YEAR

2023-4

Machine Learning-based Fake News Detection

a Machine learning model that detects fake news

Kelompok Project Machine Learning

Team Members:

- RAYES JORDAN PRADANA / 2502033102
- STEVEN ANDRIAN PRADITA / 2501996764
- NOVITA ARYANTI / 2502029484

NEXT

Topics

01

Problem Background

Describe the issue the research aims to address

02

Purpose & Benefit

Explain the main objective and potential impact

03

Literature Review

Provide an overview of relevant previous research and key findings

04

Methods

Describe the research methodology used to achieve the objectives

05

Experiments

Briefly summarize any ongoing or completed experiments

06

Progress

Progress of experiments

NEXT

PART I

Problem Background

The rise of social media and digital platforms has made it easier for fake news to spread quickly, making it challenging for people to distinguish between real and fake news. Fake news can have severe consequences, including influencing public opinion, causing panic, and disrupting social and political stability.

NEXT

PART I

Problem Background: How Urgent?

- Infodemics and misinformation about COVID-19, vaccines, and pandemic negatively affect people’s health behaviours —WHO
- Misleading information about the conflicts between Ukraine and Russia
- The increase in fake news during the election period affects the public's preception

NEXT

PART II

Purpose & Benefit

The primary goal of this project is to combat the spread of false information by detecting fake news articles from various sources. With the increasing prevalence of fake news in today's digital age, it has become essential to develop tools that can identify and flag misleading content by detecting and preventing the spread of fake news, this project aims to promote the dissemination of accurate and reliable information to the public.

NEXT

PART III

Literature Review

Paper	Author(s)	Proposed Model	Dataset	Conclusions
Identification of Fake News Using Machine Learning	Rahul R. Mandical, Mamatha N., Shivakumar N., Monica R., & Krishna A.N.	Naïve Bayes, Passive aggressive, DNN	Jru, Pontes, ClaimsKG, Kaggle, Liar, Newsfils, Superset	Each dataset gives different results and performances. DNN gives better results for 6 out of 7 used datasets
Fake News Detection Using Machine Learning Approaches	Z. Khanam, B. N. Alwasel, H. Sirafi, & M. Rashid	XGBoost, Random Forest, Naïve Bayes, KNN, Decision Tree, SVM	Liar	XGBoost gives the highest accuracy with 75% followed by SVM and Random Forest with the accuracy of approximately 73%.
Fake News detection Using Machine Learning	Nihel Fatima Baarir and Abdelhamid Djeffal	SVM	Fake news dataset called "Getting Real about Fake News" and real news dataset called "All the news"	Using the combination of both datasets, SVM proves to give a good result. Parameter Cost C, gamma γ, and epsilon ϵ influences the accuracy result.
Supervised Learning for Fake News Detection	Julio C. S. Reis, Andre Correia, Fabricio Murai Adriano Veloso, & Fabricio Benevenuto	KNN, Naïve Bayes, Random Forest, SVM, XGBoost	BuzzFeed news articles	Random Forest and XGBoost give the best performance based on the F1 score.
Rapid detection of fake news based on machine learning methods	Barbara Probierz, Piotr Stefański, & Jan Kozak a	CART, SVM, Random Forest, AdaBoost, Bagging	ISOT Fake News Dataset	The bagging model gives the highest accuracy rate and F1 score for both real and fake news while SVM gives the worst result.
Effective prediction of fake news using two machine learning algorithms	M. Sudhakar & K.P. Kaliyamurthie	Logistic Regression, Naïve Bayes	Dataset containing fake news from the political information	The logistic regression model gives better performance compared to Naïve Bayes with 98.7080 accuracy result.
Automatic Identification of Fake News Circulation in Social Media using Logistic Regression over Naïve Bayes and Xg Boost Algorithm to Improve Accuracy	C. Balaji, A.Prabhu Chakkaravarthy	Logistic Regression, Naïve Bayes, XGBoost	Two datasets called "TRUE" and "FAKE" obtained from Kaggle	The logistic regression model showed that it gives the highest accuracy with 93.68% followed by XGBoost.

