



ANALYSIS OF FAKE DATA ON SOCIAL MEDIA

Tanisha Majumder¹, FarooqueAzam², Kajal³, Vijay B⁴, Asmita Rani⁵
School of Computing and Information Technology, REVA University
Bangalore, Karnataka

tanishamajumder80@gmail.com, farooque53786@gmail.com, kajalsandal006@gmail.com,
appuv098@gmail.com, asmitarani21@gmail.com

Abstract—Fake news is continuously spreading through social media. One such forum is Facebook. Identifying the originality of news has been the focus of this research article. In this paper, we are going to propose LSTM based model for the authenticity of the news. We have used ANN for classification. We have made a satisfactory improvement against the methods used before and found the accuracy of our model as 93%.

Keywords: Facebook, Fake news, LSTM, ANN

I. INTRODUCTION

Nowadays, data is the most significant resource in this century. The current issue to be tackled is to assess whether the data is fake or real. The fake information has a huge impact on individuals and associations that may even panic the individuals.

However, fake news on social media has been proliferated for personal or social benefits [2]. Fake news is commonly a bit of false data in nature, where its main role is to deceive or misdirect readers. It has many similarities with spam messages since they share common features such as grammatical mistakes, false information, using similar limited set of words that affects the reader's opinion [3].

As we have referred to the previous year paper, we come across that detecting of fake news was performed in various criteria. So, it is difficult to identify that the post is fake or real.

Whereas in our work we don't need any specific means to detect the fake news, we can just find whether the news is fake or real from the entire page. The news (content) in our concern is

inherently serialized. So the work we are using is LSTM to predict the data is real or fake.

The paper is organized as: Section II deals with related work. Section III, deals with method such as Long Short Term Memory (LSTM) Section IV, we observe that LSTM gives the best result of accuracy 93.47%. Finally, Section V concludes our work that our method scales an accuracy of 93% which is more than that of the other models.

II. RELATED WORK

The problem of fake news is well-known and many have already studied and given their interpretation.

A. Jain et.al. [1] has used Naïve Bayes to classify fake news and conclude the news, stories are genuine news uses a similar arrangement of words. The conditional probability of finding a particular word in fake news articles and in evident news articles can be settled in a given training set, which contains lots of news articles, marked as real or fake, one can differentiate the probability of finding a particular word in any news article as the proportion of fake news articles, it contains the word to all number of fake news. This algorithm gave 72.94% accurate results.

C. M. M. Kotteti et.al. [5] Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression. They have taken the bags-of-words features where the different articles were collected and converted into a binary format by using various vector techniques based on the

specific requirements. This algorithm gave 88.42% accurate results.

Al-Saremet.al. [7] they had basically focused on a linear model which is implemented in three concealed layers. Whereas, in Tensorflow it has been considered based on neurons. Ma et al [6]Keras it has utilized few layers to avoid overfitting. For their activation function, they used the Rectified Linear Unit (ReLU), which has been found to perform well in NLP applications. The accuracy acquired for these models was 81.42% and 92.62%.

We propose Long short term memory as it is good at characterizing serialized objects since it will specifically retain the past information and use that with the current input. We are going to analyze the fake data using LSTM and using ANN as a classification of the LSTM model.

III. METHODOLOGY

To declare any information as fake information there ought to be a few checks if the information didn't pass those checks, at that point that information will be marked as the fake information. The properties are offer words, contradicting the facts, have comparative words, and the contents are not adequate dues to conflicts.

