Cross-Linguistic Fake News Detection Using Machine Learning: A Comparative Analysis on English and Filipino News Datasets

Chris Lawrence De Vera
College of Computing and
Information Technologies
National University
Manila, Philippines
deveraca@students.national-u.edu.ph

Lovely Joy Reyes
College of Computing and
Information Technologies
National University
Manila, Philippines
reyeslp@national-u.edu.ph

Jude Renwell Prodigalidad
College of Computing and
Information Technologies
National University
Manila, Philippines
prodigalidadjb@national-u.edu.ph

***Abstract*—The reliability of information is threatened by the alarming rate of spread of fake news, especially in multilingual countries such as the Philippines. This study investigates the effectiveness and cross-linguistic generalizability of the four machine learning models, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes; for fake news detection. The models were evaluated in three (3) phases using English and Filipino datasets. In Phase 1, all models demonstrated strong performance on English data, with SVM having a 0.9977 score on F1-score. When English-trained models were used to evaluate Filipino text, Phase 2 revealed a significant drop of performance, scoring an average F1-scores falling to as low as 0.31, which shows limited cross-lingual transferability. Phase 3 showed that retraining the models with Filipino data restored the models accuracy—SVM and Logistic Regression both achieved a 97.50% accuracy and 0.9739 F1-score. The findings of this study reveal that linguistic mismatch, not model inefficiency, is the main cause of the model's performance drop. Furthermore, the study underscores the importance of language-specific training data in developing a reliable and context-aware fake news detection systems suitable for multilingual contexts like the Philippines.**

***Index Terms*—Fake news detection, machine learning, cross-linguistic analysis, text classification, multilingual datasets, natural language processing (NLP), Filipino language.**

I. Introduction

The spread of fake news poses a serious threat to the accuracy of information in democracies across the globe. From a conceptual standpoint, fake news is best described as the purposeful presenting of (typically) incorrect or misleading statements as news, when these are deceptive by design, rather than just as false information [1]. In order to ensure extensive circulation and belief, the statement "by design" is essential because it redirects attention from human purpose to the systemic characteristics of the sources and dissemination channels that are designed to manipulate audience cognition, frequently by taking use of cognitive biases [1], Public relations companies, state-sponsored campaigns, and ideologically agnostic clickbait models all work together to spread widespread misinformation in the Philippine, which is a clear example of an organized "disinformation shadow economy" [2]. Young Filipinos, who are becoming more dependent on social media for news consumption, are disproportionately affected by such systems, which are frequently motivated by political influence and profit rather than ideology [3].

The study of Cruz et al. [4] highlighted that fake news classifiers developed on English-language corpora, do not perform well consistently when it is applied to Filipino news across various types of articles. Many models for detecting fake news perform well when it comes to in-domain test sets but are having a hard time when it comes to generalize to novel or regionally-specific contexts. In the study of Zhu et al. [5] identified that entity bias demonstrates significant performance declines as entity distributions alter over time. If the models are trained with older time periods, it cannot generalize well to newly created fake news, a phenomenon known as diachronic bias which is shown in the study of Murayama [6].

This study aims to investigate the effectiveness and generalizability of multiple machine learning algorithms—namely Naive Bayes, Logistic Regression, Support Vector Machine, and Random Forest—for fake news detection across multilingual datasets. The research first trains the models using English-language datasets and evaluates their accuracy when tested on Filipino-language news datasets to assess the impact of linguistic and contextual differences on classification performance. The study further examines if language mismatch, rather than model inefficiency, is the cause of the observed performance drop by retraining models using Filipino datasets. The researchers anticipates that the results will highlight the importance of language-specific data in developing a reliable fake news detection systems applicable to the Philippine context. Furthermore, the findings of the study will provide a valuable insights for policymakers, media organizations, fact-checking institutions, and social media platforms by emphasizing the necessity of adopting appropriate and context-aware fake news detection tools.

