Applied Machine Learning Group Project Final Report

Fake News Detection with Semantic Model

By:

Group - 3

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

Table of Contents

Section 1: Introduction 3

Section 2: Project Objectives, Development & Architecture 4

Section 2.1: Project Objectives

Section 2.2: Methodology

Section 2.3: Data Set

Section 2.4: Workflow diagram

Section 3: Project Results, limitations & Future work 8

Section 3.1: Project Results

Section 3.2: Project Ethics

Section 3.3: Limitations

Section 3.4: Future work or features

Section 3.5: Conclusion

References 11

Section 1: 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.

Various models are used to provide an accuracy range of 60-75%. Which comprises Naive Bayes classifier, Linguistic features based, Bounded decision tree model, SVM, etc. The parameters that are taken into consideration do not yield high accuracy. 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.

Section 2: Project Objectives, Development & Architecture

Section 2.1: Project Objectives

The goal of this project is to propose a model to identify the fake news prediction. In this project, we built 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.

Section 2.2: Methodology

In this project we are using the dataset which has text data in it. So first of all we are going to use natural language processing using the TfidfVectorizer to convert the data into a usable vector (in the form of numerical values). This combines 2 concepts: Term Frequency (TF) and the Document Frequency (DF). It checks for the count of the word by how often the word appears in the document.

Then, we will split the data into training and testing data. After this, we are going to use some Sklearn algorithms to check for the maximum accuracy we get from the testing data. So the algorithms that we are going to use are: Passive Aggressive Classifier (PAC), Gradient Boost Classifier (GB), Logistic Regression Classifier (LR), and Support Vector Classifier (SVC).

And lastly, we will calculate the accuracy score, the performance matrix, and the classification report for every prediction model. The model with the maximum accuracy will be known.

Section 2.3: Data Set

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. Final dataset consists of 12936 fake news and 15342 real news. 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.

Section 2.4: Machine Learning Classifier

Machine learning is a type of algorithm which aims at helping the software systems to achieve higher accuracy without directly reprogramming them. Classification is one of the most significant machine learning techniques for predicting image classes using models inferred from training data (Khanam, 2021).

As usual, classification is a supervised learning method that classifies or maps the data to the labels predefined in a given dataset. Han (2011) mentioned that the Data classification model can be built in two steps. The first step is learning from the training dataset. And the second step is working on classification, which means the model is for individual testing or to predict the labels that are not visible in the dataset. Below are the machine learning classification algorithms that will be applied in this project.

Section 2.4.1: Passive Aggressive Classifier (PAC)

Passive Aggressive Classifiers are often used for large-scale learning. The input data is presented sequentially and the machine learning model is updated incrementally, as opposed to batch learning, which uses the entire training dataset at once. This is useful when there is a lot of data and the entire dataset cannot be trained due to the sheer size of the data. We can simply say that an online learning algorithm takes a training sample, updates the classifier, and then throws away the sample.

Section 2.4.2: Gradient Boost Classifier (GB)

Gradient Boosting falls under the category of boosting methods, which iteratively learn from each of the weak learners to build a strong model. The term "Gradient" in Gradient Boosting refers to the fact that you have two or more derivatives of the same function. Gradient Boosting is an iterative functional gradient algorithm, i.e an algorithm which minimizes a loss function by iteratively choosing a function that points towards the negative gradient; a weak hypothesis.

