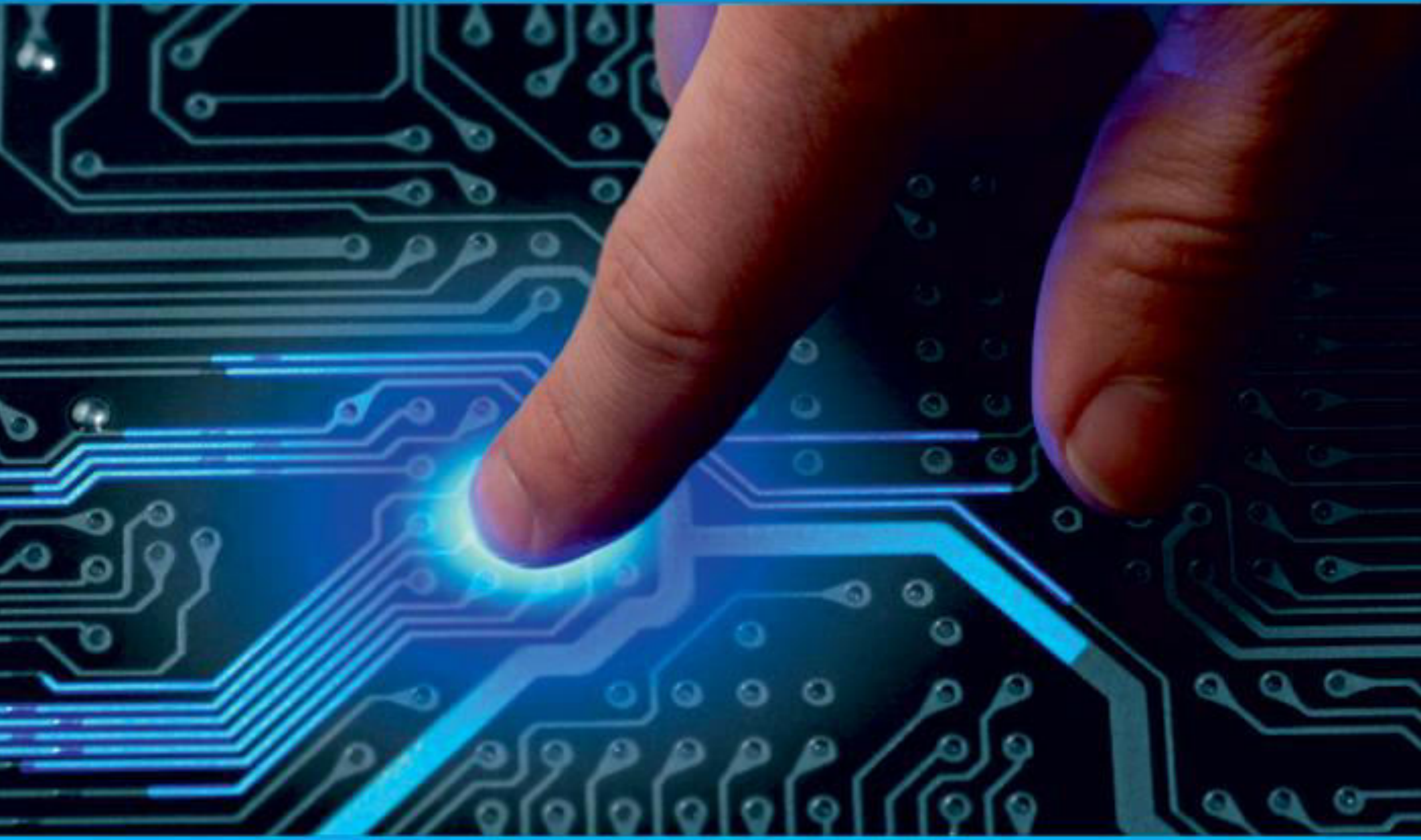
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005269 |

Fake News Detection using Machine Learning Framework

Nikhat Shaikh, Prof. Sapike N.S.

