isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 6 Issue: 3 Apr,2021,ISSN_NO: 2321-3337

FAKE NEWS DETECTOR USING MACHINE LEARNING

Mr. B. ARUNMOZHI M.E. Professor of Computer Science Department,
Mr. S. SRINIVASAN, Student of Computer Science Department,
Mr. M. SANJAY, Student of Computer Science Department,
Mr. A. KASEELAN, Student of Computer Science Department,
St. Joseph College of Engineering, Sriperumbudur, Chennai

Abstract

The advent of the World Wide Web and the rapid adoption of social media platforms paved the way for information dissemination that has never been witnessed in the human history before. With the current usage of social media platforms, consumers are creating and sharing more information than ever before, some of which are misleading with no relevance to reality. In this work, we propose to use machine learning ensemble approach for automated classification of news articles.

1. Introduction

The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. Besides other use cases, news outlets benefitted from the widespread use of social media platforms by providing updated news in near real time to its subscribers..

2. Literature Survey Indrajit Saha1, Khushboo Agarwal, [1] proposede fake news, especially on online social networks (OSNs), has become a matter of concern in the last few years. These platforms are also used for propagating other important authentic information. Thus, there is a need for mitigating fake news without significantly influencing the spread of real news. We leverage users' inherent capabilities of identifying fake news and propose a

warning-based control mechanism to curb this spread.

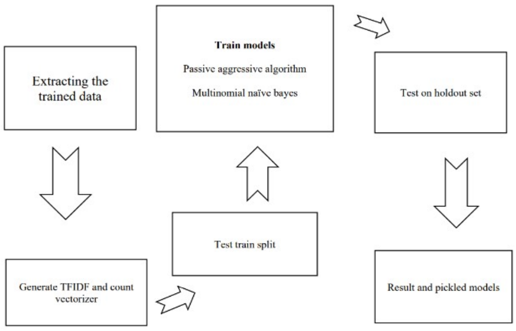
MUHAMMAD UMER, ZAINAB IMTIAZ[2] Proposed Society and individuals are negatively influenced both politically and socially by the widespread increase of fake news either way generated by humans or machines. In the era of social networks, the quick rotation of news makes it challenging to evaluate its reliability promptly. Therefore, automated fake news detection tools have become a crucial requirement. To address the aforementioned issue, a hybrid Neural Network architecture, that combines the capabilities of CNN and LSTM, is used with two different dimensionality reduction approaches, Principle Component Analysis (PCA) and Chi-Square.

Rahul R, Mandical D, [3] proposed a system to handle Fake news has been a problem ever since the internet boomed. The very network that allows us to know what is happening globally is the perfect breeding ground for malicious and fake news. Combating this fake news is important because the world's view is shaped by information. People not only make important decisions based on information but also form their own opinions. If this information is false it can have devastating consequences. Verifying each news one by one by a human

GYU SANG CHOI, AND BYUNGWON [4] proposed a The (time) asymptotic proportions of the individual populations are derived using stochastic approximation tools. These results are instrumental in deriving relevant type-1, type-2 performance measures, and formulating an optimization problem to design optimal warning parameters. We derive structural properties of the performance, which reduce the complexity of the optimization problem.

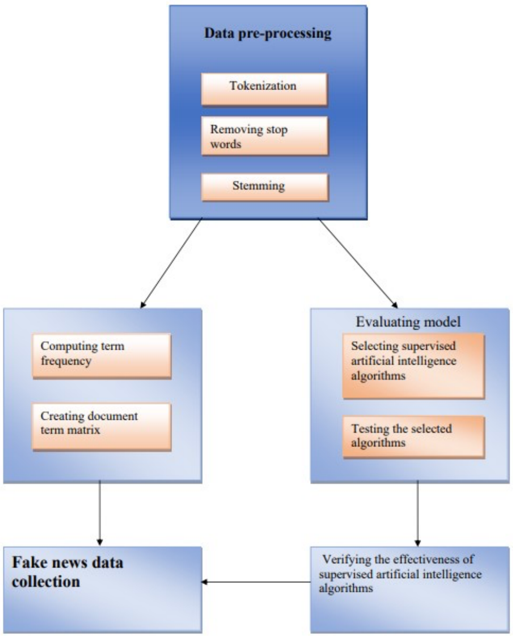
System Design

The utmost contribution of this work is to propose a feature reduction techniques along with hybrid deep learning models, involving two neural network layers, i.e., CNN and LSTM. The proposed approach produces higher predictive performance when compared to the traditional deep learning models. Test on holdout set Result and pickled models Extracting the trained data Train models Passive aggressive algorithm Multinomial naïve bayes Generate TFIDF and count vectorizer Test train split27 To analyse the relationship, four data models are developed.


