TEST Engineering & Management
May-June 2020
ISSN: 0193-4120 Page No. 4409-4412

Fake News Identification System

[1]Mrigank Satapathy, [2]Nitin C Hegde, [3]Pankaj Jajoo, [4]Adarsh Narayan, [5]Ranjitha U N

[1,2,3,4,5]C & IT, REVA University, Bangalore 560064, India
[1]mrigank214@gmail.com, [2]nhnitinhegde24@gmail.com, [3]pankaj.jajoo75@gmail.com,
[5]ranjitha.un@reva.edu.in

Article Info
Volume 83
Page Number: 4409-4412
Publication Issue:
May - June 2020

Article History
Article Received:19 November 2019
Revised: 27 January 2020
Accepted: 24 February 2020
Publication: 12 May 2020

Abstract

Nowadays, there is a huge surge in data, this results in a decrease in information precision on the Internet. Especially in the social media sphere and other platforms, there is a rapid exchange of data between multiple hands. This raises an important concern. Fake news detection is an important, yet very challenging topic. Traditional methods using lexical features have only very limited success. In this paper, we propose a method to gauge the authenticity of such information using only a few attributes of news with a simple user interface. Along with it, the application will also contain day-to-day important news. All these together will help in keeping the credibility and integrity of the information intact and steer us towards a safer community. To sum it up, the main problem we have in hand is the spread of fake news/misinformation and the application tries to curb it and provide reliable information.

Keywords: *Fake News, Logistic Regression*

1. Introduction

With the widespread use of the internet and the usage of social media, retrieving information has become a very easy task. But this comes with it spros and cons. We do receive information pretty quickly, but there is no guarantee that the information is accurate or not. Misinformation has led to various incidents that have been proven very fatal to society. The Government and many social media platforms have been trying to curb the spread of misinformation but the results haven't been very promising.

Major technology giants like Google and Facebook have now begun testing out new tools to help users better spot and flag fake news and sites as many countries are facing the issue due to the dilution of fake news with original or true content, thus it won't be wrong to call it a global issue and a global challenge. Also, India is a country with a vast population with a user base of 493 million regular internet users, which makes the spread of misinformation effortless and monitoring of such a huge exchange of information is a difficult task. Even, WhatsApp tried to eradicate 'fake news forwards' by limiting the number of forwards to five per message included with the sub-heading 'forwarded' but that also resulted in limited success.

This motivates us to bring out an innovative product to help society. This application will help users to authenticate any information the users have provided it has there quired attributes. This will help us classify and verify the information and will prevent further unwanted mishaps from occurring.

This work presents a literature review in the second section which has helped in giving the basic structure to the product. Section three consists of how the current technology has been implemented to provide a better alternative. We dive into the results of our and other works as well to provide a better comparison in the fourth section. The fifth section highlights the scope of improvement and the conclusion of the project. The last section contains the acknowledgment and the list of references.

2. Literature Survey

Fake News identification is not a new topic of interest. The evolution of such applications or websites has been popping up for a decade. A work on Spam Filtering 'Naive Bayes' classifier [1] by Mykhailo Grank and Volodymyr Mesyura gave appreciable results.

Similar work 'Fake News Detection' used Web Scraping to extract the data from the internet and used the same 'Naive Bayes' classifier gave better results but was decided that it is not sufficient concerning the importance of this work [2] which was by Akshay Jain.

Automation of this process is becoming a new

May-June 2020
ISSN: 0193-4120 Page No. 4409-4412

interest. Work on automating Fake News Detection System Using the "Multi-level Voting model" by Sawinder Kaur gave a better understanding of automation [3]. Social Media platform Twitter is facing the issue of fake news, so the work of Cody Buntain on automatically "Identifying fake news in popular Twitter Threads" gave a broad idea of overcoming this issue [4].

The issue was approached by Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia using Deep Learning in SMU-2018 [13] whichgave us an overview of how neural networks work. That gave us an idea of using LSTM which takes into account neural networks and weights helping us build a model that keeps making itself effective time after time.