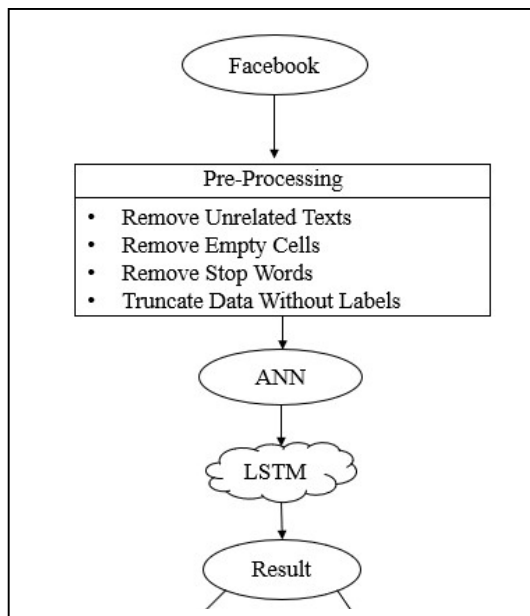


Fig 1: Flowchart of our working model

The data has been extracted from Facebook. After extraction the data has been pre-processed by removing unrelated texts, empty cells, stop words and truncating data without labels. In feature extraction the ANN algorithm carry out in three layers such as input layer, hidden layer and output layer. After this the LSTM fetches the current inputs and process the data and gives the result as real or fake.

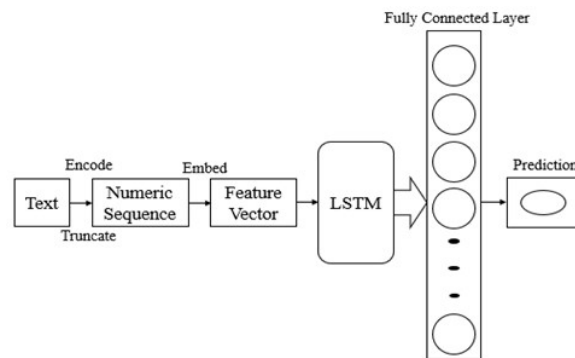


Fig. 2.Working of LSTM

The model we are using to identify the fake news is LongShort Term Memory (LSTM). We first clean the content information by removing all characters which are neither letters nor numbers. At that point we count the recurrence of each word appeared in our training dataset to discover 5000 most basic words and give everyone a one of a kind number ID. S. Setty et.al. [4]has concluded the most well-known word will have ID 0, and the second most common one will have 1, and so forth. After that we replace every regular word with its allotted ID and delete every uncommon word. At that point for that list than 500 words, we pad 0's toward the start of the list. We likewise erase the data with just a couple of words since they don't carry enough data for training. By doing this, we move the first content string to a fixed-length number vector while preserving the word order data. At last we use word embedding to move each word ID to a 32-measurement vector. The word embedding will train each word vector dependent on word similarity. If two words

much of the time show up together in the content, they are believed to be increasingly similar and the separation of their vectors is less. The pre-processing transfers every news in raw content into a fixed size matrix. At that point we feed the training data into the LSTM unit to train the model.

Firstly, we feed the text which is encoded into a numeric sequence and the unnecessary texts are truncated. Then the sequence is embedded and vectorized, fed into a long short term memory model and process the data, passes onto fully connected layer. In this layer, each layer checks the probability and predict as 0 or 1(real: 0 or fake: 1).

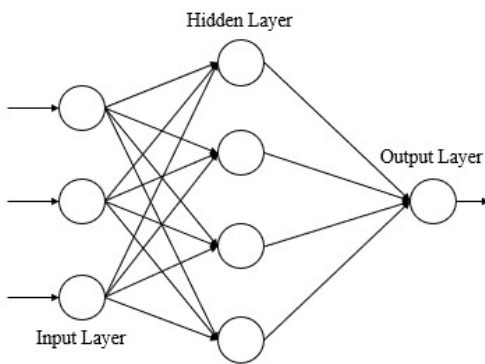


Fig .3. ANN Architecture

In artificial neural network,first we assign the random weights to all the linkages to start than by using the input linkages to find the active rate of hidden nodes and by using the active rate of hidden nodes and linkages to output we find the activation rate of output nodes. Then find the error rate at the output node and all the links between hidden in output nodes differently and by using the weights and error found at output node then cascade the error to hidden nodes recalibrate the weights between hidden node and input nodes repeat the process till criteria are met by using the final link weight score the active rate of the output nodes. In order to analyze our method we are using precision, recall and accuracy as which is mathematically defined as follows,

