Content Based Article Authenticity Detection System

Abhishek Mitra, Shaik Naseera

Abstract: The primary objective of this project is to detect the authenticity of a news article based on the content of the same. As the classification is purely based on the article's contents, the project uses various NLP techniques for pre-processing the text before performing classification. To make the project more user-friendly, Raspberry Pi's Pi camera module is used to capture images of news articles which are automatically converted to text, thereby saving the hassle of typing the whole news article for the user. A comparative analysis of various machine learning and deep learning classification models is presented. This paper also presents two new approaches for article authenticity detection using deep LSTM network and deep bidirectional LSTM network. These outperform the existing approaches for detecting article authenticity and a 3.26% improvement in the F1 score from the standard existing bidirectional LSTM model is obtained.

Keywords:Artificial Neural Network;Bag of Words; Deep Learning; Fake News; Long Short Term Memory; Machine Learning; Natural Language Processing; Recurrent Neural Network; Raspberry Pi

I. INTRODUCTION

Content based article authenticity detection system is an emerging field of research in machine learning. Recent rapid spread of fake news articles has made this field an important sector of study. There has been an exponential increase in the number of fake news articles appearing on the Internet misguiding people into believing them. It not only creates false opinions but also leads to confusion among communities, companies, religions, etc. It is believed that the recently concluded US Presidential elections was heavily affected because of the spread of fake news articles against certain target candidates and political parties. It leads to situations when users start doubting real news articles which they read on the Internet. To manually filter out such false articles is not just tough but practically impossible. Data scientists around the world have applied various classifications, deep learning algorithms and techniques to detect and separate the fake articles from the real ones. All the algorithms have varied accuracies, some being better than others but no one algorithm can be specified as correct for detecting fake news.

Revised Manuscript Received on March 25, 2019.
Abhishek Mitra, VIT University, Vellore
Shaik Naseera, VIT University, Vellore

Thus, this project aims to build an effective and efficient system for automating the process of detecting the authenticity of any article. This project provides the user with an easy to use system in which by just capturing the image of the article using Raspberry Pi's Pi camera module, the system can detect whether the article is authentic or not based on various NLP, Machine Learning and Deep Learning models. Data is scraped from nytimes.com and politifact.com to act as a base dataset. Various machine learning and deep learning architectures are applied to detect article authenticity. The different approaches are compared and multiple techniques are implemented to improve the classification performance of the system. The performance of article authenticity detection is improved using the deep LSTM and deep bidirectional LSTM networks. These models boost the performance of the ML model vastly and provide better accuracy, precision, recall and F1 scores. Also, there is a lack of a complete system for news authenticity detection. This project eases it by having the user to just capture an image of the news using Raspberry Pi's Pi camera, and the backend automatically converts the image to text, feeding it to the machine learning and deep learning models which in turn detects it as fake or not. Hence, unlike others, this project proposes an easy to use content authenticity detector with improved performance using deep LSTM and deep bidirectional LSTM architectures.

II. RELATED WORK

A closely related work is identifying fake news using stance detection as performed in [1] focusses on a Fake News Challenge (FNC) dataset and checks whether the headline article pairs agree, disagree or needs to be discussed. Using a bidirectional LSTM network, 9.7% improvement on the dataset and 22.7% improvement on the mean F1 as a whole was recorded. Also based on the Fake News Challenge, [2] uses stance detection for the problem. It uses SVM on features extracted using tf-idf, lexical linear classifier, bag of words multi-layer perceptron, along with LSTM model for better accuracy. Bidirectional conditionally encoded LSTM achieves the best accuracy. Fake News Challenge dataset also used by [3], uses neutral stance detection. Bidirectional encoder (conditional, unconditional, concatenated), attentive reader with simple and full attention and bilateral multiple perspective matching are used for accuracy improvement. Data mining perspective for fake news detection is also another related field explored by [4].

It conducts various surveys to find out about fake news characterization and detection in terms of psychology, knowledge-based, stance-based, style-based and propagation-based approach. In [5] deep learning is modified to form CSI (Capture, Score, Integrate) approach. Capture stores the response and text using RNN. Source takes into account the characteristics of the source of the article and Integrate is used to classify the news based on the scores of the previous two modules.

