**Portfolio Project: Fake News Detection**

Prudence Gemmell

Colorado State University Global

MIS 581: Capstone – Business Intelligence and Data Analytics

Dr. Steve Chung

May 14, 2023

**Contents**

[Abstract 3](#_Toc135094861)

[Introduction 4](#_Toc135094862)

[Objectives 5](#_Toc135094863)

[Overview of Study 6](#_Toc135094864)

[Research Question and Hypothesis 8](#_Toc135094865)

[Literature Review 9](#_Toc135094866)

[Research Design 12](#_Toc135094867)

[Methodology 13](#_Toc135094868)

[Methods 13](#_Toc135094869)

[Limitations 15](#_Toc135094870)

[Ethical Considerations 16](#_Toc135094871)

[Conclusion 19](#_Toc135094872)

[Recommendations 20](#_Toc135094873)

[References 21](#_Toc135094874)

# Abstract

The spread of misinformation has developed into a global crisis. It can cause harm to individuals and society by eroding trust in institutions, influencing public opinion, and leading to misguided decisions and actions that can have serious consequences, such as public health crises, social unrest, and political instability. Fake news refers specifically to false information deliberately created and disseminated with the intent to deceive and manipulate people. It is a subset of the broader category of misinformation, but it is distinct in that it is intentionally fabricated and often presented in a sensational or misleading way to attract attention and generate clicks or views. Therefore, fake news detection is needed to identify and verify the accuracy of news and information and flag any suspected news articles as false. Detecting and debunking fake news can help to mitigate these risks by promoting greater accuracy and accountability in media and information sources, and by enabling individuals to make more informed decisions based on reliable and trustworthy information. Fake news detection techniques used within this project include feature engineering and extraction along with a machine learning model, logistic regression algorithm. These techniques were used on a public dataset to predict whether a news article was fake or real based solely on the title of the article. There was a 98% accuracy of predicting fake news using the combination of stop word removal and lemmatization.

**Portfolio Project: Fake News Detection**

# Introduction

Fake news, disinformation, misinformation, propaganda, or alternative facts are a serious problem within today’s society. “Fake news refers to all kinds of false stories or news that are mainly published and distributed on the Internet, in order to purposely mislead, befool or lure readers for financial, political or other gains” (Almansa, 2020, Introduction). Fake news can spread misinformation that can mislead people and affect their decisions. If people believe false information, they may make decisions based on inaccurate data, leading to negative consequences. Fake news destroys the public trust in journalism and media, making it more difficult for people to differentiate between factual news and fabricated stories. This can ultimately weaken the reliability of democratic institutions and lead to social unrest.

Additionally, people tend to believe news stories that confirm their existing beliefs or opinions. Unfortunately, fake news can exploit this tendency, reinforcing people's biases and leading to further division in society. Fake news can be used as a tool to manipulate public opinion or influence the outcomes of elections. This is particularly concerning in the age of social media, where fake news stories can be shared quickly and widely. Overall, fake news is a significant problem because it undermines the ability of people to make informed decisions and contributes to the destruction of trust in media and democratic institutions (Raza & Ding, 2022).

“Data analyst tools is a term used to describe software and applications that data analysts use in order to develop and perform analytical processes that help companies to make better, informed business decisions while decreasing costs and increasing profits” (Datapine, 2021, para. 2). Data analytics tools first require a high-quality dataset since the data directly impact the

accuracy, reliability, and usefulness of the insights that can be derived from it. Accuracy in data is crucial since errors can lead to inaccurate insights, which has to the potential to be extremely costly for businesses. A good dataset should also be complete and consistent. The data should contain all the necessary data points required for analysis. Incomplete datasets can lead to biased results and missing out on important insights. The data should also follow a standard format and structure. Inconsistent datasets can make analysis difficult and can lead to inaccurate insights as well. Additionally, the dataset should be relevant to the analysis at hand since irrelevant data can lead to wasted time and resources and may not provide any useful insights. The dataset should be up-to-date and timely. Outdated data can lead to inaccurate insights and may not reflect current trends or situations. Analytics tools rely on good datasets to function effectively. Without quality data, analytics tools cannot provide accurate insights or predictions. In short, a good dataset is the foundation of data analytics and analytics tools, and it is essential for deriving meaningful insights and making data-driven decisions (Stedman & Vaughan, 2022).

