INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD

MINI PROJECT

Fake News Detection

Group Members:

DEBASISH DAS
IIB2019031

SUMIT BAKOLIYA
IIT2019083

SHIVANGI VERMA
IIT2019224

ABHISHEK BITHU
IIT2019199

RAJ CHANDRA
IIT2019200

Supervised By:
Dr.Manish Kumar

Abstract

Fake News has become one of the major problems in the existing society. Fake News has high potential to change opinions, facts and can be the most dangerous weapon in influencing society. In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to sway election outcomes for the benefit of certain candidates.

Introduction

Fake News have become more prevalent in recent years and with great amount of dynamism in internet and social media, differentiating between facts and opinions, relating to commercial or political upheavals has become more difficult than ever.

- Fake information is purposely or unintentionally spread throughout the internet. It is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. It has weakened public trust in governments.
- The reach of fake news was best highlighted during the critical months of the 2016 U.S. presidential election campaign. During that period, the top 20 frequently discussed fake election stories generated 8,711,000 shares, reactions, and comments on Facebook—ironically, more than the 7,367,000 for the top 20 most-discussed election stories posted by 19 major news websites [Silverman 2016]

What is Fake News?

The definition of fake news is information that pushes people down the wrong road. Fake news is spreading like wildfire these days, and people are sharing it without confirming it. This is frequently done to promote or impose specific views, and it is frequently accomplished through political agendas.

- **Broad definition of fake news:** Fake news is false news.
- **Narrow Definition of Fake News:** Fake news is intentionally false news published by a news outlet.

Type Of Fake News Detection

1. KNOWLEDGE-BASED FAKE NEWS DETECTION

When detecting fake news from a knowledge-based perspective, one often uses a process known as fact-checking. Fact-checking, initially developed in journalism, aims to assess news authenticity by comparing the knowledge extracted from to-be-verified news content (e.g., its claims or statements) with known facts. In this section, we will discuss traditional fact-checking (also known as manual fact-checking) and how it can be incorporated into automatic means to detect fake news (i.e., automatic fact-checking).

➢ **Manual Fact-Checking**

The traditional Fact-checking is known as Manual Fact-Checking.
Manual Fact-checking divided into :

❖ **Expert-based manual fact-checking**

Expert-based fact-checking relies on domain experts as fact-checkers to verify the given news contents. Expert-based fact-checking is often conducted by a small group of highly credible fact-checkers, is easy to manage, and leads to highly accurate results, but it is costly and poorly scales with the increase in the volume of the to-be-checked news contents.

❖ **Crowd-sourced manual fact-checking**

It relies on a large population of regular individuals acting as fact-checkers (i.e. the collective intelligence). Such a large population of fact-checkers can be gathered within some common Crowd-sourcing marketplaces. Such as, in Amazon Mechanical Turk, based on which CREDBANK [Mitra and Gilbert 2015], a publicly available large-scale fake news dataset, has been constructed. Compared to expert-based fact-checking, crowd-sourced fact-checking is relatively difficult to manage, less credible and accurate due to the political bias of fact-checkers and their conflicting annotations, and has better (although insufficient) scalability.

➢ **Automatic Fact-Checking**

Manual fact-checking does not scale with the volume of newly created information, especially on social media. So to address scalability, automatic fact-checking techniques have been developed, heavily relying on IR (Information Retrieval), NLP (Natural Language Processing), and ML techniques, as well as on network/graph theory. Considering that a systematic approach for automatic fact-checking of news has to the best of our knowledge never been presented before, here we prioritize organizing the related research to clearly present the automatic news fact-checking process over presenting each related study in detail.