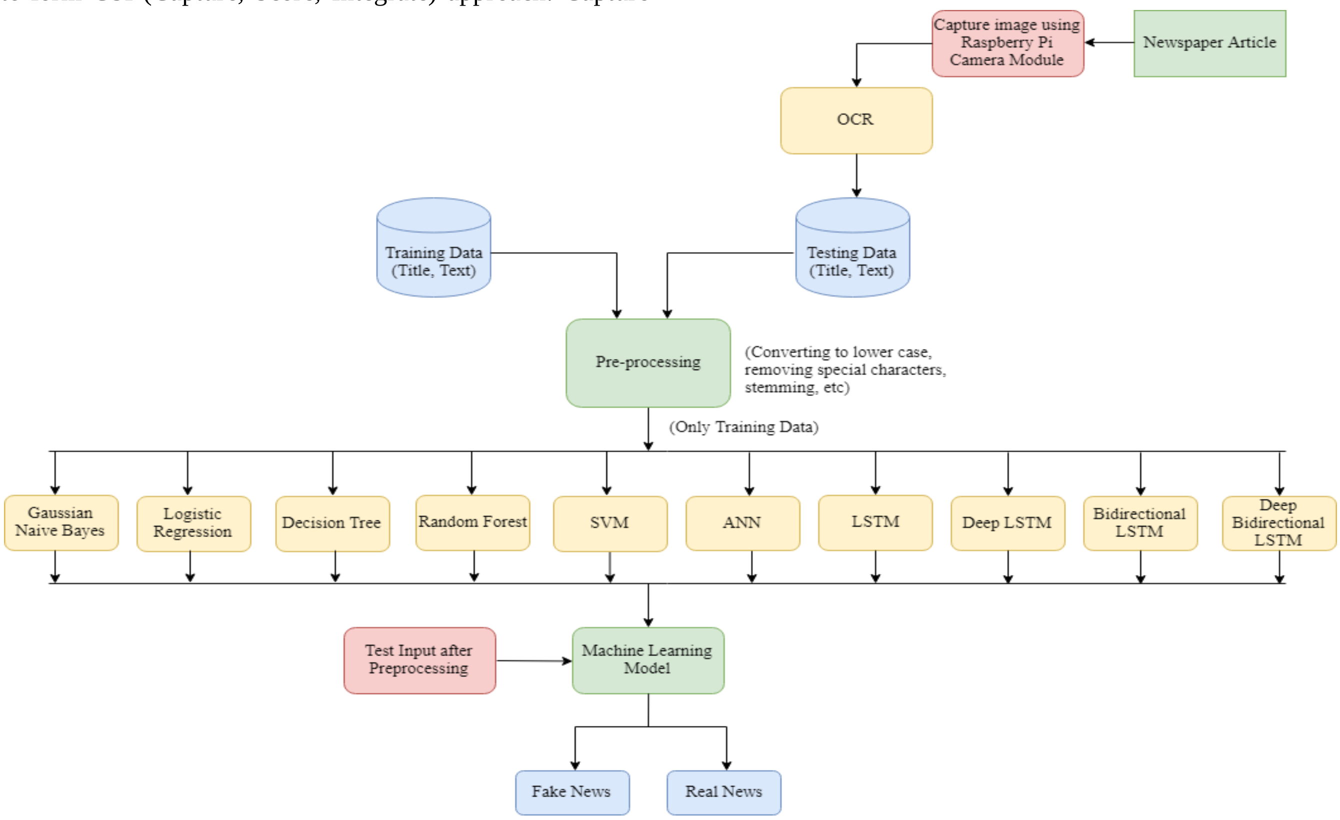

Fig. 1: System architecture

Modification of deep learning in the form of a three level hierarchical model (3HAN) in [6] enables differential attention to the various parts of a sentence because of the hierarchical layers and accuracy of 96.77% is recorded. A novel design based on attention like mechanisms in convolutional neural networks (CNN) is proposed in [7]. Another closely related task is [8] which uses the n-gram approach to determine whether the headline and the article is related. This work mainly focusses on click bait detection which is used by websites for monetary gains. Analytical study for detecting and separating the language of fake news from real ones in [9] is a unique approach to the fake news detection problem. Using this approach, the model is put under a classifier and LSTM combined with LIWC features outperforms other algorithms. The method proposed in [10] predicts whether the news article is satire, hoax, click bait and propaganda. It uses Doc2Vec and TF-IDF in logistic regression, RNN and CNN for a linguistic based classification. It predicts that tweet contents and social network interactions are the only important factors for fake news detection. The authors in [11] discusses the Fake News Challenge Database using neural networks with attention and conditional encoding variations. The model proposed in [12] uses the same FNC database but it implements the bag of vectors and RNN with attention approaches as well. Another way to classify news as fake news as in [13] suggests a SVM approach with five feature extraction module namely, absurdity, humour, grammar, negative affect and punctuation.