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahmednagar, India

ABSTRACT: it is a popular method for the broadcast of real-time news all around the world on social media. One of the reasons for its appeal is the easy and speedy dissemination of information. Social media websites are used by a large number of people of various ages, genders, and societal ideas. Despite these advantages, there is a big drawback in the form of fake news, as most individuals read and distribute information without questioning its authenticity. It is critical to investigate news authentication methods. This article offers a two-phase benchmark approach for false news detection using machine learning classification called FakeNews, which is built on word embedding (WE) over linguistic variables. The first phase uses linguistic features to preprocess the data set and assess the veracity of news content. In the second phase, the linguistic feature sets are combined with WE and classification is used. This study also meticulously creates a fresh FakeNewsdata set with thousands of articles to validate its approach, which includes many data sets to give an unbiased categorization output.

KEYWORDS: Fake News, User Profile, Trust Analysis, machine learning, Social Media

I. INTRODUCTION

The reliability of information diffused on the World Wide Web (WWW) is a central issue of modern society. In particular, in the recent years the spreading of misinformation and fake news on the Internet has drawn increasing attention, and has reached the point of dramatically influencing political and social realities. As an example, showed the significant impact of fake news in the context of the 2016 US presidential elections; analyzed the most viral tweets related to the Boston Marathon blasts in 2013, finding that the share of rumors and fake content was higher than the share of true information. As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to seek out and consume news from social media rather than traditional news organizations. The reasons for this change in consumption behaviors are inherent in the nature of these social media platforms: (i) it is often more timely and less expensive to consume news on social media compared with traditional news media, such as newspapers or television; and it is easier to further share, comment on, and discuss the news with friends or other readers on social media. For example, 62 percent of U.S. adults get news on social media in 2016, while in 2012; only 49 percent reported seeing news on social media1. It was also found that social media now outperforms television as the major news source2. Despite the advantages provided by social media, the quality of news on social media is lower than traditional news organizations.

However, because it is cheap to provide news online and much faster and easier to disseminate through social media, large volumes of fake news, i.e., those news articles with intentionally false information, are produced online for a variety of purposes, such as financial and political gain. It was estimated that over 1 million tweets are related to fake news ”3 by the end of the presidential election. Given the prevalence of this new phenomenon, news” was even named the word of the year by the Macquarie dictionary in 2016. The extensive spread of fake news can have a serious negative impact on individuals and society. First, fake news can break the authenticity balance of the news ecosystem. For example, it is evident that the most popular fake news was even more widely spread on Facebook than the most popular authentic mainstream news during the U.S. 2016 president election. Second, fake news intentionally persuades consumers to accept biased or false beliefs. Fake news is usually manipulated by propagandists to convey political messages or influence. For example, some report shows that Russia has created fake accounts and social bots to spread false stories. Third, fake news changes the way people interpret and respond to real news. For example, some fake news was just created to trigger people's distrust and make them confused; impeding their abilities to differentiate what is true from what is not6. To help mitigate the negative effects caused by fake news-both to benefit the public and the news ecosystem-It's critical that we develop methods to automatically detect fake news on social media.

Detecting fake news on social media poses several new and challenging research problems. Though fake news itself is not a new problem-nation or groups have been using the news media to execute propaganda or influence operations for centuries-the rise of web-generated news on social media makes fake news a more powerful force that challenges tradi-tional journalistic norms. There are several characteristics of this problem that make it uniquely challenging for automated

detection. First, fake news is intentionally written to mislead readers, which makes it nontrivial to detect simply based on news content. The content of fake news is rather diverse in terms of topics, styles and media platforms, and fake news attempts to distort truth with diverse linguistic styles while simultaneously mocking true news. For example, fake news may cite true evidence within the incorrect context to support a non-factual claim. Thus, existing hand-crafted and data-specific textual features are generally not sufficient for fake news detection. Other auxiliary information must also be applied to improve detection, such as knowledge base and user social engagements. Second, exploiting this auxiliary information actually leads to another critical challenge: the quality of the data itself. Fake news is usually related to newly emerging, time-critical events, which may not have been properly verified by existing knowledge bases due to the lack of corroborating evidence or claims. In addition, users' social engagements with fake news produce data that is big, incomplete, unstructured, and noisy. Effective methods to differentiate credible users extract useful post features and exploit network interactions are an open area of research and need further investigations.

A. Motivation

Acute myocardial infarction, commonly referred to as Ar-rhythmia diseases is the most common cause for sudden deaths in city and village areas. It is one the most dangerous disease among men and women and early identification and treatment is the best available option for the people.