# Objectives

Data analytics plays a crucial role in the detection of fake news by enabling researchers to identify patterns and trends in data that might indicate the presence of misinformation. One main way in which data analytics helps in the detection of fake news is through text analysis. “Text analysis is the process of using computer systems to read and understand human-written text for business insights. Text analysis software can independently classify, sort, and extract information from text to identify patterns, relationships, sentiments, and other actionable knowledge” (AWS, 2023, para. 1). Additionally, Artificial Intelligence (AI) is vital in fake news detection through natural language processing (NLP) and machine learning (ML) algorithms. AI is used in labeling of fake news datasets due to its capability of handling large datasets and the huge amount of human effort required to complete this task. However, labeling incorrectly would have a colossal effect on the performance of a supervised learning model (Almansa, 2020). The objective of this research is to analyze a fake news dataset in order to determine if a ML algorithm can accurately predict whether a news article is fake or real based solely on the article’s title since other research has only been done on detecting fake news through the body of the news article. Additionally, five different preprocessing methods are used as inputs for the ML algorithm to see which one will provide the best fit.

# Overview of Study

A public dataset of fake and real news (Bisaillon, 2020) found on Kaggle, a repository of public datasets, was chosen for this research project since it only contained a few columns specific to the research question. The dataset provides detailed information about almost two thousand articles, such as the title, the text, the subject, and the date. The dataset will be analyzed to train and test a ML algorithm with different inputs in order to predict fake news. The dataset comes in two comma-separated valued (CSV) files: fake news and true news. The fake news dataset consists of four columns to describe a list of 21.4 thousand articles, while the true news dataset consists of four columns to describe a list of 23.5 thousand articles. The four columns represent the article title, the article text, the subject of the article, and the date the article was published. None of the columns have missing values, which will reduce the amount of time required for preprocessing. The title, text, and subject columns are all string data types, while the date is of data type datetime. The two files are combined to create one dataset with a fifth column added to indicate whether the article is fake or true. This column has a data type of Boolean since it is a yes or no column.

The visual model below represents a data dictionary for the fake news csv files, shown in Figure 1, and the true news csv files, shown in Figure 2. The columns are broken down into the title of the article, the text of the article, the subject of the article, and the date the article was published. Figure 3 represents a data dictionary for the combined dataset of the fake news and true news datasets.

**Table 1**

*Data dictionary of fake news dataset*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| title | string | The title of the article |
| text | string | The text within the article |
| subject | string | The subject of the article, such as news, politics, other |
| date | datetime | The publish date of the article; Date range of articles are from January 12, 2016 to December 30, 2017 |

**Table 2**

*Data dictionary of true news dataset*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| title | string | The title of the article |
| text | string | The text within the article |
| subject | string | The subject of the article, such as worldnews and politicsnews |
| date | datetime | The publish date of the article; Date range of articles are from March 30, 2015 to February 18, 2018 |

**Table 3**

*Data dictionary of combined fake and true news datasets*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Description** |
| title | string | The title of the article |
| text | string | The text within the article |
| subject | string | The subject of the article, such as news, politics, other |
| date | datetime | The publish date of the article |
| fake | Boolean | Fake news = 1; real news = 0 |

The next section discusses the research question and hypothesis statements related to detecting fake news by analyzing the title of a news article. After which, an overview of related works is presented to show how processing tools and ML techniques were used to analyze the body of news articles to determine whether the articles were fake. The research design section discusses the research architecture along with the specific requirements needed to conduct the experiment. The methodology, methods, limitations, and ethical considerations are presented within this section along with the related diagrams of the data workflow. The findings section details the experimental results based on the ML model used with the five different feature extractions. The final sections close the paper with the conclusion and recommendations for future work.