Extracting the trained data
Train models
Passive aggressive algorithm
Multinomial naïve bayes
Test on holdout set
Test train split
Generate TFIDF and count vectorizer
Result and pickled models

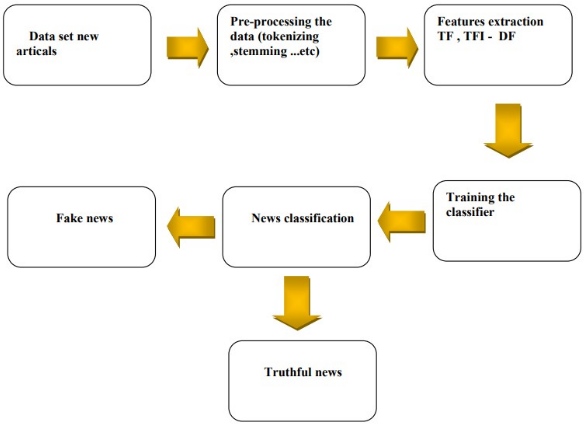
In the first model, all the features are used without pre-processing for classification. In the second model Model third and fourth is developed by using dimensionality reduction techniques including PCA and Chi-square.

This work further investigates which of these models are most suited for being used in conjunction with hybrid CNN and LSTM model when dealing with text data. After the features are selected by any of the four models discussed above, the selected features are fed to the CNNLSTM architecture.


Data pre-processing
Tokenization
Removing stop words
Stemming
Computing term frequency
Creating document term matrix
Evaluating model
Selecting supervised artificial intelligence algorithms
Testing the selected algorithms
Fake news data collection
Verifying the effectiveness of supervised artificial intelligence algorithms

The first layer of the model is the embedding layer that accepts the input headlines and article bodies and converts each word into a vector of size 100. The number of features is 5000, thus, this layer will output a matrix of size 5000 $*$ 100. The output matrix will contain weights that we get through matrix multiplication, to produces a vector for each word. These vectors are passed to the CNN layer to extract contextual features.

Principal Component Analysis (PCA) is a widely used technique that uses a linear transformation to reduce the dimensions of a feature set. The resulting dataset is simplified but it retains the characteristics of the original data set. The new dataset might have an equal or lesser number of features than the original dataset. The covariance matrix is used to compute the principal components.


Data set new articals
Pre-processing the data (tokenizing ,stemming ...etc)
Features extraction TF , TFI - DF
Training the classifier
News classification
Fake news
Truthful news

These components are arranged in decreasing order of importance. On the other hand, in feature extraction methods, a new assume that the original matrix comprises ’a’ dimensions and ’b’ observations and it is required to reduce the dimensionality into a ’t’ dimensional subspace then its transformation can be given by the following equation.

3. Implementation

There is increasing evidence that consumers have reacted absurdly to news that later proved to be fake. One recent case is the spread of novel corona virus, where fake reports spread over the Internet about the origin, nature, and behaviour of the virus. The situation worsened as more people read about the fake contents online. Identifying such news online is a daunting task.

