Fake News Detector

Vatsal Chaudhary (2020549),
Tanuj Khatri (2020578)
Rajat (2020568)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

Nowadays, social media platforms act as news sources, Anyone can post to these platforms and due to this independence there is a rise in fake news.

Fake news is a very serious issue as most of the people that are reading the news don’t check the source and authenticity of it, and then this fake news is shared via social media and it goes viral.

If a controversial news goes viral, it could lead to unrest in public and distrust in social media.

There are news channels that organize special programs just for checking the authenticity of the news, but as this process is manual and there is a lot of news going around the world this is not efficient. Therefore, we proposed for Fake News detection using Machine Learning.

Literature Review – 1

In the paper entitled “A survey on Natural Language Processing For Fake News Detection”, Fake News detection is a critical problem. The rapid rise of social media platforms yielded a vast increase in information accessibility and accelerated the spread of fake news, and thus threatening public safety. In this paper we introduced the challenges in the detection of fake news and how researchers formulate the ML solutions to tackle this problem.

In this paper, Machine Learning models that are used are, Non - Neural Network models, like Support Vector Machine (SVM) and Naive Bayes Classifier are some of the frequently used classification models. Both of these structures are basically used in baseline models. Decision Tree Such as Random Forest Classifier and Logistic regression are also used occasionally.

Literature Review – 1

In this survey, initially they discussed on the definitions of fake news detection. Then they discussed about some datasets and experimental results for the different methods.

Then they talked about some of the Machine Learning algorithms like,

Logistic regression, Decision trees, Support Vector Machines(SVM) and Naive Bayes Classification etc used in the task of Fake News Detection.

Link to the Research Paper.

https://aclanthology.org/2020.lrec-1.747.pdf

Literature Review – 2

In the paper entitled “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. William Yang Wang, Says that automatic Fake News detection is a challenging problem and has a huge real world political and social impacts.

As in this past election for 45th president for United States, the world witnessed a epidemic of fake news. Fake news not only poses threats to journalism but also effects the political world, and also it also seems to create real like fears.

LIAR - A new benchmark dataset, one of the most obvious application of the dataset is to help in development of the Machine Learning models for Fake News detection. We used baselines like , Regularized Logistic Regression (LR), A support vector machine classifier (SVM),

Literature Review – 2

A bi-directional long short term memory model, Convolutional Neural Network Model (CNN).

As a conclusion, We introduced LIAR a new dataset for automatic fake news detection, LIAR is larger in magnitude enabling the development of

Statistical and computational approach for the Fake News detection.

Link to Research Paper :

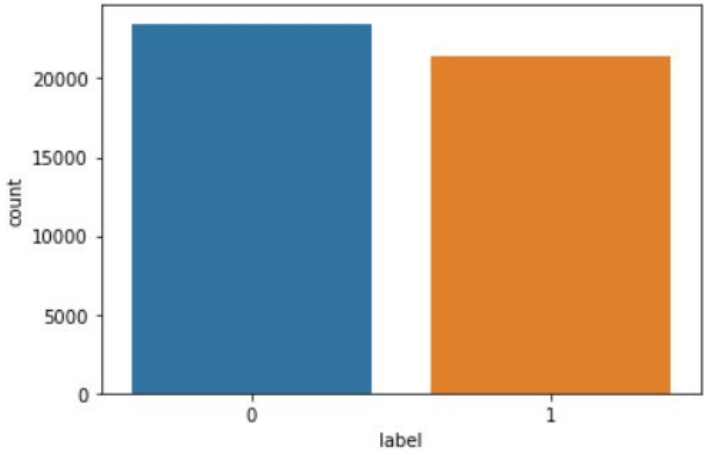
https://aclanthology.org/P17-2067.pdf

Dataset Description

- We are using “fake and real news dataset” from kaggle.(link)
- It contains data of fake and real news in two separate csv files with approximately 21 k news entries each.
- Both files have 4 columns or attributes:
  - Title (of article)
  - Text (article)
  - Subject (like political news, world news)
  - Date

Data Preprocessing

- A new column “label” is added to both dataframes(real and fake news) such that label is ‘0’ for fake news and ‘1’ for real news.
- Both the dataframes of fake news and real news are merged and shuffled to form a new dataframe.
- As you can see in the given figure, the number of fake news and real news are equal in our data.



Data Preprocessing

- For now we will just be using the ‘text’ column for training our ML models.

- We made following changes to the text in ‘text’ column to improve performance of our ML models:
  - Removing Stop Words also known as un-informative words such as (so, and, or, the).
  - Converting all the text to lowercase.
  - Removing numbers, punctuations and symbols from text.

Methodology: Extracting Features

- ML models cannot use textual data for training. So we need a way to convert this textual data to quantitative data.
- We have used ‘TfidfVectorizer’ function of ‘feature_extraction’ library of sklearn for this conversion.
- Using ‘TfidfVectorizer’ whole text column is converted to term-document matrix.
- A term-document matrix represents texts using the frequency of terms or words that appear in a set of documents.
- This term-document matrix can be used by ML models.

Methodology

- We split this term-document matrix and corresponding labels into training set and testing with test size of 0.25.
- As our problem is a classification problem, for now we are using multinomial naive bayes and binary logistic regression models.
- We are using frequency of top 350 words in the news texts as features for our ML model.

Results

Naive bias:

	precision	recall	f1-score	support
0	0.85	0.90	0.87	7042
1	0.88	0.82	0.85	6428
accuracy			0.86	13470
macro avg	0.86	0.86	0.86	13470
weighted avg	0.86	0.86	0.86	13470

Logistic Regression:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	5794
1	0.97	0.88	0.92	5431
accuracy			0.93	11225
macro avg	0.93	0.92	0.93	11225
weighted avg	0.93	0.93	0.93	11225

- As we can see from the results both the models show high accuracy over testing set.
- Also both the models have high precision and recall meaning we will have low number of false positives and false negatives respectively.
- Here Logistic regression is more accuracy than Naive bias over the testing set.

Results

- Accuracy of Some other ML models we used:
  - SVM- 0.88
  - Random Forest- 0.65
  - Decision Tree- 0.47

Results

- We tried Ensembling several models by taking their average predictions and using sklearn's VotingClassifier but it didn't increase the accuracy.

- We also tried using MLP model with hyper parameters that were obtained by grid search but that also didn't increase the accuracy.

- After trying every ML model taught in class for classification problem, Logistic regression gave the best results over the testing set.

Contribution

Each member of group has contributed to each task and no task has been done entirely by one member.

Thank You