International Journal of Current Research and Modern Education (IJCRME)
Impact Factor: 6.925, ISSN (Online): 2455 - 5428
(www.rdmodernresearch.com) Volume 8, Issue 1, 2023

FAKE NEWS DETECTION USING LSTM DEEP LEARNING

A. Dhivya*, S. S. Sneha Priscilla, K. Ganga**, R. Sabatham** & R. Susika****

* Assistant Professor, Department of Information Technology, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur, Tamilnadu

** UG Student, Department of Information Technology, Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur, Tamilnadu

Cite This Article: A. Dhivya, S. S. Sneha Priscilla, K. Ganga, R. Sabatham & R. Susika, "Fake News Detection Using LSTM Deep Learning", International Journal of Current Research and Modern Education, Volume 8, Issue 1, Page Number 21-25, 2023.

Abstract:

Newspapers are the primary source of news for people worldwide. However, offlate, due to the significant growth and updates in technologies, there has been astupendous rise in the popularity of social media. As a consequence, social networks such as social media, websites, blogs, etc. have emerged as relevant platforms togather all kinds of news. People rely more on social networks than newspapers thesedays. This survey paperdescribes the various methods and models used for the detection of fake news. Our project aims to use Natural Language Processing to directly detect fake news, basedon the text content of news articles. The model building and testing are done using Jupyter Notebook 6.4.11 and the news article is classified by using website which isdone in HTML5, CSS3 and Flask 2.1.2. Our aim is to find a reliable and accurate model that classifies given news article as either fake or true using machine learningor deep learning techniques.

Key Words: Fake News, Social Media, NLP, News Articles, Accuracy, Deep Learning Techniques, Tensor Flow, Python Flask.

1. Introduction:

In the modern era, the spread of fake news has become very evident. Fake news is being used for both economic and political benefits.The need of the hour is to prevent the spread of fake news. The first thing that needs to be done to achieve this is todetect fake news. Our project aims to develop a robust model to identify a news source whether they are fake or not.A corpu so flabe ledrea land fake articles is used to build a classifier that can make decisions about information based on the content from thecorpus. Our model focuses on identifying sources of fake news, based on multiplearticles originating from a source. Once a source is labeled as a producer of fake news,prediction of all future articlesfrom the same source are also a producer of fake news takes place.Computational strategies have demonstrated helpful incomparable setting. Moreover, regularities in Bot conduct and monetarily.

2. Litterature Survey:

Alkhodair, S. A., Ding, S.et al. [1] have proposed a method The problem of understanding incorrect information spreading on social media has gotten very little attention. One among the early examples of this type of study is seen in Zhao, Resnick, and Mei's (2015) publication. It was suggested by the authors tos start by identifying "signal tweets" using a manually compiled list of regular expressions. Instead of employing a list of predetermined, manually produced regular expressions, our suggested model learns the features automatically. The context of an event can be learned from the sequence of tweets viewed thus far and used to classify the current tweets, according to a sequential classifier model recently suggested (Zubiaga, Liakata, et al., 2016). Our methodology determines the micro-post's class purely from its text. Historical information is not required.

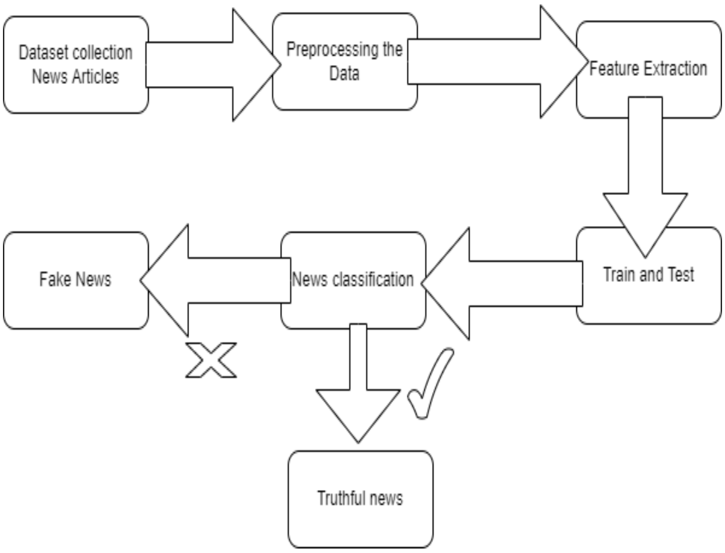
Allcott, H., & Gentzkow, M [2], With 14% of Americans citing social media as their "most important" source of election news, it was important but not the only source; 2) Of the known false news stories that surfaced in the three months leading up to the election, those in favour of Trump had a total of 30 million Facebook shares, while those in favour of Clinton had 8 million; 3) The typical American adult saw one or possibly several fake news stories in the months resulting up to the election.

