UNIVERSITY OF ECONOMIC AND LAW

**FACULTY OF INFORMATION SYSTEMS**



**FINAL PROJECT REPORT**

**COURSE: MACHINE LEARNING**

**Topic:**

**TABLOID NEWS DETECTION**

**USING MACHINE LEARNING**

**Lecturer: M.S. Truong Hoan Phan**

*Ho Chi Minh City, May 5, 2023*UNIVERSITY OF ECONOMIC AND LAW

**FACULTY OF INFORMATION SYSTEMS**



**FINAL PROJECT REPORT**

**COURSE: MACHINE LEARNING**

**Topic:**

**TABLOID NEWS DETECTION**

**USING MACHINE LEARNING**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Member of group** | | | | | | | |
| **NO.** | **Name** | **ID** | | **Class** | **Role** | **Phone** | **Email** |
| 1 | Vu Van Son | | K204160669 | K20416C | Member | 0944244857 | sonvv20416c  @st.uel.edu.vn |
| 2 | Chau Nguyen Huong Thao | | K204162000 | K20416C | Member | 0919772802 | thaocnt20416c  @st.uel.edu.vn |
| 3 | Huynh Hong Uyen | | K204162006 | K20416C | Member | 0934190565 | uyenhh20416c  @st.uel.edu.vn |
| 4 | Nguyen Tuong Vy | | K204162009 | K20416C | Leader | 0764187037 | vynt20416c  @st.uel.edu.vn |

**Lecturer: M.S. Truong Hoai Phan**

*Ho Chi Minh City, May 5, 2023*

**ACKNOWLEDGEMENTS**

During the research period, we got a lot of devoted help and advice from the lecturer. Besides, we also referenced a number of authors. Thanks to your distribution, we can fully complete our final project.

Despite our best effort, mistakes are inevitable. Therefore, we are glad to receive your judges and comments to improve our research. Those will be our enormous motivation to develop our project on the horizon.

**COMMITMENT**

We commit that our midterm project is unique due to the whole team’s research. There are still some documents we referenced, which we have listed and cited particularly in the report. If all of the above is wrong, we will take all responsibility.

Ho Chi Minh City, May 5, 2023.

**Committed person**

**TEAM**

# TABLE OF CONTENT

[TABLE OF CONTENT 6](#_Toc134309943)

[LIST OF FIGURES 8](#_Toc134309944)

[LIST OF TABLES 9](#_Toc134309945)

[LIST OF ACRONYMS 10](#_Toc134309946)

[CHAPTER 1: INTRODUCTION 11](#_Toc134309947)

[CHAPTER 2: THEORETICAL BASIS AND RELATED WORKS 13](#_Toc134309948)

[2.1. Related definitions 13](#_Toc134309949)

[2.1.1. Fake news and Tabloid news 13](#_Toc134309950)

[2.1.2. Types of tabloid news detection 15](#_Toc134309951)

[2.2. Related works 16](#_Toc134309952)

[2.3. Feature Extraction 18](#_Toc134309953)

[2.3.1. Bag of Words (BOW) 18](#_Toc134309954)

[2.3.2. Term Frequency–Inverse Document Frequency (TF-IDF) 19](#_Toc134309955)

[2.4. Algorithm 20](#_Toc134309956)

[2.4.1. Naïve Bayes classification 20](#_Toc134309957)

[2.4.2. Linear Regression 21](#_Toc134309958)

[2.4.3. Support Vector Machine 22](#_Toc134309959)

[2.5. Model evaluation 23](#_Toc134309960)

[2.5.1. Confusion matrix and Accuracy 23](#_Toc134309961)

[2.5.2. F1 score 23](#_Toc134309962)

[CHAPTER 3: EXPERIMENT 25](#_Toc134309963)

[3.1. Dataset 25](#_Toc134309964)

[3.2. Data preparation 26](#_Toc134309965)

[3.3.1. Build train/test set 27](#_Toc134309966)

[3.3.2. Naïve Bayes 28](#_Toc134309967)

[3.3.3. Logistic Regression 29](#_Toc134309968)

[3.3.4. SVM 30](#_Toc134309969)

[3.5. Evaluation 32](#_Toc134309970)

[3.6. Applications 32](#_Toc134309971)

[CHAPTER 4: CONCLUSION 35](#_Toc134309972)

[4.1. Conclusion 35](#_Toc134309973)

[4.2. Limitations and Future Works 35](#_Toc134309974)

# LIST OF FIGURES

[Figure 1: Confusion matrix 22](#_Toc134309930)

[Figure 2: Dataset overview 24](#_Toc134309931)

[Figure 3: Distribution of two-kind news in the dataset. 25](#_Toc134309932)

[Figure 4:After removing punctuation 26](#_Toc134309933)

[Figure 5: Text after cleaning. 26](#_Toc134309934)

[Figure 6: Main screen of application 32](#_Toc134309935)

[Figure 7: Input text area 32](#_Toc134309936)

[Figure 8: Tabloid message 33](#_Toc134309937)

[Figure 9: Original message 33](#_Toc134309938)

# LIST OF TABLES

[Table 1: Example of BOW 18](#_Toc134309916)

[Table 2: Results of Naive Bayes model 30](#_Toc134309917)

[Table 3: Results of SVM and Logistic Regression models 31](#_Toc134309918)

[Table 4: Comparison of different algorithm results 31](#_Toc134309919)

# LIST OF ACRONYMS

|  |  |
| --- | --- |
| **ACRONYMS** | **FULL NAME** |
| SVM | Support Vector Machine |
| TF-IDF | Term Frequency–Inverse Document Frequency |
| GUI | Graphical User Interface |
| BOW | Bag of Words |
| NLP | Natural Language Processing |

# CHAPTER 1: INTRODUCTION

Tabloid news, with its sensational headlines and often misleading content, has become increasingly popular in today's media landscape. This type of journalism has the potential to spread false information and contribute to the spread of fake news, which can be harmful to society. Therefore, it is important to develop effective methods to detect and filter out tabloid news.