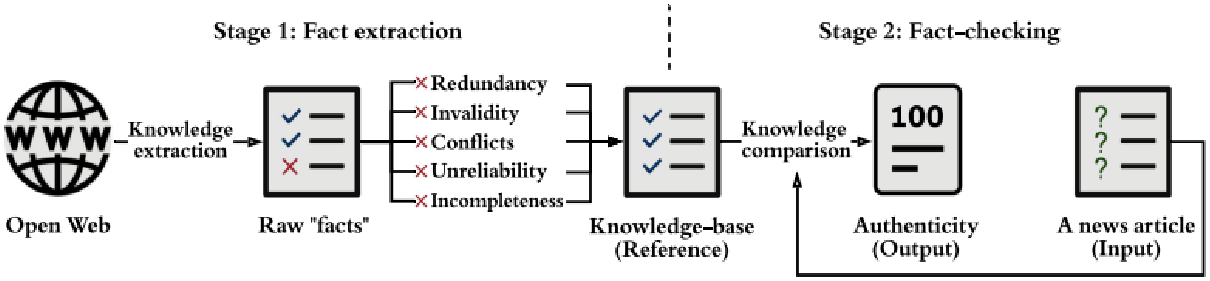


Fig. 3. Automatic news fact-checking process.

It is divided into two stages :

❖ **Fact extraction**

To collect facts and construct a Knowledge Base, knowledge is first extracted from the open web as raw “facts” that need further processing. And to form a Knowledge Base from the extracted raw “facts”, they need to be further cleaned up and completed by addressing some issues.

#Knowledge Base : Set Of Facts

Knowledge Extraction classified into :

Single Source Knowledge Extraction	Open Source Knowledge Extraction
➔ It relies on one comparatively reliable source (e.g., Wikipedia) to extract knowledge ➔ It is relatively efficient but often leads to incomplete knowledge .	It aims to fuse knowledge from distinct sources. It is less efficient than single-source knowledge extraction but leads to more complete knowledge.

- ❖ **Fact-checking**

To assess the authenticity of news articles, we need to compare the knowledge extracted from to-be-verified news content (i.e., SPO triples) with the facts (i.e., true knowledge),. KBs are suitable resources for providing ground truth for news fact-checking.

- ➔ A set of (Subject, Predicate, Object) (SPO) triples extracted from the information by Fact Extract that well represent the given information.
- ➔ A fact is a knowledge (SPO triple) verified as truth.

2. STYLE-BASED FAKE NEWS DETECTION

Style-based fake news identification analyses news material in the same way as knowledge-based false news detection does. Instead of analysing the accuracy of news information, this method assesses the writer's intent to deceive the public. Publishers of fake news frequently aim to influence big audiences by disseminating twisted and misleading information. To make the bogus titles more appealing, they utilise nearly all capitalised words, a higher proportion of proper nouns, and fewer stop words. To detect false news, style-based techniques capture the distinguishing features of writing styles between legitimate users and anomalous accounts. reports on the research of hyperpartisan news writing styles in relation to false news.

3. PROPAGATION-BASED FAKE NEWS DETECTION

When detecting fake news from a propagation-based perspective, one can investigate and utilize the information related to the dissemination of fake news, such as how users spread it. Similar to style-based fake news detection, propagation-based fake news detection is often formulated as a binary (or multi-label) classification problem as well, but with a different input. Broadly speaking, the input to a propagation-based method can be either a (I) news cascade, a direct representation of news propagation, or a (II) self-defined graph, an indirect representation capturing additional information on news propagation. Hence, propagation-based fake news detection boils down to classifying,

(I) *news cascades* or **(II)** *self-defined graphs.*

Fake News Detection Using News Cascades

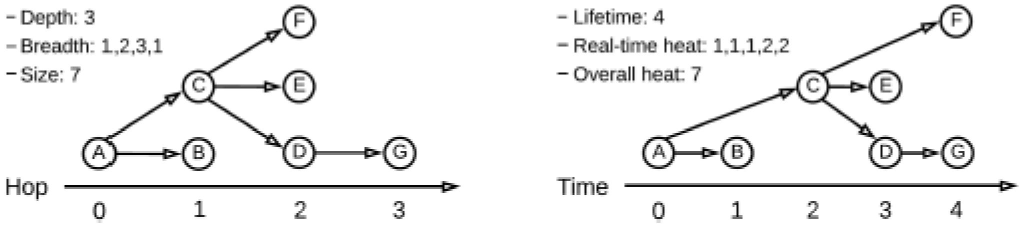


Fig. 9. Illustrations of news cascades.