# Research Question and Hypothesis

The purpose of hypotheses in statistical analysis is to help organizations make informed conclusions about the population based on a sample of data. A hypothesis is a statement about a population parameter, such as the mean or the proportion, that tests a sample of data. There are two types of hypotheses in statistical analysis: the null hypothesis and the alternative hypothesis. The purpose of testing hypotheses is to determine whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. This is done using statistical tests, such as t-tests or logistic regression, which calculates the probability of obtaining sample results to determine if there is enough evidence to fail to reject the null hypothesis. If this probability is very low (typically below a pre-specified threshold, such as 0.05 or 0.01), the null hypothesis is rejected in favor of the alternative hypothesis. By testing hypotheses in this way, organizations can make more accurate and informed conclusions about the population based on a sample of data. It allows better understanding into whether any observed differences between the sample and the population are due to chance or are statistically significant, which is important for making decisions in any industry.

The public fake and real news dataset published by Bisaillon (2020) will be used to classify news articles as fake or real along with detecting patterns on their linguistic features to find a correlation between the title of the article and the likelihood of the article being fake.

RQ: Can the title of an article be used to predict whether an article is fake or real?

H0: There is no significant relationship between the title of an article and its classification as either fake or real.

Ha: There is a significant relationship between the title of an article and its classification as either fake or real.

# Literature Review

Ahmad et al. (2020) primarily focused on using machine learning ensemble methods in fake news detection. Their proposed framework expanded the current literature in classifying news articles as true or fake by introducing novel ensemble techniques using a variety of linguistic features from three different datasets taken from public repositories and ISOT Fake News Dataset and the fourth dataset was the combination of the other three. Firstly, the datasets were preprocessed and cleaned before inputting them into their models for training. The feature sets were created using the Linguistic Inquiry and Word Count (LIWC) tool which extracted 93 different features from any given news article. The linguistic features converted text into a numerical form for use of training different machine learning models. For training and testing the models, each dataset used the 70/30 split and then shuffled to ensure an even distribution of fake and true news articles. The ensemble techniques used in examining the performance of the models included bagging, boosting, and voting classifier. Additionally, performance metrics, such as accuracy, precision, recall, and F1, were used to evaluate the performance of each model. Their reasoning for using ensemble learners is because they “tend to have higher accuracies, as more than one model is trained using a particular technique to reduce the overall error rate and improve the performance of the model” (Ahmad et al., 2020, Section 2). The LR accuracies for the different datasets ranged from 87% to 97%.

Shoemaker (2019) suggests that ISOT Fake News Dataset is not a suitable dataset train with since the classifiers pick up on the pattern easier than other datasets. She recommends using the ISOT dataset for testing purposes only. Shoemaker used the Natural Language Toolkit (NLTK) to preprocess three datasets used within their research instead of the LIWC tool. The three datasets, which included the ISOT dataset, the FakeNewsNet dataset, and an original dataset, used a train-test split of 80/20. The purpose of their research was to identify the effectiveness of different features for different classifiers, such as Random Forest (RF) and Naïve Bayes (NB). The classification accuracies for the ISOT news dataset averaged 93% for the RF model and 88% for the NB model.

Bharadwaj et al. (2020) analyzed multiple datasets found online and shortlisted it to one since all other datasets lacked information required for their research. The features that were considered included author, published date, language, site\_url, image\_url, type, label, title, text, and hasImage. The dataset included eight different labels for classifying the types of fake news, such as bias, bs, hate, state, conspiracy, satire, junksci, and fake. Their primary goal was to present comprehensives steps used for classifying fake news through different machine learning models in conjunction with feature engineering and extraction processes. The dataset was split in three different ways: 60/40, 70/30, and 80/20 train-test. They compared Term Frequency – Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe with different machine learning models. The Word2Vec technique was used for feature extraction to train words against other neighboring words, while GloVe is a technique used for word representation through global vectors. The LR model using TF-IDF on a 70/30 split resulted in 75% accuracy.

