ISSN 2320 - 2602

Volume 10 No.6, June 2021

International Journal of Advances in Computer Science and Technology

Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst021062021.pdf

https://doi.org/10.30534/ijacst/2021/021062021

Fake News Detection System Using machine Learning

Anagha S Anand[1], Aneeta Solaman[2], Berin John[3] ,Vineeta Samson[4],Asst Prof Teenu Jose[5]

[1]Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Kerala, India,anandanagha110@gmail.com

[2]Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Kerala, India,aneetasolaman1999@gmail.com

[3] Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Kerala, India,berinjohn98@gmail.com

[4]Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Kerala, India,vineetasamson@gmail.com

[5]Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Kerala, India, teenujose@aisat.ac.in

ABSTRACT

Expansion of deluding data in ordinary access news sources, for example, web-based media channels, news web journals, and online papers have made it testing to distinguish reliable news sources, hence expanding the requirement for computational apparatuses ready to give bits of knowledge into the unwavering quality of online substance. In this paper, every person center around the programmed ID of phony substance in the news stories. In the first place, all of us present a dataset for the undertaking of phony news identification. All and sundry depict the pre-preparing, highlight extraction, characterization and forecast measure in detail. We've utilized Logistic Regression language handling strategies to order counterfeit news. The prepreparing capacities play out certain tasks like tokenizing, stemming and exploratory information examination like reaction variable conveyance and information quality check (for example invalid or missing qualities). Straightforward pack of-words, n-grams, TF-IDF is utilized as highlight extraction strategies. Strategic relapse model is utilized as classifier for counterfeit news identification with likelihood of truth.

Key words: Fake news detection, Logistic regression, TF-IDF vectorization.

1. INTRODUCTION

Counterfeit news spread can't be undermined as it ready to convey adverse consequences to the general population for a since a long time ago run. Tricky issues may emerge from counterfeit news, for example, defamations, disarray and confusions and provocative untruths until up to a level conclusion issues played by flippant gatherings or people who love to spread disdain and devastation among one after another. A news that has been controlled or created in its substance with things that are random, completely or somewhat bogus is classified as phony news. Identifying the news is being phony or on the other hand not phony is truly troublesome. However, within the ability of man-made brainpower in AI, this is made conceivable to recognize counterfeit news. A few nations have shown their responsibility in managing counterfeit news because of its expected effect on the society. At the end of the day, counterfeit news is truly deluding and impacting individuals to accept on something that isn't valid and most likely have been controlled. Sometime in the past on the off chance that anybody required any news, the person in question would sit tight for the following day paper. Nonetheless, with the development of online papers who update news quickly, individuals have discovered a superior and quicker approach to be educated regarding the issue of his/her advantage. These days informal communication frameworks, online news entryways, and other on the web media have become the primary wellsprings of information through which fascinating and breaking news are shared at a fast pace. Not with standing, numerous news entries serve exceptional interest by taking care of with mutilated, in part right, and at times fanciful news that is probably going to draw in the consideration of an objective gathering of individuals. Counterfeit news has become a significant concern for being dangerous once in a while spreading disarray and purposeful disinformation among individuals. The term counterfeit news has become a popular expression nowadays. Notwithstanding, a concurred meaning of the expression "counterfeit news is still to be found. It may very well be characterized as a kind of sensationalist reporting or publicity that consists of purposeful falsehood or lies spread by means of conventional print and broadcast news media or online web-based media. These are distributed for the most part with the goal to misdirect to harm a local area or individual, make tumult, and gain monetarily or strategically. Since individuals are frequently unfit to invest sufficient energy to cross-check references and make certain of the validity of information, mechanized discovery of phone news is essential. Accordingly, it is accepting incredible consideration from the examination local area.

2. LITERATURE REVIEW

In general, Fake news might be categorized into three groups. the primary group is fake news, which is news that is completely fake and is formed up by the writers of the articles. The second group is fake satire news, which is fake news whose main purpose is to supply humour to the readers.