B. Objectives

1. The first one is fake news where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult.

2.For example, on Social media rumors, scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread fake news and obscure the truth.

3.The second challenge is data sparsity or the long-tail phenomenon where a majority of sources only contribute a small number of claims, providing insufficient evidence to determine those sources trustworthiness.

II. REVIEW OF LITERATURE

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers. "In this section, we briefly review the related work on fake news detection system and their different techniques.

In this paper [1], the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Text blob, Natural Language, and SciPy Tool kits were used to develop a novel fake news detector. Advantages-1. Used natural language processing2. Fake news detection based on attribute classification Disadvantages-Time consuming process.

This paper [2] introduce the data sets which contain both fake and real news and conduct various experiments to organize fake news detector.Advantages is 1. Used Natural Language Processing, Machine learning and deep learning techniques to classify the data sets 2.Accuracy is better and disadvantages is use Limited data set.

This paper [3] proposed a distributed framework to imple-ment the proposed truth discovery scheme using Work Queue in an HTCondorsystem.Advantages is 1. Find trustworthy in-formation on Social media2.Proposed truth discovery scheme using Work Queue in an HTCondor system and disadvantages is Accuracy is low

This Paper [4] Studied various detection techniques i.e. con-tent based, social context based and hybrid based. Advantages is Proposed content-based, social context-based and hybrid-based methods and disadvantages is only survey state of the methods.

This paper [5] Present a new fake news detection model using unified key sentence information which can efficiently perform sentence matching between question and article by us-ing key sentence retrieval based on bilateral multi perspective matching model. Advantages is Implement natural language processing using key sentence retrieval and disadvantages is Fake news detection accuracy is low.

This Paper [6] classifies fake news messages from Twitter posts using hybrid of convolutional neural networks and long-short term recurrent neural network models. Advantages is Implement hybrid CNN and RNN Models and Accuracy is much better. Disadvantages is only consider tweet headlines.

This paper [7] Compare news to other sources in 2016 year. Advantages is 1.detect 2016 election fake news spread through social media2. Goal in this paper is to offer theoretical and empirical background to frame this debate. Disadvantages is1. Limited dataset used2. Limited to 2016 news only.

This paper [8] shows a new approach for fake news detec-tion using naive Bayes classifier. Use Implement na¨ıve bayes machine learning algorithm but accuracy is low.

This paper [9] introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, we reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction. Advantages is in this paper, they explored the fake news problem by reviewing existing literature in two phases i.e. characterization and detection but on Use static data.

This study [10] contributes to the scientific knowledge re-garding the influence of the interaction between various types of media use on political effects. Advantage is Used multiple news sources for fake news detection and disadvantage is Focus on only political data.

III. PROPOSED METHODOLOGY

A. Methodology to solve the task

We propose a new method called Fake News exclusively focused on text data in stages are Fake news prediction using linguistic feature sets(LFS).WE over LFS for improved fake news detection over a WELFake data set and Comparative analysis of the linguistic features based results with state-of-the-art CNN and BERT methods. Proposed method improves accuracy using a novel method based on four stages that:

- Creates a larger data set with improved generalization.
- Identifies the most significant twenty linguistic features and creates three unique LFS based on categories.
- Applies two WE methods to train various ML models.
- Generates the final prediction using a two-stage voting classification.

