STAR LION COLLEGE OF ENGINEERING AND TECHNOLOGY

Date	11/10/23
Name	K.Ezhilkaviya
Department	Computer Science and Engineering
Project	Fake News Detection Using NLP
Register Number	822021104009
Naan Mudalvan ID	au822021104009

FAKE NEWS DETECTION USING NLP

Problem definition:

The major problem in detecting fake news is the lack of a massive dataset and a label benchmark dataset with ground-truth labels. For example, some of the datasets are constructed only with political statements like Fact, LIAR, Weibo, etc.

Design thinking:

- DATA SOURCE:

Abstract Fake news detection is a critical yet challenging problem in Natural Language Processing (NLP). The rapid rise of social networking platforms has not only yielded a vast increase in information accessibility but has also accelerated the spread of fake news. Thus, the effect of fake news has been growing, sometimes extending to the offline world and threatening public safety. Given the massive amount of Web content, automatic fake news detection is a practical NLP problem useful to all online content providers, in order to reduce the human time and effort to detect and prevent the spread of fake news. In this paper, we describe the challenges involved in fake news detection and also describe related tasks. We systematically review and compare the task formulations, datasets and NLP solutions that have been developed for this task, and also discuss the potentials and limitations of them. Based on our insights, we outline promising research directions, including more fine-grained, detailed, fair, and practical detection models. We also highlight the difference between fake news detection and other related tasks, and the importance of NLP solutions for fake news detection.

- DATA PREPROCESSING*:*

What is pre-processing?

To process your text simply means to bring your text into a form that is predictable and analyse for your task. The goal of pre-processing is to remove noise. By removing unnecessary features from our text, we can reduce complexity and increase predictability (i.e. our model is faster *and* better). Removing punctuation, special characters, and 'filler' words (the, a, etc.) does not drastically change the meaning of a text.

- FEATURE EXTRACTION:

Abstract

Following the advancement of the internet, social media gradually replaced the traditional media; consequently, the overwhelming and ever-growing process of fake news generation and propagation has now become a widespread concern. It is undoubtedly necessary to detect such news; however, there are certain challenges such as events, verification and datasets, and reference datasets related to this area face various issues such as the lack of sufficient information about news samples, the absence of subject diversity, etc. To mitigate these issues, this paper proposes a two-phase model using natural language processing and machine learning algorithms. In the first phase, two new structural features, along with other key features are extracted from news samples. In the second phase, a hybrid method based on curriculum strategy, consisting of statistical data, and a k-nearest neighbor algorithm is introduced to improve the performance of deep learning models. The obtained results indicated the higher performance of the proposed model in detecting fake news, compared to benchmark models

- MODEL SELECTION:

This article aims at describing the model selection and hyper parameter tuning process to perform fake-news detection.

My previous article, in which I explained how the **model selection** is executed within the field of machine learning on an **Amazon Kindle reviews** dataset, has been a success among readers. The project presented a limit, though, given the size of the dataset containing roughly 1.000.000

reviews, we couldn't deploy the algorithm on a good portion of unseen data. It was done on 50.000 records, approximately 5%, because of computing limitations.

In this article, the goal is to identify and tune the **most performing** model. Training and testing will be done on all the records this time and the resulting model will be deployed on a different dataset to test its accuracy and generalization ability.

- MODEL TRAINING:

This tutorial will create a **natural language processing** application from scratch and deploy it on Flask. In the end, you will have a Fake news detection web app running on your local machine. See the teaser here.

The tutorial is organized in the following structure:

Step1: Load data from to Google lab.

Step2: Text processing.

Step3: Model training and validation.

Step4: Pickle and load model.

Step5: Create a Flask APP and a virtual environment.

Step6: Add functionalities.

Conclusion.

Step1: Load data from to Google Co lab

There are two separate CSV files in the folder, *True* and *False*, corresponding to Real and Fake news. Let's have a look at what the data look like:

Step2: Text pre processing

The datasets have four columns, but they have no label yet, let's create labels first. Fake news as label 0 and Real news label 1.

```
true['label'] = 1
fake['label'] = 0
```

Step3: Model training and validation

You can try multiple classification algorithms here: Logistic Regression, SVM, XG Boost, Cat Boost or Neural Networks. I am using the *Online Passive-Aggressive Algorithms*.

Step4: Pickle and load model

Now, time to pickle (save) the model and vector so you can use them elsewhere.

Step5: Create a Flask APP and a virtual environment

Flask is a lightweight WSGI web application framework. Compared with Django, Flask is easier to learn, whereas it's inappropriate for production use because of security concerns. For the purpose of this blog, you will learn Flask. Instead, feel free to follow my other tutorial on how to deploy an app using Django.

From the terminal or command line, create a new directory:

Step6: Add functionalities

To start, let's create a new file in the same directory with the following content and name it app.py, and we will add the functionalities in this file. Move the pickled model and vector in the previous step to the same directory.

We are going to build four functions: *home* is for returning to the home page; *predict* is for getting the classification result, whether the input news is fake or real; *web app* is for returning the prediction on the web page; is to convert the classification result to JSON file to build

You can see an *index.html* file in the previous section, and it's the home page of the application. Create a folder named *"Templates"* in the root folder, and create a file *"index.html"* inside. Now let's add some content to the page.

Conclusion

In this tutorial, you built a machine learning model to detect fake news from real ones from scratch and saved the model to build a web application using Flask. The web application is running in your local machine, and you can try to make it public using cloud services such as Hero , AWS or Digital Ocean.

- **EVALUATION**

Evaluating the Spread of Fake News and its Detection Techniques on Social Networking Sites

The phenomenon of fake news has become a much contentious issue recently. The controversy regarding this issue has further been intensified by the openness of social media platforms. Via a systematic review, this paper offers a discussion on the spread and detection techniques of fake news on Social Networking Sites (SNSs). A total of 47 articles eventually fulfilled the inclusion criteria and were coded for the literature synthesis. The overall findings from the literature on fake news and social media have been extracted and synthesized to explore the creation, influence and popular techniques and dimensions used for fake news detection on SNSs. The results showed that various entities are involved in the creation and spread of fake news on SNSs, including malicious social and software agents. It was also found that early registered users, old people, female users, delusion-prone persons, dogmatic persons, and religious fundamentalists are more likely to believe in fake news than other groups of individuals. One of the major problems of the existing techniques is their deficiency in datasets. Therefore, future studies on fake news detection should focus on developing an all-inclusive model with comprehensive datasets. Social media users require fake news detection skills especially using linguistic approach. This study provides the public with valuable information about the spread and detection of fake news on SNSs. This is because SNSs are an important avenue for fake news providers.

Keywords: Fake News, Social Media, Detection Techniques, News Content, Social Network