Kumari (2022) proposes using supervised learning to predict fake new by calculating the accuracy of different models. Kumari compares logistic regression, naïve-bayes, decision tree and passive-aggressive classifier. Supervised learning is a machine learning technique that uses labelled datasets to train algorithms. Each model evaluates the probably of an event occurring given independent variables from the datasets. The purpose of the dataset is to train an algorithm to determine the difference between fake and real news articles. Once the algorithm is trained, it could then be applied to new and real-time news stories to detect any false claims. Ideally, any media organization could then use this algorithm and apply it to their own historical data in order to flag potential fake news and put those offenders on a watch list. This combination produced a 91% accuracy using TF-IDF with the LR model.

Sudhakar and Kaliyamurthie (2022) used the LR algorithm and the NB algorithm to train and test the data, which was stated to be confidential. There was no processing of the data mentioned, only that the data was split in four different ways. The first way was a 90% training dataset with a 10% test dataset. The subsequent splits were 80/20, 70/30, and 60/40. The best accuracy for the NB algorithm was with the 80/20 split with a 95% accuracy prediction. The best accuracy for the LR algorithm was with the 70/30 split with a 99% accuracy prediction.

# Research Design

Data analytics can be a powerful tool for detecting fake news because it allows for large amounts of data from various sources to be analyzed, such as social media, news articles, and online forums, to identify patterns and anomalies that may indicate the presence of misinformation or propaganda. Some specific reasons why data analytics is useful for fake news detection include scalability, speed, accuracy, and integration. With the sheer volume of data available online, it is impossible for humans to manually monitor every piece of content. Data analytics can help process large volumes of data quickly and efficiently to identify potential instances of fake news. Additionally, fake news can spread rapidly online, making it critical to detect and address it quickly. Data analytics can help identify and track the spread of fake news in real-time, allowing for a more rapid response. By using data-driven methods, fake news detection can be more accurate and objective than relying solely on human judgement, which can be prone to bias and errors. Data analytics can be integrated with other tools, such as natural language processing and machine learning, to improve the accuracy and effectiveness of fake news detection. Overall, data analytics can help the industry better understand the spread and impact of fake news online and enable organizations to take proactive steps to combat it.

The motivation for this research paper is due to other research on the detection of fake news only being tested on the body of the news article. By testing the title of the articles first, it could ideally eliminate extra processing time on the articles are identified as real news. A powerful analytics tool that is commonly used in the detection of fake news is Python, since it has a variety of packages that can be used to handle historical data, to provide insights on trends, and to offer data visualizations for any audience.

# Methodology

Since “qualitative research refers to research which focuses on collecting and analyzing words (written or spoken) and textual or visual data, whereas quantitative research focuses on measurement and testing using numerical data” (Jenson & Warren, 2023, Section 2), both qualitative and quantitative methodology will be used within the research of detecting fake news based on an article’s title. Qualitative content analysis focuses on investigating written works, such as articles, to provide insights into communication trends. Additionally, frequency analysis will be performed on the data to assess the frequency of the terms within the title, which will be part of the data preprocessing. Quantitative research methods will be used to test the null hypotheses and alternative hypotheses associated with the research questions proposed to detect fake news. Quantitative research is used to find patterns, test relationship, and make predictions by analyzing numerical data. Statistical analysis through the use of inferential statistics, will be used to determine the accuracy of the LR algorithm in the prediction of fake news articles (Bhandari, 2021). Lastly, the dataset was readily available through a public dataset repository. Bisaillon (2020) did not specify the methods used in the collection of the news articles.

# Methods

One way of detecting fake news involves analyzing news article. There are several data analytics tools and techniques that can be used to detect fake news, but this paper will focus on text analysis, which will include natural language processing (NLP) and ML algorithms. NLP tools, such as the Natural Language Toolkit (NLTK), will analyze the text within the title of the news articles which will then be used to help identify patterns to indicate whether an article is fake. NLTK is a toolkit used for the preprocessing of the dataset, such as stop word removal, lowercase conversion, stemming, and lemmatization (Lemma). Next feature engineering will be performed using the TF-IDF (term frequency-inverse document frequency) technique, which is “a weighting metric often used in information retrieval and natural language processing. It is a statistical metric used to measure how important a term is to a document in a dataset” (Bharadwaj et al., 2020, Section 2).