B. Architecture

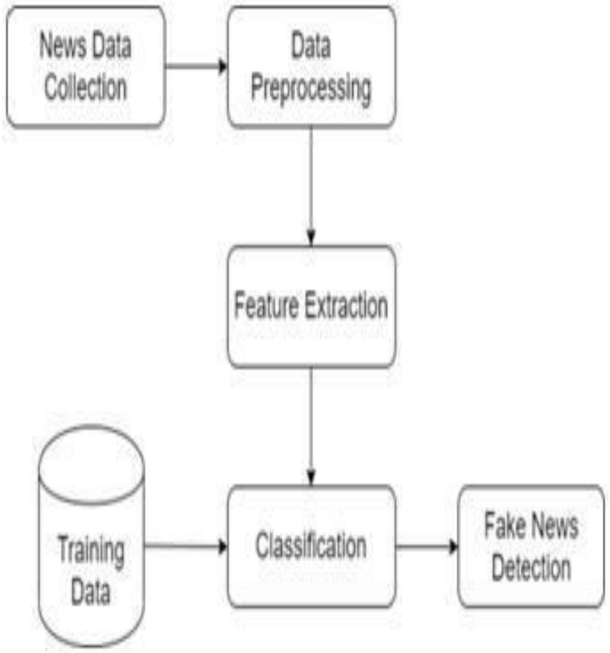


Fig. 1. System Architecture

C. Algorithm

1.Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

This is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The Random Forests Algorithm handles some of the limitations of Decision Trees Algorithm, namely that the accuracy of the outcome decreases when the number of decisions in the tree increases. So, in the Random Forests Algorithm, there are multiple decision trees that represent various statistical probabilities. All of these trees are mapped to a single tree known as the CART model. (Classification and Regression Trees). In the end, the final prediction for the Random Forests Algorithm is obtained by polling the results of all the decision trees.

2.Decision Tree

Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this.

Decision tree learning is a method commonly used in data mining.] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the ”classification”. Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes, signifying that the data set has been classified by the tree into either a specific class, or into a particular probability distribution (which, if the decision tree is well-constructed, is
skewed towards certain subsets of classes).

A tree is built by splitting the source set, constituting the root node of the tree, into subsets - which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

IV. CONCLUSION

Growing popularity of social media, more and more people consume social media news instead of traditional media. However, social media have also been used to disseminate false news, which has strong negative impacts on individual users and the wider society. Here to explore the problem of false news by reviewing existing literature in two phases: characterization and detection. In the characterization phase, we introduce the basic concepts and principles of false news in both traditional media and social media. In the detection phase, we reviewed the current false news detection approaches from a machine learning perspective, including feature extraction and model building. We also discuss evaluation metrics, and future promising directions in fake detection research and expand the field to other applications.

REFERENCES

[1] Terry Traylor, Jeremy Straub, Gurmeet, Nicholas Snell” Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator”2019.

[2] Rohit Kumar Kaliyar” Fake News Detection Using A Deep Neural Network”2018.

[3] Daniel (Yue) Zhang, Dong Wang, Nathan Vance, Yang Zhang, and Steven Mike” On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications” 2018.

[4] ZaitulIradahMahid,SelvakumarManickam,ShankarKaruppayah” Fake News on Social Media: Brief Review on Detection Techniques” 2018.

[5] Namwon Kim, DeokjinSeo, Chang-Sung Jeong” FAMOUS: Fake News Detection Model based on Unified Key Sentence Information” 2018.

[6] OluwaseunAjao, DeepayanBhowmik, ShahrzadZargari” Fake News Iden-tification on Twitter with Hybrid CNN and RNN Models”2018.

[7] Hunt Allcott Matthew Gentzkow” SOCIAL MEDIA AND FAKE NEWS IN THE 2016 ELECTION” 2017

[8] MykhailoGranik, VolodymyrMesyura” Fake News Detection Using Naive Bayes Classifier”2017

[9] Kai Shu , Amy Sliva , Suhang Wang , Jiliang Tang , and Huan Liu” Fake News Detection on Social Media: A Data Mining Perspective”2016

[10] MeitalBalmas” When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism” 2014

[11] Daniel (Yue) Zhang, Chao Zheng, Dong Wang, Doug Thain, Chao Huang, Xin Mu, Greg Madey .”Towards Scalable and Dynamic Social Sensing Using A Distributed Computing Framework.” Department of Computer

Science and Engineering Department of Aerospace and Me-chanical Engineering University of Notre Dame Notre Dame, IN, USA IEEE 2017.

[12] Daniel (Yue) Zhang, Dong Wang, HaoZheng, Xin Mu, Qi Li , Yang Zhang." Large-scale Point-of-Interest Category Prediction Using Natural Language Processing Models." Department of Computer Science and Engineering Department of Aerospace and Mechanical Engineering Uni-versity of Notre Dame Notre Dame, IN, USA IEEE 2017.

[13] Daniel (Yue) Zhang, Rungang Han, Dong Wang, Chao Huang." On Robust Truth Discovery in Sparse Social Media Sensing." Department of Computer Science and Engineering University of Notre Dame Notre Dame, IN, USA IEEE 2016

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com