

Applied Machine Learning Group Project Proposal

Fake News Detection with Semantic Model

Manavi Ghorpade

Jennifer Wu

Rohit Raut

CSCI 611

Spring 2022

Professor: Dr. Renee Renner

Department of Computer Science

California State University, Chico

April 11, 2022

Introduction

In the world of rapidly increasing technology, information sharing has become an easy task. There is no doubt that the internet has made our lives easier and access to lots of information. This is an evolution in human history, but at the same time, it unfocuses the line between true media and maliciously forged media. Today anyone can publish content – credible or not – that can be consumed by the world wide web. Sadly, fake news has quickly become a social problem. People get deceived and don't think twice before circulating such misinformation pieces to the world. It has been shown that the propagation of fake news has had a non-negligible influence on 2016 US presidential elections. The social media sites like Facebook, Twitter, and Whatsapp play a major role in supplying this false news. Many scientists believe that counterfeited news issues may be addressed through machine learning and artificial intelligence.

Problem Statement

The goal of this project is to propose a model to identify the fake news prediction. In this project, we build a classify model to recognize the fake news based on the title and text by using machine learning, semantics and natural language processing. The proposed system contains a Vectorization for finding the relationship between the words and with the obtained information of the existing relations, the articles are categorized into fake and real news. The motive of this project is to increase the accuracy of detecting fake news more than the presently available results to accurately predict the fake news.


Dataset

The data sources used for this project are from Kaggle. We have used 4 datasets and integrated them in a single dataset. This final dataset includes three features and 28278 tuples. The labels in this dataset are binary, which is True or False. We will split the dataset into 80% for training and 20% for testing.

| df.head() | | | |
|-----------|--|---|-------|
| | title | text | label |
| 0 | Donald Trump Sends Out Embarrassing New Year'... | Donald Trump just couldn t wish all Americans ... | 0 |
| 1 | Drunk Bragging Trump Staffer Started Russian ... | House Intelligence Committee Chairman Devin Nu... | 0 |
| 2 | Sheriff David Clarke Becomes An Internet Joke... | On Friday, it was revealed that former Milwauk... | 0 |
| 3 | Trump Is So Obsessed He Even Has Obama's Name... | On Christmas day, Donald Trump announced that ... | 0 |
| 4 | Pope Francis Just Called Out Donald Trump Dur... | Pope Francis used his annual Christmas Day mes... | 0 |

