**1.Data Source: Choose the fake news dataset available on Kaggle, containing articles titles and text, along with their labels (genuine or fake).**

Fake news arouses to be untrue with the point of deceiving it openly which is now viewed as the greatest threat to society by cultivating the political division and doubts in government. Since this kind of news is disseminated in sheer volume through social media, driving the improvement of strategies for the recognizable proof of false news is necessary. Therefore, this study focuses on text analytics to derive the hidden properties of stylistic content to detect fake and real news. An erudite literature study of fake news detection diverted towards issues such as attention, context, and parallelization. In this same direction, the assessment evaluates the sequential memory-based deep learning model in comparison to the parallel memory-based deep learning model. For sequential, Long Short-Term Memory (LSTM), Bi-Directional LSTM, and Attention-based Bi-directional LSTM are taken into consideration. Besides, for parallel, the transformer-based BERT model is examined. To identify the efficacy of applied approaches, four datasets are taken from diverse domains such as political news, entertainment news, satire news, conspiracy news, and global pandemic news. The experimental analysis of real-world information demonstrates that the pre-trained transformer encoder-based BERT model outperforms with a quite significant margin of improvement. Also, as inspected Attention-based Bi-directional approach provides state-of-the-art results with good training accuracy.

**2.Data Preprocessing: Clean and preprocess the textual data to prepare it for analysis.**

Nowadays, it is very important to quickly recognize the false information referred to as fake news. This is especially important in the case of news appearing on the Internet because of its wide and rapid spreading. It is equally important to be able to initially classify news as fake or true based on the title itself. In this paper, we propose an approach to classifying news based on the title without analyzing the other aspects. The obtained results will be compared with classification based on the whole news text. The goal of this work is to propose a method that balances between data analysis time and quality of classification in fake news prediction. We use natural language processing (NLP) to describe the title and text of the news. This is a complex process, requiring good analysis to be applied to classification. Therefore, the use of complex classifiers – in this case, classical ensemble methods – has been proposed in order to achieve a high quality of classification (measured by popular measure). In this paper, we present analyses of a real data set and results of news classification using the proposed model – including an ensemble of classifiers and single classifiers.

**3. Feature Extraction: Utilize techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings to convert text into numerical features.**

Vulnerability classification is an important activity in software development and software quality maintenance. A typical vulnerability classification model usually involves a stage of term selection, in which the relevant terms are identified via feature selection. It also involves a stage of term-weighting, in which the document weights for the selected terms are computed, and a stage for classifier learning. Generally, the term frequency-inverse document frequency (TF-IDF) model is the most widely used term-weighting metric for vulnerability classification. However, several issues hinder the effectiveness of the TF-IDF model for document classification. To address this problem, we propose and evaluate a general framework for vulnerability severity classification using the term frequency-inverse gravity moment (TF-IGM). Specifically, we extensively compare the term frequency-inverse gravity moment, term frequency-inverse document frequency, and information gain feature selection using five [machine learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm) on ten vulnerable software applications containing a total number of 27,248 security vulnerabilities**.** The experimental result shows that: (i) the TF-IGM model is a promising term weighting metric for vulnerability classification compared to the classical term-weighting metric, (ii) the effectiveness of feature selection on vulnerability classification varies significantly across the studied datasets and (iii) feature selection improves vulnerability classification.