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

The extent of right positive classification(genuine positives) from cases that are predicted as positive. True Positive (TP) means is predicted as true which is actually true. False Positive (FP) means individuals incorrectly as false.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

The extent of the right positive classification(genuine positives) from cases that is really positive. It is predicted as true which is actually true. False Negative (FN) means fake incorrectly labeled as not fake.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \tag{3}$$

The extent of the right classification (genuine positive negatives) from the overall number of cases. True Positive (TP) means is predicted as true which is actually true. False Positive (FP) means individuals incorrectly as false. It is predicted as true which is actually true. False Negative (FN) means fake incorrectly labeled as not fake.

id	title	author	text	label
0	House Dem Aide: \Darrell Lucus		House Dem Aide:	1
1	FLYNN: Hillary Clin Daniel J. Flynn		Ever get the feeling	0
2	Why the Truth Mig Consortiumnews.com		Why the Truth	1
3	15 Civilians Killed J Jessica Purkiss		Videos 15 Civilians	1
4	Iranian woman jall Howard Portnoy		Print	1
5	Jackie Mason: Holl Daniel Nussbaum		In these trying times	0
6	Life: Life Of Luxury nan		Ever wonder how	1
7	Benoit Hamon W. Alissa J. Rubin		PARIS "France ch	0
8	Excerpts From a Dr. nan		Donald J. Trump is sc	0
9	A Back-Channel Pls Megan Twohey and Scott		A week before Mich	0
10	Obama's Organ Aaron Klein		Organizing for Actio	0
11	BBC Comedy Sketc: Chris Tomlinson		The BBC produced sp	0
12	Russian Researche: Amando Flavio		The mystery	1
13	US Officials See Nc Jason Ditz		Clinton Campaign	1
14	Re: Yes, There Are AnotherAnnie		Yes, There Are Paid	1
15	In Major League So Jack Williams		Guillermo Barros Sch	0
16	Wells Fargo Chief J Michael Corkery and Stacy		The scandal engulfin	0
17	Anonymous Donor Starkman		A Caddo Nation	1
18	FBI Closes In On Hi The Doc		FBI Closes In On	1
19	Chuck Todd: "Bj Jeff Poor		Wednesday after De	0
20	News: Hope For Th nan		Email	1
21	Monica Lewinsky, (Jerome Hudson		Screenwriter Ryan M	0
22	Rob Reiner: Trump Pam Key		Sunday on MSNBC	0
23	Massachusetts Cop nan		Massachusetts	1

Fig.4. Sample CSV file from Kaggle dataset[8]

The datasets utilized for this paper were drawn from Kaggle[8]. The preparation dataset has around 16600 lines of information from different

articles on Facebook. A full training dataset has the following attributes: id, title, author, text, and label.

id involves the unique id for a news article, title involves the minimal information required to understand the news article similar to the heading of the newspaper which describes the content. Author entails the author of the news article; text entails a detailed description of the news article embedded with peculiarities like location, details, people involved and their background etc. Label is basically a tag which tells whether the news articles are Fake or Real so it is indicated in the binary format. (Fake: 1, Real:0).

IV. RESULTS

This experiment has being carried out with Intel i7 8th generation processor, 8GB of ram to run the anaconda software and the datasets we have taken from Kaggle [8].