Once the data has been preprocessed, it will provide the input for ML model, which are then trained and tested with a train-test split of 70/30, respectively, to detect patterns in news articles that are characteristic of fake news. The LR algorithm is the statistical model used since it predicts the probability of an event occurring, such as fake news or real news. The dependent variable is a binary output used for classification. The LR algorithm is a supervised learning model. Once the model has been trained, it is able to provide accurate findings when new data is inputted. The five different transformed data inputs include count-word, TD-IDF, stopwords, lemmatization, and a combination of stopwords and lemmatization.

**Figure 1**

*Data flow diagram for classification process*

A picture containing text, screenshot, diagram, font

Description automatically generated

*Note*. Adapted from “Detection of online fake news using n-gram analysis and machine learning techniques” by H. Ahmed, I. Traore, & S. Saad, 2017, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 10618*(2017),127-138. Copyright 2017 by Springer, Cham.

# Limitations

Limitations of data exist for different reasons. One limitation is that someone else collected the data so therefore that person determined what information was important to collect. The impact of this limitation has the potential to produce biased conclusions. However, specific limitations regarding the dataset used for this research paper includes missing information about the types of fake news, the source of the article, and the author of the article. Fake news can be split up into categories for better classification, such as fabrication, hoax, and satire. Fabrication is the conscious omission of information within an article, where the author is fully aware of the story’s error. Hoax is a style of reporting used to mislead the general public into believing the story is real. Satire is a humorous fake story shared to the public, but if people are unaware of the source, they might believe it (Mishra et al., 2022). Additional labels could also include bias, hate, state, conspiracy, and junk science since these categories also fall under the scope of untrustworthy news. Knowing the source and author of the article can be used in deciding the credibility of a news source. The credibility depends on a few different factors: the dependability of the media, the source of information, and the message. In the overall topic of fake news, additional labels will need to be classified since they also have an impact in influence the public’s opinion.

Network analysis can be used to understand how information spreads across social media platforms and other online channels. By examining the networks of individuals and organizations that share content, we can identify patterns of behavior that might be indicative of fake news. Data analytics can be used to cross-reference claims made in news articles with other sources of information. This dataset does not follow the path of any news articles therefore, network analysis could not be used. However, multiple datasets could have been incorporated to include the missing information to provide a more thorough investigation into the detection of fake news.

# Ethical Considerations

There are several ethical considerations that should be considered when handling data from any industry, such as privacy, bias, misrepresentation, data ownership, responsibility, informed consent, data security, and transparency. Particularly, the key considerations in data-driven fake news detections also includes privacy, bias, transparency, accuracy, and free-speech. It is essential to respect the privacy of individuals and organizations whose data is being analyzed so any data collection for fake news detection needs to adhere to privacy regulations and guidelines. This includes obtaining informed consent from individuals whose data is being collected and processed. Additionally, any data that can be used to identify individuals or organizations should be anonymized or removed. It is also important to ensure that the data and algorithms used for fake news detection are not biased against certain groups of people. This requires careful selection of training data and ongoing monitoring to detect and address any biases that may arise. The analysis should be conducted in a transparent manner to prevent any biases from being introduced (Deepak, 2020).

The methods and algorithms used for fake news detection should be transparent and explainable to ensure accountability and trust. This includes providing clear information about the data sources, the criteria for detecting fake news, and the results of any automated decisions made. Any data that is used to support research findings must be accurately represented. Fake news detection systems should be evaluated for accuracy and reliability to avoid the potential for false positives or false negatives that could harm individuals or organizations. Fake news detection should not be used as a tool to suppress free speech or silence dissenting voices. It is important to balance the need for accurate information with the need to protect free speech and open debate. Overall, ethical considerations are crucial in developing effective and trustworthy fake news detection systems that uphold individual rights and societal values (Deepak, 2020). Additionally, ethical considerations should constantly be revisited throughout the lifecycle of a project since they are an integral part of research.

