Fake News Detection using LSTM and Logistic Regression

Utkarsh Midha[1[0009-0003-2161-3439]], Ujjwal Kumar[1[0009-0000-2826-7282]], Tripti Saloni[1[0009-0006-5600-9428]] and Hriday Kumar Gupta[1[0000-0001-6115-225X]]

[1] Department of Computer Science and Engineering, KIET Group of Institutions, Ghaziabad, India

Abstract. In this informational age, whether we examine a bit of writing or watch the news on television, we look for a sinceresupply. Internet plays a major role for providing any information or news, people rely on this very much, but the internet and social media are both full of false information. Misinformation or news that has been edited and posted on social media with the intention of harming a person, business, or enterprise is referred to as fake news. This false information can affect a person life adversely. Disasters might result from the spread of false information in urgent circumstances. The need of fake news identification is a must. The spread of fake news necessitates the developmentof computer algorithms to identify it. We have therefore included Machine Learning algorithms and techniques like NLTK, LSTM in order to prevent the harm that can be caused by spread of false information through technology.

Keywords: LSTM; Logistic Regression; Stemming; Lemmatization; Tokenization; NLP.

1 Introduction

Fake news is information that has been deliberately misinformed or falsely [22] claimed to be real. These tales aretypically written to sway people's evaluations, forward a political agenda, or create confusion, and they can regularlyconvey in cash for internet courses. This issue was chosen since it is turning into a significant social problem. It is ensuing in a poisonous on-line environment and riots and lynchings on the streets. Examples consist of political fakenews, stories about sensitive subjects like faith, religion, rumour that salt and garlic may treat corona, and every othercomparable messages we encounter on social media. We can all see the harm that can result from fake information, which is why there is a determined need for a tool which can affirm sure news, whether it is faux or real, and offer individuals with a sense of authenticity primarily based on which they can decide whether to act. With so much noise from fake news and fake statistics, if people lose faith in records, they will no longer be able to access even the most important information, which can even occasionally be life-changing or even fatal. Hence identification of information becomes a necessary step or process to prevent any harm that can be caused by faux news.

2 Related Work

Due to social media, spreading false records has end up pretty simple in the contemporary era. To counteract this,we are going to expand a model that uses ML and NLP ideas and algorithms to identify whether or not a particular piece of information is genuine or not in order to determine whether the news is real or not.

2.1 Stemming

Along with a pre-processing stage in Text Mining applications [2], stemming is a common precondition for Natural Language Processing (NLP) [15] activities. In truth, it plays a crucial function in almost all information retrieval systems. of a section or subsection is not indented. The simple aim of [13] the stage stemming is generally to lessen word's various grammar structures, which are as its verb, adverb, adjective, and noun forms, to its most basic form. We are able to mention stemming goals[24] to lessen a phrase's morphological patterns and once in a while derivationally related forms to a fundamental form that is shared with the aid of all.

2.2 Lemmatization

Locating a word's normalized shape is referred to as lemmatization [8]. Lemmatization is a normalization technique [5], to find a transformation to use to a phrase on the way to achieve its normalized form has similarities to this. One of the approach specializes inword ends, particularly which word suffix should be saved or introduced to supply the normalized form. This examination contrasts the outcomes of two-word lemmatization algorithms, one based on if-then regulations and the other on ripple down regulations induction strategies. It discusses why the Ripple Down Rules (RDR) [25] approach is ideal for the motive and then addresses the problem of lemmatization of terms from Slovene free text. While studying from a corpus of lemmatized Slovene words, the RDR approach yields rules that are simpler to comprehend and feature better classification accuracy than the ones obtained through rule learning in earlier research.