In this research report, we aim to investigate the effectiveness of three popular machine learning algorithms: Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) in detecting tabloid news. Our model will be tested on a dataset of Vietnamese tabloid news articles, which is an important consideration as Vietnamese is a complex and rich language system with many different regional variations.

Tabloid news detection using machine learning algorithms is an inheritance of research on fake news and clickbait detection. While several studies have been conducted in this field, there has been limited research on detecting tabloid news specifically.

Therefore, our research will contribute to the development of more effective methods for detecting tabloid news in Vietnamese, as well as provide a comparison between the performance of different machine learning algorithms on this task. All models will be developed in Python, a popular programming language for machine learning applications.

We will begin by exploring related works on fake news and clickbait detection using machine learning algorithms. We will then describe our methodology for data collection, pre-processing, feature extraction, and model building. Finally, we will present our experimental results and discuss the strengths and limitations of our approach. In addition, we aim to develop a user-friendly Graphical User Interface (GUI) application for our model to make it more accessible to end-users.

The development of a GUI is an important aspect of our research, as it will allow users to easily input a news article and receive an output indicating whether the article is a tabloid news or not. Our GUI will be designed with a user-friendly interface that can be easily navigated by users with varying levels of technical expertise.

In summary, our research focuses on the detection of tabloid news in Vietnamese using machine learning algorithms. The use of Vietnamese data sets and development of a GUI are unique features of our research. Our ultimate goal is to provide a practical solution that can be used by individuals, organizations, and social media platforms to identify and filter out tabloid news.

# CHAPTER 2: THEORETICAL BASIS AND RELATED WORKS

Tabloid news detection using machine learning algorithms is a challenging task that requires several steps, including data collection, pre-processing, feature extraction, and model building. In this chapter, we will explore the theoretical basis of each of these steps, as well as provide an overview of the machine learning algorithms used in our research: Naive Bayes, Logistic Regression, and Support Vector Machines (SVM).

## Related definitions

### Fake news and Tabloid news

* **Fake news**

Fake news is a term that has gained popularity and controversy in recent years, especially in the context of the 2016 US presidential election and the COVID-19 pandemic. However, there is no consensus on the definition of fake news, as different sources may use different criteria, perspectives, and purposes to define and identify fake news. Therefore, it is important to review some of the existing definitions of fake news and their implications for research and practice.

According to Dictionary.com, fake news is “false news stories, often of a sensational nature, created to be widely shared or distributed for the purpose of generating revenue, or promoting or discrediting a public figure, political movement, company, etc.” [[1](#a1)]. This definition emphasizes the intention and effect of fake news, as well as its format and medium. However, this definition may not capture the complexity and diversity of fake news, as some fake news may not be entirely false or sensational, but rather distorted or misleading. Moreover, this definition may not account for the role of the audience and the context in interpreting and evaluating fake news.

According to Cambridge Dictionary, fake news is “false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke” [[2](#a2)]. This definition focuses on the content and purpose of fake news, as well as its source and channel. However, this definition may not distinguish between different types and degrees of falsity, such as fabrication, manipulation, misinformation, disinformation, malinformation, etc. [[3](#a3)]. Furthermore, this definition may not consider the challenges and limitations of detecting and verifying fake news.

According to Reporters Without Borders (RSF), fake news is “a form of information disorder that consists of deliberate disinformation or hoaxes spread via traditional news media or online social media” [[4](#a4)]. This definition highlights the nature and scope of fake news, as well as its origin and dissemination. However, this definition may not address the various factors and actors that contribute to the production and consumption of fake news, such as cognitive biases, emotional reactions, social networks, algorithms, etc. [[5](#a5)]. Additionally, this definition may not deal with the ethical and legal issues of defining and regulating fake news.

These are some of the existing definitions of fake news that have been proposed by different sources. They reflect different aspects and dimensions of fake news, such as intentionality, veracity, format, medium, purpose, effect, etc. However, they also have some limitations and gaps that need to be addressed by further research and discussion. In this report, we adopt a broad and inclusive definition of fake news as any form of information that is intentionally or unintentionally false or misleading and that is presented or perceived as news by some audience or context. We use this definition to guide our research on tabloid news detection using machine learning techniques such as TF-IDF, word embedding, and deep neural networks.

* **Tabloid news**

Tabloid news is a term that refers to a type of fake news that contains sensationalized or exaggerated information that may mislead or manipulate readers. Tabloid news can be found in tabloid newspapers, which are newspapers with a compact page size and a popular style of journalism that emphasizes features such as sensational crime stories, astrology, gossip columns, and celebrity scandals [[6](#a6)]. Tabloid news can also be found in online platforms, such as websites, blogs, social media, etc., that use similar techniques and topics to attract and engage readers [[7](#a7)].