The data potential that goes beyond this paper would be to detect fake news in real-time. Fact-checking in real-time plays a significant role to help in the prevention of the spread of false information and misinformation since it can quickly gain traction and reach a large audience before being fact-checked. Real-time detection of fake news can be achieved using a combination of artificial intelligence and natural language processing techniques. These methods can analyze the content and source of the news and compare it to a database of known false information. Some platforms have already implemented real-time fake news detection systems, which can flag potentially false information and alert users to the potential inaccuracies. However, these systems are not foolproof and can sometimes miss false information or incorrectly flag accurate information.

**Findings**

Table 4, below, summarizes the predicted accuracies of the LR algorithm classifier using five different inputs with 70% of the data being used for training purposes and the remaining 30% of the data being used to test the algorithm. The input features included the 1) count-word, 2) TF-IDF, 3) NLTKStop + count-word, 4) Lemma + count-word, and 5) NLTKStop + Lemma + count-word. The count-word, a count of all the words in title, produced an accuracy of 95.97%. The TFIDF-word, a numerical representation of the words’ importance within the title, produced an accuracy of 95.17%. The NLTKStop + count-word, a count of the words in the title but with stop words removed, produced an accuracy of 94.89%. The Lemma + count-word, a count of the root form of the words, produced an accuracy of 97.50%. The final feature extraction combined the stop word removal and the root form of the word produced the best accuracy of 98.66%.

Since the accuracies of using the five different features as inputs for the LR algorithm are all above 94%, the general conclusion is that the title can be used to predict whether the article is fake. Therefore, the null hypothesis was rejected. The best feature extraction was the combination of removing stop words and lemmatizing which produced the highest predicted accuracy. However, given that the accuracies are so close, additional tests would need to be made within the body of the article to fully grasp if testing the title of an article can be the first line of defense in combating fake-news.

**Table 4**

*Logistic Regression Classification Results*

|  |  |
| --- | --- |
| **Logistic Regression – Classification accuracies** | |
| **Count-word** | 95.97% |
| **TFIDF-word** | 95.17% |
| **NLTKStop + Count-word** | 94.89% |
| **Lemma + Count-word** | 97.50% |
| **NLTKStop + Lemma + Count-word** | 98.66% |

# Conclusion

News is consumed through several different forms daily; hence, why it is so important to be able to discern the difference between fake and authentic. When fake news is broadcast, it can affect the public’s perception of the world and change a person’s thoughts. The spreading of false information has extreme consequences, and hence, concentrated efforts should be taken to reduce its influence. An ideal situation would be to eradicate the fake news sources altogether. The primary goal of this research is to utilize machine learning and artificial intelligence tools and techniques for the classification of fake and real news articles. Additionally, the effectiveness of real-time fake news detection depends on the quality of the detection algorithms and the accuracy of the data used to train them. As technology continues to evolve and improve, we can expect more sophisticated real-time detection systems to be developed, which will help to combat the spread of fake news.

Creating research questions is an essential step in data analytics because it helps to guide the entire research process, from data collection to analysis and interpretation. Research questions define the purpose of the study and help to focus the analysis on specific topics or areas of interest. Additionally, they help to focus the analysis, provide clarity and structure, ensure rigor, and ensure relevance. Hypothesis statements are used to test the research questions through statistical analysis using the quantitative research method. However, careful precautions must be taken when drawing conclusions since any false positives has the potential to ruin a news agency’s reputation.

The experiment presented in this paper analyzed a fake news dataset by first dividing the data into 70% for training and 30% for testing. Then five different feature extraction methods were used as inputs for the LR algorithm to create the learning models that were then used to predict the labels for the testing data. The proposed model achieves the highest prediction accuracy of 98% when the words within the title of the article were reduced to the root words and the stop words were removed using the Natural Language Toolkit Library in Python.