Using Georgia M.Kapitsaki and Marios Christou's work, Scrum [5] methodology a better application can be made that focuses on the overall development and deployment of the software. Scrum is an Agile process framework that helps in managing complex knowledge with the initial emphasis being on the software development section. To get an overview of what the product can look like the reference was gathered from a UK based website [14]. It gave usa base of what the product should look like. The dataset on which our model is based is from Kaggle, it gave us an idea of how to separate the available data into different attributes. This data set was changed to fit our needs which reduced the computational complexity[15].

3. Proposed Technology

The application will be providing the public with their daily news requirement from various sources around the world. The application will also help the authenticity of the information the user has to get checked.

The application is built on the ANDROID STUDIO platform. The User Authentication system uses JAVA code and FireBase to authenticate user's credentials and act as a database respectively. The news retrieval is performed using the NEWS API that is built on JAVA.

Android Studio has been used as the Integrated Development Environment (IDE) for Android application development. It is based on the IntelliJ IDEA which is a Java integrated development environment for software. It incorporates its code editing and developer tools. Android Studio contains a Gradle-based build system, emulator, code, templates, and GitHub connection that supports the application development within the Android operating system. The projects in Android Studio contain one or more modules with source code and resource files. These modules accommodate Android app modules, Library modules, and Google App Engine modules. This application is inscribed in Java. The articles in the application have been retrieved from the News API. The News API is an HTTP REST API that searches or retrieves live articles of different categories from all over cyberspace. It helps you answer questions like:

- What articles are being covered by CNN at this moment?
- What updates were released on the latest SAMSUNG mobilephone?
- Has my college or institution been interviewed by anyreporter?

The articles can be searched using particular keywords or complete content. The articles being searched must be in the English language. The network call along with the JSON or XML parsing is handled by Retrofit. For example, Gson handles the JSON parsing.

This allows the application to make synchronous or asynchronous HTTP requests to the remote web servers.

Google FireBase has been used for the authentication of the application. For the authentication to work, the user's data i.e, his identity is saved securely by the Firebase into its cloud which helps in facilitating the personalized experience and ease of use.

Backend services, easy-to-use SDKs, and ready-made UI libraries are provided by the Google Firebase to authenticate users to the application.

The authentication is provided using passwords. A Google sign-in has also been enabled which makes it very user friendly. Google Firebase Authentication has tight integration with other Firebase services, and it leverages industry standards like OAuth 2.0 and OpenID Connect, which helps in easy integration with the customized backend.

To ease out the usage, an OCR has been provided within the application which can be used to directly enter any piece of information they have from their camera or any other source to verify its authenticity. OCR (Optical Character Reader) is implemented using Google's ML toolkit. A restful service is created using FLASK (Flask is a micro web framework that is written in Python).

Logistic Regression algorithm is used to label the news and give it a score to decide its authenticity. LSTM classification algorithm has been used to compare it with the accuracy of Logistic Regression, The logistic regression model is used to model the possibility or probability of a specific class or event, such as hot/cold, sweet/bitter, true/false ,pass/fail.

This technique addresses the situation of qualitative independent variables. Logistic Regression is used mainly to predict binary relations like a product will be success or failure, or pass/fail, etc. Not only this task, but a logistic regression model can also be used in proper prediction or classification of image data i.e. if an image is of a cat or dog, etc. The result will be assigned a probability between 0 and 1. Usually, a model has a dependent variable with two possible values, such as true/false or pass/fail. This is represented by a variable called indicator variable, where the values are labeled "1" or "0". The main advantage of such models is, it is simple and efficient, it has low variance and it provides a probability score for each observation. With the given probability score for each observation, a different setup could be developed to have other options than just pass/fail or true/false. When using categorical data of huge dimensionality logistic regression brings up errors. Also, nonlinear features need to be transformed before using logistic regression.[6]

LSTM which stands for Long short-term memory is an algorithm based on the Recurrent Neural Network in the field of deep learning. Deep learning is part of machine learning.