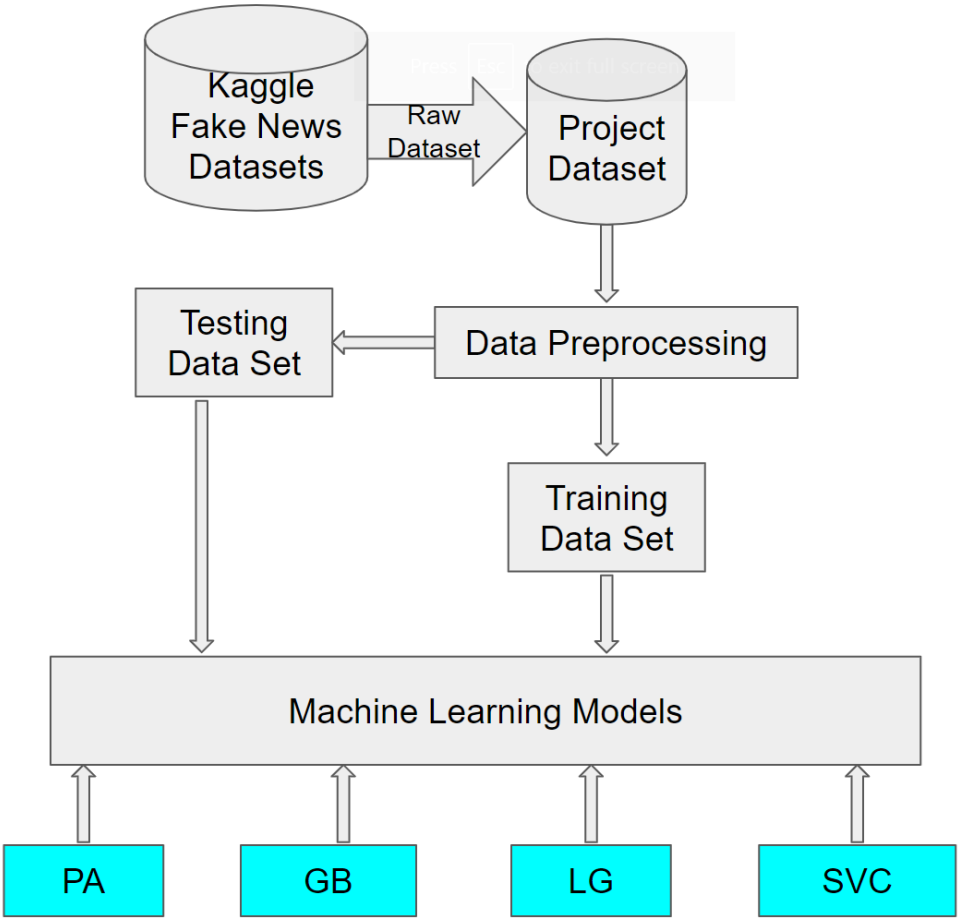
Section 2.4.3: Logistic Regression Classifier (LG)

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The intention behind using logistic regression is to find the best fitting model to describe the relationship between the dependent and the independent variable.

Section 2.4.4: Support Vector Classifier (SVC)

The SVC algorithm is due to the distribution of each data point in the range of dimensions, and the value is the count of the specified coordinates. By giving a set of n features, It will map each data into an n-dimensional feature space at the beginning. Then, It will categorize the hyperplane that the data items into two categories when maximizing the marginal distance for the minimization of the classification error and the actual classes. The class marginal distance is the distance between the decision hyperplane and its nearest instance which is a member of that class (Uddin, 2019).

Section 2.4: Workflow diagrams



Section 3: Project Results, limitations & Future work

Section 3.1: Project Results

We used four machine learning algorithms to build fake news models and compared them based on the accuracy of the models. According to the results in Table 1, SVC has the highest accuracy rate of 95%.