The third group is poorly written news articles, which have a point of real news, but they're not entirely accurate. In short, it's news that uses, for example, quotes from political figures to report a totally fake story. Usually, this type of stories is meant to market certain agenda or biased opinion [1]. In the article published by Kai Shu, Amy SlivaSuhang Wang, Jiliang Tang, and Huan Liu [2], they explored the fake news problem by reviewing existing literature in two phases: characterization and detection. In the characterization phase, they introduced the essential concepts and principles of faux news in both traditional media and social media. within the detection phase, they reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction.

Hadeer Ahmed, Issa Traore, and Sherif Saad [3] proposed in their paper, a fake news detection model that uses n-gram analysis and machine learning techniques. They investigated and compared two different features extraction techniques and 6 different machine classification techniques. Experimental evaluation yields the best performance using Term FrequencyInverted Document Frequency (TF-IDF) as feature extraction technique, and Linear Support Vector Machine (LSVM) as a classifier, with an accuracy of 92%. Perez-Rosas, Veronica & Kleinberg, Bennett and Lefevre Alexandra and Rada Mihalcea [4] in their publication “Automatic detection of faux news” specialise in the automatic identification of faux contents in online news. For this they introduced two different datasets, one obtained through crowd sourcing and covering six news domains (sports, business, entertainment, politics, technology and education) and another one obtained from the web covering celebrities. They developed some classification models using linear sum classifier and fivefold cross verification with accuracy, precision and recall and FI measures averaged over the five iterations that rely on the mixture of lexical, syntactic and semantic information also as features representing text readability properties which are like human ability to identify fakes. E.M Okoro, B.A Abara, A.O. Umagba, A.A. Ajonyeand Z. S. Isa [5] in their publication _A Hybrid approach to fake news detection on social media employing a combination of both human-based and machine-based approaches. Since traditional and machine based approaches have some limitations and can’t single handedly solve the problems like human literacy and cognitive limitations and the inadequacy of machine based approach. To solve all these problems, they proposed a Machine Human (MH) model for fake news detection in social media. This model combines the human literacy news detection tool and machine linguistic and network-based approaches. This way, the 2 parallel approaches of detection are at work, each helping to supply a balance for the opposite . The existing system and research work reveal that the majority classification algorithms perform well to detect or predict the fakeness of a news story . Though the logistic regression serves well for this purpose, our system is based on this information and thus we focus to figure with classification algorithms just like the logistic regression.

3. METHODOLOGY

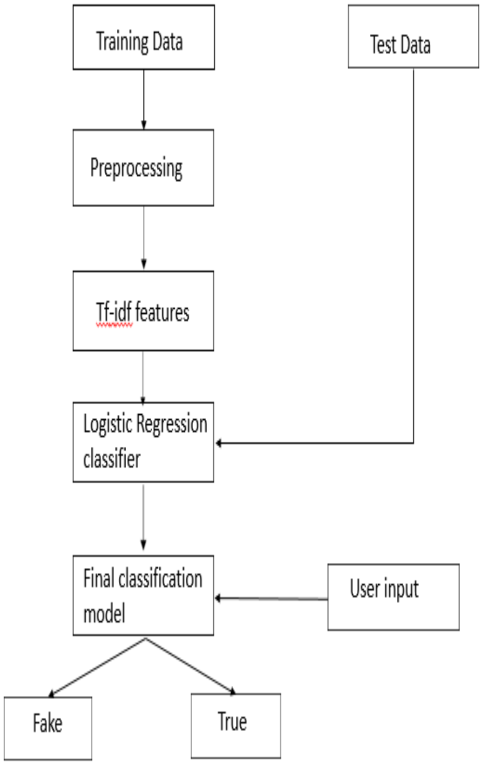


Figure 1:flowchart of the proposed system

3.1 Data Preprocessing

This module contains all the pre processing functions needed to process all the input documents and texts. First we read the train, test and validation data files then perform some pre processing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

3.2 Stemming

In linguistic morphology and knowledge retrieval, stemming is that the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a word form. The stem needn't be just like the morphological root of the word; it's usually sufficient that related words map to an equivalent stem, even if this stem isn't in itself a legitimate root.