2.3 Tokenization

The process of adding linguistic annotation [23] to certainly occurring text can be seen of as a chain of adjustmentsmade to the authentic textual context, each of which removes the surface distinctions. It means that the input text is split into tokens that are feasible for further analysis.[3] One of the preliminary stages of this transition in the course of nlp is tokenization. It refers to division of text which is taken as input, it is to a computer only a single characters' string into tokens. Then, these tokens are utilized in later stages [16] of natural language processing such morphological analysis, wordclass tagging, and parsing. Tokenization frequently includes identifying both sentence and token boundaries [11] due to the fact these subsequent treatments , are usually designedto work on particular sentences. A number of challenging issues are raised by tokenization when it is in an automated [9] TPS or text processing system demanding situations, handful of them have complete great solutions, no matter being occasionally inspected and quickly discarded. It is not the initial step in the abstraction [19] process. Most differences in typesetting, when are compared to an genuine printed piece of document, the texts are filtered to remove page layout ,font size, graphics, photos and font style.

2.4 Logistic Regression

Logistic regression is a statistical analysis method [4] to predict a binary outcome, such as yes or no, based on prior observations of a data set. Odds ratios are obtained using logistic regression [12] when there are multiple explanatory variables. Apart from the reaction (response) parameter [17] given that the process is a binomial, is noticeably similar to the multiple linear regression. The result is the effect of each and every variable on the probability ratio of the observed significant occurrence. The main advantage is that it is possible to prevent confusing effects of variables by examining how all the variables are related.

2.5 Word Cloud

A "word cloud" is a depiction of word frequency [10]. It shows the most common words in a text document. [1] The word size in the accompanying graphic increases with the phrase's frequency [22] of occurrence in the studied text. Using word clouds as a quick method to decide the primary idea of written content is growing. For instance, they have been used to visualize the content of politicalspeeches in business, education, and politics.

2.6 Natural Language Processing

The collection of techniques known as "natural language processing" enables computers to understand human language. Natural language processing has permeated our daily lives over the past ten years in a number of ways: On the internet and in social media, automatic machine translation is widely used; text classification prevents our email inboxes from overflowing with spam; Search engines have developed to a high level of linguistic complexity beyond string matching and network analysis; dialog systems offer an increasingly popular and efficient means of obtaining and exchanging information [19].

3 Proposed System

3.1 LSTM Model

A recurrent neural network's layers are constructed from long short-term memory (LSTM) units (RNN). A cell,an output and input gate along with a forget gate makes up an unit of long short-term memory [20]. The in charge of "remembering" values over a long period of time such that the relationship between words at the beginning of text can affect how those words appear later in the sentence is the cell. Traditional neural networks seem to have a significant flaw in that they cannot remember or keep track of the whole thing that occurs before they are put into action, which prevents the desired influence of words from the previous phrase from having any impact at the finalwords.

Dataset Overview. The data was obtained through the Kaggle platform. It possesses the subsequent attributes: title: the title of a news article, text: the article's material, subject: the article's subject, and date: the article's publication date. 21417 news articles in all make up the dataset used for model testing and training. A blend of true and fake news is used to createthe dataset.
The following image depicts LSTM cell structure[26].

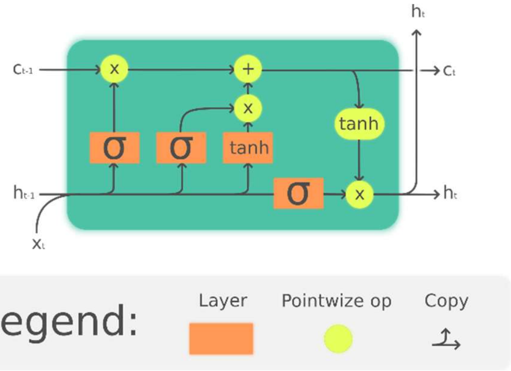


Fig. 1. LSTM Architecture(LSTM Model)

Pre- processing. The data set needs to be processed before it can be transformed into the appropriate format. As the raw data had no labels, so we classified real and fake news, datasets and merged them. Firstly, we eliminated special characters and NULL field data values from our datasets.