A news cascade is a tree or tree-like structure that directly captures the propagation of a certain news article on a social network . The root node of a news cascade represents the user who first shared the news article (i.e., initiator); other nodes in the cascade represent users that have subsequently spread the article by forwarding it after it was posted by their parent nodes, which they are connected to via edges. A news cascade can be represented in terms of the number of steps (i.e., hops) that the news has traveled (i.e., hop-based news cascade) or the times that it was posted (i.e., time-based news cascade).

Hop-based news cascade, often a standard tree, allowing natural measures such as

—Depth: the maximum number of steps (hops) that the news has traveled within a cascade;

—Breadth (at hop k): the number of users who have spread the news k steps (hops) after it was initially posted within a cascade; and

—Size: the total number of users in a cascade.

Time-based news cascade, often a tree-like structure, allowing natural measures such as

—Lifetime: the longest interval during which the news has been propagated;

—Real-time heat (at time t): the number of users posting/forwarding the news at time t; and

—Overall heat: the total number of users who have forwarded/posted the news.

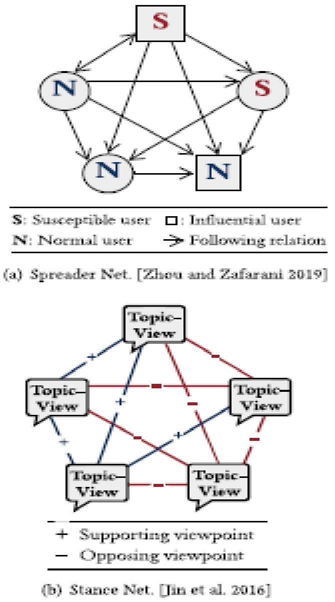
To perform this classification, some proposed methods rely on
(I) *traditional ML*, whereas others utilize (II) *(deep) neural networks*.

Traditional ML models : Within a traditional ML framework, to classify a news cascade that Traditional ML has been represented as a set of features, one often relies on supervised learning methods such as SVMs , decision trees], decision rules , naive Bayes , and RF .

DL models : Within a DL framework, learning the representation of news cascades often relies on neural networks, where a softmax function often acts as a classifier.

Fake News Detection Using Self-Defined Propagation Graphs

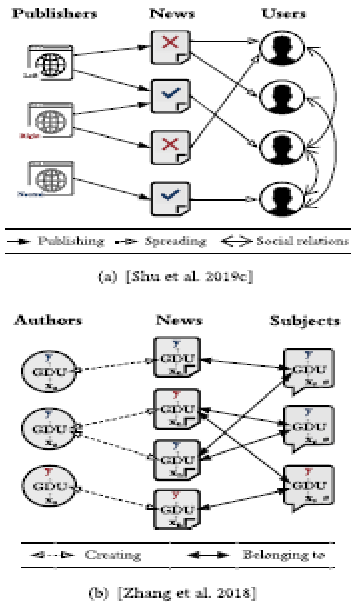
When detecting fake news using self-defined propagation graphs (networks), one constructs flexible networks to indirectly capture fake news propagation. These networks can be homogeneous, heterogeneous, or hierarchical.



(a) Spreader Net. [Zhou and Zafarani 2019]

(b) Stance Net. [Jin et al. 2016]

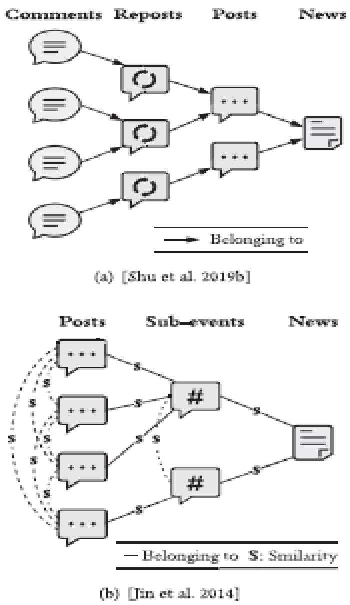
Fig. 12. Homogeneous networks.



(a) [Shu et al. 2019c]

(b) [Zhang et al. 2018]

Fig. 13. Heterogeneous networks.



(a) [Shu et al. 2019b]

(b) [Jin et al. 2014]

Fig. 14. Hierarchical networks.

Homogeneous network:

Homogeneous networks are networks containing a single type of node and a single type of edge.