II. Review Of Related Literature

According to the research written by Kudari et al., [7], the only influence fake news may have on individuals and society is negative. With enhanced technology, fake news is more easily accessible and spreads through traditional and social media, with the goal of misleading readers and viewers. It can persuade viewers to change their minds and make judgements based only on what they see. Fake news was believed to be propagated by social bots and populist political websites. On the bright side, the goal of this study is to develop a model that can accurately determine whether an article is fake news.

The selection of a model and feature extraction is important for fake news detection. According to the studies [7], [8], on some datasets, conventional machine learning models combined with TF-IDF vectorization can perform noticeably better than complex deep learning architectures. For instance, Support Vector Machine (SVM) and Naive Bayes classifiers, have outperformed DL models like Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, which scored 74% and 76%, respectively, with accuracy score as high as 89.34% [7], [8]. This implies that smaller easier to understand models can be quite successful for some applications and datasets. Strict text preprocessing, including stopword removal, is part of the standard pipeline for these models. Other steps include feature extraction using TF-IDF or Count Vectorizer to measure word importance and classification using algorithms like Naive Bayes or Passive-Aggressive classifiers, with confusion matrices used to assess performance [7], [8].

In accordance with research conducted by Figuera, A., artificial and simulated information has a substantial impact on enterprises and society [9]. In real-world situations, it quickly causes a great deal of casualties and has the power to destroy lives. Considering that it spreads quickly, methods are developed to mitigate the severity of the problem. In order to propose an algorithmic solution that would automatically detect fake news from web services, the study divides the solution into three categories: content-based, source-based, and diffusion-based [9]. It demonstrates that while younger generations are more proficient in technology than their older generation, they nonetheless blend in with society at large since 44% of them lack the ability to think critically to distinguish between real and false information. The fact that new publishers have less control over the distribution of news is one of the issues that increases the likelihood of fake news being published. In addition, social media platforms and search engines that allow them to publish and make money are being overtaken by new news markets. The most important issue at the moment is keeping track of information online, which calls for sustainability, control, and attention [9].

In the study of Souresh el al., three factors should be addressed in order to create a trustworthy model for fake news algorithms: linguistic style, human characteristics, and network propagation dynamic [10]. The Hidden Markov Model was used to create the model, which was able to identify the truth of 75% of rumors more quickly than other models. The Facebook Task Team method uses human classification to determine whether a post is spam and, ultimately, whether it is not. Using "social graph" Facebook obtains intelligence if the post is directly connected to the person who posted it. Therefore, Facebook created an algorithm to detect fake news based on how users interact with each other [10].

To shorten it up, the machine model cannot judge the value of information on its own because news accessibility is not limited to the conventional methods. Every user has the right to post what they are able and permitted to post, and freedom of speech is still at risk. The best method for guaranteeing the legitimacy of news that algorithms follow is still fact-checking and confirming the credibility of the source. Algorithms like existing APIs will demonstrate accuracy when implemented, though not always flawlessly. Even if there are currently only check blacklists of unreliable websites for detecting fake news, they have overlooked the need of cross-checking, history monitoring, and dynamic reputation updating; this still needs to be improved while taking freedom and transparency into account [10].

In today's world, social media has emerged as a valuable source of information. However, looking at the bad sides of social media reveals that fake news is one of the most severe problems confronting society today. Fake news is being used to propagate fake information on numerous social media platforms such as Facebook, Twitter, Instagram, and WhatsApp [11]. Fake news identification is a significant area of study in the discipline of Natural Language Processing (NLP). Patel et al. [11] conducted a comparative investigation of well-known machine learning algorithms for detecting fake news, with the goal of determining which classifier performs the best at identifying false news from legitimate news articles.

