DETECTION OF FAKE NEWS ON CAMEROONIAN NEWS SITES

Guefa Nguimnang Valdes and Fomekong Tchoffo Frank Borel
frankboreltchoffo@gmail.com , valdesguefa@gmail.com

Abstract

In our modern era where the internet is ubiquitous, everyone relies on various online resources for news such as news sites. We propose à system of detection of fakes news on Cameroonian news sites such as Actucameroun, ... This is done by automatically retrieving information from the latter and comparing them with those of other news sites with great international and national credibility such as EcoMatin, Cameroon tribune, LeBledParle, Stopintox, Stopblabla, 237check, Ccousp. This work is done in order to fight against the disinformation of the Cameroonian society and to measure the credibility of other news sites such as those mentioned (ActuCameroon, ...).

I. Introduction and Problem statement

As we spend more and more of our lives interacting online via social media platforms, more and more people tend to seek out and consume social media social media instead of traditional news organizations. traditional. Explanations for this change in consumption behaviours are inherent to the nature of these of these social media platforms: it is often faster and cheaper to consume information on social media compared to traditional journalism, such as newspapers or newspapers or television; and it is easier to continue sharing, discussing. Despite the advantages offered by the social media of information, they can nevertheless disseminate on their platform of false news for reasons such as: the increase of the audience rate, the search for financial and political gain, etc... Given the prevalence of this new phenomenon, ―Fake news" was even named the word of the year by the Macquarie dictionary in 2016 [1]. The extensive spread of faux news can have a significant negative impact on individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance and make people suspicious and confused. Fake news has the effect of amusing, mocking, hurting, shocking, enlisting, influencing or even killing. In political matters, a fake news concerning a politician, once relayed in mass on the internet, can easily tarnish the image of the latter. On the social level, in the context of a tense social climate, the sharing of fake news on social networks can lead to controversy and even to violence. As examples of fake news relayed by Cameroonian news site we have: [2] and [3] talking about a new covid19 variant that has been detected in Cameroon; [4] that talks about a pastor having tied up the populations to preach them the gospel.

The fundamental problem here is to be able to retrieve articles from a news site, each article being characterized by its title, content and author, and from the real articles in the dataset to tell if it is a fake or not. Given the lack of fake news in our dataset, we are in the presence of a single class classification problem based on the identification of outliers (in this project it is fake news) in the data. We will therefore use unsupervised learning algorithms that attempt to model normal examples (real information) in order to classify the new examples as normal or abnormal. The use of the anomaly detection algorithm is due to the lack of sample of the minority (fake news, those of class 1). on the other hand, this classification will consist in fitting a model on the 'normal' data (real news, those of class 0) and predicting whether the new data are normal or abnormal. the minority class (of label 1) will be used to test the model and

II. Related work

- Mykhailo Granik et. al. in their paper [5] shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

- Himank Gupta et. al. [6] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of Words model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately 18%.
- Utsav aggarwal [7] presented a framework based on the use of a dataset having as content words of American politicians [8] the proportion of fake news and real information is almost the same (27000). From this dataset he detects the recurrent words in real news and those in fake news then from the proportion of these words in the new sample predicted the class if it is a fake or not. As a result, he got about 0.98 % as accuracy (using a logistic regression model); the same procedure was followed by Dhanya Mary Biju and Mohak Lall [9]

  The state of the art achieved on the detection of fakes on information sites led us to the works cited above. We found that none of them dealt with single class classification, the language used in these is English, in addition we did not find any dataset related to the African context in general and Cameroon in particular in apart from those mentioned above almost all use the dataset [8].

III. Methodology

III.1 Pre-processing Data

a) Data Collection, cleaning and annotation

In this section, we describe how we collected data and annotated it, and some data statistics are shown. To constitute our data set, we first identified the news sites that are very followed internationally and in Cameroon, then we identified those on which it was possible to do web scraping (sites whose page structure is not dynamic and whose HTML code can be accessed automatically) for this we used the python library beautiful soap. On a page, the collection was done in 2 steps: first we collected the title and the name of the author on the presentation page. the sites on which we have applied

These are: Cameroon tribune, Ecomatin, LeBledParle, stopIntox, 237check, Ccousp. after the collection we cleaned the dataset obtained this by removing the samples in English, the samples resulting from advertisement, of publicity on the site where the sample was taken, of JavaScript. the features such as the title, the theme of the article, the label was according to the link of the page where the article is filled automatically, has these articles have assigned the label 0. the collection of fake news is done manually on sites of fact checking such as Stopblablacam [10]; Stopintox [11]; 237check [12]; Ccousp [13]; fact check [14].