# Recommendations

Recommendations would include to use different datasets to train and test on different algorithms to see which combination would produce the best accuracy on predicting fake-news from the title of an article. A goal for future use would be to use combination of algorithms. Additionally, differentiating and classifying different types of fake news, such as satire and hoax should be explored. One more recommendation would be to filter out the articles from the same year in order to focus the scope of the articles.

# References

Ahmed, H., Traore, I., & Saad, S. (2017, October 11). Detection of online fake news using n-gram analysis and machine learning techniques. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 10618*(2017),127-138. <https://doi.org/10.1007/978-3-319-69155-8_9>

Ahmad, I., Yousaf, M., Yousaf, S., & Ovais Ahmad, M. (2020, October 17). Fake news detection using machine learning ensemble methods. *Complexity, 2020*(8885861), 1-11. <https://doi.org/10.1155/2020/8885861>

Almansa, E. (2020, November 26). *An overview of textual and visual content to detect fake news*. Medium. <https://medium.com/swlh/an-overview-of-textual-and-visual-content-to-detect-fake-news-8d3cf076dce9>

AWS. (2023). *What is text analysis?* <https://aws.amazon.com/what-is/text-analysis>

Bhandari, P. (2021, February 15). *An introduction to quantitative research.* Scribbr. <https://www.scribbr.com/methodology/quantitative-research/>

Bharadwaj, A., Ashar, B., Barbhaya, P., Bhatia, R., & Shaikh, Z. (2020, June). Source Based Fake News Classification using Machine Learning. *International Journal of Innovative Research in Science, Engineering and Technology* (IJIRSET), 9(6), 1-7. <http://www.ijirset.com/upload/2020/june/115_4_Source.PDF>

Bisaillon, C. (2020). *Fake and real news dataset: Classifying the news*. Kaggle. <https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset>

Brownlee, J. (2016, March 16). *Supervised and unsupervised machine learning algorithms*. <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/>

Datapine. (2021). *Essential data analyst tools: Discover a list of the 12 best data analyst software & tools on the market.* <https://www.datapine.com/articles/data-analyst-tools-software#:~:text=Data%20analyst%20tools%20is%20a,decreasing%20costs%20and%20increasing%20profits>.

Deepak, P. et al. (2020, November 20). Ethical considerations in data-driven fake news detection. In: *Data Science for Fake News* (Vol. 24, pp 205–232). Springer, Cham. <https://doi.org/10.1007/978-3-030-62696-9_10>

Kumari, K. (2022, September 6). *Detecting fake news with natural language processing*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/>

Jenson, D. & Warren, K. (2023, April). *What is research methodology? A plain-language explanation & definition (with examples)*. <https://gradcoach.com/what-is-research-methodology/>

Mishra, S., Shukla, P., & Agarwal, R. (2022, March 21). Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets. *Wireless Communications and Mobile Computing, 2022*(1575365), 1-18. <https://doi.org/10.1155/2022/1575365>

Rastogi, S. & Bansal, D. (2022, November 15). A review of fake news detections 3T’s: Typology, time of detection, taxonomies. *International Journal of Information Security, 22*(2023), 177-212. <https://doi.org/10.1007/s10207-022-00625-3>

Raza, S. & Ding, C. (2022, January 30). Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(2022), 335–362. <https://doi.org/10.1007/s41060-021-00302-z>

Shoemaker, E. (2019). *Using data science to detect fake news* [Senior Honors Project, James Madison University]. JMU Scholarly Commons. <https://commons.lib.jmu.edu/honors201019/714>

Stedman, C. & Vaughan, J. (2022, December). *Data quality*. Tech Target Network. <https://www.techtarget.com/searchdatamanagement/definition/data-quality>

Sudhakar, M. & Kaliyamurthie, K. P. (2022, September 27). Effective prediction of fake news using two machine learning algorithms. *Measurement: Sensors, 24*. <https://doi.org/10.1016/j.measen.2022.100495>