3.3 Tokenizing

Tokenization is that the process of replacing sensitive data with unique identification symbols that retain all the essential information about the info without compromising its security. Tokenization, which seeks to attenuate the quantity of knowledge a business must keep it up hand, has become a well-liked way for little and mid-sized businesses to bolster the security of mastercard and e-commerce transactions while minimizing the value and complexity of compliance with industry standards and government regulations.

3.4 Feature Selection

In this module we've performed feature extraction and selection methods from sci-kit learn python libraries.

3.5 Count Feature

The CountVectorizer provides an easy thanks to both tokenize a set of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. you'll use it as follows: 1. Create an instance of the CountVectorizer class. 2. Call the fit() function so as to find out a vocabulary from one or more document. 3. Call the transform() function on one pr more document as required to encode each as vector.

An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared within the document. Because these vectors will contain tons of zeros, we call them sparse. Python provides an efficient way of handling sparse vectors in the scipy.sparse package. The vectors returned from a call to transform() are going to be sparse vectors,and you'll transform them back to numpy arrays to seem and better understand what's happening by calling the to array() function.

3.6 Classifier

In this module everybody build all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. One and all used Logistic Regression classifier from sklearn. Each of the extracted features were utilized in the classifier.Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, two best performing models were selected as candidate models for fake news classification.Finally selected model was used for fake news detection with the probability of truth. additionally to this, also extracted the highest 50 features from our term-frequency tfidf Vectorizer to ascertain what words are most and important in each of the classes. All of us have also used Precision-Recall and learning curves to see how training and test sets perform once everybody increases the quantity of knowledge in our classifiers.

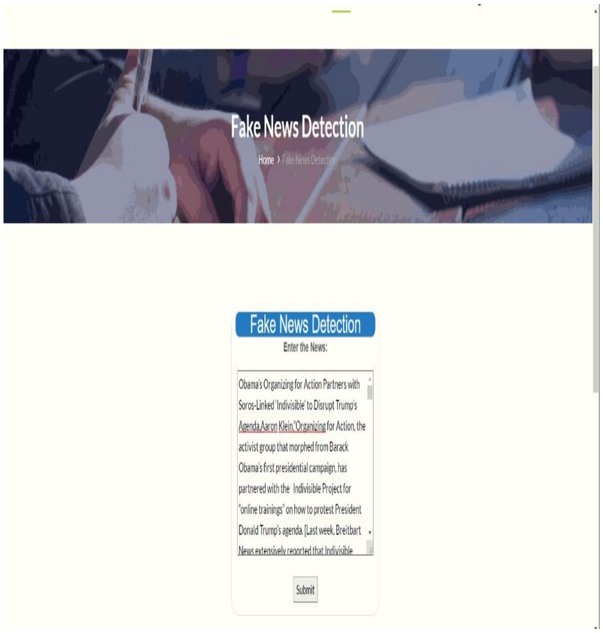
3.7 Logistic Regression

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable may be a binary variable that contains data coded as 1 (yes, success, etc) or 0 (no, failure, etc.).In other words, the logistic regression model predicts P(Y=1) as a function of X.

4. RESULT AND CONCLUSION

In this paper, each person used Logistic Regression classifier which can serve the model and work with the user input. Here, ourselves presented a detection model for fake news using TFIDF analysis through the lenses of different feature extraction techniques. Everyone have investigated different feature extraction and machine learning techniques. The proposed model achieves accuracy of roughly 92% when using TF-IDF features and logistic regression classifier. After testing the data, the result will be an exactness of 0.92% and F1 score of 0.923.