We obtained 4406 usable samples.

b) Remove punctuation

Punctuation can provide a grammatical context for a sentence that sentence, which helps our understanding. But for our vectorizer, which counts the number of words and not the not the context, it does not add value, so we remove all special characters. special characters. e.g.: Bien Dormi? → Bien Dormi moreover, we remove the special characters and the line break characters which do not impact on the meaning of the sentence. after that we put all the text in capital letters. after that we concatenate the title, the text and the author's name to generate a new feature that we called 'total'.

c) Tokenization

Tokenizing separates text into units such as sentences or words. It gives structure to previously unstructured text. eg: bonjour monsieur → 'bonjour', 'monsieur '

d) Remove stopwords

Stopwords are common words that will likely appear in any text. They don't tell us much about our data so we remove them. eg: Je vais au zoo ce matin →'je', 'vais', 'zoo', 'matin'.

e) Lemmatisation

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meanings to one word.eg: rochers → rocher

III.2 Feature Generation

a) Vectorizing Data

Vectorizing is the process of encoding text as integers i.e. numeric form to create feature vectors so that machine learning algorithms can understand our data. In this project we are use Bag of Words (BoW) and term frequency-inverse document frequency (TF-IDF). Bag of Words (BoW) or CountVectorizer describes the presence of words within the text data. It gives a result of 1 if present in the sentence and 0 if not present. It, therefore, creates a bag of words with a document matrix count in each text document. It computes —relative frequency that a word appears in a document compared to its frequency across all documents TF-IDF weight represents the relative importance of a term in the document and entire corpus.

TF stands for Term Frequency: It calculates how frequently a term appears in a document. Since, every document size varies, a term may appear more in a long-sized document that a short one. Thus, the length of the document often divides Term frequency.

$$TF(t,d) = \frac{number\ of\ times\ t\ occurs\ in\ document\ d}{Total\ word\ count\ of\ document\ d}$$

IDF stands for Inverse Document Frequency: A word is not of much use if it is present in all the documents. Certain terms like —a‖, —an‖, —the‖, —on‖, —of‖ etc. appear many times in a document but are of little importance. IDF weighs down the importance of these terms and increase the importance of rare ones. The more the value of IDF, the more unique is the word

$$IDF(t,d) = \frac{Total\ number\ of\ documents}{Number\ of\ document\ with\ term\ t}$$

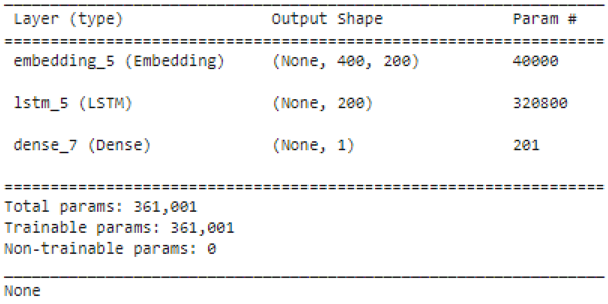
TF-IDF is applied on the body text, so the relative count of each word in the sentences is stored in the document matrix.

$$TFIDF(t,d) = TF(t,d) * IDF(t,d)$$

III.3 Building model

To carry out this task we have set up 3 models OneClassSvm; a convolutional neural model CNN1 having: an Embedding layer, an LSTM layer, a Dense layer and another CNN2 having an Embedding layer, an LSTM layer, a GlobalMaxPool1D layer, a Dropout layer and a Dense layer.

when using the OneClassSvm model we divided our dataset in 2 parts : one part formed only of real information which will be used for training and another one formed of fake news for testing. For the CNN models, given the imbalance between real news and fake news, we have modified the class_weight parameter which is by default equal to 1 for each of the classes to 0.01 for class 0 (majority/real news) and 600 for class 1 (minority/fake news). Moreover, for these models, 80% of the dataset will be used for training and 20% for testing.



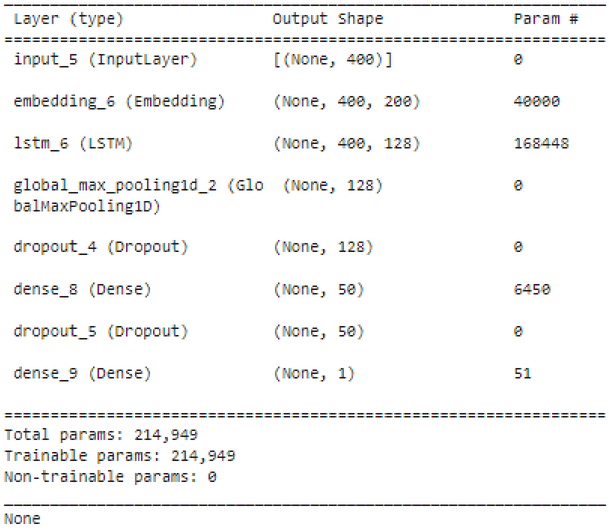
```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_5 (Embedding)     (None, 400, 200)          40000

 lstm_5 (LSTM)               (None, 200)               320800

 dense_7 (Dense)             (None, 1)                 201

=================================================================
Total params: 361,001
Trainable params: 361,001
Non-trainable params: 0
_________________________________________________________________
None
```