In addition, this type of news is a problem that deserves attention and action because it can have negative impacts on individuals, organizations, and society. For individuals, tabloid news can affect their knowledge, beliefs, attitudes, and behaviors by providing them with inaccurate or biased information that may influence their decision making and actions [[8](#a8)]. For organizations, tabloid news can damage their reputation, credibility, and trustworthiness by spreading false or negative information about them or their products or services. For society, tabloid news can undermine the quality and diversity of public discourse and democratic processes by creating polarization, confusion, distrust, and hostility among different groups and stakeholders.

Therefore, tabloid news detection is a task that aims to identify and prevent tabloid news from harming individuals, organizations, and society. Tabloid news detection is a challenging task because it involves various aspects of natural language processing, such as text representation, feature extraction, sentiment analysis, topic modeling, etc. [7]. Tabloid news detection also requires a deep understanding of the context, source, and intention of the news articles, which may not be easily captured by machine learning algorithms [8]. Moreover, tabloid news detection faces some practical difficulties, such as the lack of labeled data, the dynamic and evolving nature of tabloid news, the ethical and legal issues of detecting and removing tabloid news, etc.

In this report, we propose a novel method to detect tabloid news in Vietnamese using machine learning techniques such as TF-IDF and some classification algorithms (Naive Bayes, Logistic Regression and SVM). The detailed description of the method will be discussed in the folliowing chapters.

### Types of tabloid news detection

In general, tabloid news detection is a type of fake news detection that focuses on identifying and filtering out sensationalized, exaggerated, or fabricated news stories that are typically found in tabloid newspapers or magazines. Some possible types of tabloid news detection are:

* *Content-based methods:* These methods analyze the text or image or both within the news article to detect tabloid news. For example, they may use linguistic features, such as sentiment, emotion, exaggeration, or sensationalism, to identify tabloid news [[9](#a9)] They may also use image analysis techniques, such as face detection, manipulation detection, or celebrity recognition, to verify the authenticity of the images in tabloid news [[10](#a10)]
* *Social context-based methods:* These methods use the information from the social media platforms where the tabloid news is shared, such as user comments, likes, shares, etc. For example, they may use social network analysis, user credibility assessment, or opinion mining to detect tabloid news. They may also use temporal analysis, such as burstiness or trendiness, to identify tabloid news that exploits timely events or topics[[11](#a12)].
* *Hybrid methods:* These methods combine both content and social context features to improve the detection performance of tabloid news. For example, they may use machine learning or deep learning models, such as classifiers, neural networks, or transformers, to integrate multiple sources of information and learn complex patterns of tabloid news. They may also use ensemble methods, such as voting or stacking, to combine the outputs of different models and increase the accuracy of tabloid news detection.

## Related works

Tabloid news detection is a challenging task that aims to identify sensationalized or exaggerated news articles that may mislead or manipulate readers. Machine learning techniques have been widely applied to this problem, using various features and models to classify news articles as tabloid or non-tabloid. In this section, we review some of the related works that use similar machine learning methods as our proposed approach.

**Nguyen Thi Thanh Thuy and Nguyen Thi Thanh Nhan (2021)** [[12](#a12)]propose a method to detect tabloid news using machine learning techniques such as TF-IDF, word embedding, and deep neural networks. The method consists of three steps: preprocessing, feature extraction, and classification. They conduct experiments on two datasets of Vietnamese news articles and achieve an accuracy of 93.75% and 94.17% respectively. They also analyze the impact of different word embedding methods and neural network architectures on the performance of tabloid news detection. Some of the advantages and disadvantages of the above research are:

* *Advantages :*
* The method uses a combination of TF-IDF and word embedding to capture both the frequency and the semantic information of the words in the news articles.
* The method uses a deep neural network with multiple layers and dropout to learn complex and non-linear patterns from the features and prevent overfitting.
* The method achieves high accuracy on both datasets, which shows its effectiveness and robustness for tabloid news detection.
* *Disadvantages :*
* The method does not consider other features that may be useful for tabloid news detection, such as headline, sentiment, source, or metadata.
* The method does not provide any explanation or interpretation of the results, which may limit its applicability and trustworthiness for users or stakeholders.
* The method does not compare its performance with other existing methods or baselines, which may limit its validity and generalizability.

**Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake News Detection Using Machine Learning Approaches** [[13](#a13)]This paper aims to analyze the existing research on fake news detection and propose a supervised machine learning model that can classify fake news as true or false using Python scikit-learn and natural language processing (NLP) tools. The paper follows a systematic approach of feature extraction, feature selection, and model evaluation using confusion matrix and accuracy metrics. The paper also compares different machine learning models, such as logistic regression, decision tree, random forest, and support vector machine, and selects the best one based on the highest accuracy. Some of the advantages and disadvantages of this paper are:

* *Advantages:*
* It provides a comprehensive literature review of the existing methods and techniques for fake news detection and their challenges and limitations.
* It uses a well-known dataset of fake news articles from Kaggle that contains both text and label information for each article.
* It applies various NLP techniques for feature extraction, such as tokenization, lemmatization, stop words removal, etc., using Python scikit-learn library.
* It performs feature selection using chi-square test to select the most relevant features for fake news detection.
* It evaluates and compares different machine learning models using confusion matrix and accuracy metrics and selects the best one based on the highest accuracy.
* *Disadvantages*
* It does not consider the social context information of the fake news articles, such as user comments, likes, shares, etc., which can provide additional clues for fake news detection.
* It does not use any advanced or deep learning models, such as neural networks or transformers, which can capture more complex and semantic features of the text data.
* It does not perform any cross-validation or testing on unseen data to validate the generalization ability of the model.
* It does not discuss the ethical or social implications of fake news detection and its potential applications or challenges in real-world scenarios.