Heterogeneous network:

Heterogeneous networks have multiple types of nodes or edges.

Hierarchical network:

In hierarchical networks, various types of nodes and edges form set subset relationships (i.e., a hierarchy).

4. SOURCE-BASED FAKE NEWS DETECTION

Overview of our project

☐ Here, We use various NLP and preprocessing methodologies like tokenization, stop words removal, stemming and machine learning classification algorithms - logistic regression, naive bayes, svm, and LSTM to build a model that analyzes the performance of these various classification methodologies to choose the best classifier for a dataset.

☐ Now to detect source we are creating a social network and then using depth first search algorithm traversing nodes to find the node and interconnection between them.

Implementation - I

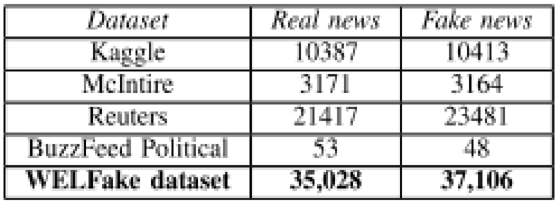
(Working With Dataset)

A. About Dataset

There are a large number of data sets available for the study of fake news and some of the popular dataset are

In this project, we are working with ***WELFake Dataset***. This data set is designed larger to prevent overfitting of classifiers and enable better ML training. For this purpose, it merged four popular news datasets (i.e. ***Kaggle, McIntire, Reuters, and BuzzFeed Political***) and prepared a more generic data set of 72 134 news articles with 35 028 real and 37 106 fake news.

WELFake Data Set

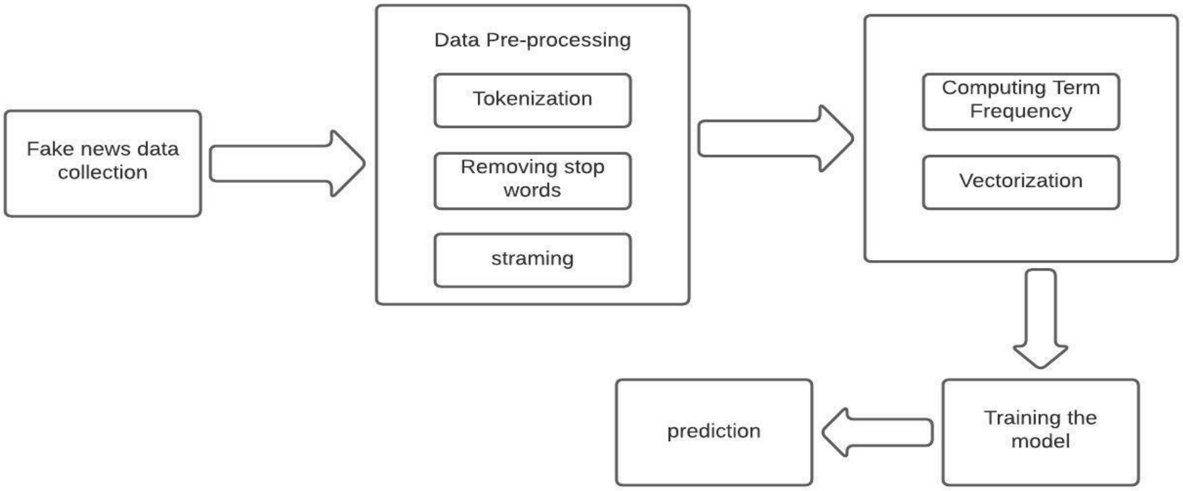


Dataset	*Real news*	*Fake news*
Kaggle	10387	10413
McIntire	3171	3164
Reuters	21417	23481
BuzzFeed Political	53	48
WELFake dataset	**35,028**	**37,106**

Dataset contains four columns:

- Serial number (starting from 0);
- Title (about the text news heading);
- Text (about the news content); and
- Label (0 = fake and 1 = real).

B. Working Methodology



❖ News Data Collection

This is essential for a balanced and unbiased data set and the key to providing high quality training data and delivering good results. Although there exist an important number of open data sets for the study of fake news, we are working with WELFake data set that combines four data sets, Kaggle, McIntire, Reuters, and BuzzFeed, for two reasons. First, they have a similar structure with two categories (i.e., real and fake news). Second, combining the data sets reduces the limitations and the bias of each individual data set.