4.1 Input

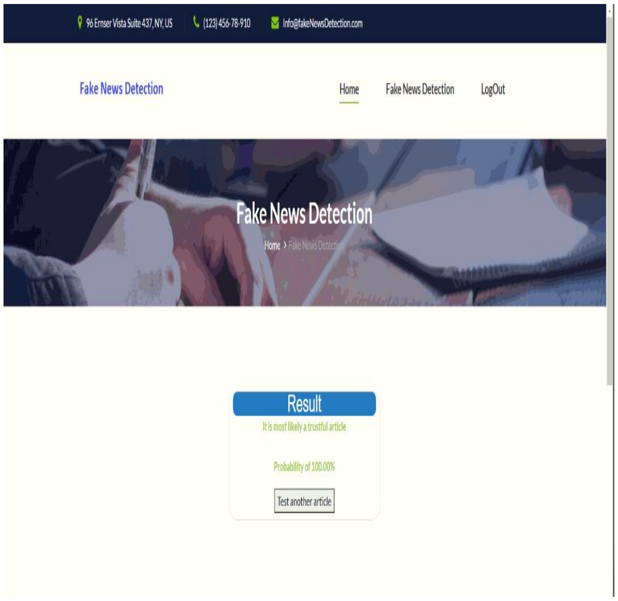


Obama's Organizing for Action Partners with Soros-Linked 'Indivisible' to Disrupt Trump's Agenda,Aaron Klein,"Organizing for Action, the activist group that morphed from Barack Obama's first presidential campaign, has partnered with the Indivisible Project for "online trainings" on how to protest President Donald Trump's agenda. [Last week, Breitbart News extensively reported that Indivisible

Figure 2: user entering a news

Figure 2 shows the user interface where input values are submitted to the system.

4.2 Output



It is most likely a trustful article

Probability of 100.00%

Figure 3: Result shown after prediction

Figure 3 is the snapshot of output screen which gives the output values to user.

REFERENCES

[1]. Schow**, "A:The 4 Types of 'Fake News' "**. observer (2017). http://observer.com/2017/01/fake- news- russia hacking-clinton-loss/.

[2]. **"Fake News Detection on Social Media: A Data Mining Perpective".** Kai Shu, Amy Sliva, Jiliang Tang,and Huan Liu Computer Science & Engineering, Arizona State UniversityTempe, AZ, USA CharlesRiver Analytics, Cambridge,MA, USA Computer Science & Engineering ,Michigan State University,East Lansing, USA .

[3].**"Detection of Online Fake News Using N-Gram Analysis And machine learning technique".**Hadeer Ahmed, Issa Traore, and Sherif Saad ECE Department, University of Victoria, Victoria, BC, Canada School of Computer Science, University of Windsor, Windsor, ON, Canada.

[4]. Verónica Pérez - Rosas, Kleinberg Bennett, Alexandra Lefevre and RadaMihalcea, **"Automatic detection of Fake news",** ||proceedings of the 27th International Conference on Computational Linguistics, pp. 3391–3401,Santa Fe, New Mexico, USA, 2018.

[5]. E. M. Okoro, B. A. Abara, A. O. Umagba, A.A. Ajonye, And Z.S Isa**,—"Hybrid Approach to Fake news detection on social media"**,‖ vol. 37, no. 2, pp. 454-462, 2018.

[6]. Metz C (2016), **"The bittersweet sweepstakes to build an AI destroys fake news''**, Dec 2016 (Online). Availablehttps://www.wired.com/2016/12/bittersweet-sweepstakesbuild-aidestroysfake-news/.

[7]. Granik M, Mesyura V (2017) ,**" Fake news detection Using Naïve bayes classifier".** In: 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON), Kiev, Ukraine.

[8].Bhowmik D, Zargari S, Ajao O (2018), **" Fake news areidentification twitter with hybrid CNN and RNN models"**. In: Proceedings of the 9th international conference on social media and society.

[9]. Zheng L, Zhang J ,CuiQ, LiZ, Yang P S, YangY(2018). **"TICNN :convolutional neural networks for fake news detection".**Arxiv preprint.

[10]. Lakshmanarao A, Swathi Y, Kiran TSR (2019). **" An efficient fake news detection system using machine learning".** Int J InnovTechnol Exploring Eng (IJITEE) 8(10).