## Feature Extraction

### Bag of Words (BOW)

Bag of Words (BOW) is a popular technique used in natural language processing (NLP) for feature extraction. It is a simple and effective approach that converts textual data into a numerical representation that can be easily used by machine learning algorithms. In BOW, a document is represented as a bag of its constituent words, disregarding grammar and word order but keeping track of the frequency of each word. This approach allows us to represent the textual data as a matrix where each row represents a document and each column represents a unique word in the corpus.

The BOW technique has been extensively used in various NLP applications such as sentiment analysis, spam filtering, and topic modeling. One of the key advantages of BOW is that it can handle large datasets and high-dimensional feature spaces efficiently. Moreover, it can capture the essential information about the text while ignoring the irrelevant details.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **about** | **bird** | **heard** | **is** | **the** | **word** | **you** |
| About the bird, the bird, bird bird bird | 1 | 5 | 0 | 0 | 2 | 0 | 0 |
| You heard about the bird | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| The bird is the word | 0 | 1 | 0 | 1 | 2 | 1 | 0 |

Table 1: Example of BOW

### Term Frequency–Inverse Document Frequency (TF-IDF)

The TF-IDF approach is related to the Bag of Words (BOW) approach in that they both involve feature extraction from text data. However, while the BOW approach considers the frequency of occurrence of each word in a document, the TF-IDF approach goes further by assigning weights to each word based on its frequency and rarity across the documents in the corpus.

Particularly, this technique involves two main components: term frequency (TF) and inverse document frequency (IDF). The term frequency component measures the frequency of a word in a document and normalizes it by the total number of words in the document. This helps to give more weight to words that occur frequently in a document relative to those that occur rarely.

IDF (Inverse Document Frequency) measures the significance of a term () in a collection of documents () by dividing the total number of documents by the number of documents that contain the term, and then taking the logarithm of that quotient. This measure helps to determine how important a particular term is in the collection of documents[[14](#a14)].

The inverse document frequency component measures the rarity of a word in the corpus by taking the logarithm of the ratio of the total number of documents in the corpus to the number of documents that contain the word. This helps to give more weight to words that occur rarely across the corpus relative to those that occur frequently.

The TF-IDF approach combines the term frequency and inverse document frequency components to produce a weight for each word in each document. This weight represents the importance of the word in the document relative to its importance in the corpus. The resulting matrix can be used as a feature representation for machine learning algorithms.

The TF-IDF technique has several advantages over the BOW approach. Firstly, it considers the importance of words in the corpus by assigning weights to each word based on its frequency and rarity across the documents. This helps to capture the essential information about the text while ignoring the irrelevant details. Secondly, it can handle large datasets and high-dimensional feature spaces efficiently. Lastly, it is a robust approach that is not affected by the presence of stopwords, which are common words that do not carry much meaning.

## Algorithm

### Naïve Bayes classification

The Naïve Bayes algorithm is a commonly used method for categorizing documents, with two main models used for text classification: the multinomial Naive Bayes model and the multivariate Bernoulli Naive Bayes model. This report focuses on the multinomial model, which assumes that documents are ordered sequences of word events drawn from the same vocabulary. It also assumes that document length is independent of class and that the probability of each word event is independent of the word's context and position in the document. This leads to the "Bag-of-Words" representation for documents, where each document is represented as a set of words drawn from a multinomial distribution.

We can estimate the probability of a word given its class as following:

where is the number appearances of word wt in the document di, and refer to the vocabulary size and the data set size respectively.

As a result the conditional probability can be estimated following:

At last, we can write down the Bayes’ rule for classification decision as following,

### Linear Regression

Linear regression is a statistical analysis method used to model the linear relationship between an independent variable and a dependent variable. It is one of the most popular techniques in the field of prediction and model building. Linear regression can be used to predict the value of a dependent variable given the value of an independent variable. It can be applied to both numerical and continuous data, as well as categorical data. Linear regression can be performed in many ways, including the least squares method, maximum likelihood method, and matrix method. Linear regression is an important tool for analysis and prediction in data science, economics, finance, and many other fields.

There are two types of linear regression: simple linear regression and multiple linear regression. Simple linear regression uses one independent variable to predict the value of a dependent variable. Multiple linear regression uses more than one independent variable to predict the value of a dependent variable. The formula for multiple linear regression is represented as follows:

where is the value of the dependent variable. , ..., are the values of the independent variables. is the intercept of the model. , , ..., are the coefficients of the independent variables and is the random error.

Multiple linear regression models can be used to analyze the relationships between multiple independent variables and a dependent variable. This allows for the evaluation of the effect of each independent variable on the dependent variable while controlling for the effects of other independent variables. Multiple linear regression is a powerful tool for predicting outcomes and understanding the relationships between variables in many fields such as data science, economics, and social sciences.

### Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm that classifies data by determining the optimal boundary between data classes. SVM is used to solve binary classification problems or multi-class classification problems by finding the best hyperplane in high-dimensional space to separate the classes. SVM can handle linear or nonlinear data, and its accuracy is guaranteed by the Maximum Margin Optimization principle. SVM has been widely applied in many fields, including handwriting recognition, image classification, disease prediction and detection, and data science applications.

Support Vector Machine (SVM) uses the method of classification by hyperplanes to find an optimal boundary between data points of different classes. The classification formula of SVM is represented as follows:

where is the classification function, which determines the class of the data point . is the feature vector of the data point being classified. is the weight vector of the hyperplane, which determines the direction of the boundary and is the bias, which determines the position of the hyperplane on the vertical axis.

The sign function assigns the class label based on the location of the data point with respect to the hyperplane. Data points on one side of the hyperplane belong to one class, while data points on the other side belong to the other class. SVM seeks to maximize the margin between the hyperplane and the nearest data points of each class, which leads to better generalization and improved classification accuracy.

## Model evaluation

### Confusion matrix and Accuracy

* **Confusion matrix:** This matrix shows true label and result predicted label by model. The confusion have four values True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN).
* **True Positive(TP):** A test result that correctly indicates the presence of a condition or characteristic.
* **True Negative(TN):** A test result that correctly indicates the absence of a condition or characteristic.
* **False Positive(FP):** A test result which wrongly indicates that a particular condition or attribute is present.
* **False Negative(FN):** A test result which wrongly indicates that a particular condition or attribute is absent.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 1: Confusion matrix

### F1 score

* **Precision:** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

* **Recall:** Recall is the ratio of correctly predicted positive observations to the all observations in actual class

The F1-score is a performance metric that considers both Precision and Recall and calculates their weighted average. This score takes into account both false positives and false negatives and is particularly useful when dealing with uneven class distributions. While accuracy is easier to comprehend, the F1-score is often a more informative measure. Accuracy is most effective when false positives and false negatives have comparable costs. However, if the costs of these errors vary significantly, it is advisable to examine both Precision and Recall to get a better assessment of the model's performance.

# CHAPTER 3: EXPERIMENT

## Dataset

The group uses the fake news dataset [[15](#a15)] in Vietnamese created by a group of authors on github, after processing the data set, the group conducts classification, comparison and comparison to change it to the label "official news" and " tabloid" based on the source domain of the article's link to current Vietnamese online news sites.

The dataset used in our experiment consists of news articles from various online sources and includes 5 features such as title, text, source domain, and label.

* **title**: The title of the news article.
* **text**: The content of the news article.
* **source\_domain**: The domain of the website where the article was published.
* **label**: A binary label indicating whether the article is Tabloid news (F) or Official news (T).
* **full**: A concatenation of the title and text of the article. This feature can be used as an alternative input to the model instead of using title and text separately.

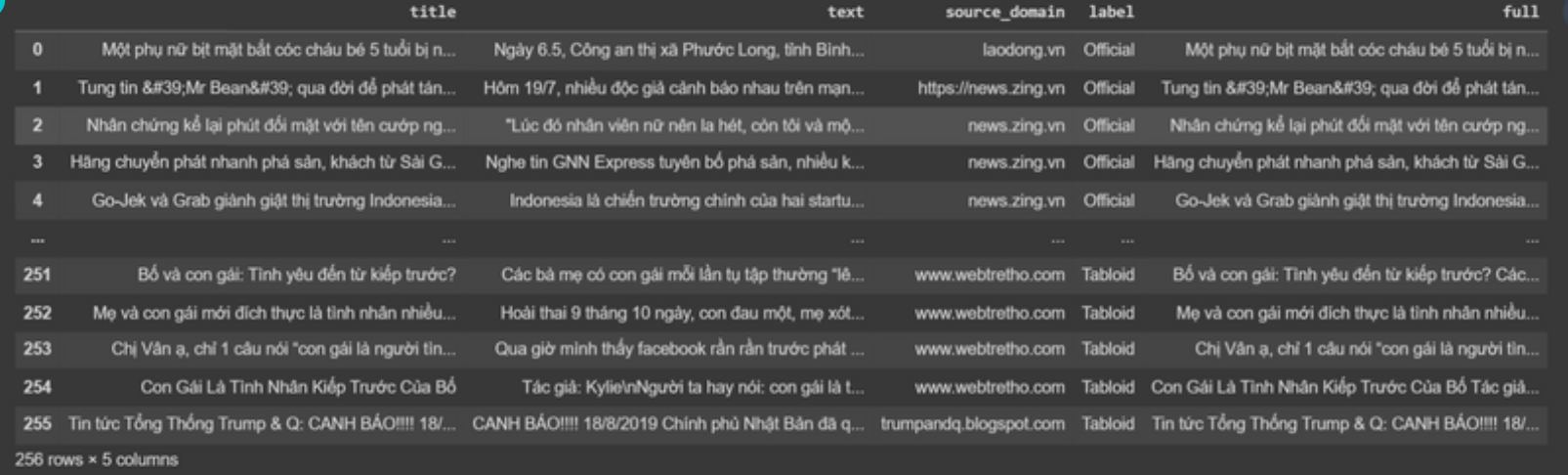
****

Figure 2: Dataset overview

The figure illustrates 256 records of news, where tabloid (F) and official news (T) are distinguished by the label. The number of different types of news is tied to each genre.