To test the model that had been supervised by a machine learning approach dataset. A total of 44,921 items from Kaggle, including 21,418 true news and 23,503 fraudulent news [11], [12]. Tokenizing the data is essential. It entails breaking down phrases into words and symbols, reducing words to base form, removing common words and features like n-grams, and speech-tagging. The technique used for confusion matrices, measurement of accuracy, precision, recall, and F1-score metrics is a comparison of six machine learning classifiers: Naive Bayes, SVM, Random Forest, Decision Tree Classifier, Multinomial Naive Bayes, and Logistic Regression.

The comparison results showed that SVM surpassed all other machine learning classifiers with the accuracy rating of 94.93%, preciseness of 93.98%, recall of 96.04%, and F1-Score of 94.99%, alongside Random Forest fell 92.37% and Multinomial Naive Bayes scored 92.98%. The Decision Tree Classifier and Naive Bayes perform at 89.70% and 88.58%, respectively. To summarize, the study compares the efficacy of SVM in fake news detection to other algorithms [11].

Conforming to the study of Cruz et al. [4], developed fake news detection models perform well in languages with abundant resources but frequently fall short in languages with limited resources. Thus, they established a benchmark combining transformer-based transfer learning and multitask learning to train the model to adjust based on writing style and general language patterns in order to make this more cross-multilingual. A total of 3,206 news stories that are half phony and half real are created using this method. They attain an impressive 91% accuracy rate; but by utilizing multitask augmentation, they were able to enhance their performance by up to 96% and decrease errors by up to 14% in comparison to current models. The approach is also adaptable to a variety of news subjects, including politics, entertainment, sports and more.

The studies reviewed by the researchers explored at different perspectives to improve and broaden their models for detecting misinformation. The gaps, however, remain clearly visible. The vast majority of machine learning models in use today were trained solely in English, resulting in poor performance in multilingual environments. Furthermore, it ignored intellectual and interpersonal variables that influence people's susceptibility to counterfeit news. Feature extraction using existing machine learning techniques failed to capture text's deeper lexical and contextual meaning. Particularly in the patterns for identifying false information, this needs to be improved. Ultimately, without delving further into other facets and formats of news, the research concentrates on achieving high accuracy. Enhanced adaptability and reduced diachronic bias to create an algorithm that will help the users distinguish and differentiate the real from fake news.

III. Methodology

The research project analyzes online news articles in both English and Filipino to detect fake news using supervised machine learning techniques. The study employs two primary algorithms, Naive Bayes and Logistic Regression, to evaluate and compare in identifying real and fake news. Before model training, the data underwent several data preprocessing and feature engineering steps. Each process is outlined and expounded as follows:

A. Data Collection

This initial phase and preliminary processing phase is important in fake news detection. The datasets utilized in the study were collected using a wide range of data from an open-source website called Kaggle for Data Science. The dataset, which was compiled in the United States, is separated into two categories: fake news and true news. Of the 44,898 data it contains, 21,417 are real news and 23,481 are fake [12]. Additionally, it includes the 3,206 equally divided Filipino-based dataset from the HuggingFace website [13].

During the preliminary processing phase, data cleaning is mandatory; this is done to ensure that all duplicates, missing data are eliminated before proceeding to data pre-processing to guarantee consistency and quality. Deploying visualization of news subjects using pie charts and using figures to show frequency words used.

B. Data Exploratory

Before pre-processing the data, exploratory data analysis was conducted to understand the structure, completeness, and balance of the dataset collected that will be used in this study. The datasets provided from Kaggle were separated—Fake.csv and Real.csv, both were concatenated into a single DataFrame. An additional column named ”Label” was added to distinguish between Fake (F) and Real (R) news articles.

Upon checking the combined dataset using *.shape* function, this showed that the dataset contains 44,898 rows and 5 columns. The structure of the dataset was confirmed using *info()* function.

TABLE I: English Dataset Structure

Column	Non-Null Count	Dtype
title	44,898	object
text	44,898	object
subject	44,898	object
data	44,898	object
Label	44,898	object

This shows that the dataset contained no missing values. Further verification using *isna().sum()* function revealed no null entries across the datasets. However, using *duplicated().sum()* to check for any duplicate values, revealed 209 duplicate rows, which were subsequently removed to maintain data integrity.