Word Embedding. Word to Vector representation: The method cannot accept input in text format, so we ought to convert it into numeric form. To do this, we use word to vector encoding. In word to vector representation the dimension is set to be 100. Each word vector undergoes tokenization. Word tokenization adds text to a list that is then referred to as a vocab list. The list of all the terms used in the narrative serves as the output for this stage.

Model. The model is given the word embedding's output. The machine learning model used in this instance is a sequential version with embedding as the first layer and values for vocabulary size, the number of features, and sentence length. Next comes the LSTM, which has 128 neurons per layer, then the dense layer with sigmoid activation function sincewe only need one output at the end. To prevent overfitting, we added a drop out layer among the Adam optimizer for adaptive estimation and binary cross entropy to calculate loss. The model is then trained and tested.

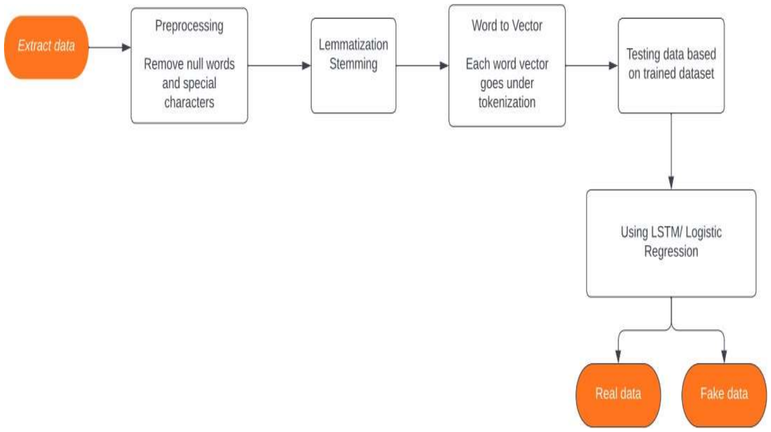
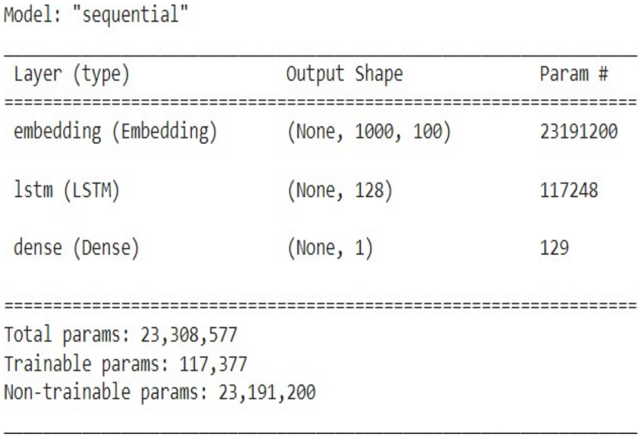

Extract data
Preprocessing
Remove null words and special characters
Lemmatization Stemming
Word to Vector
Each word vector goes under tokenization
Testing data based on trained dataset
Using LSTM/ Logistic Regression
Real data
Fake data

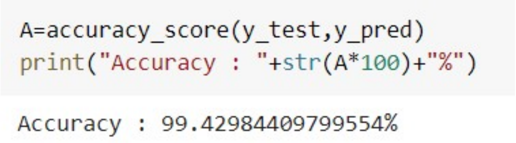
Fig. 2. Proposed System Module



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 1000, 100)         23191200

 lstm (LSTM)                 (None, 128)               117248

 dense (Dense)               (None, 1)                 129

=================================================================
Total params: 23,308,577
Trainable params: 117,377
Non-trainable params: 23,191,200
_________________________________________________________________
```

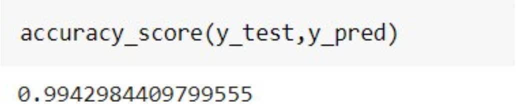
Fig. 3. LSTM Model

Accuracy. The accuracy of the model is 99.4 %.