❖ Data Preprocessing

This solves different problems in the collected data, like typographic errors, unstructured data format, and other limitations, using several methods, depending on the data set and objectives.

*(In our dataset, there are columns **'title'** & **'text'** , and in our project we are working with **'title' column** because content in 'text' column are very long which will take more time to process)*

- ***Missing data*** handles *undefined* (NaN) and *blank values* (NULL) present in the data set, which hinder the feature engineering process. Since deleting the data entries containing missing values may cause the loss of important information, we performed a missing value imputation process that estimates missing values and then analyzes the complete data set as if these values were the actual observed ones.

- Irrelevant data removes ***stop words*** (and other noise) that make the sentence grammatically complete, but do not have semantic significance in news classification operations. Removing stop words and keeping relevant tokens only significantly increases the model performance.
- ***Stemming*** converts the text into its root word by applying the Porter–Stemmer algorithm on text features for accuracy improvement. In case it cannot recognize the root word, it generates the canonical form of a corresponding word.
- ***Tokenization***, where the text is divided into a set of meaningful pieces & these pieces are called tokens.

❖ **Vectorization**

Since Machine Learning cannot understand *textual data*. So, we need to convert the *textual data* into *numerical data* which we did by vectorization.

❖ **Split The Data**

To train a model, we need to first split the data into training data and text data. We experimented with each ML model on random samples of the WELFake data set with training-testing data combinations: 80%–20%.

➔ ***80% in Training Data***

➔ ***20% in Test Data***

❖ **Training Model**

After splitting the data into Training and Test Data, we train different models and find out the best model for our WELFake dataset.Also, we train a Neural network and find out the difference between the accuracies.

C. Machine Learning Classification Methods

We review in this section a few ML methods used for fake news classification in the WELFake model.

- **Logistic Regression:** Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.
- **Support Vector Machine:** This is a supervised learning algorithm that works for both classification and regression problems. The algorithm finds the best line for set separation and predicts the correct set for new data values.
- **Naive Bayes:** This is a supervised learning algorithm based on Bayes' theorem that gives fast predictions with better accuracy in the domain of sentiment analysis, spam filtration, and text classification?

D. Text Classification Methods

This section reviews the state-of-the-art LSTM classification methods used in our experimental evaluation

- **Neural Network (LSTM):** LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. It is a type of recurrent neural network.In LSTM we can use a multiple word string to find out the class to which it belongs. This is very helpful while working with Natural language processing. If we use appropriate layers of embedding and encoding in LSTM, the model will be able to find out the actual meaning in input string and will give the most accurate output class.

E. Experimental Evaluation

(Evaluation Metrics)

We define four evaluation parameters, true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN), based on the relation between the predictive news classification and the actual one. Based on these parameters, we evaluated the WELFake model on four performance metrics

- **Accuracy:** This is the ratio between the number of correct predictions and the total number of predictions.

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

- **Precision:** This measures the positive predicted value, as the ratio between the number of correct positive predictions to the total number of positive predictions Precision = TP TP + FP.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- **Recall:** R measures the sensitivity of the model as the ratio between the number of correct positive predictions to the total number of correctly predicted results

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- **F1-Score:** F1 measures the testing accuracy of the model as the harmonic mean of the precision

$$F1 - \text{score} = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}}.$$

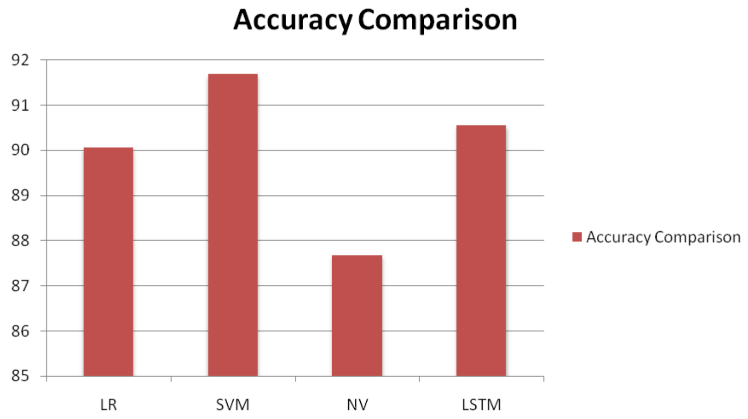
F. Results

The Result we get after training Logistic Regression Model; Naive Bayes Model; Support Vector Machine; and LSTM model with our WELFake Dataset:

