The proposed novel deep learning architectures offer state-of-the-art results on all tested datasets, i.e., an accuracy of 88.66 on Kaggle [29], 74.80 on Fakeddit [32], and 72.30 on GossipCop [35]. We note that the longer texts enable the bidirectional LSTM layers with and without a CNN layer on top to detect meaningful patterns for fake news detection. Even though the experimental results show that 3BiLSTM architecture obtains good results, the CNN-3BiLSTM architecture achieves the best scores on FNC, Kaggle, and GossipCop. This result arises from the combination of a CNN layer with a MaxPooling layer, which generates new features and selects the best features among those generated. Moreover, on the Kaggle data, CNN-3BiLSTM outperforms 3BiLSTM with all embedding strategies we use. Furthermore, we observe that, on large datasets, a large-size padding strategy does not necessarily improve performance. It follows that an ample analysis w.r.t. size and word embedding should be performed on each dataset to determine the best padding approach.

### B. FAKE NEWS MITIGATION

The proposed ranking function takes into account 3 network-specific components for mitigating harmful nodes: (1) $H(n)$, the length of the diffusion path, (2) $A(n)$, the spread of information, and (3) $f_i(n)$, the speed of information diffusion.

In our experiments on a real-world dataset, we observe that: (1) regardless of graph size, $f_i(n)$ has the largest contribution in the function, (2) for a small number of nodes to mitigate, the length of the diffusion path $H(n)$ adds more weight to the ranking function than the number of nodes that can be infected directly $A(n)$, and (3) for a large number of nodes to mitigate, both $H(n)$ and $A(n)$ have a similar impact on the ranking.

As $f_i(n)$ has the largest contribution of the three components to the score value of a node, it is, in effect, the most important factor in mitigating fake news in a network, according to our empirical findings. The comparison of different calculation strategies of $f_i(n)$ indicates the impact of the potential to spread fake news quickly, while the individual evaluation of the three components of the ranking function indicates that network structure is also important for the immunization task.

### VII. CONCLUSION

In this paper, we proposed (1) two novel deep learning architectures for fake news detection, (2) a real-time algorithm for building the minimum-cost weighted directed spanning tree (i.e., MCWDST), and (3) a function that ranks harmful nodes in real-time, to combat the spread of fake news on social media.

The proposed deep learning architectures use convolutional and bidirectional LSTM layers that capture semantic, syntactic, and context information. These models offer state-of-the-art performance on the tested datasets. Our mitigation algorithm greedily constructs a minimum-cost weighted directed spanning tree for a given source node. The novel ranking function takes into account network information (i.e., the length of the diffusion path, the spread of information, and the speed of information diffusion) as well as its harmful content to mitigate the spread of fake news in a network-aware, real-time manner. Ample benchmarking on five real-world datasets shows that the proposed models for fake news detection and the mitigation strategy are effective in stopping the spread of harmful content online.

For future research directions, we will consider using: (1) transformers as a way of generating word embeddings, and (2) graph embedding techniques to develop a more complex network-aware model that identifies harmful sources in real-time.

### ACKNOWLEDGMENT

(Ciprian-Octavian Truică, Elena-Simona Apostol, and Radu-Cătălin Nicolescu contributed equally to this work.)

### REFERENCES


### TABLE 5. Mitigation scalability tests.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Edges</th>
<th>MCWDST ($)</th>
<th>Ranking ($)</th>
<th>DAVA</th>
<th>NeiShield ($)</th>
<th>SparseShield</th>
</tr>
</thead>
<tbody>
<tr>
<td>978</td>
<td>10217</td>
<td>2.01</td>
<td>0.66</td>
<td>1.78</td>
<td>0.29</td>
<td>0.07</td>
</tr>
<tr>
<td>5210</td>
<td>49124</td>
<td>53.68</td>
<td>20.69</td>
<td>60.75</td>
<td>12.02</td>
<td>0.42</td>
</tr>
<tr>
<td>10210</td>
<td>89124</td>
<td>411.96</td>
<td>138.44</td>
<td>386.99</td>
<td>102.13</td>
<td>1.33</td>
</tr>
</tbody>
</table>
ELENA-SIMONA APOSTOL was a Postdoctoral Researcher with Microsoft Research Center in collaboration with the INRIA Institute, where she worked on state-of-the-art big data analysis, multisite cloud computing, and bioinformatics. She also held research positions with the Department of Information Technology, Uppsala University, and the Fraunhofer FOKUS Institute. She is currently an Associate Professor of computer science with the Computer Science and Engineering Department, Faculty of Automatic Control and Computers, National University of Science and Technology Politehnica Bucharest, Romania. Her current research interests include big data analytics, data science, text and data mining, and cloud computing.

RADU-CĂTĂLIN NICOLESCU received the bachelor’s degree in computer engineering from the University Politehnica of Bucharest, Romania, in 2021. He is currently pursuing the M.Sc. degree in advanced software services with the National University of Science and Technology Politehnica Bucharest, Romania. The bachelor’s thesis was on the subject of fake news detection using deep learning methods. His current research interests include natural language processing, computer vision, and data mining. He aims to find new approaches and improvements to current challenges in artificial intelligence through his research.

PANAGIOTIS KARRAS received the M.Sc. degree in electrical and computer engineering from the National Technical University of Athens and the Ph.D. degree in computer science from The University of Hong Kong. He is currently a Professor of computer science with the University of Copenhagen. His current research interests include the design of robust, scalable, and versatile methods for data access, mining, analysis, exploration, and representation. He has been awarded the Hong Kong Young Scientist Award, the Singapore Lee Kuan Yew Postdoctoral Fellowship, the Rutgers Business School Teaching Excellence Fellowship, and the Skoltech Best Faculty Performance Award.