```
A=accuracy_score(y_test,y_pred)
print("Accuracy : "+str(A*100)+"%")
```

```
Accuracy : 99.42984409799554%
```



```
accuracy_score(y_test,y_pred)
```

```
0.9942984409799555
```

3.2 Logistic Regression (LR) model

In the supervised learning space, one of the most widely used machine learning algorithm is logistic regression. Itis used to interpret a collection of predetermined independent variables [7] and a categorical dependent variable. To forecast the state of a categorical dependent variable, one make use of logistic regression [14]. The outcomes must therefore be distinct or categorical values. Instead of an exact value between 0 and 1, it promises a probabilistic valuebetween 0 and 1. True or false, 0 or 1, yes or no, and so on are all possible outcomes [6].
The following is the model curve between logistic and linear regression[27].

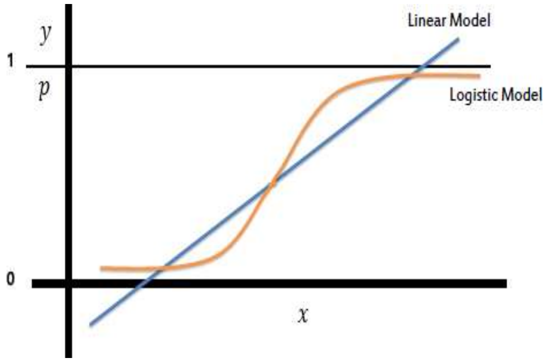
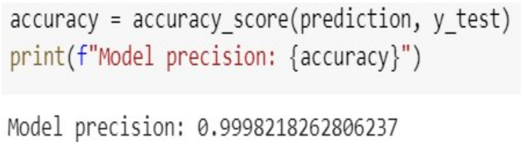


Fig. 4. Model Curve of Logistic VS Linear Regression

Implementation details. The data set needs to be pre-processed to be able to be converted into the precise format. First, we purged the dataset of all NAN and NULL values. Then the process of stemming is completed, followed by vectorization.

Model. The model used is the logistic regression. This model offers greater accuracy, it offers an accuracy of 99.98%.



```
accuracy = accuracy_score(prediction, y_test)
print(f"Model precision: {accuracy}")
```