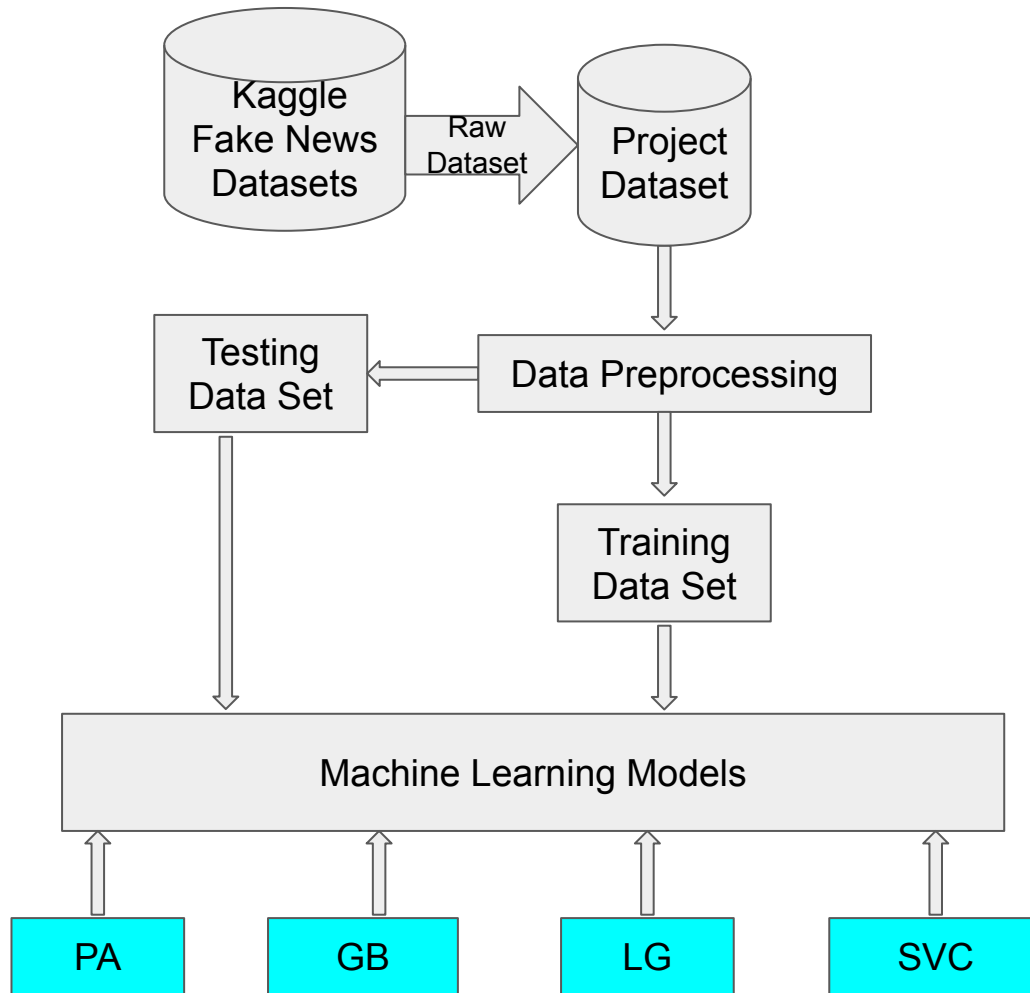
Methodology

- `TfidfVectorizer()`: For finding the relationship between the words and with the obtained information of the existing relations, the articles are categorized into fake and real news. `TfidfVectorizer` turns a collection of raw documents into a matrix of TF-IDF features.
- Map: For mapping the target column having textual data into numerical data.

```
 # Init the TfidfVectorizer function  
vectorizer = TfidfVectorizer()  
  
# Transform training data set  
X_train = vectorizer.fit_transform(X_train)  
  
# Transform testing data set  
X_test = vectorizer.transform(X_test)
```

Machine Learning Algorithms

- Passive Aggressive Classifier (PAC)
- Gradient Boost Classifier (GB)
- Logistic Regression Classifier (LG)
- Support Vector Classifier (SVC)



DEMO

Support Vector Classifier

```
In [17]: from sklearn import svm
SVM = svm.SVC(C=1.9, kernel='linear')
SVM.fit(X_train, y_train)

svm_prediction = SVM.predict(X_test)

# Display the accuracy score
svm_score = accuracy_score(y_test, svm_prediction)

print('The confusion matrix is: ')
print(confusion_matrix(y_test, svm_prediction))
print(f'\nThe accuracy score is: {round((svm_score)*100, 2)}%')
print('\nThe classification report is:\n', classification_report(y_test,svm_prediction))
```

The confusion matrix is:

```
[[2471  110]
 [ 152 2923]]
```

The accuracy score is: 95.37%

The classification report is:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.94 | 0.96 | 0.95 | 2581 |
| 1 | 0.96 | 0.95 | 0.96 | 3075 |
| accuracy | | | 0.95 | 5656 |
| macro avg | 0.95 | 0.95 | 0.95 | 5656 |
| weighted avg | 0.95 | 0.95 | 0.95 | 5656 |