**4. Model Selection: Select a suitable classification algorithm (e.g., Logistic Regression, Random Forest, or Neural Networks) for the fake news detection task.**

Fake news poses a grave threat with devastating consequences in this information-centric age. While advances in data science undeniably hold the key to accurately detecting and curtailing the unfettered spread of fake news, guidance on the selection of algorithms and models that are best suited to a specific fake news scenario leaves much to be desired. Most studies have focused on fake news in a specific domain and employed a limited range of algorithmic techniques. In contrast, the thematic diversity of fake news raises questions over the comprehensiveness of such techniques, whose performance drops when exposed to fake news from a different domain. The current study responds to this call for guidance by focusing on thematically diverse datasets, applying a series of complex algorithms, and performing topic modeling on them. The results demonstrate that ensemble techniques outperform other algorithms, achieving high levels of accuracy of over 98 percent and 95 percent on thematically diverse and pandemic-related datasets, respectively. The study also demonstrates that neural networks are not a panacea for all situations, while topic modeling helps illustrate the lack of coherence in fake news articles. The study offers a distinct perspective on the accuracy of a diverse set of algorithmic approaches and their ability to adapt to an ever-evolving multi-domain world of fake news. A key implication of the study is the unique and comprehensive view of classification performance when exposed to diverse datasets, including pandemic-related news and data from other disciplines, as opposed to its performance on pandemic-related data alone. Our practical contribution is truly the comparative perspective we offer to practitioners when a choice of algorithm is to be made to accurately detect fake news with thematic heterogeneity.

**5. Model Training: Train the selected model using the preprocessed data.**

In this paper, we investigate the impact of pattern leakage during [data preprocessing](https://www.sciencedirect.com/topics/computer-science/data-preprocessing) on the reliability of [Machine Learning](https://www.sciencedirect.com/topics/computer-science/machine-learning) (ML) based [intrusion detection systems](https://www.sciencedirect.com/topics/computer-science/intrusion-detection-system) (IDS). Data leakage, also known as pattern leakage, occurs during [data preprocessing](https://www.sciencedirect.com/topics/computer-science/data-preprocessing) when information from the testing set is used in training, leading to overfitting and inflated accuracy scores. Our study uses three well-known [intrusion detection](https://www.sciencedirect.com/topics/computer-science/intrusion-detection) datasets: NSL-KDD, UNSW-NB15, and KDDCUP99. We preprocess the data to create versions with and without pattern leakage and train and test six ML models: Decision Tree (DT), Gradient Boosting (GB), K-neighbours (KNN), [Support Vector Machine](https://www.sciencedirect.com/topics/computer-science/support-vector-machine) (SVM), [Random Forest](https://www.sciencedirect.com/topics/computer-science/random-decision-forest) (RF), [Logistic Regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression) (LR). Our results show that building IDS models with data leakage leads to higher accuracy but is unreliable. Additionally, we find that some algorithms are more sensitive to data leakage than others, as seen by the drop in model accuracy when built without leakage. To address this problem, we provide suggestions for mitigating data leakage in the training process and analyzing the sensitivity of different algorithms. Overall, our study emphasizes the importance of addressing data leakage in the training process to ensure the reliability of ML-based IDS models.

**6. Evaluation: Evaluate the model's performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.**

Link prediction is a paradigmatic and challenging problem in network science, which attempts to uncover missing links or predict future links, based on known [topology](https://www.sciencedirect.com/topics/physics-and-astronomy/topology). A fundamental but still unsolved issue is how to choose proper metrics to fairly evaluate [prediction algorithms](https://www.sciencedirect.com/topics/mathematics/prediction-algorithm). The area under the receiver operating characteristic curve (AUC) and the balanced precision (BP) are the two most popular metrics in early studies, while their effectiveness is recently under debate. At the same time, the area under the precision–recall curve (AUPR) becomes increasingly popular, especially in biological studies. Based on a toy model with tunable noise and predictability, we propose a method to measure the discriminating ability of any given metric. We apply this method to the above three threshold-free metrics, showing that AUC and AUPR are remarkably more discriminating than BP, and AUC is slightly more discriminating than AUPR. The result suggests that it is better to simultaneously use AUC and AUPR in evaluating link [prediction algorithms](https://www.sciencedirect.com/topics/mathematics/prediction-algorithm). At the same time, it warns us that the evaluation based only on BP may be unauthentic. This article provides a starting point towards a comprehensive picture about effectiveness of evaluation metrics for link prediction.