Fake news detection

Student: Teodor-Andrei POPOVICI

Contents

Abstract

The idea of this application is to detect fake news using LSTM (Long-Short Term Memory) recurrent neural network.

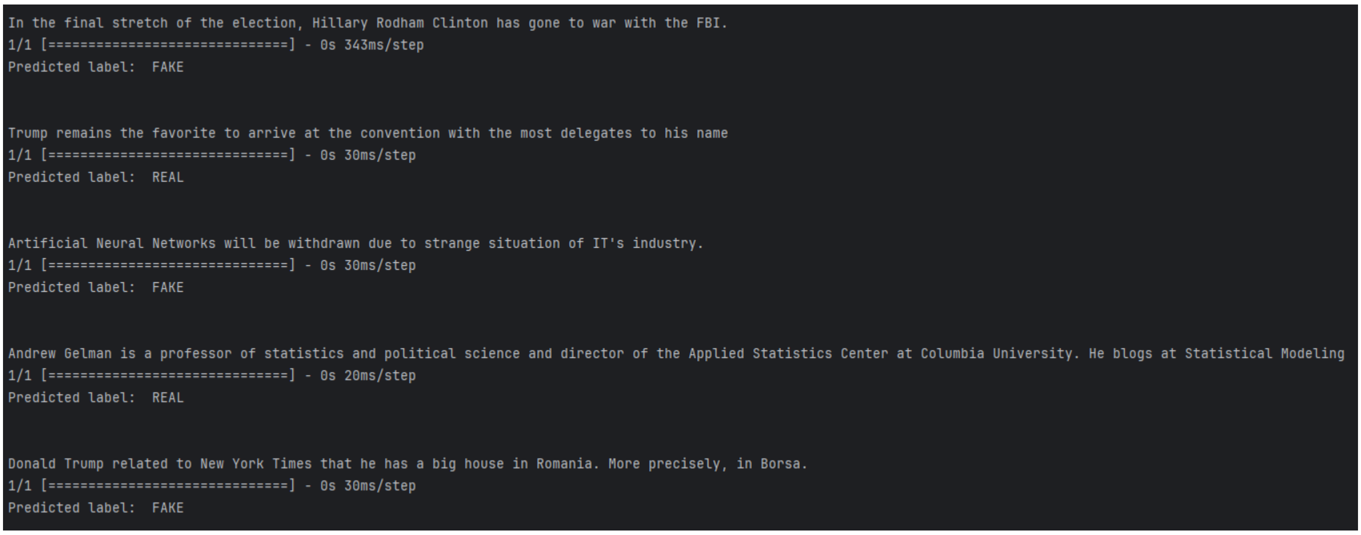
For detecting these fake news, we have some train data, more precisely an Excel file with .csv extension provided from **Kaggle** website. **Kaggle** is a platform for data science competitions, where data scientists and machine learning engineers can compete with each other to create the best models for solving specific problems or analyzing certain data sets [1]. Beside Excel file, we have a JSON file with some keys (of **FN1**, **FN2**, . . ., **FNn** type) and values inside them (inputs for verifying if based on the trained data they are categorized as fake news or not).

The Excel file, called *news.csv*, contains the following columns:

- **Title** - Title of the article
- **Text** - Text of the article
- **Label** - Category of the article, which is **FAKE** or **REAL**

After loading the Excel file, we must parse the most important fields, which are the contents from **Text** and **Label** columns. For analyzing the inputs, we will take only the words without any keywords (e.g. websites, spaces, empty lines), encode the information and train datas with 10 epochs and batch size to 128. After almost 2 minutes of training, we will show the sentence and the prediction for that sentence. Of course, the aproximated time nay vary in function of computer's specifications and number of epochs, batch size and / or length of the sentences for building the model.