A picture containing text, screenshot, font, logo

Description automatically generated

Figure 3: Distribution of two-kind news in the dataset.

## Data preparation

The research model will perform tabloid detection via the content of the article, therefore 2 variables are selected for the model: “full” - a priori and “label” - posterior. At the same time, the study data is balanced for both classifications, so there is no need to deal with data imbalance.

* **Cleaning experimental text**

The first step is to remove any unnecessary information such as special characters, punctuation, and numbers. This can be achieved by using regular expressions or built-in string manipulation functions in Python. Additionally, stop words, which are commonly used words such as "the" and "and", can be removed from the text as they do not contribute to the meaning of the text.

* Standardize Unicode encoding (bring back building Unicode)
* Standardize Vietnamese accent typing (using “**ò**a úy” instead of “o**à** uý” )
* Perform Vietnamese word splitting (using word splitting libraries like pyvi, underthesea, vncorenlp, ...)
* Return text lower (lowercase)
* Remove special characters: “.”, “,”, “;”, “)”, …

A screenshot of a computer

Description automatically generated with low confidence

Figure 4:After removing punctuation

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5: Text after cleaning.

After cleaning, compound words will be connected by dashes as shown above.

* 1. **Model training**

### 3.3.1. Build train/test set

The data has been separated into two groups: the training set (X\_train, y\_train) and the testing set (X\_test, y\_test). The proportion of data in the training set and the testing set is 80% and 20%, respectively. Additionally, the label data has been transformed into a vector form using LabelEncoder to simplify the calculation process.

### 3.3.2. Naïve Bayes

To train the Logistic regression model to classify documents we use the following pipeline:

*text\_clf = Pipeline([('vect', CountVectorizer(ngram\_range=(1,1),*

*max\_df=0.8,*

*max\_features=None)),*

*('tfidf', TfidfTransformer()),*

*('clf', MultinomialNB()) ])*

The given code sets up a machine learning pipeline for text classification using the scikit-learn library. The pipeline consists of three main steps:

* **CountVectorizer:** This step converts the text data into a matrix of token counts. It takes in raw text data as input and outputs a matrix where each row corresponds to a document and each column corresponds to a word in the vocabulary. The ngram\_range parameter specifies the range of n-gram sizes to consider when tokenizing the text (in this case, unigrams only). The max\_df parameter specifies the maximum document frequency (in this case, 0.8) for a word to be included in the vocabulary. Words that appear in more than 80% of the documents are ignored. The max\_features parameter specifies the maximum number of features (i.e., words in the vocabulary) to keep. If set to None, all features are kept.
* **TfidfTransformer:** This step applies the term frequency-inverse document frequency (TF-IDF) weighting scheme to the count matrix produced by CountVectorizer. This scheme assigns weights to the words in the matrix based on their frequency in the documents and their rarity in the corpus. The resulting matrix has rows corresponding to documents and columns corresponding to words, but the cell values represent the TF-IDF weights rather than raw counts.
* **MultinomialNB:** This step trains a Naive Bayes classifier on the TF-IDF weighted count matrix. The MultinomialNB classifier is a variant of the Naive Bayes algorithm that is suitable for handling discrete count data, such as the count matrix produced by CountVectorizer and TfidfTransformer. The classifier is trained on the training data (X\_train and y\_train).

Finally, the fit() method is called on the pipeline object with the training data as inputs (X\_train and y\_train).

### 3.3.3. Logistic Regression

To train the Logistic regression model to classify documents we use the following pipeline:

*text\_clf = Pipeline([('vect', CountVectorizer(ngram\_range=(1,1),*

*max\_df=0.8,*

*max\_features=None)),*

*('tfidf', TfidfTransformer()),*

*('clf', LogisticRegression(solver='lbfgs',*

*multi\_class='auto',*

*max\_iter=10000))])*

The pipeline would have the method to transform the raw input data into features that have the same as the above pipeline used as input to the logistic regression classifier. So we just explain the final step:

The LogisticRegression classifier is a type of linear classifier that uses a logistic (sigmoid) function to model the probability of each class.

The solver parameter specifies the algorithm used to solve the optimization problem that underlies the logistic regression model. In this case, the 'lbfgs' solver is used, which is a popular algorithm for optimization problems with many variables.

The multi\_class parameter specifies the strategy used to extend the logistic regression model to multiple classes. In this case, the value 'auto' is used, which selects the best strategy based on the nature of the problem.

The max\_iter parameter specifies the maximum number of iterations allowed for the optimization algorithm to converge. In this case, a very large value of 10000 is used, indicating that the algorithm should run until convergence is achieved or until this maximum number of iterations is reached.

The LogisticRegression classifier is set up as the final step in a machine learning pipeline, where it is responsible for making predictions based on the input features. The previous steps in the pipeline would have transformed the raw input data into features that can be used as input to the logistic regression classifier.

### 3.3.4. SVM

Support Vector Machines (SVM) are a type of supervised learning algorithm that can be used for classification or regression analysis. In the context of tabloid news detection, SVMs can be used to classify news articles as either real or fake based on their text content. The basic idea behind SVMs is to find the hyperplane that separates the two classes of data with the largest margin.