(All the other evaluation results are available in Code)

MODEL	Accuracy
Logistic Regression	90.06%
Support-Vector-Machine	91.69%
Naive Bayes	87.97%
LSTM(Long Short Term Memory Network)	90.56%

G. Why is SVM Accuracy better than Neural Network LSTM?



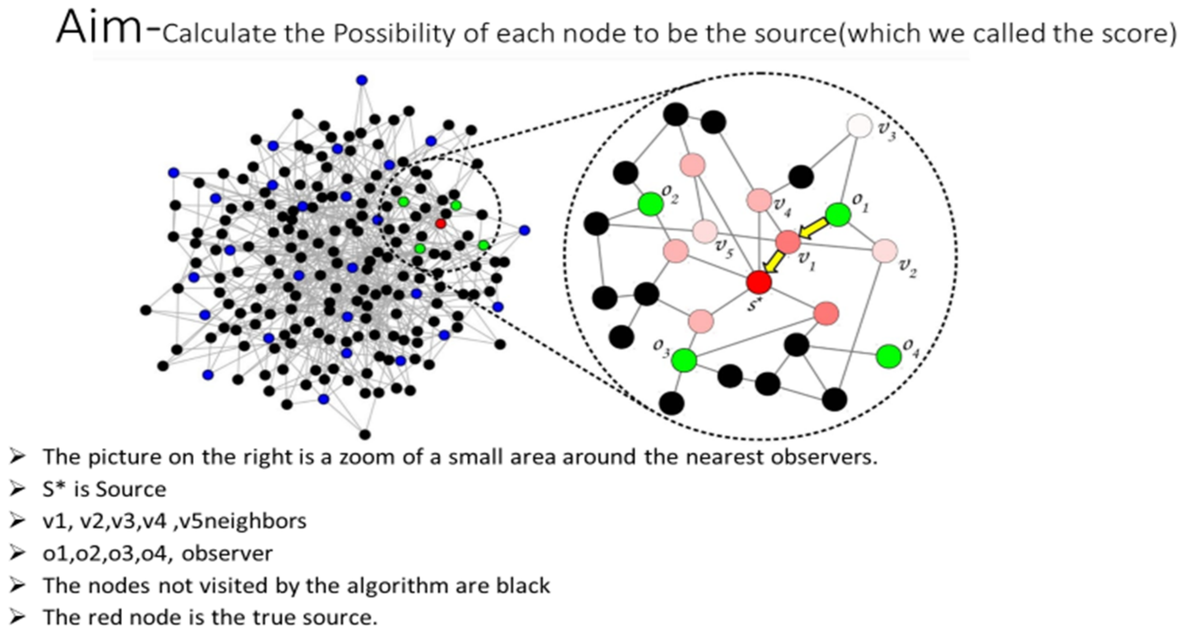
Generally the accuracy we get from the Neural Network model is better compared to other Classification models. But, When we compare the accuracy and F1-score of the SVM model with the LSTM model. The SVM model achieved a 91.69% accuracy, while LSTM achieved a maximum accuracy only up to 90.56%. Similarly, SVM also shows a better F1-score compared to LSTM due to its ***better generalization***. While Kaliyar et al. achieved a 98.36% accuracy using a deep NN on a single data set, its accuracy reduced to 92.48% on the WELFake data set. Similarly, LSTM is a pretrained model which works well with labeled data, while its performance gets compromised in a generalized data set where testing data are independent of the training data.

Implementation - II

(Working To Find Source)

Fake News Detection Using Self-Defined Propagation Graphs

When detecting fake news using self-defined propagation graphs (networks), one constructs flexible networks to indirectly capture fake news propagation. These networks can be homogeneous, heterogeneous, or hierarchical.