NEXT

PART III

Literature Review: Key Findings

- 2 out of 7 papers concluded that XGBoost gives the highest accuracy or F1 Score among others proposed machine learning models. [2][4]
- 2 out of 7 papers concluded that Logistic Regression gives the highest accuracy result among others proposed machine learning models. [6][7]
- Compared to XGBoost, a paper showed that Logistic Regression gave better performance. [7]
- Random Forest model could be considered as one of the models to use. [2][4]
- Many paper (5 out of 7 papers) used Naive Bayes as a comparison model. [1][2][4][6][7]
- SVM could be considered one of the models to use if the hyperparameter tuning is done. [3]
- Deep learning model such as DNN could be use to do fake news classification. [1]

NEXT

PART III

Hypothesis

The logistic regression model improves the accuracy result compared to XGBoost for fake news classification.

NEXT

PART IV

Methods

The project will use a supervised learning approach to develop a machine learning model. A dataset of news articles from various categories, such as general news, politics, government news, and others, will be used and labeled as real or fake news.

Based on the literature review that we conducted, it can be concluded that XGBoost was found to be the most accurate method in 2 out of the 7 papers that we reviewed, with Random Forest following closely behind. therefore in this experiment we will use 2 algorithms which is the following

NEXT

PART IV

Methods

Doc2vec

Doc2vec is a vectorizer that index each unique word and use neural network model to learn relationship between word. after that it use PV-DBOW (Paragraph Vector - Distributed Bag of Words) to train the document vectors to predict the probability distribution of words in the document so that the model can create high dimensional vector that represents the overal meaning of the documents

NEXT

PART IV

Methods

01

XGBoost

XGBoost is a popular and a efficient open-source implementation of the gradient boosted trees algorithm, which is a supervised learning algorithm that attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

we use this model because it is known for its high accuracy rate, speed, and its ability to handle high dimensional and imbalanced data

02

Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set

we use this model because of its efficiency, ability to handle nonlinear relationships between the independent variables and the outcome, and suitable choice for handling imbalanced datasets

NEXT

PART IV

Methods

Dataset

The dataset is a combination of four popular news datasets: Kaggle, McIntire, Reuters, and BuzzFeed Political. The dataset has 72,134 news articles with 35,028 being real and 37,106 being fake where the real dataset is labeled as 1 and the fake news is labeled as 0. There are 4 columns, the index, the title, the text, and the label.

NEXT

PART IV

Methods

Typical Data in the Dataset – Real vs fake news

- If there is more than 1 data containing the same topic then that data is most likely real
- the data that comes from Reuters (a news o) is always fake, so maybe if the model sees the word Reuters the data is immediately considered as fake
- but in general the data that is either fake or real is difficult to detect without looking at the source and checking it

693	White House official says North Korea is test for U.S.-China relations	WASHINGTON (Reuters) - U.S. President Donald Trump will discuss how to rein in North Korea's nuclear...	0
694	Mexico says upcoming U.S. execution of national is 'illegal'	MEXICO CITY (Reuters) - Senior Mexican diplomats on Monday condemned the upcoming execution of a Mex...	0
695	Khamenei says Iran will 'shred' nuclear deal if U.S. quits it	ANKARA (Reuters) - Iranian Supreme Leader Ayatollah Ali Khamenei said on Wednesday Tehran would stic...	0
696	Myanmar says U.S. sanctions against general based on 'unreliable accusations'	(Reuters) - Myanmar feels sad over a U.S. decision to sanction a military general, a government sp...	0

