International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Fake News Detection using Machine Learning

Pallavi Petkar and Dr. S. S. Sonawan

Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

The extensive spread of fake news (low quality news with intentionally false information) has the potential for extremely negative impacts on individuals, society and particular in the political world. Therefore, fake news detection on social media has recently become an emerging research which is attracting tremendous attention. The Fake News Challenge was encourage for development of machine learning- based classification system that perform stance detection that is to identify whether a particular news headline is related or is unrelated to a particular news article to allow journalists and others to easily investigate possible instance of fake news. This is technically challenging for several reasons. Use of various social media tools, content is easily generated and quickly spread, which lead to a large volume of content to analyze. Online information is very wide spread, which cover a large number of subjects, which contributes complexity to this task. More attention needed on real time detection as rumors and bad news are particularly hard to fix once they spread. The application of machine learning techniques are explored for the detection of fake news that come from non- reputable sources which mislead real news stories. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information.

Keywords: *Fake news detector, Stance detection, Fake news categorization, content modeling, Machine learning, Social media, online fake news, twitter.*

Introduction

Fake news identification has attracted growing public and researcher attention as the dissemination of misinformation online increases, especially on social media such as media feeds, news forums, and online news articles. Fake news can be any content that is not true and generated to convince its readers to believe in something that is untrue. Fake news intentionally misleads people into believing false information and shifting people's response to real news [4]. Fake news identification from online social media is extremely challenging due to various factors. First, the collection of fake news data is difficult, and it is also difficult to manually label fake news. Second people write fakenews.Many liars have a tendency to deliberately use their words to avoid being caught. Because of the booming trends of online social networks, fake news has emerged in large numbers and in the online world for various commercial and political purposes. Online social network users can quickly get infected with this digital fake news with misleading words, which has already had enormous effects on the offline community.

A. Need and Motivation

The viral spread of misinformation may result in affecting the reliability of the news ecosystem, damaging the reputation of any personal or organization and cause the fear among general public which will weaken the social stability [3]. Generated fake news are very hard to detect based on their content of the news because the language used in fake news is similar to the language used in true news, as the fake news are created with the intention to be trust the public. Hence there is need of fake news detection.

B. Challenges

The main challenge of fake news detection is to identify the part of information is a fact or not. Fact is a simple conception which is composed of something that has happened at some time, somewhere, eventually with someone. It seems clear to recognize the value of information should not be easy to perform by machines in the case they are given total control to decide which information is displayed to whom, when and which To capture the essence of the journalistic criteria to find out the information to report on and it makes a difference because so many posts on social media follow general idea of reporting on something [7].

C. Fake news Detection

1. Definition

Fake news is the false information which is intentionally written to mislead people [1]. There are main two key features of this definition that is authenticity and intent. The very first, fake news contain false information that can be verified as such and second, fake news is created with dishonest intention to mislead readers. The viral spread of misinformation can result in serious problem such as affecting the reliability of the news ecosystem, damaging the reputation of any personal or organization, or causing fear among the public that can weaken the social stability [9].

Given a set of "m" news article which contain "text" and "headline", so that the data can be represent as set of headline and text tuples A= {(A^H , A^T)}. In the fake news detection problem, it predicts whether the given piece of information is fake or not [3]. The following three type of fake news detection which are mention as follows:

Fabrication: Fabrication news is an intentional lie that does not usually go beyond one source. The source is probably aware that the story is false. Fabrication news heavily depends on clickbait.

Hoax: This type of news use more sophisticated methods of fooling audience. Hoax news are spread by multiple sources. Some of which may believe that the story is true. This type of news seen on various sources, for example the false news of Donald trump during election was spread through various social media like twitter, facebook, blogs etc., so public can easily trust on such type of news

Satire: A False news story that the source presents as true as a joke. When satire is shared with people that not familiar with the source. There is always a chance someone will think it is real.

2. Approaches

The study of fake news detection performed by various researched by using different approaches [9] which are discussed as follows:

Knowledge-based: Knowledge based approaches based on fact-checking by using external sources which proposed claims in news content.

The main goal of fact-checking is to assign a true value to a claim in a particular context of the news article. Fact checking approaches can be categorized as crowdsourcingoriented, expert-oriented and computational-oriented.

Style-based: Style-based approaches detect fake news by capturing the manipulators in the writing style of news content. There are mainly two categories of style-based methods first is Deception-oriented and second is Objectivity-oriented [11]. In Deception-oriented capture the deceptive statements or claims from news content. Objectivity-oriented approaches capture style signals that indicate a decreased objectivity of news content.

Stance-based: Stance-based approaches utilize users viewpoint from relevant post contents to check the veracity of original news articles. Stance of users can be represented as explicitly or implicitly. Stance based approaches used on social context.