We assume that our JSON file contains 5 keys with different sentences. The result is the following:



```
In the final stretch of the election, Hillary Rodham Clinton has gone to war with the FBI.
1/1 [==============================] - 0s 343ms/step
Predicted label:  FAKE


Trump remains the favorite to arrive at the convention with the most delegates to his name
1/1 [==============================] - 0s 30ms/step
Predicted label:  REAL


Artificial Neural Networks will be withdrawn due to strange situation of IT's industry.
1/1 [==============================] - 0s 30ms/step
Predicted label:  FAKE


Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He blogs at Statistical Modeling
1/1 [==============================] - 0s 20ms/step
Predicted label:  REAL


Donald Trump related to New York Times that he has a big house in Romania. More precisely, in Borsa.
1/1 [==============================] - 0s 30ms/step
Predicted label:  FAKE
```

Introduction

In a world which is dependent of technology, people have the possibility to access in a easy and friendly way all the informations regarding some interest topics or general news about what's happening in our world. But exists some situations when these informations are spreaded in a wrong way (more precisely, telling some false situations which are not posted by an official website) and can appear some confusions regarding them or, in the worst case, that news or topics are believed. From this point, it can appear panic situations and the "victims" will spread the information to all the known persons. Happily, some people can prove that the spreaded inforamtion are, actually, false. But, what we can do for someome which is not yet familiar with the technology? How should prevent them from such situations? From this cause, we will talk about the **fake news** phenomenon.

Fake news represent a tactics in which are spreaded false informations or wrong proved data as news. This term sppeared on 2017 as a a neologism [2]. Of course, we can inlcude more types of fake news, such as satire or parody, false connecion or context, fabricated content.

Research has shown that fake news hurts social media and online based outlets far worse than traditional print and TV outlets [2]. After a survey was conducted, it was found that 58% of people had less trust in social media news stories as opposed to 24% of people in mainstream media after learning about fake news [2].

To avoid this phenomenon, it appeared meanwhile some machine learning tactics which help us to know what news are fake or not. In this case, we used Long-Short Term Memory, an extend and improved recurrent neural network (RNN). The idea is that to train this neural network with a lot of datas (which are in this case texts and labels) and based on what this algorithm "learned" and some inputs from a news sentence, we will make the desired prediction. With the help of this prediction we have the possibility to avoid easily the people from the fake news.

Methods

In this project, we used the LSTM (Long-Short Term Memory) recurrent neural network.

Recurrent Neural Network (RNN) represents a type of Neural Network which is used for handling sequential data, which involved variable length inputs or outputs. This neural network has two inputs:

- Data being fed into the RNN (x_t)
- Accumulated information (i.e. memory) of all the previous words in that sentence (h_t)

The motivaton of using this algorithm was that this RNN has the ability to give arbitrary length inputs and outputs

LSTM (Long-Short Term Memory) represents a improved branch of RNN (Recurrent Neural Network) which is used often for prediction tasks ans excels in capturing long-term dependencies [3].

LSTMs use a cell state to store information about past inputs. This cell state is updated at each step of the network, and the network uses it to make predictions about the current input. The cell state is updated using a series of gates that control how much information is allowed to flow into and out of the cell [3].

Some advantages regarding using this algorithm are regarding their memory cell which is capable of long-term information storage [3] and enables the model to capture and remember the important context, even when there is a significant time gap between relevant events in the sequence [3].

Also, the disadvantage of using this algorithm is that in case of training more LSTM networks, it will be time consuming. A good example can be this application, because during development, in case we want to take the maximum length of a sentence for padding sentences (in this case, 10725), some time and memory will be spended or, in the worst case, to crash the computer while training.

For developing this application, the used algorithm is the following:

1. Load the .csv file
2. Parse JSON file, put the content in a variable and treat it as a dictionary
3. Clean spaces, symbols and new lines from parsed Excel file
4. Extract text and label