```
Model precision: 0.9998218262806237
```

4 Results

The accuracy of the Logistic Regression model is 99.98%. Maximum of the time, while we enter news textual content into the interface, it accurately detects the news. We placed this to the test using articles from The Onion. A parody "news" internet site called The Onion publishes made-up, a laugh testimonies. Some of the news that we copied and pasted from the website was appropriately detected as bogus. However, while we tried to verify the newsfrom the BBC or the New York Times, those have been recognized as genuine. Since the LSTM model's accuracy was 99.42%.

5 LSTM vs Logistic Regression Model

LSTM gives the accuracy of 99.42% whereas Logistic Regression has 99.98% accuracy. We may infer that theLogistic Regression Model is more accurate than the LSTM model.

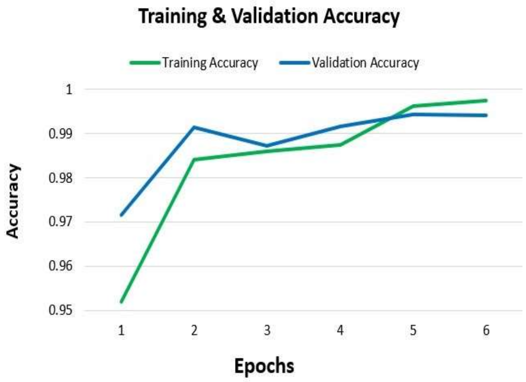


Fig. 5. Model Accuracy Chart

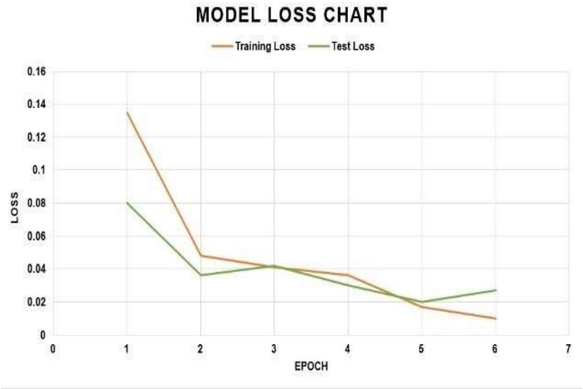


Fig. 6. Model Loss Chart

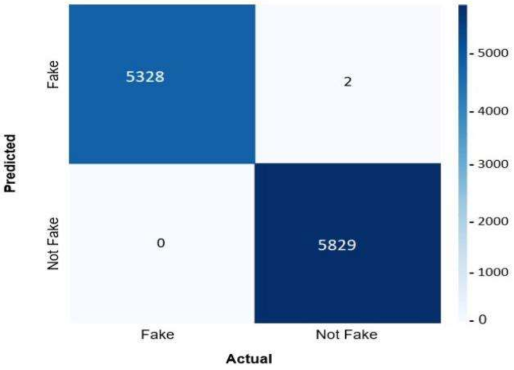


Fig. 7. Logistic Regression Confusion Matrix

6 Limitations

While the data shown here imply a model, some external factors, such as the news's author, location of origin, and time stamp, which may additionally have an impact on the model's output, have been now not taken into considered. There aren't many datasets and research papers available for detecting fake news. The dataset may be tainted, that may lead to detrimental findings. The accuracy of the final results is impacted by the shorter length of the headline or overall report. Training time in faux news grows as module layers increase.

7 Applications

Journalism: Since newspapers and news channels are the primary sources of reliable information, this detection can be used toconfirm the news before it is broadcast.

Social Media: It is simple to distort any news or information in modern's social media environment. Such fabricated information misleads the target audience. Knowing if news is phoney or authentic is essential. This document offers a spread of facts detection and information categorization procedures.

8 Conclusion

With the help of our mission, bogus news can be detected early. If the piece is cleverly written and without any sensationalization, the model produces worse outcomes. We made every effort to address this problem, in spite of how hard it's miles. We think that the consumer-friendly interface makes it simpler for the everyday consumer to confirm the veracity of a information object. To forestall the propagation of false information on social media, initiatives like this one with extra sophisticated features should be carried out. This project has a whole lot of room for future improvement. One approach is to add a go-checking functionality to the machine learning model so it is able to examine the information inputs with the reliable news sources. It desires to be carried out online and in real time, which will be very difficult. We additionally intend to integrate various machine learning techniques in the future, as well as decorate the model accuracy making use of extra and better datasets.

9 Future Scope

This paper has demonstrated the effectiveness of LSTM and NLP techniques for detecting fake news. However, there are several areas for future research that could expand upon this work. Firstly, model accuracy could be improved by exploring ways to fine-tune the models or incorporate additional features. Secondly, the robustness of LSTM and NLP models to adversarial attacks could be evaluated to ensure that they are effective in real-world scenarios. Thirdly, multimodal features could be incorporated to develop more comprehensive fake news detection models. Fourthly, the potential of transfer learning using pre-trained models could be investigated to further improve performance. Lastly, the real-world applications of fake news detection using LSTM and NLP techniques could be explored, such as integration into social media platforms or news websites. Pursuing these future research directions could significantly advance the field of fake news detection and contribute to the fight against misinformation.

