Capstone Proposal Abdallah El Ghamry

Machine Learning Engineer Nanodegree September 2020

Fake News Detector Using Text Classification

1. Domain Background

Natural Language Processing (NLP) is one of the most important fields in machine learning. It's extremely useful in real-life applications such as: question-answering, spam-detection, language translator, grammar checkers, and more. One of the most important applications of NLP is fake news detector. Rubin et al. [1] discuss 3 types of fake news:

1. Serious fabrications
2. Large-Scale hoaxes
3. Humorous fake news

The hardest type of fake news to detect is humorous fake news. It can deceive the detector system. It is a real challenge for NLP.

Due to the importance of automatic detecting fake news, several researchers have presented proposed models to solve it like using recurrent neural network (RNN).

2. Problem Statement

Fake news and articles have become a dangerous effect on online users. Fake News Detection is a binary (0/1) classification problem. Due to the rumors, it is important to detect the fake articles before they spread. The main goal of the model is to find out whether an article is fake or real. The article is the input and the output is fake or true. This problem is similar to mail filtering and sentiment analysis with some differences.

3. Datasets and Inputs

We will use fake and real news dataset which is available on *Kaggle* platform [2]. It can be obtained at https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. It's a large dataset that will be suitable to train the model. The article's data is the input.

Each entry in the dataset has the following labels:

- Title
- Text
- Type (news, politics or others)
- Date (from Mar 2015 to Feb 2018)
- Subject (fake or true)

There are two csv files: *Fake.csv* which represents of the face news, and *True.csv* which represents the real news.

Fake.csv File

- 4 Attributes (title, text, type, date)
- 23481 Records

Fig. 1 Fake.csv File Contents

True.csv File

- 4 Attributes (title, text, type, date)
- 21417 Records

title	text	subject	date
The title of the article	The text of the article	The subject of the article	The date that this article was posted at
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	13Jan16 – 31Dec17
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017

Fig. 2 True.csv File Contents

4. Solution Statement

We want to determine the top features that can identify whether an article is fake or true. Natural Language Toolkit (NLTK) library will be used for tokenization, stemming, parsing, cleaning, removing stop words, and semantic reasoning of the data. Bag of words could be a good model to represent the data.

There are many good classifiers such as: Logistic Regression, K-NN, SVM, Naive Bayes, Decision Tree Classification, and Random Forest Classification. A random forest classifier is an ensemble learning method for classification. It would be a good classifier to choose to fit the data with high accuracy. The classifier with the highest accuracy in scikit-learn will be chose to fit the data.

Finally, the model will be deployed on Amazon SageMaker.

5. Benchmark Model

As mentioned, there are many solutions for this problems on Kaggle platform [2] with high accuracy. So, it will be a challenge to train the model with at least 95% accuracy. The problem will be “Fake News Detection Using RNN” on kaggle [3]. The problem listed in “Detecting opinion spams and fake news using text classification” [4] could be included too.

6. Evaluation Metrics

Dividing the correct classifications on the dataset size is a good quantify the performance of a binary classification model. An accuracy of at least 95% would be appropriate for this problem.

$$accuracy = \frac{correct\ classifications}{cleaned\ dataset\ size} = \frac{true\ positives + true\ negatives}{cleaned\ dataset\ size}$$

TP	FP
FN	TN

False classifications can appear for several reasons including:

1. Satirical news.
2. Ambiguity of the text or uncleaned text.

7. Project Design

1. Data preprocessing and data visualization.
2. Feature extraction.
3. Splitting the dataset:
    a. 70% Train set.
    b. 30% Test set.
4. Modeling using random forest classifier or any appropriate classifier.
5. Testing the results to evaluate the performance and comparing models.
6. Deploying the model on SageMaker with the best determined accuracy.

References

[1] Victoria L. Rubin, Yimin Chen, and Niall J. Conroy. 2015. Deception detection for news: three types of fakes. In Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST '15). American Society for Information Science, USA, Article 83, 1–4.

[2] Fake and real news dataset https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

[3] Fake News Detection Using RNN https://www.kaggle.com/therealcyberlord/fake-news-detection-using-rnn

[4] Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.