Higher accuracy can be produced with more data to train on. LSTM can process single data points, like images or text, and entire sequences of data such as audio or video. LSTM is usually used in the field of Natural Language Processing.

Most of the neural networks have feed forward connections, LSTM has a feedback connection. It's usually composed of a cell, an input gate, an output gate, and a forget gate. Over arbitrary time intervals, the cell remembers the values, and the flow of information is regulated by the three gates in and out of the individual cells.

LSTM networks are well suited for classification, processing, and making well-suited predictions based on the time series data since there can be lags of unknown duration between important events in a time series. This algorithm was developed to deal with the common problem of exploding and vanishing gradient which were encountered in the traditional Recurrent Neural Network.[7]

Dataset is collected from two sources, from Kaggle which was used for class prediction of news, and articles collected through News API by google. Using a custom-built article scraping python script which omits the unwanted columns, a CSV file of huge dimensionality was built overtime. The dataset contains various news articles from mainly India and USA of different fields on which the model has been trained on. Datasets from the USA and India are focused so that the degree of noise can be reduced.

The Training data set has attributes like Headline, Author, and label. A unique id is given to them while extraction. The testing data set has the same attributes but no label. The label is being predicted by the algorithm. Label '0' says that the headline is true and '1' makes it false.

4. Results

The results, that were received from two different algorithms are as follows:-

The classification accuracy for true news articles and false news articles is roughly the same, but classification accuracy for fake news is slightlyless compared to the true news. This may be caused by the skewness of the dataset.

Another major issue to address is the difference between the algorithm used one is a machine learning algorithm i.e. Logistic Regression and other, an ANN algorithm i.e. LSTM (Long Short- Term Memory). Eventhough the machine learning algorithm outperforms LSTM, it is highly dependent on the way news articles are available.[Table1]. Also, no increase in accuracy is seen with an increase in the number of articles or rows in the dataset. But with LSTM there is an increase in the accuracy with an increase in the dataset. Also, LSTM can determine on its own if a prediction is accurate or not through its neural network.

This data result is completely based on limited news articles. Larger data would result in better accuracy or agreeable accuracy.

With different combinations of attributes, the accuracy was compared and the combination with higher accuracy was taken and the rest was discarded. The attributes taken were headline and the name of the author. If the body attribute is taken the accuracy in LR and LSTM increases but due to the restrictions in the free data body which wasn't available for every article.

Table 1: Accuracy percentage

Models	**Output Predicted**			
	Number of news Articles in test data.	**True**	**False**	**Accuracy**
1. LR	5000	2518	2482	91.3%
2. LSTM	5000	2261	2739	83.9%
Actual Labels	5000	2663	2337	

Work done in this field yielded results which when compared to LR and LSTM algorithm gave a lesser accuracy. The maximum accuracy of 83.16% [Table 2] was obtained on Naïve Bayes with Lidstone smoothing, which is a machine learning algorithm.

Machine learning algorithms shunt on large data giving the same accuracy after increasing the degree of the dataset since the data is time-series data and a huge amount of news article is generated everyday.

Table 2: Result from previously done work.

Sr. No.	**Model Used**	**Accuracy**
1.	Naïve Bayes with Lidstone smoothing	83.16%
2.	Support Vector Machine	81.65%

Even though it only has a small difference in accuracy compared to LSTM (83.9%), the benefit of using LSTM shows up when the limit of data is taken out. Due to the scarcity of free structured data available on the internet the accuracy of LSTM stops at 83.9%. Another observable variation was the fact that the change in attributes during training. Three different combinations were used i.e. from every news article the combination formed was

i) Headline only ii) Headline + Body iii) Body only and a small variation in accuracy was observed.

May-June 2020
ISSN: 0193-4120 Page No. 4409-4412

Generally, a simple algorithm performs better with fewer variant data(LR). Using deep learning classification algorithms (LSTM), the high dimensionality.