The SVM model is trained on a set of labeled data, where each news article is labeled as either fake or news. The SVM algorithm then finds the optimal hyperplane that maximizes the margin between the two classes of data. The margin is defined as the distance between the hyperplane and the nearest data points from each class.

*('clf', SVC(gamma='scale')) ])*

The code above set up a Support Vector Machine (SVM) classifier as the final step in a machine learning pipeline.

The SVC (Support Vector Classification) function is a class in the scikit-learn library that sets up the SVM classifier. The gamma parameter is set to 'scale', which is a built-in parameter that automatically sets the gamma value based on the number of features in the input data. Gamma is a hyperparameter that controls the shape of the decision boundary in SVM and has a significant impact on the performance of the classifier.

The SVC function is set up as the final step in a machine learning pipeline, where it is responsible for making predictions based on the input features. The previous steps in the pipeline would have transformed the raw input data into features that can be used as input to the SVM classifier.

Finally, we will save the model as a pickle file so that it can be easily re-evaluated later, as well as used in the demo on Tkinter with Pycharm.

* 1. **Results**

To evaluate the above classification machine learning models, the group compared the evaluation indicators including F1 score, accuracy, Precision, ..

|  |  |  |  |
| --- | --- | --- | --- |
| **Predict/Actual** | | **Actual** | |
| **Tabloid (Positive)** | **Official (Negative)** |
| **Predict** | **Tabloid (Positive)** | 21  True Positives | 3  False Positives |
| **Official (Negative)** | 3  False Negatives | 25  True Negatives |

Table 2: Results of Naive Bayes model

|  |  |  |  |
| --- | --- | --- | --- |
| **Predict/Actual** | | **Actual** | |
| **Tabloid (Positive)** | **Official (Negative)** |
| **Predict** | **Tabloid (Positive)** | 22  True Positives | 2  False Positives |
| **Official (Negative)** | 2  False Negatives | 26  True Negatives |

Table 3: Results of SVM and Logistic Regression models

|  |  |  |
| --- | --- | --- |
| **ALGORITHM** | **F1-SCORE** | **ACCURACY** |
| Naive Bayes | 0.89 | 0.88 |
| Logistic Regression | 0.93 | 0.91 |
| SVM | **0.93** | **0.92** |

Table 4: Comparison of different algorithm results

## Evaluation

After comparing algorithms, we found that SVM has the highest F1 score and Accuracy of 0.93 and 0.92, respectively. So we decided to use this algorithm to develop a small app for news detection.

## Applications

Fake news is a growing problem in our society, as it can have serious consequences for individuals and organizations. To help reduce this problem, we developed a small GUI-based news detection application that helps users check on tabloid or original news.

After running the application, the main screen with the title “**KIỂM TRA TIN TỨC**” will appear.

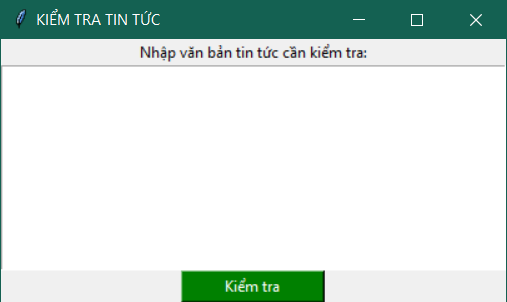


Figure 6: Main screen of application

Users copy a title or a piece of news that needs to check on and paste it into the input text box below the label “**Nhập văn bản cần kiểm tra:**”

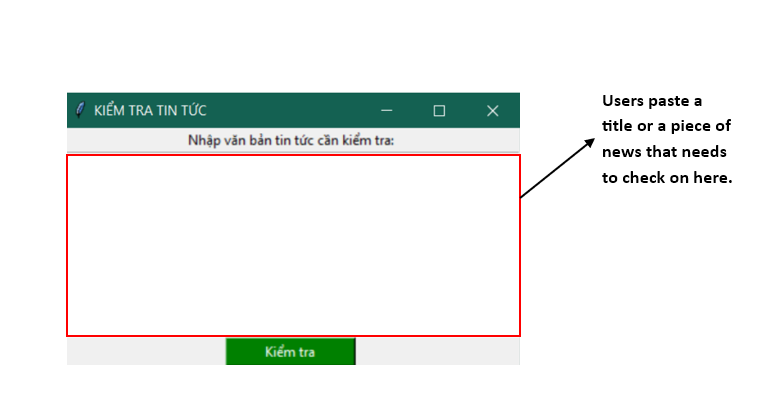


Figure 7: Input text area

After that, Users press the button “**Kiểm tra**” to detect the news. Just wait a minute the result will be returned. If the news is tabloid the message below will display in your screen.

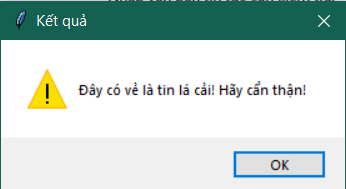


Figure 8: Tabloid message

Else, if the news is original the message below will display.

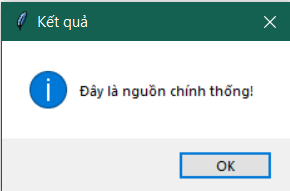


Figure 9: Original message

After all, this app only helps us to check the news partially. The application cannot give completely accurate results, so we should still consider carefully when viewing daily news to avoid losing time and avoid being fooled by fake news.