Propagation-based: Propagation-based approaches focused on how fake news spread and interrelations of social media posts to predict news credibility. Propagation-based approaches used on social context.

Literature Review

1] The study of literature survey show the use of different machine learning algorithm using different parameter which helps to increase the accuracy of research. The following two papers [1][6] used the classification algorithm.

Bilal Ghanem and Paolo Rosso [1] presented an approach which combines lexical, word embedding and n-gram features for detecting the stance in fake news. This paper used SVM classifier to extract most important category to each classification class and NN for extracting important feature. Paper investigated the importance of different lexicons in the detection of the classification labels .

2] The following paper [2][3][4][5][7] used machine learning with deep learning algorithm to increase the accuracy of the model.

Ali K. Chaudhry and Darren Baker [2] developed several deep neural network-based models to tackle the stance detection problem, ranging from relatively simple feed- forward networks to elaborate recurrent models featuring attention and multiple vocabularies.

Proposed model used to identify whether a particular news headline "agrees" with, "disagrees" with, "discusses," or is unrelated to a particular news article.

Yang Yang and Lei Zheng [3] proposed a model named as TI-CNN (named as Text and Image information based Convolutional Neural Network) which combine the text and image information with the explicit and latent features. The proposed model h can easily absorb other features of news. TI-CNN model is trained with both the text and image information simultaneously.

Gayathri Rajendran and Bhadrachalam Chitturi [4] employed deep neural networks for feature extraction and stance classification. RNN models with its extensions showed the significant variations in the classification of detailed class. Bidirectional LSTM model give the best accuracy for broad as well as detailed class classification. The research work extended on stance classification with extra features like user or topic or sentiment based features.

Sherry Girgis and Eslam Amer [5] used deep learning with the large dataset which increase the learning and get best result using word embedding for extracting features or cues that distinguish relations between words in syntactic and semantics. Mykhailo Granik and Volodymyr Mesyura [6] used simple approach for fake news detection using Naïve Bayes classifier.

The main idea used in this paper is to treat each word of the news article independently. Rachana Kunapareddy and SriRohitha Madala [7] proposed the model which utilizes the machine learning and deep learning to recognize fake news problem by testing against an informational index of newspost.

Proposed result showed that phony news identification can be tended by using SVM, Random forest and CNN.

3] The paper [8] used hybrid approach with some inference parameter.

Oluwaseun Ajao and Deepayan Bhowmik [8] Proposed framework that consider inference of tweet geo-location and origin of fake news and author which quickly identifies the fake news stories.
Bilal Ganem and Paolo Rosso [10] used LSTM neural network model which infused-emotionally to detect false news. This research shows that fake information has different emotional patterns of various news.

Kashyap Popat and Subhabrata Mukherjee [11] present neural network model that aggregate with signal of external evidence, language and trustworthiness of the sources.This research experiment on four different dataset which overcome the limitations.

Fatemeh Torabi Asr and Maite Taboada [12] introduced introduce the Misinfotext repository and performed topic modeling experiment to elaborate the gaps. This work used natural language processing to detect automatically the false information.

The above literature survey used various parameter such as:

- Context-based feature
- Intelligent feature extraction
- Location feature extraction Sentiment-based feature extraction

System Architecture

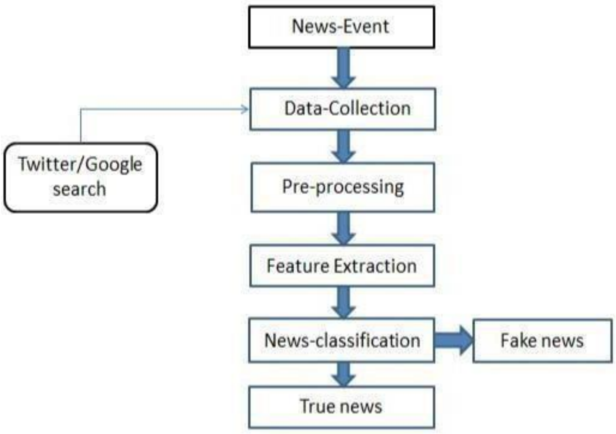

Fig.1 General System Architecture

The above figure [1] is the brief introduction of system architecture which involve various task for analyzing the text of information is fake or not:

A. News Content Feature

Source: Author and publisher of the news article

Headline: Short headline text which catch the attention of readers and it describes the main topic of the news article.

Body Text: Main text that elaborates the details of the news story; there is usually a major claim that is specifically highlighted and that shapes the angle of the publisher.

Social Context Feature User-Based: User-based features provide useful information for fake news detection problem. This represent the characteristics of user who have interactions with social media news. User based feature categorized into two level, first is individual and second is group level.