Bondielli, A., & Marcelloni, F [3], The structure of the essay is as follows. In Section 2, we present some definitions and give an overview of the many kinds of incorrect information that can be obtained online. The methods for gathering and pre-processing data for detecting false information and rumours are covered in Section 3. In Section 4, we go over the many features that have been retrieved and applied to detection in the literature.

Chang, C. C., & Lin, C. J [4], The goal is to help users to easily apply SVM to their applications. LIBSVM has gained wide popularity in machine learning and many other areas. In this article, we present all implementation details of LIBSVM. Issues such as solving SVM optimization problems theoretical convergence multiclass classification probability estimates and parameter selection are discussed in detail.

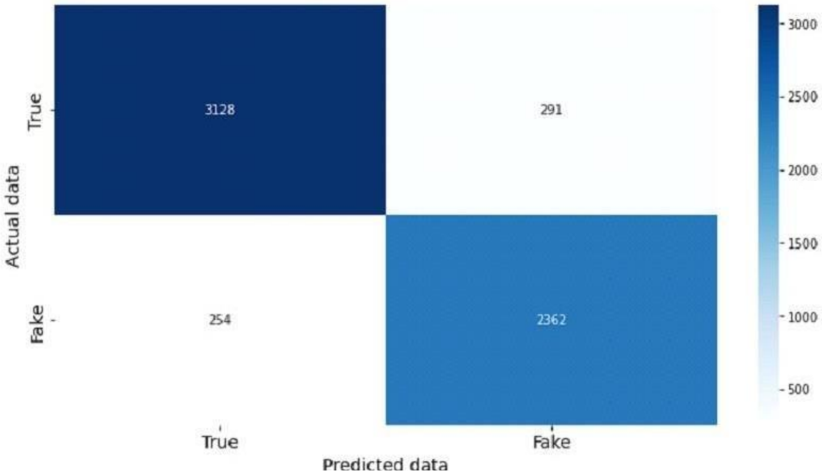
5. Guan, M.et al. [5] In this study, we studied the effectiveness and actions of genetic testing in supporting cancer patient therapy adjustment and evaluated how word embedding and deep learning algorithms might help to efficiently extract information from free-text EHR documents (such as progress notes.

3. Methodology:



- Machine learningdataset is defined as the collection of data that isneeded to train the model and make predictions.
- These datasets are classified as structured and unstructured datasets, where the structured datasets are in tabular format in which the row record and column corresponds to the features, and unstructured datasets corresponds to the images, text, audio, etc.
- Dataset preprocessing is a step in the data mining and data analysis process that transforms raw data into a format that computer and machine learning algorithms can understand and analyze. The actual raw data in the form of text, images, videos, etc.
- Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data. • To measure if the model is good enough, a method called Train / Test are used. It is called Train/Test because you split the data set into two sets: a training set and a testing set. To make a valuable training set, the issue needs to be comprehended for which it is being settled.

4. Flow Chart:



There are certain parameters involved in confusion matrix, they are

- Condition Positive (P): The number of real positive cases in the data
- Condition Negative (N): The number of real negative cases in the data
- True Positive (TP): A test result that correctly indicatesthepresenceofaconditionor characteristic.
- True Negative (TN): A test result that correctly indicates the absence of a condition or characteristic.
- False Positive (FP): A test result which wrongly indicates that a particular condition

- False Negative (FN): A test result which wrongly indicates that a particular condition or attribute is absent.