III.IMPLEMENTATION

The coding of the project was done in Python 3.6 using the tesseract, theano, tensorflow, keras, numpy, sklearn, and pandas python libraries. Appropriate dropout rates were used for the RNN based LSTM implementation. The following paragraphs depict the whole implementation procedure for the project. Fig. 1 represents the complete system architecture for the project. As shown in Fig 1, the first step involves collecting the training and testing data for the project. The training data is obtained by web scraping of nytimes.com and politifact.com websites. These websites have pre-classified tags of articles as real and fake and hence helps in preparing an appropriate dataset for this project. The sequence is as in Fig. 2, the testing data is collected by the user by capturing an image of the article using Pi cam. For proper focussing of the image before taking the article's image, MotionEye is used. It is a python extension which helps provides live feed to the computer and thus can be used to focus on the article before capturing. The image is then converted to text using tesseract python library.

The next step involved is pre-processing the data. Both the training and the test set undergo pre-processing, i.e conversion to lower case, removing special characters, lemmatization and stemming (if words are ‘playing, ‘played, ‘playable’ then it is converted to root word, ‘play’) and a BoW model is formed. Thereafter, various classification models are applied and the news is classified as fake or real.

A. Models Used

Bag of Words (BoW)- Feature extraction from textual datasets is mostly done with the help of Bag of Words approach. After the pre-processing steps, the BoW model is applied. It not only creates a bag of words of known vocabulary but also extracts the measure of the presence of the known words. In this project, we have used TF-IDF (Term Frequency-Inverse Document Frequency) and count-vectorizers for this purpose. Count vectorizers create a sparse matrix of integer values while TF-IDF uses real numbers. Also, cosine similarities are calculated for the similarities between words in corpus. TF-IDF measures the importance of each word in the document over a BoW corpus. It does so by comparing the frequency of the word appearing in the document and also the measure of the frequency of the word over the whole dataset.

$TF(w)_d = n_w(d)/|d|$, where $n_w(d)$ is the frequency of word w in document d. (1)
$IDF(w)_D = 1+\log(|D|/|\{d:D|w\in d\}|)$, where D is the corpus (2)
$TF\text{-}IDF = TF(w)_d * IDF(w)_D$ (3)

Gaussian Naïve Bayes- It is a ML predictive classifier based on probabilistic Bayesian theorem from statistics. The formulas are as follows:
$P(h/d)=(P(d/h)*P(h))/P(d)$ (4)
where h is the hypothesis and d is the new data entered. After calculating the posterior probability for a number of hypothesis, the maximum is chosen as the next prediction. Therefore,
$MAP(h)=\max(P(d/h)*P(h))$ (5)

Logistic Regression- It is another predictive classifier which takes one or more independent variables as input and the output variable should be dichotomous.
$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$ is one such equation, (6)
where y is the output, b0 is the bias and b1 is an input. It is best suited for binary classification as output values are either 0 or 1.

Decision Tree- This ML model provides a tree like structure. If a given condition has considerable chances, then a particular route is followed else a different is followed. Based on the conditions in the internal nodes, the tree splits are determined. Leafs represent class labels while branches are the outcome of tests. Since, this is a classification problem, decision tree fits in.

Random Forest- Mostly used in supervised learning, it is a stable extension of decision trees. The model creates multiple versions of decision trees and then merges them together to present better, accurate and stable classifications.

SVM- Support Vector Machine (SVM) is based on constructing a hyperplane for classification. The hyperplane that has the largest distance to the nearest training-data point of any class. It is because larger the margin the lower the generalization error of the classifier.
$b0 + (b1 * x1) + (b2 * x2) = 0$, is typical hyperplane equation (7) b0 is the intercept while b1 and b2 represent the slope in a 2D plane. Classification is done by the fact whether the new point lies above or below the hyperplane.

ANN- Artificial Neural Networks (ANN) is another type of classification based on deep learning. A set of features is taken as input and fed into the neural network. These inputs have varied weights depending upon the feature extraction used and fed into a hidden layer with Relu activation. Depending upon the number of hidden layers, the accuracy of the model can be determined. The hidden layers lead to the output layer using a sigmoid activation function. The ANN used is shown in Fig. 3.
$z = f(b+x_i*w_i)$, where z is output, b is bias, x_i is input neuron and w_i is associated weight with the neuron (8)

LSTM- Long Term Short Memory (LSTM) networks are special kind of RNN (Recurrent Neural Networks) capable of learning long term dependencies. The embedding for each word is made using word2glove vectors. They are useful for storing sentence sequences. Composed of three gates, input, output and forget gates, LSTM networks are capable of predicting the next word in any sentence, hence able to predict fake news. The LSTM used in this project is similar to Fig. 4 excluding the Deep LSTM hidden layer and uses 64 LSTM nodes for classification.