5. Separate all the words and lowercase them

6. Encode the informations based on the words from the Excel

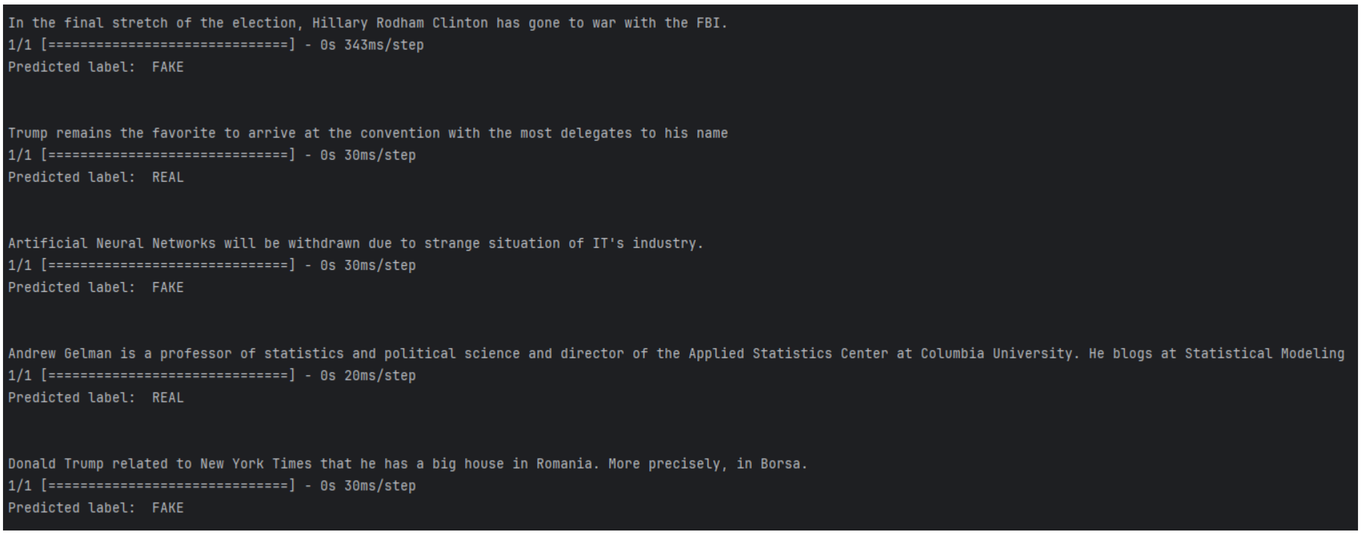
7. Build the Sequential model, including LSTM recurrent neural network

8. Train the model

9. Take every item from JSON file and make the prediction based on the trained data

Results

An example for the results can be shown here:



```
In the final stretch of the election, Hillary Rodham Clinton has gone to war with the FBI.
1/1 [==============================] - 0s 343ms/step
Predicted label:  FAKE


Trump remains the favorite to arrive at the convention with the most delegates to his name
1/1 [==============================] - 0s 30ms/step
Predicted label:  REAL


Artificial Neural Networks will be withdrawn due to strange situation of IT's industry.
1/1 [==============================] - 0s 30ms/step
Predicted label:  FAKE


Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He blogs at Statistical Modeling
1/1 [==============================] - 0s 20ms/step
Predicted label:  REAL


Donald Trump related to New York Times that he has a big house in Romania. More precisely, in Borsa.
1/1 [==============================] - 0s 30ms/step
Predicted label:  FAKE
```

In case on the trained data we have sentences which were labeled as real, obviously the result will be a REAL one.

On the other hand, in case when the sentence were from the trained data or some words are not precised on the trained data or were categorized as FAKE data, the FAKE prediction will be shown.

Conclusions

We got succesfully to train this neural network to make the prediction of the fake news based on trained data (in this case, the text of the news and his label (category)).

A possible development direction is that to insert for each news some specific websites. In case of the link has a relatable format (e.g. https://www.cnn.com), we will consider that the news are real, but in case the format may be a strange one (e.g. http://my-name-is-alexa.com/welcome/to/my-page), the neural network will consider that the link is a dubious one, not necessary considered as fake. From this point, we will distinguish following cases:

- Bad format, real news: the neural netowrk may consider as FAKE. To be considered this as REAL, that news from the bad link must contain the source from a well-known website.
- Bad format, fake news: the neural network will consider as FAKE
- Good format, good news: the neural network will consider as REAL

Bibliography

[1] https://www.coursera.org/articles/kaggle

[2] https://en.wikipedia.org/wiki/Fake_news

[3] https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory