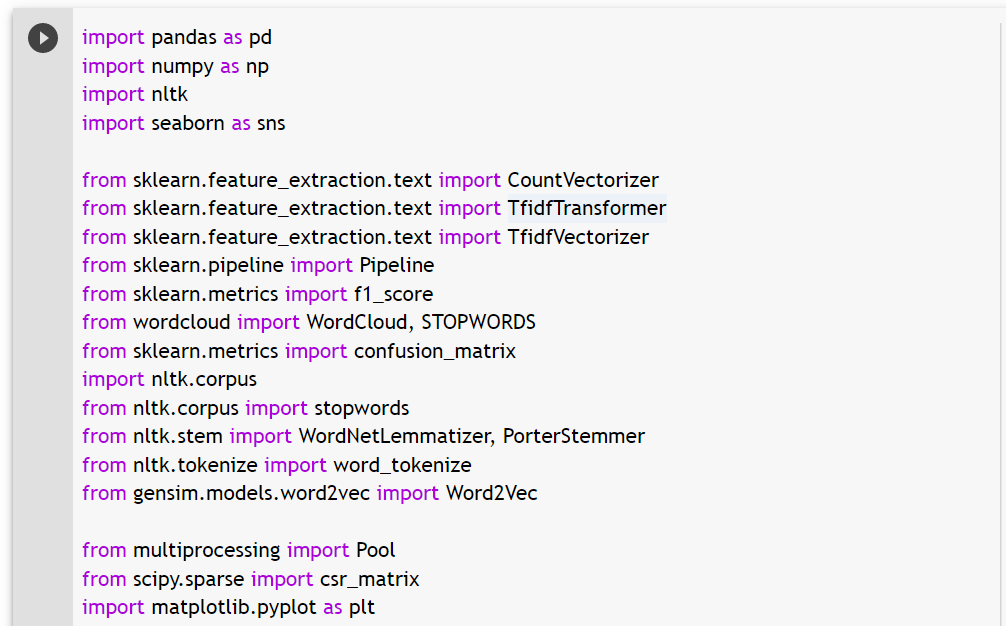
# Appendix A

## Issues and Solutions

1. Pycharm and Visual Studio Code was also used to test pieces of coding during this project. It was downloaded from the Internet free of charge but not demonstrated in the project due to the limitations of Pycharm (importing frameworks and libraries).
2. Python Flask framework was attempted to build the final product but due to time constraints, I decided to completely work on the project using Google Colab environment.

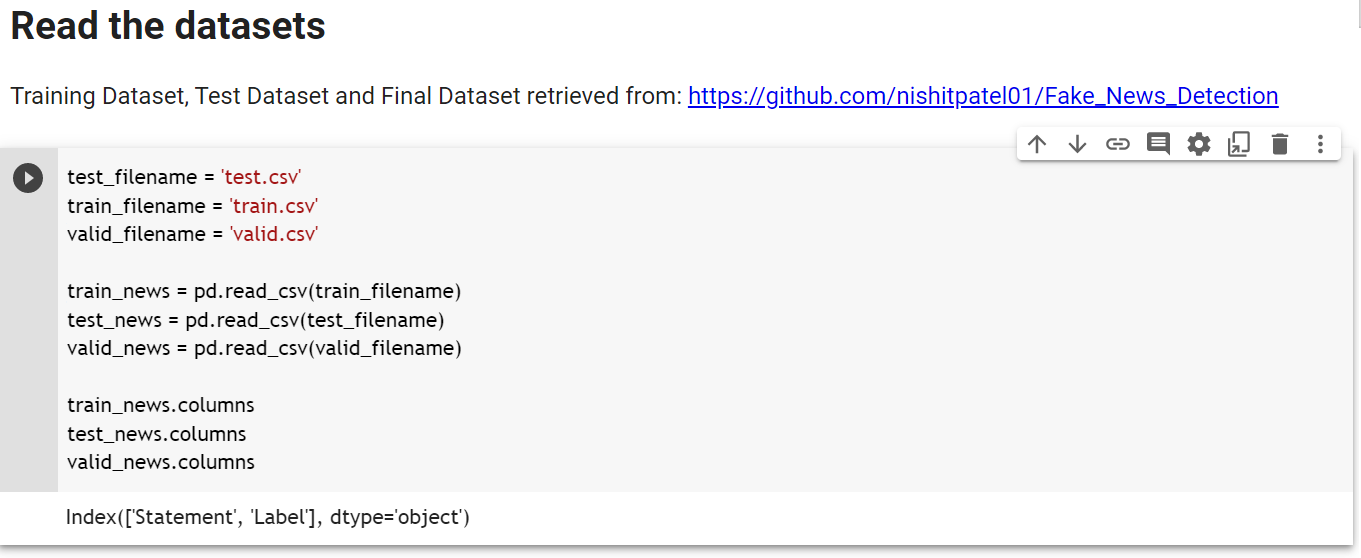
## Coding and Output

1. Import Python Libraries



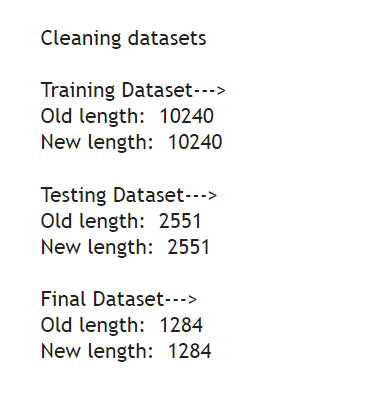
**Figure 1: Importing Python libraries**

1. Read the datasets



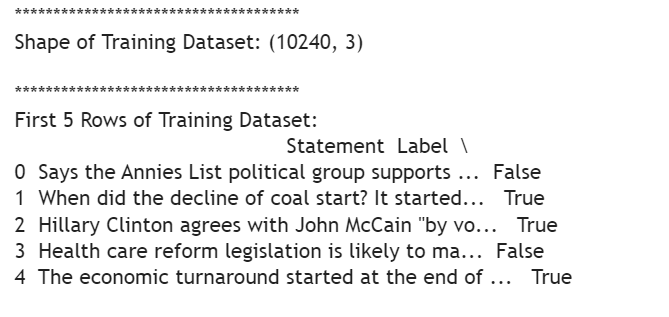
**Figure 2: Read the dataset**

1. Data Pre-processing



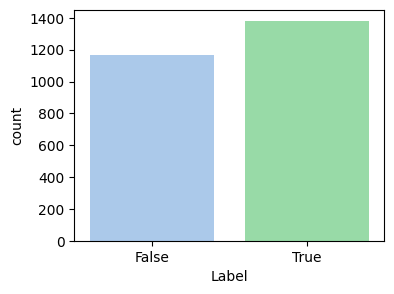
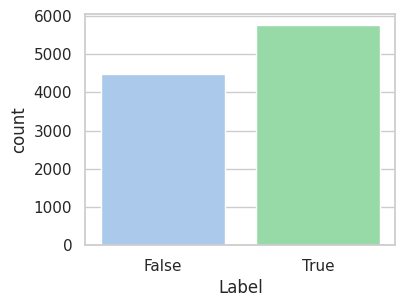
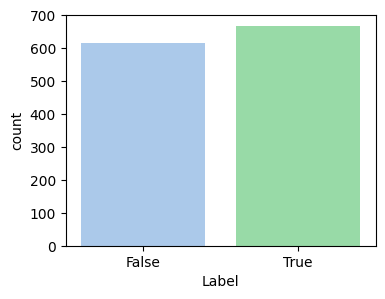
**Figure 3: Pre-processed dataset**

1. View the Datasets



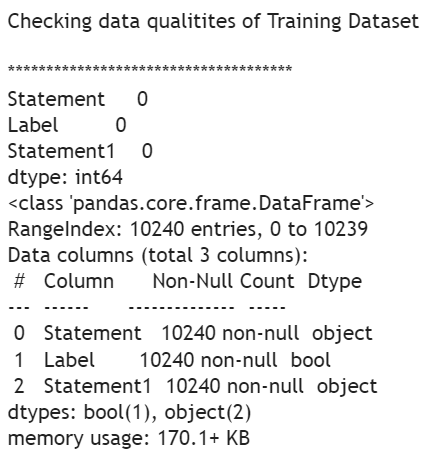
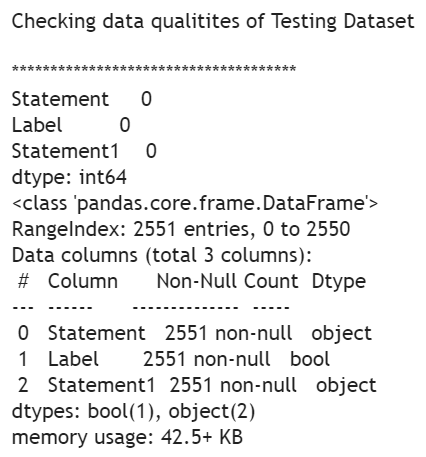
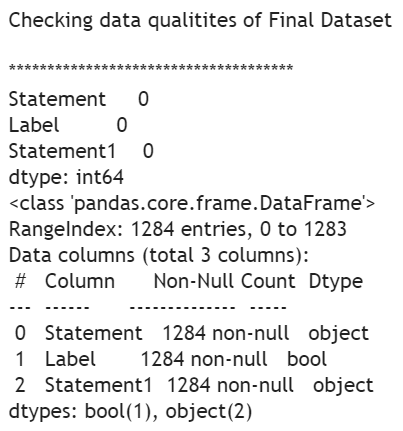
**Figure 4: Shape of training dataset**

1. Data Distribution

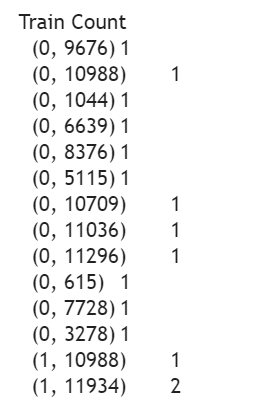
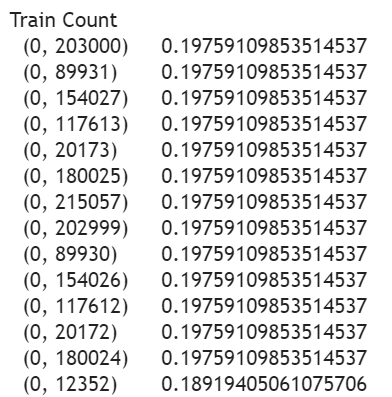
***Figure 5: Data distribution of Training, Testing and Valid datasets***

1. Data integrity check

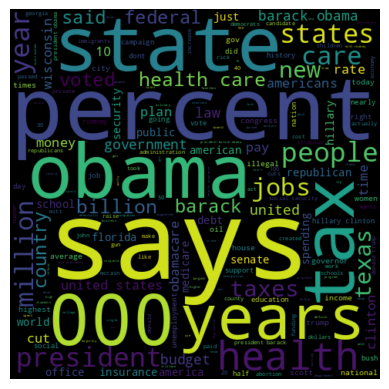
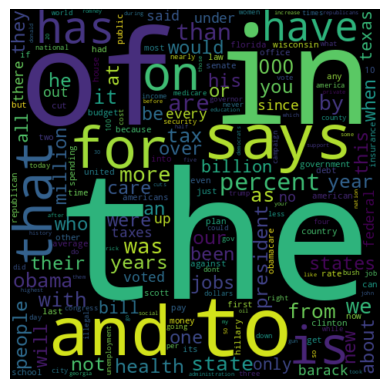
  

***Figure 6: Dataset quality check - Training, Testing and Final datasets***

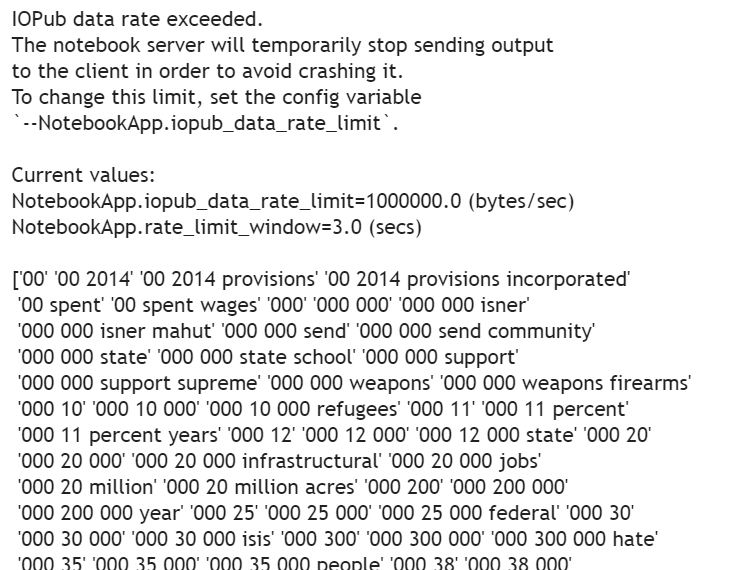
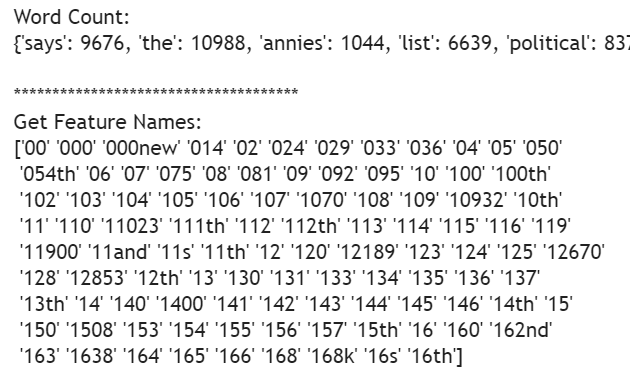
1. Count Vectorizer and Tfidf Vectorizer:

***Figure 7: Train count for CountV & TfidfV***



***Figure 8: Word Clouds for CountV & TfidfV***



***Figure 9: Word Count and Feature names for CountV and TfidfV***

1. Building Models using Bag of Words Technique (Count Vectorizer)

|  |  |
| --- | --- |
| **Figure 10: Classification Report - Model - Naive Bayes** | **Figure 11: Classification Report - Model - Logistic Regression** |
| **Figure 12: Classification Report - Model - SVM** | **Figure 13: Classification Report - Model - SGD** |
| **Figure 14: Classification Report - Model - Random Forest** | **Figure 15: Classification Report - Model - Passive Aggressive** |

1. Building Models using N-Grams Technique (TfidfVectorizer)

|  |  |
| --- | --- |
| **Figure 16: Classification Report - Model - Naive Bayes** | **Figure 17: Classification Report - Model - Logistic Regression** |
| **Figure 18: Classification Report - Model - SVM** | **Figure 19: Classification Report - Model - SGD** |
| **Figure 20: Classification Report - Model - Random Forest** | **Figure 21: Classification Report - Model - Passive Aggressive** |

1. K-fold Cross Validation for all classifiers

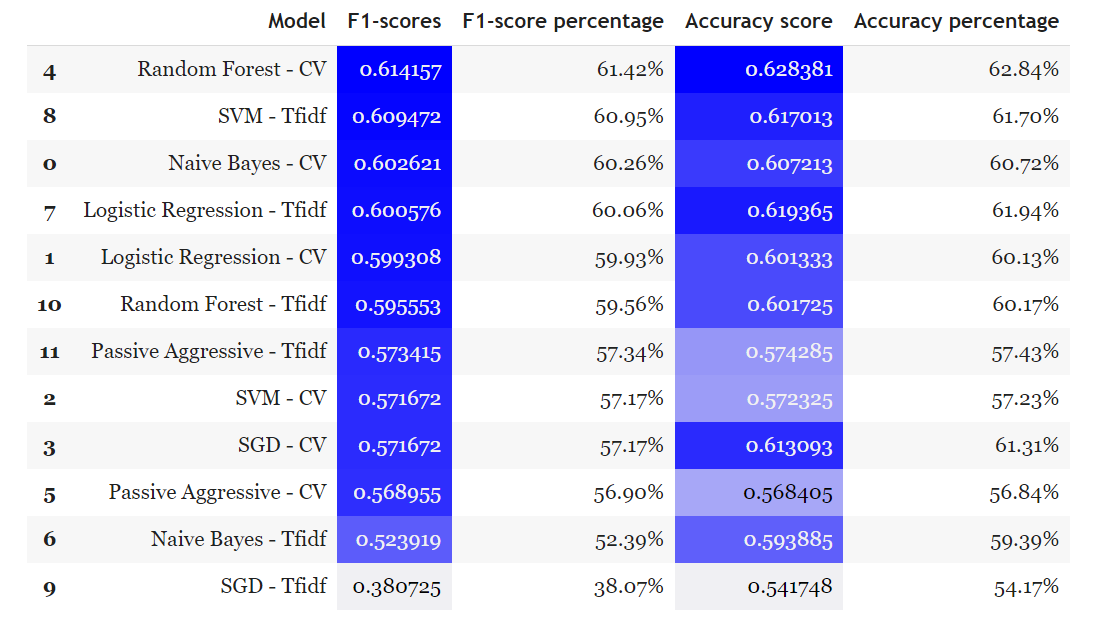
|  |  |
| --- | --- |
| **Graphical representation of Confusion Matrices generated for different Models with Count Vectorizer** | |
| **Figure 22: Confusion Matrix - Model - Naive Bayes - CountV** | **Figure 23: Confusion Matrix - Model - Logistic Regression - CountV** |
| **Figure 24: Confusion Matrix - Model - SVM - CountV** | **Figure 25: Confusion Matrix - Model - SGD - CountV** |
| **Figure 26: Confusion Matrix - Model - Random Forest - CountV** | **Figure 27: Confusion Matrix - Model - Passive Aggressive - CountV** |

|  |  |
| --- | --- |
| **Graphical representation of Confusion Matrices generated for different Models with Tfidf Vectorizer** | |
| **Figure 28: Confusion Matrix - Model - Naive Bayes - TfidfV** | **Figure 29: Confusion Matrix - Model - Logistic Regression - TfidfV** |
| **Figure 30: Confusion Matrix - Model - SVM - TfidfV** | **Figure 31: Confusion Matrix - Model - SGD - TfidfV** |
| **Figure 32: Confusion Matrix - Model - Random Forest - TfidfV** | **Figure 33: Confusion Matrix - Model - Passive Aggressive - TfidfV** |

# Evaluation, Testing and Results

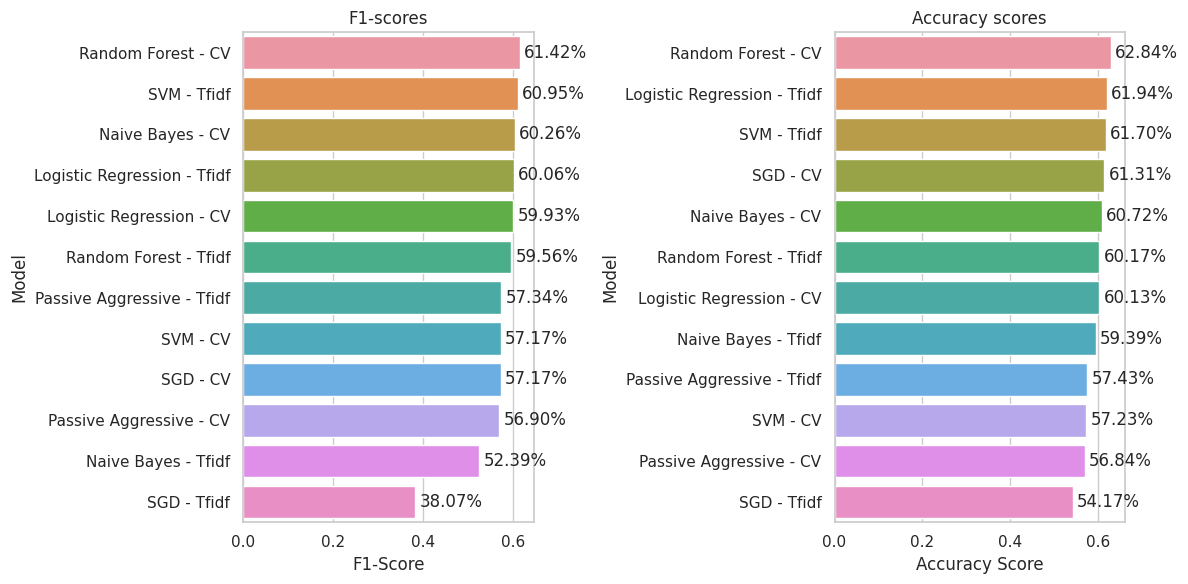
## Model Comparison

### Tabular comparison



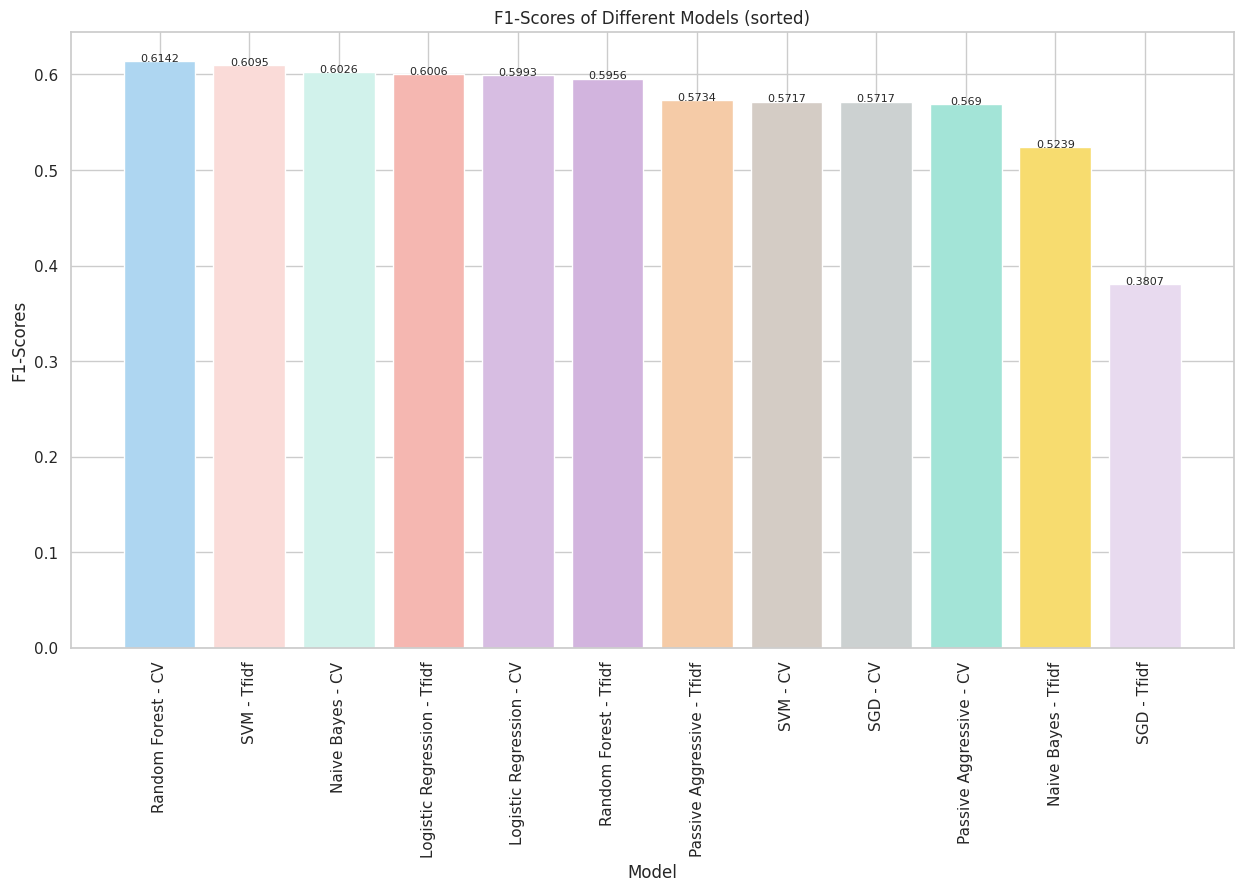
**Table 3: Evaluation - Sort using F1-score**

### Graphical comparison

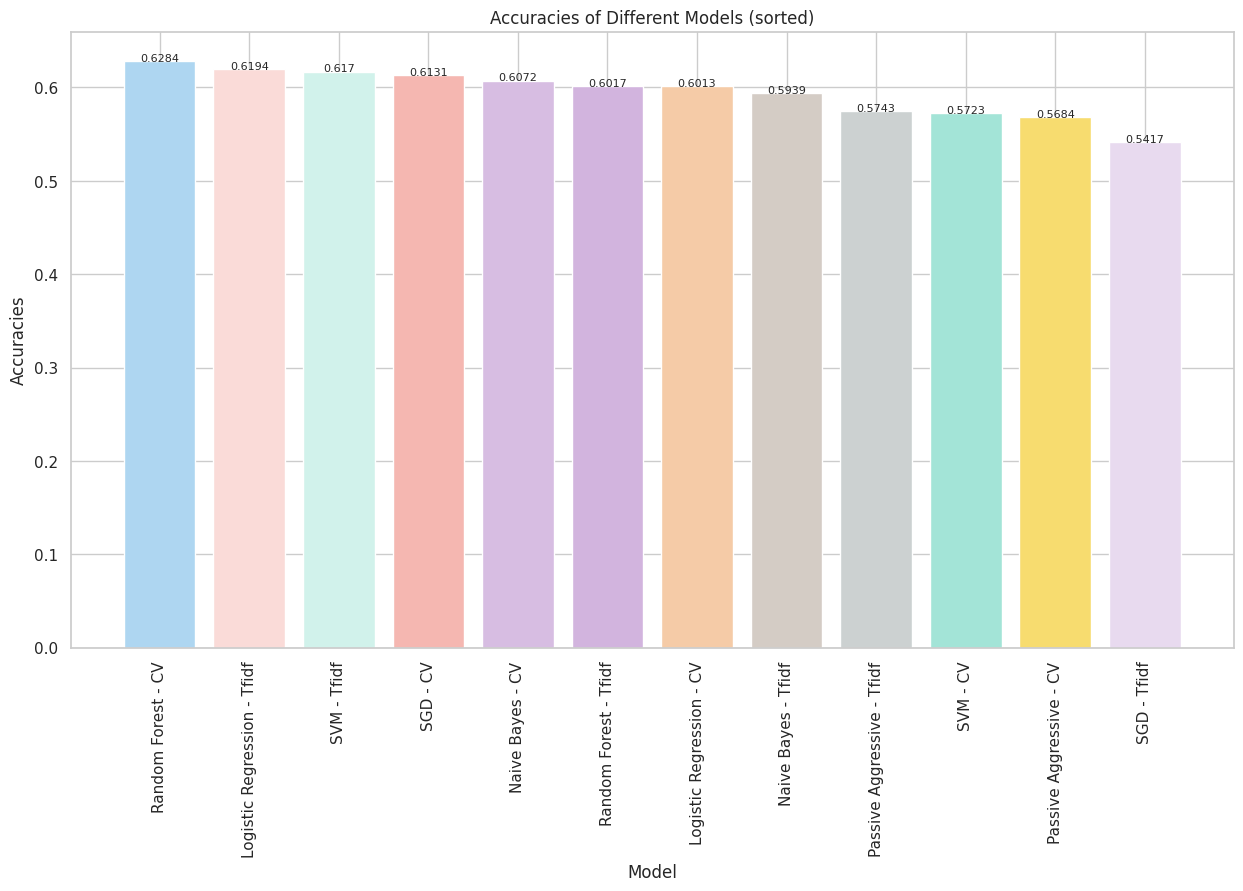


**Figure 34: Evaluation - F1-score and Accuracy score**

### Bar chart comparison



**Figure 35: F1-Scores of Different Models (sorted)**



**Figure 36: Accuracies of Different Models (sorted)**

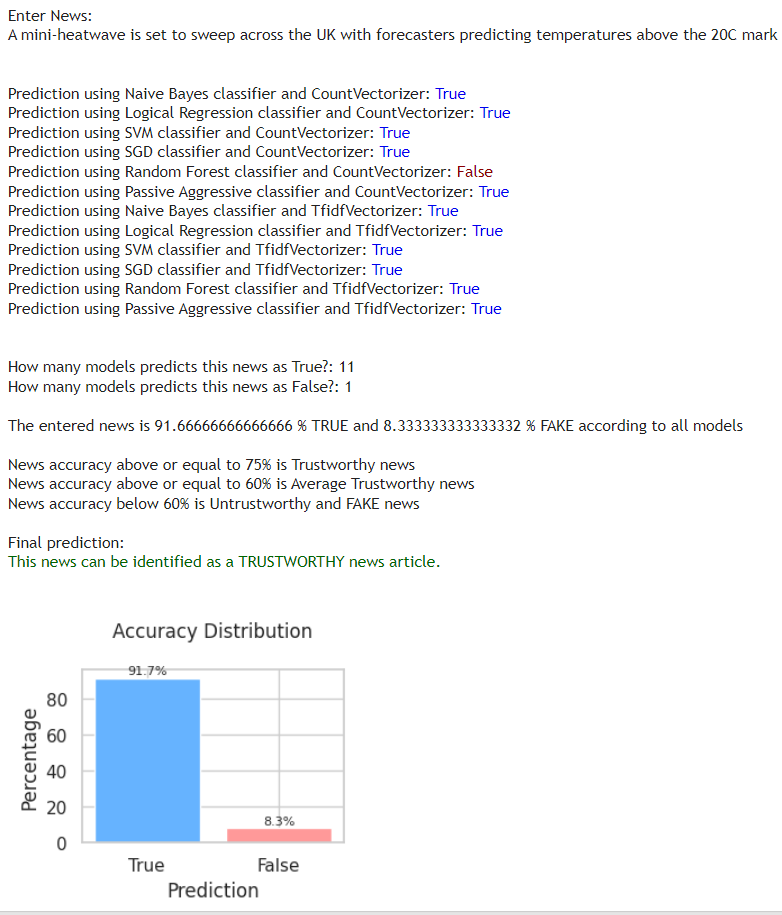
## Testing using Real World News

### News 1

*Title*: UK weather: Mini-heatwave to see Britons bake in 20C heat as hot air sweeps in from Europe

*News*: A mini-heatwave is set to sweep across the UK with forecasters predicting temperatures above the 20C mark in just a matter of weeks…………

*URL*: <https://www.gbnews.com/weather/uk-weather-latest-heatwave-warm-temperatures-from-europe>



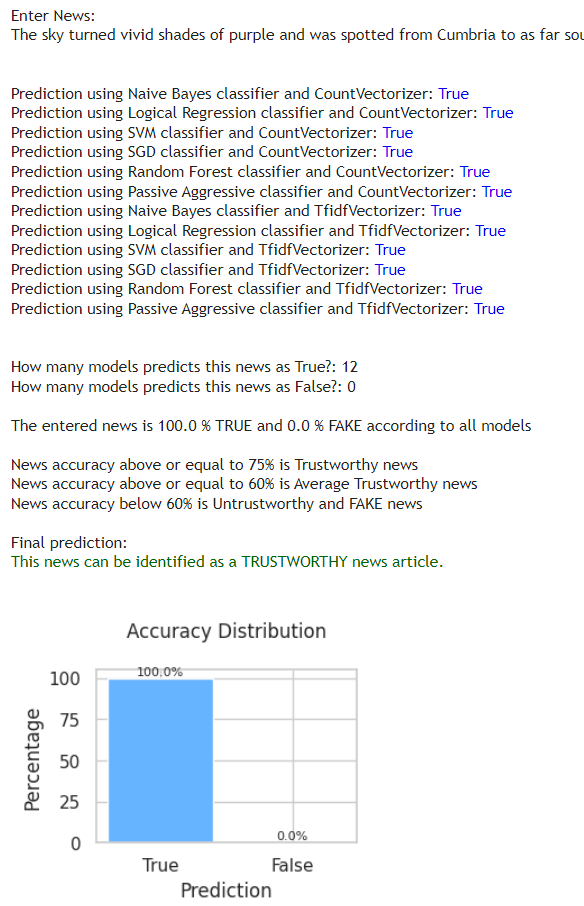
**Figure 37: Predictions for all models - News 1**

### News 2

*Title*: In pictures: Northern lights display sends the skies purple

*News*: The sky turned vivid shades of purple and was spotted from Cumbria to as far south as the Isles of Scilly…………

*URL*: <https://www.bbc.co.uk/news/uk-england-65371512>



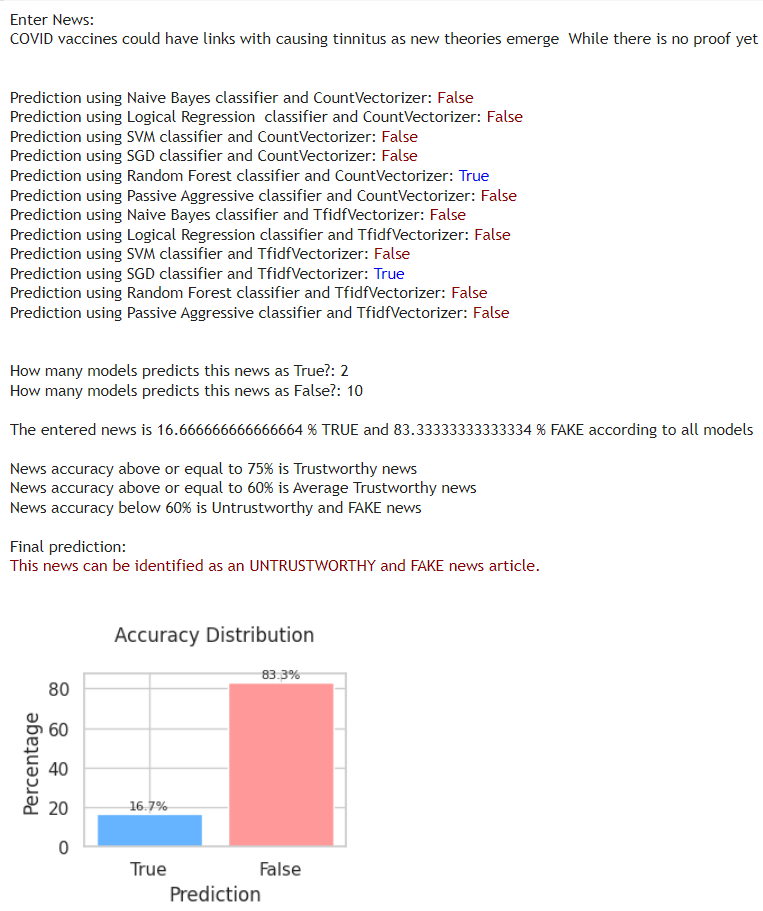
**Figure 38: Predictions for all models - News 2**

### News 3

*Title*: COVID vaccines could have links with causing tinnitus as new theories emerge

*News*: While there is no proof yet that vaccines cause the hearing condition, there are now theories which have surfaced among researchers - here's what you need to know about tinnitus and the causes of the condition.

*URL*: <https://news.sky.com/story/covid-vaccines-could-have-links-with-causing-tinnitus-as-new-theories-emerge-12864899>



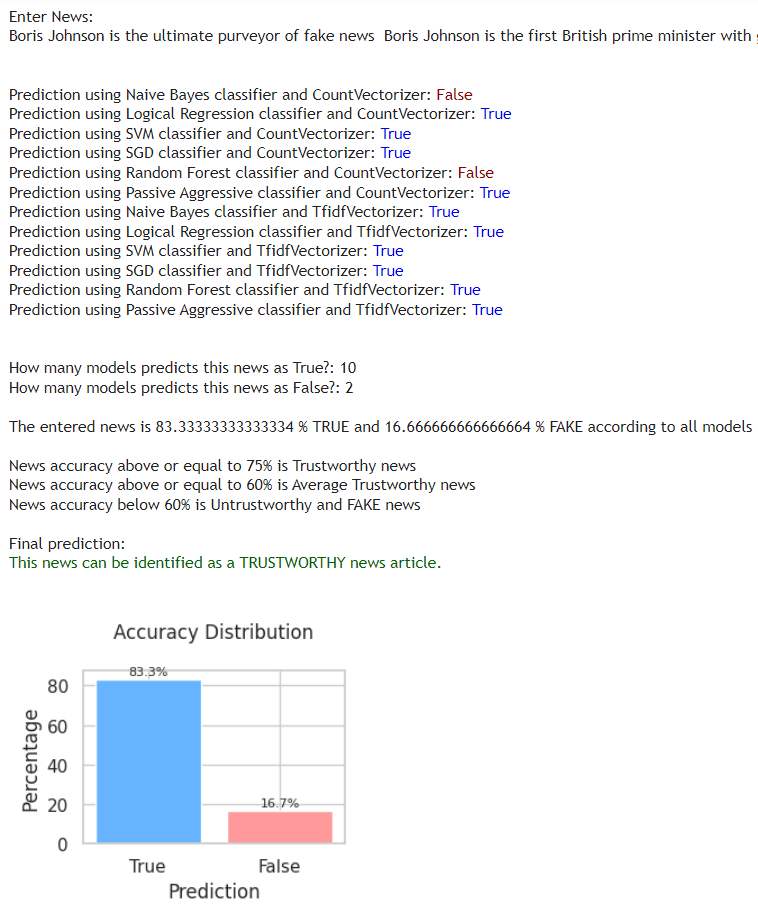
**Figure 40: Predictions for all models - News 3**

### News 4

*Title*: Boris Johnson is the ultimate purveyor of fake news

*News*: Boris Johnson is the first British prime minister with genuine journalistic experience, having been a reporter, columnist and editor. He knows from the inside what the press can achieve. From early in his career, he learned the dark arts……………

*URL*: <https://www.theguardian.com/politics/2020/feb/23/boris-johnson-is-the-ultimate-purveyor-of-fake-news>



**Figure 41: Predictions for all models - News 4**

# All References and Reading Material

All the material I referred and used are listed in this section.

## General information and reading material

DataFlair. (2019). Advanced Python Project - Detecting Fake News with Python. [online] Available at: https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/

www.mindtools.com. (n.d.). MindTools | Home. [online] Available at: https://www.mindtools.com/a0g6bjj/how-to-spot-real-and-fake-news

BBC Bitesize. (n.d.). Fact or fake. [online] Available at: https://www.bbc.co.uk/bitesize/tags/zr2yscw/fact-or-fake/1

www.learndatasci.com. (n.d.). Binary Classification. [online] Available at: https://www.learndatasci.com/glossary/binary-classification/.

de Beer, D. and Matthee, M. (2020). Approaches to Identify Fake News: A Systematic Literature Review. Integrated Science in Digital Age 2020, [online] 136, pp.13–22. doi:https://doi.org/10.1007/978-3-030-49264-9\_2.

de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V. and Mattos, D.M.F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information, [online] 12(1), p.38. doi:https://doi.org/10.3390/info12010038.

Yufeng G (2017). The 7 Steps of Machine Learning. [online] Medium. Available at: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

Alokesh (2020). *Passive Aggressive Classifiers*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/passive-aggressive-classifiers/>.

Arthanayake, N. (2021). *Tips for spotting fake news online*. [online] BBC Bitesize. Available at: <https://www.bbc.co.uk/bitesize/articles/zrprrj6>.

Arya, N. (2022). *Logistic Regression for Classification*. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2022/04/logistic-regression-classification.html>.

B, H.N. (2020). *Confusion Matrix, Accuracy, Precision, Recall, F1 Score*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd>.

de Beer, D. and Matthee, M. (2020). Approaches to Identify Fake News: A Systematic Literature Review. *Integrated Science in Digital Age 2020*, [online] 136, pp.13–22. doi:<https://doi.org/10.1007/978-3-030-49264-9_2>.

de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S.V. and Mattos, D.M.F. (2021). Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. *Information*, [online] 12(1), p.38. doi:<https://doi.org/10.3390/info12010038>.

Dinesh (2021). *කෘතීම බුද්ධිය | Artificial Intelligence in Sinhala*. [online] www.youtube.com. Available at: <https://www.youtube.com/watch?v=A7d_yQ4ECNE&list=PL495mke12zYDHN9ONfcal1eQfo8VqmOgu> [Accessed 25 Apr. 2023].

Gandhi, R. (2018a). *Naive Bayes Classifier*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>.

Gandhi, R. (2018b). *Support Vector Machine — Introduction to Machine Learning Algorithms*. [online] Towards Data Science. Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>.

Johnson, D. (2020). *Stemming and Lemmatization in Python NLTK with Examples*. [online] www.guru99.com. Available at: <https://www.guru99.com/stemming-lemmatization-python-nltk.html>.

Kapl, D., coding, ylan K. has years of experience as a S.D.S.H. enjoys, teaching and everyone, has created this website to make M.L. accessible to (2022). *Machine Learning 101: CountVectorizer Vs TFIDFVectorizer» EML*. [online] enjoymachinelearning.com. Available at: <https://enjoymachinelearning.com/blog/countvectorizer-vs-tfidfvectorizer> [Accessed 25 Apr. 2023].

Kiely, E. and Robertson, L. (2016). *How to Spot Fake News - FactCheck.org*. [online] FactCheck.org. Available at: <https://www.factcheck.org/2016/11/how-to-spot-fake-news/>.

Koo Ping Shung (2018). *Accuracy, Precision, Recall or F1?* [online] Towards Data Science. Available at: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>.

Krasadakis, G. (2021). *How to stop Fake News and misinformation using digital technologies*. [online] The Innovation Mode. Available at: <https://www.theinnovationmode.com/the-innovation-blog/misinformation-online-a-solution-powered-by-state-of-the-art-tech>.

Pandian, S. (2022). *K-Fold Cross Validation Technique and its Essentials*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/>.

Peiris, S. (2023). *MSc Project - Finding Fake News - Process*. [online] Google Docs. Available at: <https://docs.google.com/drawings/d/1weZghLELVhQCpKcIhwlqKIovowDQsG19WNy7l_dSWy0/edit> [Accessed 25 Apr. 2023].

Radhika (2020). *Mathematics Behind SVM | Math Behind Support Vector Machine*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/>.

Saini, A. (2021). *Random Forest Algorithm for Absolute Beginners in Data Science*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/>.

Scikit-learn (2018). *3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.

scikit-learn.org. (n.d.). *1.5. Stochastic Gradient Descent — scikit-learn 0.23.2 documentation*. [online] Available at: <https://scikit-learn.org/stable/modules/sgd.html>.

Tate, S. (2023). *Tackling Fake News with Machine Learning*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2023/02/tackling-fake-news-with-machine-learning/> [Accessed 25 Apr. 2023].

www.tutorialspoint.com. (n.d.). *Scikit Learn - Stochastic Gradient Descent - Tutorialspoint*. [online] Available at: <https://www.tutorialspoint.com/scikit_learn/scikit_learn_stochastic_gradient_descent.htm>.

Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5). doi:<https://doi.org/10.1145/3395046>.

## Fake News Detection Analysis

www.youtube.com. (n.d.). Project 4. Fake News Prediction using Machine Learning with Python | Machine Learning Projects. [online] Available at: https://www.youtube.com/watch?v=nacLBdyG6jE&t=1980s [Accessed 25 Apr. 2023].

www.youtube.com. (n.d.). Fake News Detection Project | Data Science with Python | Machine Learning session 2. [online] Available at: https://www.youtube.com/watch?v=xyq-zYr1cnI&t=6330s [Accessed 25 Apr. 2023].

## Dataset & Coding references

nishitpatel01 (2018). nishitpatel01/Fake\_News\_Detection. [online] GitHub. Available at: https://github.com/nishitpatel01/Fake\_News\_Detection [Accessed 29 May 2019].

### Pipeline function

scikit-learn.org. (n.d.). sklearn.pipeline.Pipeline — scikit-learn 0.24.1 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html.

### Word Cloud

Holtz, Y. (n.d.). Custom python wordcloud. [online] The Python Graph Gallery. Available at: https://www.python-graph-gallery.com/261-custom-python-wordcloud?utm\_content=cmp-true [Accessed 25 Apr. 2023].

### Documentation reference

Shoemaker, E. (n.d.). JMU Scholarly Commons Using data science to detect fake news. [online] Available at: https://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1754&context=honors201019.

FAKE NEWS DETECTION WITH REAL NEWS GENERATION A PROJECT REPORT BACHELOR OF TECHNOLOGY DEPARTMENT OF INFORMATION TECHNOLOGY FACULTY OF ENGINEERING AND TECHNOLOGY SRM INSTITUTE OF SCIENCE AND TECHNOLOGY KATTANKULATHUR -603203. (2021). Available at: http://dspace.srmist.edu.in/jspui/bitstream/123456789/44561/1/P11595.pdf [Accessed 25 Apr. 2023].

## Coding references

Badole, Mayur. “Difference between Fit(), Transform(), Fit\_transform() Methods in Scikit-Learn (with Python Code).” *Analytics Vidhya*, 30 Apr. 2021, www.analyticsvidhya.com/blog/2021/04/difference-between-fit-transform-fit\_transform-methods-in-scikit-learn-with-python-code/.

BSD. “Sklearn.feature\_extraction.text.CountVectorizer — Scikit-Learn 0.20.3 Documentation.” *Scikit-Learn.org*, 2018, scikit-learn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html.

---. “Sklearn.pipeline.Pipeline — Scikit-Learn 0.24.1 Documentation.” *Scikit-Learn.org*, scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html.

Dickinson, Bryan. “Generate Meaningful Word Clouds in Python.” *Medium*, 12 Feb. 2021, towardsdatascience.com/generate-meaningful-word-clouds-in-python-5b85f5668eeb.

Dinesh. “Python Tutorial - 55 | Pickle Module (Serializing) | Sinhala.” *Www.youtube.com*, 2021, www.youtube.com/watch?v=D-zT4tqbebQ&t=2s. Accessed 25 Apr. 2023.

Harrison. “Python Programming Tutorials.” *Pythonprogramming.net*, pythonprogramming.net/python-pickle-module-save-objects-serialization/.

Khanna, Chetna. “Text Pre-Processing: Stop Words Removal Using Different Libraries.” *Medium*, 10 Feb. 2021, towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a.

---. “What and Why behind Fit\_transform() vs Transform() in Scikit-Learn !” *Medium*, 25 Dec. 2020, towardsdatascience.com/what-and-why-behind-fit-transform-vs-transform-in-scikit-learn-78f915cf96fe.

Kharwal, Aman. “Classification Report in Machine Learning.” *Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript*, 7 July 2021, thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/.

Nantasenamat, Chanin. “Google Colaboratory.” *Colab.research.google.com*, 29 Mar. 2023, colab.research.google.com/drive/1Tmqs4934VEE2xMdQlCiYwY-GhQsakL\_n#scrollTo=-JbHzHy6oEix. Accessed 25 Apr. 2023.

Nishit Patel. “Fake News Detection.” *GitHub*, 23 Apr. 2023, github.com/nishitpatel01/Fake\_News\_Detection/blob/master/final-fnd.ipynb. Accessed 25 Apr. 2023.

Rene D. “ANSI Color Codes in Python.” *Gist*, 2020, gist.github.com/rene-d/9e584a7dd2935d0f461904b9f2950007.

SciKit-Learn. “3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.21.3 Documentation.” *Scikit-Learn.org*, 2009, scikit-learn.org/stable/modules/cross\_validation.html.

Sphinx. “Pandas.DataFrame.dropna — Pandas 1.3.1 Documentation.” *Pandas.pydata.org*, pandas.pydata.org/docs/reference/api/pandas.DataFrame.dropna.html.

---. “Seaborn.color\_palette — Seaborn 0.12.2 Documentation.” *Seaborn.pydata.org*, seaborn.pydata.org/generated/seaborn.color\_palette.html.

---. “Styling — Pandas 1.1.2 Documentation.” *Pandas.pydata.org*, pandas.pydata.org/pandas-docs/stable/user\_guide/style.html.

StudyTonight. “How to Print Colored Text in Python - Studytonight.” *Www.studytonight.com*, www.studytonight.com/python-howtos/how-to-print-colored-text-in-python?utm\_content=cmp-true. Accessed 25 Apr. 2023.

Vasanth. “How to Combine Single Pkl File? · Issue #1431 · Open-Mmlab/Mmaction2.” *GitHub*, 2022, github.com/open-mmlab/mmaction2/issues/1431. Accessed 25 Apr. 2023.

Waskom, Michael. “Seaborn.countplot — Seaborn 0.9.0 Documentation.” *Pydata.org*, 2012, seaborn.pydata.org/generated/seaborn.countplot.html.