Aim-Calculate the Possibility of each node to be the source(which we called the score)

- The picture on the right is a zoom of a small area around the nearest observers.
- S* is Source
- v1, v2,v3,v4 ,v5neighbors
- o1,o2,o3,o4, observer
- The nodes not visited by the algorithm are black
- The red node is the true source.

APPROACH

- STEPS

1. Creating social network of n nodes by taking n as user input.
2. Now creating edges by random function as connection between nodes.
3. Now we are propagating from a random node taking input from user as a source
4. Now we are implementing depth first search algorithm to traverse between node and to find which node is traversed how many times
5. Then we will calculate the delay time for each node.

DEPTH FIRST SEARCH ALGORITHM

APPROACH

- Depth-first search is an algorithm for traversing or searching tree or graph data structures. The algorithm starts at the root node (selecting some arbitrary node as the root node in the case of a graph) and explores as far as possible along each branch before backtracking.

- So the basic idea is to start from the root or any arbitrary node and mark the node and move to the adjacent unmarked node and continue this loop until there is no unmarked adjacent node.
- Then backtrack and check for other unmarked nodes and traverse them.
- Finally, print the nodes in the path.

ALGORITHM

Create a recursive function that takes the index of the node and a visited array.

- Mark the current node as visited and print the node.
- Traverse all the adjacent and unmarked nodes and call the recursive function with the index of the adjacent node.

PSEUDO CODE

```
int curr_min=INT_MAX,total_count=0,mntemp=0;
void dfs(vector<vector<int>> &G,int src,char prop,vector<int> &vis,vector<char> &propv)
{
    if(vis[src]==true)
        return;
    vis[src]=true;
    total_count++;mntemp++;
    cout<<"Current value at "<<src<<" is "<<prop<<endl;
    propv[src]=prop;
    for(int i=0;i<G[src].size();i++)
    {
        dfs(G,G[src][i],prop,vis,propv);
    }
}
```

CODE LINK

https://onlinegdb.com/jVvqkbAc4

OUTPUT

```
Hey, please enter number of nodes:
6
Please enter number of edges:
12
Generating random 12 edges for nodes 1 to 6...
1 edge: 2 5
2 edge: 4 2
3 edge: 6 2
4 edge: 5 1
5 edge: 4 2
6 edge: 3 2
7 edge: 3 2
8 edge: 6 5
9 edge: 1 1
10 edge: 5 5
11 edge: 6 3
12 edge: 4 4


-------------------------------------
Hey, please enter the source node from where to propagate:
5
please enter a message to propagation :
fake_news

These are the all nodes which can be propagated by entered source node 5 :
Current propagated message at node 5 is |fake_news| and the time interval is 0s.
Current propagated message at node 2 is |fake_news| and the time interval is 1s.
Current propagated message at node 4 is |fake_news| and the time interval is 2s.
Current propagated message at node 6 is |fake_news| and the time interval is 2s.
Current propagated message at node 3 is |fake_news| and the time interval is 3s.
Current propagated message at node 1 is |fake_news| and the time interval is 1s.

Total 5 nodes are directly connected from given node 5
----
We can propagate total 6 nodes from given source 5
----
The minimum value is : 5
----
All propagated nodes :
2 1 6 5 5


-------------------------------------
```

COMPLEXITY ANALYSIS

- **Time complexity:** O(V + E), where V is the number of vertices and E is the number of edges in the graph.
- **Space Complexity:** O(V), since an extra visited array of size V is required.

CONCLUSION

- This survey extensively reviews and evaluates current fake news research by defining fake news.
- We can detect source using dfs algorithm
- Its time complexity O(V+E)

Reference

- ➢ https://arxiv.org/abs/1812.00315
- ➢ https://ieeexplore.ieee.org/document/9395133
- ➢ https://zenodo.org/record/4561253#.Yn1J9lxBzIU
- ➢ https://www.nature.com/articles/s41598-018-20546-3

Important Links

- ➢ PPT - https://docs.google.com/presentation/d/1Wd7zBXvniGZZNnPVr1S2_KKZljczSZxRkQTn9LpQ6xw/edit?usp=sharing
- ➢ Code https://colab.research.google.com/drive/1YgL9sxX8ieWs_BXIwWMd70_FRQTBZE37?usp=sharing

THANK YOU