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 500, 32)	160064
lstm_1 (LSTM)	(None, 100)	53200
dense_1 (Dense)	(None, 1)	101
Total params: 213,365		
Trainable params: 213,365		
Non-trainable params: 0		
None		

Fig. 5.Output of LSTM

In embedding_1 the output shape has three arguments where the first argument is none which is always by default. Next argument consists of 500 nodes which are abstracted and passed into long short term memory model. In parameters, a total of 500 nodes is iterated 160064 times. Similarly, lstm_1 the output shape passes two arguments first is none which is by default, and the next argument is 100 nodes which are iterated 53200 times and the result is shown in the dense_1 where 0 indicates as real, and one indicates as fake which has been found after iterating 101 times.

```
scores = model.evaluate(X_test, y_test, verbose=0)
print("Accuracy= %.2f%%" % (scores[1]*100))
```

Accuracy= 93.47%

Fig. 6.Snippet for finding the Accuracy

With the above mentioned technique we are able to achieve 93.47% of accuracy.

V. CONCLUSION

We have carried out our analysis on Kaggle[8] dataset using LSTM and ANN. Our method scales an accuracy of 93% again the method discussed in which was 72% accurate [1]. Hence we can say that our method can be the best choice for the detection of fake news on social media.

As a part of future work, we plan to build models on Deep Learning algorithms for more accurate results.

ACKNOWLEDGEMENT

We would like to express our special thanks to our Guide, Prof. FarooqueAzam, Assistant Professor,REVA University, for continuous support and guidance.

We would also like to extend our sincere thanks to Dr. SunilkumarManvi, Director, REVA University, for providing us with all the support.

REFERENCES

[1] Jain and A. Kasbe, "Fake News Detection,"in Proc of 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, pp. 203-223,2018

- [2] M. M. Kotteti, X. Dong, N. Li, and L. Qian, "Fake News Detection Enhancement with Data Imputation," in Proc of 2018 IEEE 16th International Conference on Dependable Systems, Services, and Applications, pp. 224-252, 2018
- [3] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevento, "Supervised Learning for Fake News Detection," in Proc of IEEE Intelligent Systems, pp. 423-434, March-April 2012,
- [4] S. Setty, R. Jadi, S. Shaikh, C. Mattikalli and U. Mudenagudi, "Classification of Facebook news feeds and sentiment analysis," in Proc of IEEE International Conference on Advances in Computing, Communications and Informatics, pp. 445-446, 2014.
- [5] P. B. P. Reddy, M. P. K. Reddy, G. V. M. Reddy, and K. M. Mehta, "Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm," in Proc of the Third International Conference on Computing Methodologies and Communication (ICCMC) pp. 891-892, 2019,
- [6] Jing Ma et al. "Detecting Rumors from Microblogs with Recurrent Neural Networks". in Proc of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16) . In :(), p. 7-14, 2018
- [7] Mohammad Al-Sarem, Wadii Boulia in Proc of "Deep learning based rumour detection on fake profile identification". In: systematic review pp. 3-5, 2019
- [8] Datasets, Kaggle, <https://www.kaggle.com/c/fake-news/data>, February 2018
- [9] M. Balmas, "When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism," *Communic. Res.*, vol. 41, no. 3, pp. 430-454, 2014
- [10] D. Berkowitz and D. A. Schwartz, "Miley, CNN, and the Onion," *Journal. Pract.*, vol. 10, no. 1, pp. 1-17, Jan. 2016
- [11] C. Kang, "Fake News Onslaught Targets Pizzeria as Nest of Child-Trafficking," *New York Times* pp. 49-56, 21-Nov-2016,
- [12] C. Kang and A. Goldman, "In Washington Pizzeria Attack, Fake News Brought Real Guns," *New York Times*, pp. 1-5, Dec-2016,
- [13] R. Marchi, "With Facebook, Blogs, and Fake News, Teens Reject Journalistic 'Objectivity'" *J. Commun. Inq.*, vol. 36, pp. 246-262, 2012
- [14] C. Domonoske, "Students Have 'Dismaying' Inability to Tell Fake News From Real, Study Finds," *Natl. Public Radio Two-w*, pp. 1-3., 2016
- [15] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *J. Econ. Perspect*, vol. 31, no. 2, pp. 211-236, 2017