Upon further data exploratory, with the use of *value_counts()* function showed the distribution of labels. The results revealed that the dataset was relatively balanced with 23,481 Fake (F) and 21,417 Real (R) news article.

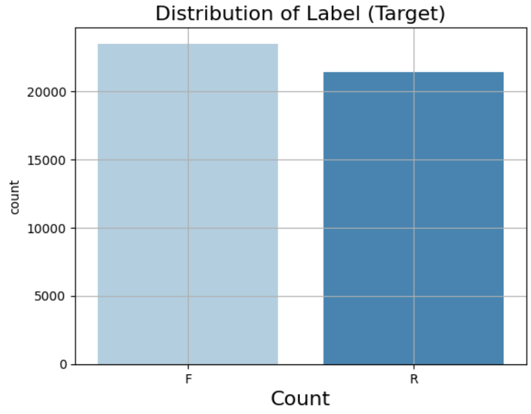


Fig. 1: Class Distribution of Labels

This slight imbalance was visualized in Figure 1, showing the Distribution of Label (Target), where both categories are nearly equal in representation, ensuring unbiased training across models.

To visually comprehend the topic diversity within the dataset, the ”subject” column was analyzed. As shown in Figure 2, the pie chart reveals that ’politicsNews’ and ’worldnews’

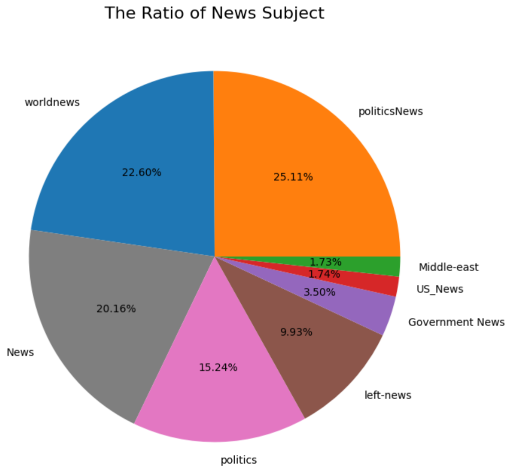


Fig. 2: Ratio of News Subject

dominate the dataset, representing approximately 25.11% and 22.60% of the total samples, respectively. Other subjects such as 'News' (20.16%), 'politics' (15.24%), and 'left-news' (9.93%) also contributed substantially, while topics like 'US_News', 'Government News', and 'Middle-east' had smaller representations.

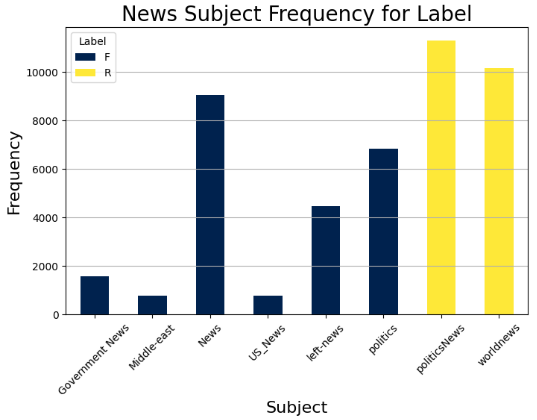


Fig. 3: News Subject Frequency by Label

To have further understanding about the dataset, the frequency of each subject was analyzed by label category (Fake or Real). As shown in Figure 3, fake news articles were more frequent in subjects such as politics, left-news, and News; while real news was primarily about subjects politicsNews and worldnews. This imbalance across topics suggest potential domain bias, where models might learn topic-specific cues rather than purely linguistic or factual distinctions.