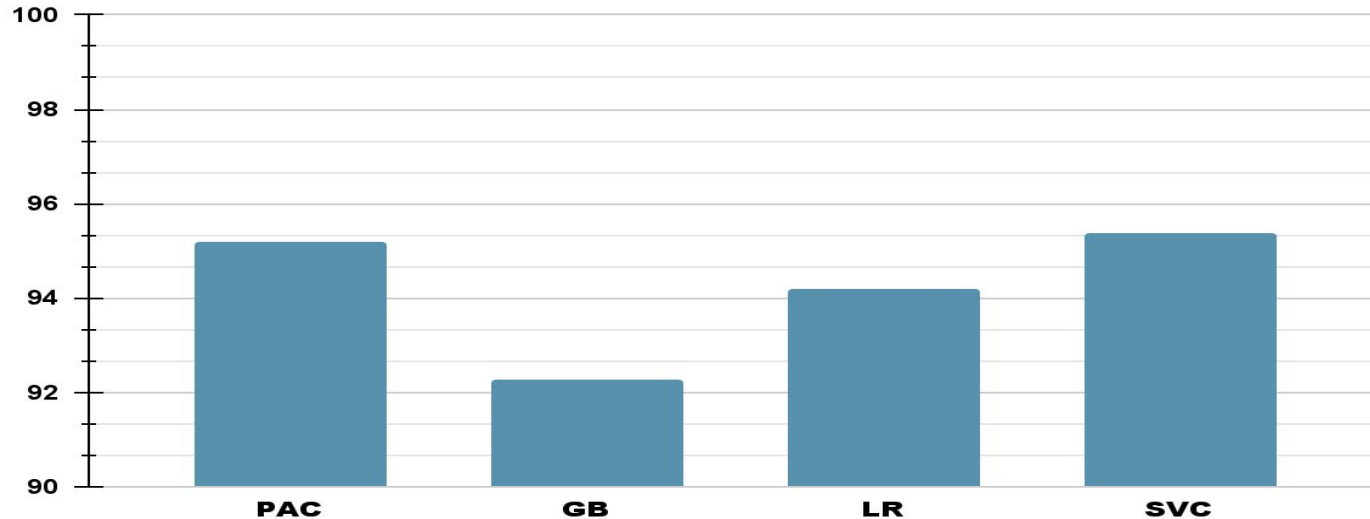
Model Analysis

Fake news recognition models accuracy table

| | PAC | GB | LR | SVC |
|----------|--------|--------|--------|--------|
| Accuracy | 0.9521 | 0.9226 | 0.9418 | 0.9537 |

Fake news recognition models accuracy bar plot

Accuracy



Summary

In the 21st century, the majority of the tasks are done online. The growing problem of fake news only makes things more complicated and tries to change or hamper the opinion and attitude of people towards use of digital technology. We have developed our Fake news Detection system that will detect whether the news is fake or true. 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. The best model, i.e. the model with highest accuracy is used to classify the news headlines or articles. Our best model came out to be Support Vector Classifier with an accuracy of 95.37%.

References

- [1] Xinyi Zhou., Atishay Jain, Vir V. Phoha., & Reza Zafarani. (2020). Fake News Early Detection: A Theory-driven Model. *Digit. Threat.: Res. Pract.* 1, 2, Article 12 (June 2020), 25 pages. <https://doi.org/10.1145/3377478>
- [2] Braşoveanu, A.M.P. & Andonie, R. (2021). Integrating Machine Learning Techniques in Semantic Fake News Detection. *Neural Process Lett* 53, 3055–3072 (2021). <https://doi.org/10.1007/s11063-020-10365-x>
- [3] Barbara Probiez., Piotr Stefański. & Jan Kozak. (2021). Rapid detection of fake news based on machine learning methods. *Procedia Computer Science*, Volume 192. Pages 2893-2902, ISSN 1877-0509. <https://doi.org/10.1016/j.procs.2021.09.060>.
- [4] A. Bani-Hani., O. Adedugbe., E. Benkhelifa., M. Majdalawieh & F. Al-Obeidat. (2020). A Semantic Model for Context-Based Fake News Detection on Social Media. *IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, Antalya, Turkey, 2020 pp. 1-7. doi: 10.1109/AICCSA50499.2020.9316504
- [5] Vicario, M.D., Quattrociochi, W., Scala, A., Zollo, F.: Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2), pp.1-22 (2019).

Thank you