# 

# CHAPTER 4: CONCLUSION

## 4.1. Conclusion

In today's digital age, the proliferation of social media and online news sources has led to a significant increase in the spread of fake news. As a result, there has been growing interest in developing effective methods for detecting fake news to prevent its dissemination and its potentially harmful effects. In this context, machine learning models have emerged as a promising solution for detecting fake news, as they can automatically learn patterns and features in news articles that can distinguish between genuine and fake news.

In this research project, we developed a machine learning model for detecting tabloid news in Vietnamese language using Naive Bayes, Logistic Regression, and SVM algorithms. We used a dataset of news articles labeled as either real or fake, which we pre-processed by tokenizing, stemming, and removing stop words. We then applied three different machine learning algorithms to the pre-processed data and evaluated their performance using cross-validation and other performance metrics. Our results showed that all three algorithms achieved high accuracy in detecting news, with the SVM algorithm performing slightly better than the other two.

To make the model more accessible to non-technical users, we also created a user-friendly GUI application using Python's Tkinter library. The application enables users to input a news article's title, text, and label, and the model then predicts whether the article is tabloid or original.

## 4.2. Limitations and Future Works

Despite the success of our research project, there are some limitations that need to be addressed in future work.

Firstly, the dataset used in this research project was relatively small and may not represent the full spectrum of news articles in Vietnamese. A more extensive and diverse dataset could improve the model's accuracy in detecting fake news.

Secondly, the machine learning models used in this research project were relatively simple, and more advanced models could yield better results. However, more advanced models can also require more computational power, which could be a limiting factor for some applications.

Thirdly, the model's accuracy may be affected by the limitations of the feature representation technique used, which was based on word frequency. Other feature representation techniques could be explored in future work to improve the model's accuracy.

Finally, while the GUI was designed to be user-friendly, its accessibility to people with disabilities or those not familiar with technology may be limited. Future work could focus on improving the accessibility of the GUI, for example, by incorporating text-to-speech or other accessibility features.

In conclusion, the above research project demonstrates the potential of machine learning models for detecting tabloid news in Vietnamese language and provides a user-friendly GUI for non-technical users. However, further research is needed to address the limitations of this approach and improve the model's accuracy and accessibility.

# 

# REFERENCES

[1] *Dictionary.com*. (2016). Www.dictionary.com. <https://www.dictionary.com/browse/fake-news>

[2] Cambridge Dictionary. (2023, May 3). *fake news*. @CambridgeWords. <https://dictionary.cambridge.org/dictionary/english/fake-news>

[3] Wardle, C., & Derakhshan, H. (2019). Thinking about ‘information disorder’: Formats of misinformation, disinformation, and mal-information. In Media and information literacy in critical times: Re-imagining ways of learning and information environments (pp. 43-52). UNESCO.

[4] *2023 World Press Freedom Index – journalism threatened by fake content industry*. (2023). Rsf.org. <https://rsf.org/en/2023-world-press-freedom-index-journalism-threatened-fake-content-industry>

[5] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36.

[6] Wikipedia Contributors. (2023, March 28). *Tabloid (newspaper format)*. Wikipedia; Wikimedia Foundation. <https://en.wikipedia.org/wiki/Tabloid_%28newspaper_format%29>

[7] Wikipedia Contributors. (2023, April 5). *Tabloid journalism*. Wikipedia; Wikimedia Foundation. <https://en.wikipedia.org/wiki/Tabloid_journalism>

[8] Ahmed, S., Hinkelmann, K., & Corradini, F. (2022). Development of fake news model using machine learning through natural language processing. arXiv preprint arXiv:2201.07489.

[9] Raza, S., & Ding, C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, *13*(4), 335–362. <https://doi.org/10.1007/s41060-021-00302-z>

[10] Alonso-Bartolome, S., & Segura-Bedmar, I. (2021). Multimodal Fake News Detection. arXiv preprint arXiv:2112.048312

<https://arxiv.org/abs/2112.04831>

[11] Bhushan, D., Agrawal, C., & Yadav, H. (2020). Fake News Detection: Tools, Techniques, and Methodologies. *Algorithms for Intelligent Systems*, 347–357. <https://doi.org/10.1007/978-981-15-4936-6_38>

[12] Trương, Q., Định, Thị, P., & Kiều, T. (2022). Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á Nghiên cứu giải pháp phát hiện tin giả trên mạng xã hội bằng ngôn ngữ tiếng Việt Mã phân loại JEL: C61; C63; C67. *Tạp Chí Nghiên Cứu Kinh Tế và Kinh Doanh Châu Á Năm Thứ*, *33*, 26–46. <http://jabes.ueh.edu.vn/Content/ArticleFiles/9918f8e1-fdaa-44eb-8f55-5fc4c759428d/JABES-2021-2-V27.pdf>

[13] Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040

[14] *Implementation of TF-IDF in JAVA*. (2015, February 17). Guendouz Mohamed; Guendouz Mohamed. <https://guendouz.wordpress.com/2015/02/17/implementation-of-tf-idf-in-java/>

[15] Ho Quang Thanh and ninh-pm-se, “thanhhocse96/vfnd-vietnamese-fake-news-datasets: Tập hợp các bài báo tiếng Việt và các bài post Facebook phân loại 2 nhãn Thật & Giả (228 bài)”. Zenodo, 27-Feb-2019.

**APPENDIX**