Predicting with the model:

[ ] def pad_to size(vec,size):

Zero =[o]+(size-len(vec))

Vec.extend(zeros) return

vec

Running code; #import statements from

sklearn.feature_extraction.text import

Tfidfvectorizer

##Declaring a vectoriser

Tfidf_vect=Tfidfvectorizer(analyzer=clean_text)

##'Fitting the vectorizer

Tfidf_vect_fit=tfidf_vect.fit(X_train['text'])

Splitting Dataset:

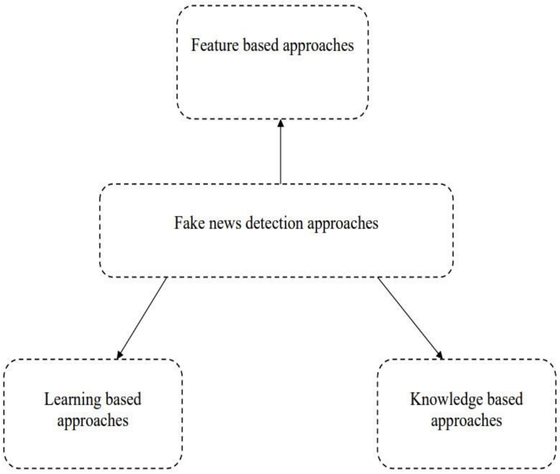
#import statetments

From sklearn>model_selection import train_test_split

Vectorizing The Corpus using thifdvectorizer; sklearn. Feature_extraction.text import Tfidfvectorizer

Tfidf_vect=TfidfVectorizer(analyzer=clean_text)

Tfidf_vect_fit=tfidf_vect.fit(x_train['text']


Feature based approaches
Fake news detection approaches
Learning based approaches
Knowledge based approaches

Training Models;

Epoch ½

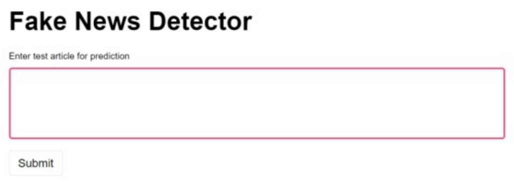
1404/1404[================]-2324s 2s/step -loss;0.0464-acc:0.9817

Epoch 2/2

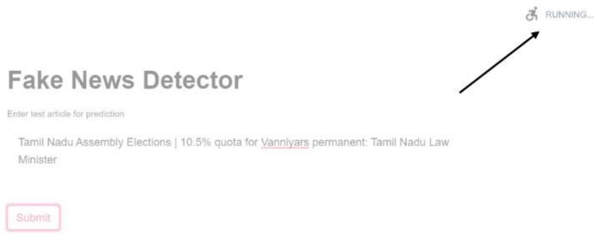
1404\1404[=================]-2341s 2s/step-loss;0.0019-acc:0.9998

To accomplish the extraction of features from the corpus, we used tool which classifies the text into different discrete and continuous variables, some of which are mentioned above. LIWC tool extracts 93 different features from any given text.

As all of the features extracted using the tool are numerical values, no encoding is required for categorical variables.



Here while running our project the new web page will get open in that, user can enter the information, which he want to know the information is real or fake.



Tamil Nadu Assembly Elections | 10.5% quota for Vanniyars permanent: Tamil Nadu Law Minister

Here our project identifying the news is fake or real

Conclusion and Future Enhancement

The task of classifying news manually requires in-depth knowledge of the domain and expertise to identify anomalies in the text. In this research, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques.

The data we used in our work is collected from the World Wide Web and contains news articles from various domains to cover most of the news rather than specifically classifying political news.

The primary aim of the research is to identify patterns in text that differentiate fake articles from true news. We extracted different textual features from the articles using an LIWC tool and used the feature set as an input to the models. The learning models were trained and parameter-tuned to obtain optimal accuracy. Some models have achieved comparatively higher accuracy than others.

References

D. M. J. Lazer, M. A. Baum, Y. Benkler et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094– 1096, 2018.View at: Publisher Site | Google Scholar

2. S. Kogan, T. J. Moskowitz, and M. Niessner, "Fake News: Evidence from Financial Markets," 2019, https://ssrn.com/abstract=3237763.View at: Google Scholar

3. J. Hua and R. Shaw, "Corona virus (covid-19) "infodemic" and emerging issues through a data lens: the case of China," International Journal of Environmental Research and Public Health, vol. 17, no. 7, p. 2309, 2020.View at: Publisher Site | Google Scholar

4. F. T. Asra and M. Taboada,

"Misinfotext: a collection of news articles, with false and true labels," 2019.View at: Google Scholar 5. H. Jawa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exBAKE: automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," Applied Sciences, vol. 9, no. 19, 2019.View at: Publisher Site | Google Scholar.