Fig 1 : CNN1



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_5 (InputLayer)        [(None, 400)]             0

 embedding_6 (Embedding)     (None, 400, 200)          40000

 lstm_6 (LSTM)               (None, 400, 128)          168448

 global_max_pooling1d_2 (Glo  (None, 128)              0
 balMaxPooling1D)

 dropout_4 (Dropout)         (None, 128)               0

 dense_8 (Dense)             (None, 50)                6450

 dropout_5 (Dropout)         (None, 50)                0

 dense_9 (Dense)             (None, 1)                 51

=================================================================
Total params: 214,949
Trainable params: 214,949
Non-trainable params: 0
_________________________________________________________________
None
```

fig 2 :CNN2

a) Machine learning algorithms

- **One Class SVM**

  One-class SVM is an unsupervised algorithm that learns a decision function for novelty detection: classifying new data as similar or different to the training set. The core of an SVM is a quadratic programming (QP) problem, separating the support vectors from the rest of the training data. The Support Vector Method For Novelty Detection by Schölkopf et al.[15] basically separates all the data points from the origin (in feature space FF) and maximizes the distance from this hyperplane to the origin. This results in a binary function which captures regions in the input space where the probability density of the data lives. Thus the function returns +1 in a “small” region (capturing the training data points) and −1 elsewhere.

The quadratic programming minimization function is slightly different from the original stated above, but the similarity is still clear:

$$\min_{w,\,\xi_i,\,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

$$\text{subject to:}$$

$$\begin{aligned}(w\cdot\phi(x_i)) \geq & \rho - \xi_i & \text{for all } i = 1,\dots,n\\ \xi_i \geq & 0 & \text{for all } i = 1,\dots,n\end{aligned}$$

In the previous formulation the parameter C decided the smoothness. In this formula it is the parameter ν that characterizes the solution. For the convolutional neural network, **the embedding layer** is one of the layers available in Keras. It is mainly used in applications related to natural language processing, such as language modeling, but it can also be used for other tasks involving neural networks. Embedding layer enables us to convert each word into a fixed length vector of defined size. The resultant vector is a dense one with having real values instead of just 0’s and 1’s. The fixed length of word vectors helps us to represent words in a better way along with reduced dimensions.

Long Short-Term Memory Network or LSTM, is a variation of a recurrent neural network (RNN) that is quite effective in predicting the long sequences of data like sentences and stock prices over a period of time.

It differs from a normal feedforward network because there is a feedback loop in its architecture. It also includes a special unit known as a memory cell to withhold the past information for a longer time for making an effective prediction.

In fact, LSTM with its memory cells is an improved version of traditional RNNs which cannot predict using such a long sequence of data and run into the problem of vanishing gradient.

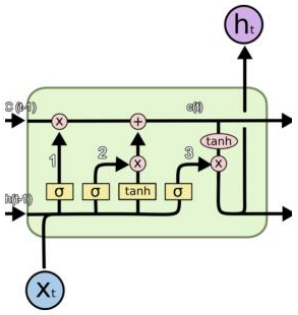


Fig3 : **Long Short-Term Memory Network**[16]

b) Training and testing

The data are trained and tested with many algorithms.

Of all of them, the OneClassSvm model gave a better score with 92%. It is therefore used for the realization of extension for Google Chrome browser and for the realization of a site https://mlpcfakenewsdetector.herokuapp.com/ of false information detection.

Model	OneClassSvm	CNN1	CNN2
Accuracy(%)	0.92	0.85	0.78

c) Conclusion and future work

The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude de of people towards use of digital technology. When a person is deceived by the real news two possible things happen- People start believing that their perceptions about a particular topic are true as assumed. Thus, in order to curb the phenomenon, we have developed our Fake news Detection system that takes input from the user and classify it to be true or fake. To implement this, various NLP and Machine Learning Techniques have to be used. The model is trained using an appropriate dataset and performance evaluation is also done using various performance measures.

References

[1]https://www.macquariedictionary.com.au/blog/article/437/#:~:text=The%20Macquarie%20Dictionary%20has%20announced,2016%20is%20%22fake%20news%22.&text=The%20Macquarie%20Dictionary%20committee%20has,enby'%20and%20'fatberg'.

[2]https://www.camerounweb.com/CameroonHomePage/NewsArchive/Covid-19-le-variant-camerounais-pourrait-faire-basculer-le-sort-de-la-CAN-2021-635026

[3]https://www.camer.be/89311/13:1/un-nouveau-variant-du-coronavirus-decouvert-en-france-probablement-dorigine-camerounaise-cameroon.html

[4]https://www.237online.com/cameroun-ils-ligotent-les-populations-pour-leur-precher-levangile/

[5] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.

[6] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383

[7] https://www.youtube.com/watch?v=5X27excCyXk&t=80s

[8] https://www.kaggle.com/competitions/fake-news/data

[9] https://www.youtube.com/watch?v=g-2-r26nDi4&t=704s

[10] https://www.stopblablacam.com/

[11] https://stopintox.cm/

[12] https://237check.org/reports/#

[13] https://www.ccousp.cm/cat_rumeur/fact-cheking/

[14] https://www.stopblablacam.com/factchecking

[15]http://scholar.google.nl/citations?view_op=view_citation&hl=en&user=DZ-fHPgAAAAJ&cstart=400&pagesize=100&sortby=pubdate&citation_for_view=DZ-fHPgAAAAJ:GFxP56DSvIMC

[16]https://machinelearningknowledge.ai/keras-lstm-layer-explained-for-beginners-with-example/