5. Conclusion and Future Scope

Some ways should improve the performance of this classification task. They are as follows:

- Get more data and use it for the training. In machine learning or deep learning problems, it is often the case when getting more data significantly improves the performance of the learning algorithm. The dataset that was described in this article contains less data or news articles compared to the news available from various timelines (around 22,000). This number is really small, and we believe that a dataset with a couple of millions of news articles would be of great help for the learning process.

Unfortunately, such a dataset is not freely available at this moment.

- Only the headline and name of the author of the news articles are used. If the entire content is used the accuracy could have been better.
- The news articles scrapped come with a lot of errors and noise. Cleaning the data set can be one big challenge, also going through them individually will take a lot of time.
- Some words that occur in the headline are rare since these words are counted or vectorized they might get recognized as false news even if the news is true. A provision can be made to recognize rare words only.

The main observation from this is, even though Logistic Regression gives more accuracy for the time being and the data available can provide only limited accuracy after a while and training. On the other hand, LSTM is an algorithm belonging to Recurrent Neural Network (RNN) which is capable of learning by itself for long term dependencies.

Even though Logistic Regression gives more accuracy, LSTM is a better option because the model created by the LSTM is smaller compared to the Logistic Regression. Also, Logistic Regression may fail in case of data with high dimensionality and volume. A sit requires structured data. In LSTM , the algorithm requires a lot of data to train on and rely on layers of ANN (Artificial Neural Network). And becomes better progressively.

The application provides the accuracy of the information. This will help to eradicate the spread of misinformation and will help the society to avoid the mishaps happening from such information. The general public will also be enlightened and keep track of the current affairs.

Acknowledgment

We would like to take this opportunity to express our gratitude to our project guide Prof. Ranjitha UN, Assistant Professor REVA University, for continuously supporting us and guiding us in our every endeavor as well for taking a keen and active interest in the progress of every phase of our project. We would also like to extend our sincere thanks to Dr. Sunil Kumar Manvi, Director, REVA University, for providing us with all the required support.

References

[1] Mykhailo Grank, Volodymyr Mesyura “Fake News Detection using Naïve Bayes Classifier”, IEEE 2017 Ukrain Conference (UKROCON).

[2] Akshay Jain, “Fake News Detection”, IEEE 2018 Maulana Azad National Institute of Technology India.

[3] Sawinder Kaur, “Automating Fake News Detection”, Precog@IIITD - IIITDelhi.

[4] Cody Buntain, “Automatically Identifying Fake news in Popular Twitter Threads”, IEEE 2017 International Conference on Smart Cloud.

[5] Georgia M.Kapitsaki and Marios Christou, “Where Is Scrumin the Current Agile World?”, SCITEPRESS (Science and Technology Publications, Lda.).

[6] Xiaoxin Chen, Rong Ye. “Identification Model of Logistic regression Analysis on listed firm frauds in China”.IEEEFebruary2009.

[7] Fazle Karim, Somshubra Majumdar, Houshang Darabi, Shun Chen. “LSTM fully Convolutional Neural Network for Time-series Classification ”, IEEE December 2017.

[8] Monther Aldwairi and Ali Alwahedi, “Detecting Fake News in Social Media Networks”, The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Network (EUSPN2018).

[9] Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, “Fake News Detection using Machine Learning and Natural Language Processing”, IEEE2019.

[10] Jiawei Zhang, Bowen Dong, and Philip S. Yu, “Fake Detector: Effective Fake News Detection with Deep Diffusive Neural Network”,2019.

[11] Sajjad Ahmed, Knut Hinkelmann and Flavio Corradini,” Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks”, CEUR-WS,2019.

[12] Prabhjot Kaur, Rajvinder Singh Boparai and Dilbag Singh,” Hybrid Text Classification Method for Fake News”, IJEAT-2019.

[13] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia,” Fake News Detection: A Deep Learning Approach”, SMU-2018.

Reference Website:

[1] https://www.logically.co.uk/

[2] The dataset from:https://www.kaggle.com/