5. Algorithms:

Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own.Some of the popular ML algorithms are:

- Logistic Regression Algorithm
- Random Forest
- Decision Tree
- Naïve Bayes

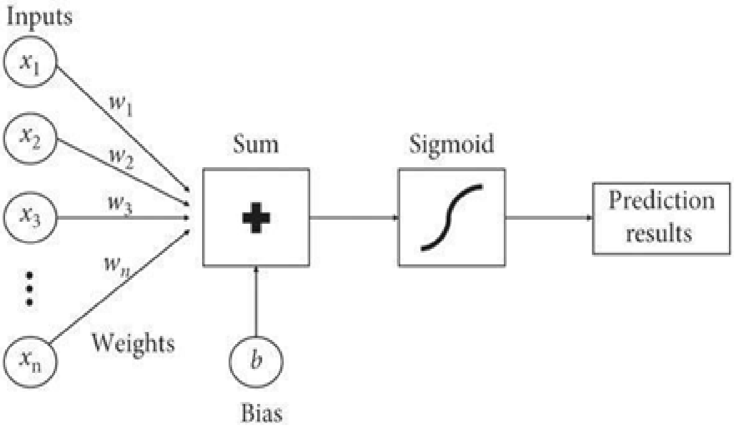
Logistic Regression Algorithm:

In logistic regression, a logit transformation is applied on the odds that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi)=1/(1+exp(-pi))

ln(pi/(1-pi))=Beta_0+Beta_1*X_1+…+B_k*K_k

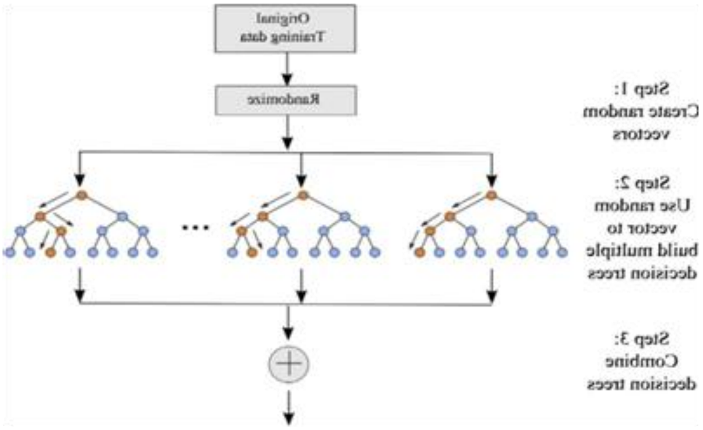
In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via Maximum Likelihood Estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds



Architecture of Logistic Regression

Random Forest Classifier:

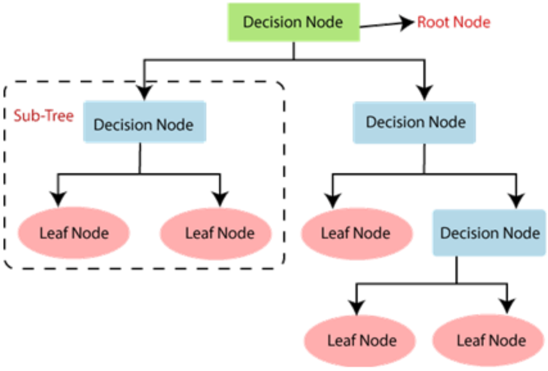
The Random Forest technique was introduced by Brieman. Predictions of several trees are combined by random forest classifiers. Many decision trees are built by the random forest algorithm. Utilizing a subset of features, each decision tree is created. Each decision tree produces one class and eventually bootstraps the votes to obtain better accuracy from the Random Forest technique.


Original Training data
Randomize
Step 1: Create random vectors
Step 2: Use random vector to build multiple decision trees
Step 3: Combine decision trees

Architecture of Random Forest Classifier

Decision Tree Classifier:

Decision tree works by recursively selecting features and splitting the dataseton those feature. These features can either be nominal or continuous in order to find the best split, it uses Gini impurity. Where p(i) is the probability of class i in the current branch. The best split is chosen as the one that decreases those os the impurity. For instance, beginning from the root, the Gini impurity is computed on the complete dataset, then the impurity of eachbranch is computed over all features, weighting it by the number of elements in eachbranch.The chosen feature is the one that has the highest impurity.