Deep LSTM- The deep LSTM model is shown in Fig. 4. 32 more LSTM nodes form a hidden layer between the 64 LSTM nodes and the classification output layer. The extra hidden layer helps improve classification F1 score though increases the computation time for the special RNN model.
Bidirectional LSTM- It functions in a similar way as Fig. 4 with no deep hidden layers and that the output is determined from the combined result of the two hidden layers functioning in the opposite direction. This helps the model to retain past information even
while working with future data. It helps in textual classifications, prediction problems and has high accuracy for content authentication.

Deep bidirectional LSTM- It is a combination of deep LSTM as Fig. 4 and bidirectional LSTM. Combining these two models provide a new hybrid approach by storing sentence sequences of past and future in a deep learning architecture of 32 hidden nodes. Although the computation time increases but the accuracy and F1 scores improve, thereby giving the best performance for determining authenticity.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

IV.RESULTS AND DISCUSSION

The BoW model mentioned in the above section is classified using the various ML models. The results of the various classifiers are summarized in Table 1. Precision= True Positive / (True Positive + False Positive) (9) Recall= True Positive / (True Positive + False Negative) (10) F1 score= 2*(Recall * Precision) / (Recall + Precision) (11)

we can clearly observe that among all the machine learning models, SVM (Support Vector Machine) is the classifier which presents the best F1 score. SVM out beats all other algorithms using both count vectorizer and TF-IDF feature extraction. The results for deep learning classifications are summarized in the above Table 2.The accuracy versus no. of epoch comparison for the various

LSTM models are shown in Fig. 5 to Fig. 8. A total of 50 epochs were performed for each algorithm since exceeding this value was leading to over-fitting of the model.

From the above figures we can observe that the Bi-LSTM and Deep Bi-LSTM models present high accuracy rates from the initial epoch onwards unlike the mono directional LSTM approaches which have moderate accuracy values in the first few epochs. A comparison of all the classification algorithms used is visualized in Fig. 9. While SVM has the best F1 score among the machine learning models, but the Deep bidirectional LSTM has the best score among all the models used. Also, while the Deep LSTMmodel has similar performance as that of bidirectional LSTM, a 3.26% improvement in the F1 score is obtained using the Deep bidirectional model.

V.CONCLUSION

The above results clearly show that the deep bidirectional LSTM network is an improvement over all other existing classifiers for article authenticity detection. A 3.26% improvement in the F1 score is observed from the previously identified best classifier. Also the initial learning rates for bidirectional LSTM models are much higher than mono directional LSTM models. Thus, in scenarios where lesser epochs are to be used, the bidirectional LSTM approach would be the best fit. The future work for this project would involve applying GRUs (Gated Recurrent Units) and improve the present accuracy levels. Also, gradient boosted decision trees can be implemented for this purpose as they are expected to perform better and provide improved accuracy levels for article authenticity detection system.

REFERENCES

1. Mrowca, D., Wang, E., & Kosson, A. Stance Detection for Fake News Identification.
2. Chopra, S., Jain, S., & Sholar, J. M. (2017). Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models.
3. Zeng, Q., Zhou, Q., & Xu, S. Neural Stance Detectors for Fake News Challenge.
4. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.
5. Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News. arXiv preprint arXiv:1703.06959.
6. Singhania, S., Fernandez, N., & Rao, S. (2017, November). 3HAN: A Deep Neural Network for Fake News Detection. In International Conference on Neural Information Processing(pp. 572-581).
7. Bajaj, S. "The Pope Has a New Baby!" Fake News Detection Using Deep Learning.
8. Bourgonje, P., Schneider, J. M., & Rehm, G. (2017). From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles.
9. Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, Yejin Choi, 'Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking'
10. Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 647-653).
11. Pfohl, S., Triebe, O., & Legros, F. (2017). Stance Detection for the Fake News Challenge with Attention and Conditional Encoding.
12. Chen, J. Y., Johnson, J., & Yennie, G. (2017). RNNs for Stance Detection between News Articles.
13. Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news.
14. Tran, K., Bisazza, A., & Monz, C. (2016). Recurrent memory networks forlanguage modeling. arXiv preprint arXiv:1601.
15. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
16. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
17. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
18. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
19. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.
20. Rajesh, M., K. Balasubramaniaswamy, and S. Aravindh. "MEBCK from Web using NLP Techniques." Computer Engineering and Intelligent Systems 6.8: 24-26.