Table 1. Fake news recognition models accuracy table

	PAC	GB	LR	SVC
Accuracy	0.9521	0.9226	0.9418	0.9537

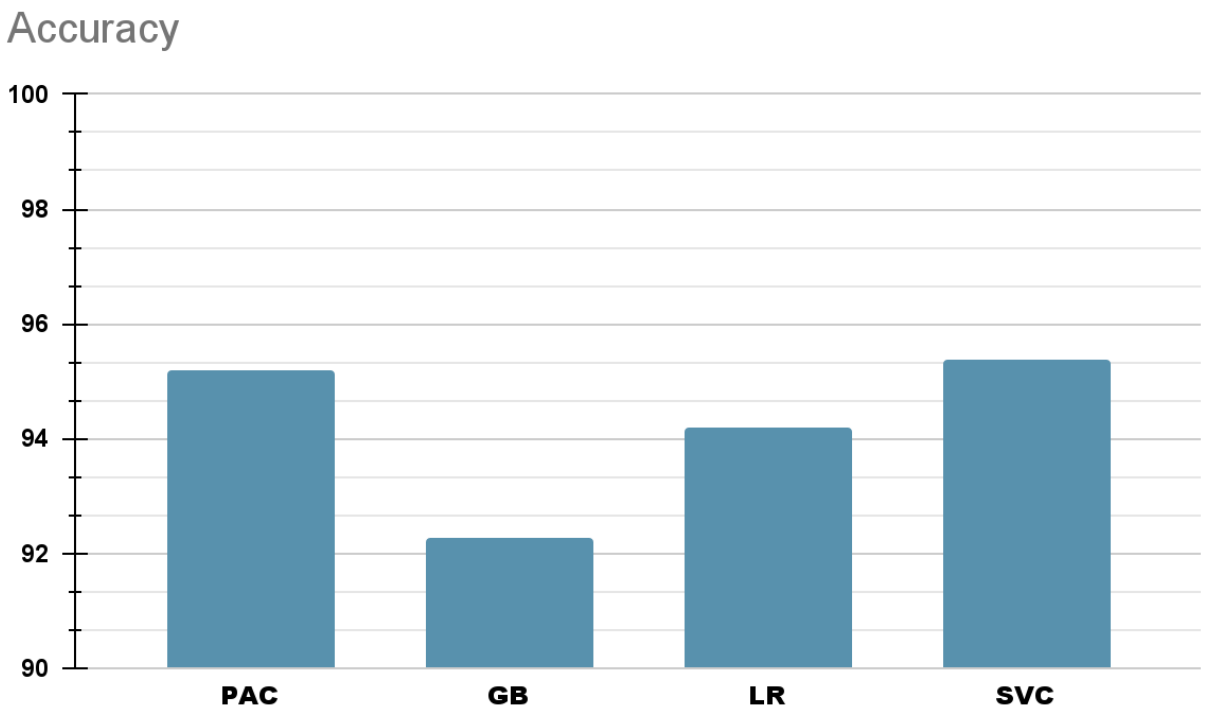


Figure 1. Fake news recognition models accuracy bar-plot

Section 3.2: Project Ethics

Data collected for our project is taken from kaggle. The proposed generalized models will not benefit any political, social, or economic institution. These models will solely be based on the relationship between the “text”, “title” and the labeled “label” provided in the training datasets.In the case of parodies or satirical news pieces, the generalized models constructed for the solution may not recognize the context. Parodies and satire may be classified as fake news.

The cost of misclassification is a concern and should be taken into consideration before making interpretations. Ignorance of the fact that the models are not perfectly accurate and can sometimes lead to misinterpretation of information, can alter credibility. Although attempts are made to avoid overfitting, generalization can be challenging sometimes.
The initiative is based on a representative sampling of all news stories published worldwide. Although sufficient attention has been taken to account for the diversity of news writing styles, it cannot be overlooked that performance may differ on articles that are not adequately represented in the training data.

Section 3.2: Limitations

- We are only working on news data from different sites.
- Time can affect the performance of the models.
- specific features and models can have better performance over time than others.
- We are only working on English language news.
- Availability of limited resources (like GPUs, memory, data storage, etc.), due to which we could not perform many experiments on other large-scale data sources.

Section 3.3: Future work or features

- We want to use web scraping and get the data from various social media and websites by yourself and use them in our system.
- We also want to work on the fake or abusive comments on social media.
- We can build an automatic fact-checking model, that is, compelling the model with some kind of knowledge base or database, the purpose of the model would then be to extract information from the text and verify the information in the database.
- We can build web applications where we can take user input and detect whether the news is fake or real.

Section 3.4: Conclusion

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 weather 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 the Support Vector Classifier (SVC) with an accuracy of 95%.

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 Probierz., 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., Quattrociocchi, 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).*

[6] Markines, B., Cattuto, C., & Menczer, F. (2009, April). Social spam detection. *In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48)*

[7] *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.*