Network - Based: The network based feature characterize the user who form different networks on social media in terms of interests, relations, and topics. Network-based features extracted by creating specific networks among the users who published related social media posts.

Post-Based: Post-based features aim on identifying useful information to work out the veracity of news from different aspects of relevant social media news posts People can express their emotions or opinions towards fake news through social media posts [12]. Thus, it help to extract post based feature to find potential fake news by reaction of the public expressed in post

Twitter data

Create the authentication with tweeter for extracting the news tweet, similar to google search topic and preprocess the tweet contains a lot of opinions about the data which are shared by different user in different ways. The twitter data is in raw form which is inconsistence and redundant. Thus extract the tweets and preprocess tweets in specific format which helps to make easy to tackle the problem.

B. Pre-processing

After fetching process, clean and preprocess the data by removing the following point which makes easy to analyze the data:

- Remove all URLs, hash tags (e.g. #topic), targets (@username)
- Correct the spellings and handled the sequence of repeated characters.
- Remove all punctuations, symbols, numbers
- Remove Stop Words
- Replace emoticons
- Non- English words

C. Classification Based on Similarity score

Once the data is clean, check the similarity (Contextual similarity, text similarity) between tweet text and google search link text and headline After calculating the similarity score, if the similarity of the text is greater than threshold , the detected news is true otherwise it is detected as false.

Result

Table1.Comparison of Table

News	News Type	Links/ Tweets	True/ False	Cosine Similarity	Jaccard cosine similarity
Sample 1	Politics	2	True	0.020	0.026
Sample 2	International politics-based news	4	True	0.071	0.0030
Sample 3	Politics-based news	5	True	0.033	0.00209
Sample 4	Social-based news	4	False	0.066	0.0

The above table [1] shows the comparison of the similarity of the various type of news. The comparison proved that the similarity is zero when the news are fake but cosine similarity give some value, hence cosine similarity not always work for any type of news.

Table 2.Test Cases

Sr. No	Test Case	Input Parameter	Expected Similarity	Actual Smilarity
1	To check whether News is Real	Real News	>0	0.020
2	To check whether News is Fake	Fake News	0.000	0.066
3	To check whether News is Fake	Fake News	0.000	0.000

The similarity is more when the large amount of text and more tweets are posted by the public for given news. The existing system used labeled train dataset which give high accuracy as compare to real time data.

Conclusion

The spread of fake news has raised concerns all over the world recently. These political fake news may have severe consequences. The identification of the fake news grows in importance. More attention needed on real time detection as rumors and bad news are particularly hard to fix once they spread.By focusing fake news problem as a text classification problem that is attempting to automatically detect whether a particular news article is fake or not. Fake mean an article that contains unverified or untrue claims, or attempts to disseminate information that is not accurate. In order to perform automatic classification of news texts, modern NLP and machine learning methods require large amounts of training data.

References

[1]. Bilal Ghanem and Paolo Rosso, " Stance Detection in Fake News: A Combined Feature Representation", Association for Computational Linguistics , November 1, 2018

[2]. Ali K. Chaudhry and Darren Baker," Stance Detection for the Fake News Challenge: Identifying Textual Relationships with Deep Neural Nets", standford.edu , 2018.

[3]. Yang Yang and Lei Zheng, " TI-CNN: Convolutional Neural Networks for Fake News Detection", arXiv:1806.00749v1 [cs.CL] 3 June 2018.

[4]. Gayatri Rajendran and Bhadrachalam Chitturi , "StanceIn-Depth deep neural approach for stance classi_cation", elsevier 2018.

[5]. Sherry Girgis and Eslam Amer, " Deep Learning Algorithms for Detecting Fake News in Online Text", ICCES2018.

[6]. Mykhailo Granik and Volodymyr Mesyura, " Fake News Detection Using Naive Bayes Classifier ", UKRCON 2017 IEEE.

[7]. Rachana Kunapareddy and SriRohitha Madala, " False Content Detection with Deep Learning Techniques ", (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019.

[8]. Oluwaseun Ajao and Deepayan Bhowmik, " Fake News Identification on Twitter with Hybrid CNN and

[9]. RNN Models" ,SMSociety, July 2018, Copenhagen, Denmark. Kai Shu and Amy Sliva, " Fake News Detection on Social Media: A Data MiningPerspective", IEEE 2016.

[10]. Bilal Ganem and Paolo Rosso "An Emotional Analysis of False Information in Social Media and News Articles" IEEE 2018.

[11]. KashyapPopat and Subhabrata Mukherjee "DeClarE: Debunking Fake News and False Claims using Evidence-Aware DeepLearning" IEEE 2018.

[12]. Fatemeh Torabi Asr and Maite Taboada " Big Data and quality data for fake news and misinformation detection" Big Data & Society January–June 2019.