10 Acknowledgement

We would like to extend our appreciation to Hriday Kumar Gupta for his invaluable support and guidance throughout the writing process. Additionally, we are grateful to the anonymous reviewers for their insightful feedback, which helped to enhance and refine the presentation of our paper.

References

1. Soesanto, A. M., Chandra, V. C., & Suhartono, D. Sentiments comparison on Twitter about LGBT. Procedia Computer Science, 216, 765-773 (2023).
2. Sahu, S. S., & Pal, S. Building a text retrieval system for the Sanskrit language: Exploring indexing, stemming, searching issues. Computer Speech & Language, 101518 (2023).
3. Carlström, Klara. "Implementation of interpretable methods for paraphrasing and text disambiguation." (2023).
4. TechTarget. (n.d.). Logistic regression. SearchBusinessAnalytics. Retrieved April 25,2023,from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression
5. Bafitlhile, K. D. A context-aware lemmatization model for setswana language using machine learning (2022).
6. K. K. Hiran, R. K. Jain, D. K. Lakhwani and D. R. Doshi, Machine Learning: Master Supervised andUnsupervised Learning Algorithms, (2021).
7. L. Yang, J. Li, W. Lu, Y. Chen, K. Zhang and Y. Li, "The influence of font scale on semantic expressionof word cloud," Journal of Visualization, (2020).
8. M. Orešković, An Online Syntactic and Semantic Framework for Lexical Relations Extraction UsingNatural Language Deterministic Model, (2019).
9. S. Ananth, D. Radha, D. Prema and K. Nirajan, "Fake News Detection using Convolution NeuralNetwork in Deep Learning," International Journal of Innovative Research in Computer and Communication Engineering, (2019).
10. B. Kuyumcu, C. Aksakalli and S. Delil, "An automated new approach in fast text classification," in International Conference on Natural Language Processing and Information Retrieval, (2019).
11. D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein and E. Thorson, "Thescience of fake news," (2018).
12. A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," International Conference on Computing for Sustainable Global Development, (2016).
13. V. Gurusamy and D. Kannan, "Preprocessing Techniques for Text Mining", (2014).
14. R. Atenstaedt, "Word cloud analysis of the BJGP," British Journal of General Practice, (2012).
15. M. A. G. Jivani , "A Comparative Study of Stemming Algorithms," Int. J. Comp. Tech. Appl., (2011).
16. S. Sperande, "Understanding logistic regression analysis," Biochemia Medica, vol. 24, (2014).
17. R. M. Yeung and W. M. Yee, "Logistic Regression: An advancement of predicting consumer purchasepropensity," The Marketing Review, vol. 11, (2011).

18. E. Clark and . K. Araki, "Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English," Procedia - Social and Behavioral Sciences, vol. Volume27, (2011).
19. Prakash M Nadkarni, "Natural language processing: an introduction," Journal of the American MedicalInformatics Association, vol. 18, no. 5, (2011).
20. A. BAYAGA, "MULTINOMIAL LOGISTIC REGRESSION: USAGE AND APPLICATION IN RISKANALYSIS," Journal of applied quantitative methods, (2010).
21. G. Laboreiro, L. Sarmento, J. Teixeira and E. Oliveira, "Tokenizing micro-blogging messages using atext classification approach," (2010).
22. L. &. S. N. &. B. M. Suanmali, "Fuzzy Logic Based Method for Improving Text Summarization.," (2009).
23. D. Palmer, "Tokenisation and sentence segmentation," in Handbook of Natural Language Processing, (2007).
24. J. Plisson, N. Lavrac and D. Mladenic, "A Rule based Approach to Word Lemmatization," in Proceedings of IS, (2004).
25. G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*, Springer, Dordrecht, (1999).
26. https://github.com/guillaume-chevalier/Linear-Attention-Recurrent-Neural-Network/tree/master/inkscape_drawings, last accessed 2023/04/27.
27. Saif, Mohammed & Hosam Raheem, Saif. (2020). Determine of The Most Important Factors That Affect The Incidence Of Heart Disease Using Logistic Regression Model. 14. 174-183, last accessed 2023/04/27