NEXT

PART IV

Methods

Typical Data in the Dataset – others

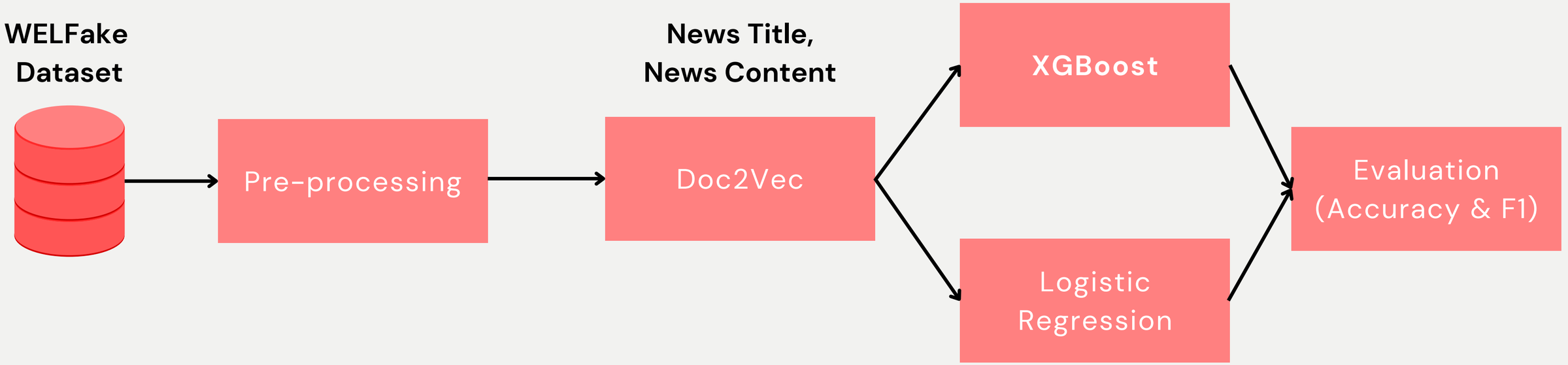
- Some of the data are not all news related, this one for example is a review of a movie

72102	Review: ‘Rogue One’ Leaves ‘Star Wars’ Fans Wanting More and Less - The New York Times	The great mystery of “Rogue One” — the big payoff, the thing people like me would be pilloried fo...	0

- The text section in the dataset doesn't include all of the news content from the source (only some parts of it)
- The data that is partly real news but also partly fake is categorized as fake news

PART IV

Methods



NEXT

PART V

Experiment Result

Model	Accuracy	F1 Score
XGBoost	0.90103	0.90099
Logistic Regression	0.90923	0.90891

Conclusion:
Logistic Regression gives better performance than XGBoost according to the accuracy and F1 score.

Hypothesis: proven (for this dataset)

References

[1] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. Krishna, *Identification of Fake News Using Machine Learning*. 2020. doi: 10.1109/conecct50063.2020.9198610.

[2] Z. Khanam, B. Alwasel, H. Sirafi, and M. Rashid, "Fake News Detection Using Machine Learning Approaches," *IOP Conference Series*, vol. 1099, no. 1, p. 012040, Mar. 2021, doi: 10.1088/1757-899x/1099/1/012040.

[3] N. Baarir and A. Djeffal, *Fake News detection Using Machine Learning*. 2021. doi: 10.1109/ihsh51661.2021.9378748.

[4] J. C. D. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto, and E. Cambria, "Supervised Learning for Fake News Detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/mis.2019.2899143.

[5] B. Probierz, P. Stefański, and J. Kozak, "Rapid detection of fake news based on machine learning methods," *Procedia Computer Science*, vol. 192, pp. 2893–2902, Jan. 2021, doi: 10.1016/j.procs.2021.09.060.

[6] M. Sudhakar and K. P. Kaliyamurthie, "Effective prediction of fake news using two machine learning algorithms," *Measurement: Sensors*, vol. 24, p. 100495, Dec. 2022, doi: 10.1016/j.measen.2022.100495.

[7] C.Balaji, & A.Prabhu Chakkaravarthy. "Automatic Identification of Fake News Circulation in Social Media using Logistic Regression over Naïve Bayes and Xg Boost Algorithm to Improve Accuracy," *Journal of Pharmaceutical Negative Results*, vol. 13, no. SO4, Jan. 2022, doi: 10.47750/pnr.2022.13.s04.073.

Thank You