C. Data Pre-processing

To ensure data consistency, quality, and readability, the text data was converted to lowercase and cleaned by removing punctuation marks, numbers, special symbols, and hyperlinks. Tokenization was applied to separate the words and the individual tokens. Additionally, stopwords were removed to reduce noise and improve the accuracy of the feature extraction. The English dataset contained 209 duplicated rows, while the Filipino dataset contained 201 duplicated rows. Given such values, the rows of duplicate values were dropped. Furthermore, for the English dataset, the categorical data labels "F" and "R" were converted into numerical values "1" and "0" for proper data modeling.

1) **Stop Word Removal (normalization)** To ensure data consistency, quality, and readability, the text data went through several preprocessing steps. Along with this, Stopword removal was applied to reduce common words such as "the", "is", "and", as well as their Filipino counterparts, "ang", "ng", "sa". Which minimally contributes to the contextual meaning of the content. Stopword removal reduced the noise in the content and allowed the models to focus on more informative and relevant terms.
   In this study, stopword removal was applied to both the English and Filipino datasets to enhance the linguistic accuracy. Additionally, A predefined list of English stopwords from the Natural Language Toolkit (NLTK) was used, while for Filipino stopwords, a custom list was created to account for the unique structure and vocabulary of the Filipino language.
2) **Lemmatization** Lemmatization is a key step in text normalization, in which it converts words into their base or dictionary form. To ensure accuracy, Lemmatization takes into consideration the word's meaning and grammatical context. This approach ensures that semantically related words are treated as a single feature, further reducing the redundancy within the dataset. By grouping word forms together, Lemmatization enhances the model's core understanding of the text. This is particularly valuable in classification tasks, which are relevant to fake news detection, where the contextual understanding of the language contributes to a higher model accuracy.

1) Tokenization: A procedure of converting text into classification tokens that can be numerically represented. Among its various strategies is Term Frequency - Inverse Document Frequency (TF-IDF). It serves as a statistical method since it can measure text inside a document environment [14].

1) **Feature Extraction**
   Inside the pipeline, the TF-IDF Vectorizer extracts and turns the text into numeric features from which the algorithm learns and predicts the feature used to categorize the text. GridSearchCV hyperparameters are used to fine-tune parameter combinations. GridSearchCV works by fitting all models to the training data and then evaluating all potential combinations using cross-validation to reach the best average accuracy. This provides unbiased outcomes. The following is the formula for the TF-IDF Vectorizer:

$$\text{TF-IDF}(t, d) = \text{tf}_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

Equation 1: TF-IDF Vectorizer

In simple terms, the formula for TF-IDF is composed of two components: the term frequency which is responsible for measuring how frequent the word appears in the document. The higher the frequency the greater the importance. The second component is Inverse document frequency, this works by increasing the weight of rare words and decreasing the common words across documents which is very helpful for text classification such as fake news detection [15].

D. Phase 1

This phase is dedicated to developing a collaborative experimental approach that will assist in the formulation of the most effective training ground for a dataset based in the United States. The experiment is carried out in Python using Jupyter Notebook or Google Colab. To achieve the goal of comparing multiple algorithms for detecting fake news.

The procedure was divided into two (2) processes: the baseline, which contains the default settings, and hyperparameter tuning, which prioritizes GridSearchCV for optimal performance. The dataset contains 44,898 records partitioned into training (80%) and testing (20%) using data splitting. Additionally, feature extraction techniques and pipeline classification algorithms are implemented. The performance of the models given below will be evaluated using key indicators such as accuracy, precision, recall, and F1-score. To guarantee consistency, the parameters are visualized using a confusion matrix, as well as fivefold cross-validation.