Structure of decision tree classifier

Gaussian Naive Bayes:

Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e., normal distribution using the Bayes theorem the naive Bayes classifier works. The naive Bayes classifier assumes all the features are independent to each other. Even if the features depend on each other or upon the existence of the other features. Continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values. The likelihood of the features is assumed to be Gaussian; hence, conditional probability is given by:

System Description:

Hardware Description:

Processor	Pentium Dual Core 2.3GHz
Hard Disk	250GB or Higher
RAM	2GB (Min)
Monitortype	15Inch VGA.
Keyboard	Keys Keyboard

Software Description:

Operating System	Windows 7 or Higher
Tools	Jupyter Notebook
Front End	HTML, CSS, Javascript, Flask
Back End	Python, Tensor Flow, Keras

Jupyter Note Book:

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, maps, plots, graphics, visualizations, and text. Jupyter Notebooks area spin-off project from the I Python project, which used to have an Ipython Note book project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the I Python kernel, which allows you to write your programs in Python, but thereare currently over 100 other kernels that you can also use. Jupyter Note book mainly used for Python because Python is used with AI, ML as well as DL. The components of Jupyter Notebook are the notebook web application, Kernels, Note book documents.

Tensor Flow:

The most famous deep learning library in the world is Google's Tensor Flow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations. To give a concrete example, Google users can experience a faster and more refined the search with AI. If the user types a keyword at the search bar, Google provides a recommendation about what could be the next word. Google wants to use machine learning to take advantage of their massive datasets to give users the best experience.

Python Environment:

- Python is available on a wide variety of platforms including Linux and Mac OS X.The following Python's standard library are used. Pandas:
- An open source library used for data manipulation. NumPy: An open source library used for processing arrays. Sklearn:
- A ML library used for classification, regression and clustering. Matplotlib:
- An open source library used for data visualization. Seaborn:
- An open source library used for making statistical graphics. NLTK:

- A suite of libraries used for statistical language processing.

Python Flask:

Flask is a webframe work that provides libraries to build light weight web applications in python. It is developed by Arm in Ronacher wholeads an international group of python enthusiasts (POCCO). It is based on WSGI toolkit andjinja 2 template engines. Flask is considered as amicro framework. Webserver Gateway Interface (WSGI) which is a standard for python web application development. It is considered as the specification for the universal interface between the web server and web application. Jinja 2 is a web template engine which combines a template with a certain data source to render the dynamic webpages.

6. Conclusion:

A computerized model for checking the verification of news extracted from news website articles which give general answers for information accumulation and expository demonstration towards fake news recognition have been analyzed. After having an idea from the supervised models and also comparing the performance accuracy of the various algorithms, a deep learning-based model is built to identify fake news. The accuracy metric presumably would be altogether improved by methods for utilizing progressively complex model. It is worth noting; that even with the given dataset, only part of the information was used. The current projectdid not include domain knowledge related features, such as entity-relationships.

7. Future Work:

For our future work, name entities from each pair of news headline and news body will be extracted and analyze their relationships through a knowledge base. The study demonstrated that even the very basic algorithms on fields like AI and Machine Learning may find a decent outcome on such a critical issue as the spread of fake news issues worldwide. Accordingly, the aftereffects of this examination propose much more, that systems like this might come very much handy and be effectively used to handle this critical issue. More experiments on other datasets in different languages will be performed. An app which detects the fake news or real news using different ML and DL models will be developed.

8. References:

1. C. C. &. L. C. J. Chang, “a library for support vector machines," IEEE, pp. 1-27., 2011.
2. H. T. I. Ahmed, "Detection of online fake news using n-gram analysis and machine learning techniques," IEEE, pp. 127-138, 2017.
3. S. A. D. Alkhodair, "Detecting breaking news rumors of emerging topics in social media," IEEE, p. 102018, 2020.
4. H. &. G. M. Allcott, "Social media and Fake News in the 2016 Election Journal of Economic Perspectives," IEEE, pp. 211-236., 2017.
5. M. F. Bondielli, "A survey on fake news and rumour detection techniques," IEEE, pp. 38-55., 2019.
6. N. K. R. V. L. Conroy, "Methods for finding fake news," IEEE, pp. 1-4, 2015.
7. G. V. A. Gravanis, "A benchmarking study for fake news detection." IEEE, pp. 201-213., 2019.
8. M. C. S. P. R. Z. W. P. B. &. T. Guan, "Natural language processing and recurrent network model" IEEE, pp. 139-149., 2019.
9. B. &. A. S. Horne, "Fake news packs a lot in title," IEEE, 2017.
10. G. &. A. T. Rampersad, "Acceptance by demographics and culture on social media" IEEE, pp. 1-11, 2020.