1) **Naive Bayes Algorithm**
Multinomial Naive Bayes is a supervised learning method for news classification that is effective at dealing with word frequencies and processing huge datasets. The model is trained using an alpha value of 0.1, which smoothens the probability of avoiding overfitting, improves generalization, and sets fit_prior to True since it assumes that feature independence is appropriate for TF-IDF outputs and accounts for imbalances by adjusting the model's probability. Grid search identifies the optimal key metrics for the model.
2) **Logistic Regression Algorithm**
Logistic regression is a supervised method that analyzes the data by finding the relationship and implementing a binary classification. Uses liblinear with a regularization parameter of C to avoid overfitting.
3) **Support Vector Machine Algorithm**
The support vector machine is a supervised algorithm due to its high performance in text classification. The algorithm utilizes a linear kernel to optimize the difference between real and fake news. Using the regularized parameter partnered with max_iterations to detect errors and enough time to optimize the model without getting a full stop.
4) **Random Forest Algorithm**
Random forest supervised methods construct multiple division trees to predict accurate and stable results. It has the ability to handle non-linear relationships that avoid overfitting, and it is good for capturing complex features that other algorithms might miss.

E. Phase 2

This phase focuses on evaluating the transferability and generalization of the models trained on the English dataset and applying them to the Filipino dataset. The goal is to determine whether the model initially trained in the English dataset can accurately classify fake and real news in another language environment. In this context the chosen language would be Filipino. This phase is for identifying language barriers and performance degradation due to cross-lingual variations.

The Filipino dataset is pre-processed using the same pipeline from Phase 1, which includes text normalization, tokenization, and TF-IDF vectorization. Furthermore, the Filipino dataset was passed through the previously trained models from Phase 1, which were the Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest, to evaluate their prediction accuracy on the Filipino dataset.

1) **Naive Bayes Algorithm**
The Naive Bayes model trained on the English dataset was then used to predict in the Filipino Dataset. Since the algorithm relies on the word frequency distribution, this test helped identify whether the statistical word relationship the model learned from Phase 1 could still apply to Filipino patterns. The output of the model revealed how well it could generalize to different linguistic structures and vocabularies without needing to retrain the model. This procedure applied the English-trained model to the Filipino dataset, and evaluation metrics such as accuracy score and classification report were used to assess prediction accuracy.
2) **Logistic Regression Algorithm**
Logistic Regression model was tested on the Filipino dataset to determine its accuracy. The model, trained using English features, attempted to separate the Filipino text into fake and real categories using its previously learned classification boundary. This helped analyze the limits of direct model transfer across the two languages
3) **Support Vector Machine Algorithm**
The SVM from Phase 1 was applied to the Filipino dataset to test its margin-based classification boundary in TF-IDF feature representations. The model's performance in this phase illustrated that the English-trained model behaves with different vocabulary distributions and cultural expressions present in Filipino texts. These commands applied the English-trained decision boundary to the Filipino feature space. Model evaluation using accuracy score and classification report measured classi-

fication performance and overall quality against linguistic differences between English and Filipino datasets.

4) **Random Forest Algorithm**
The Random Forest Classifier was also tested using the Filipino Dataset to evaluate the model's flexibility in cross-language prediction. This approach assessed how effective the tree-based learning from the English dataset could be applied to Filipino linguistic structures. Model performance was then verified using accuracy score and confusion matrix visualizations.

F. Phase 3

After analyzing the results from Phase 2, Phase 3 focuses on retraining all the models using the Filipino dataset. This phase aims to develop machine learning models that are fully optimized for the Filipino Language. Based on its linguistic characteristics and news patterns. The same pre-processing pipeline was used in Phase 1, including tokenization, lemmatization, stopword removal, and TF-IDF vectorization, but adjusted to accommodate the Filipino vocabulary. The dataset was then divided into training (80%) and testing (20%).

1) **Naive Bayes Algorithm**
The Multinomial Naive Bayes was retrained using the Filipino dataset, emphasizing the statistical frequency of the words in a local context. The model used the alpha parameter to prevent overfitting as well as to handle unique terms in the Filipino language. This allowed the model to build probability distributions that accurately represent the Filipino linguistic patterns, which enhances the fake news classification performance.

2) **Logistic Regression Algorithm**
Logistic Regression was trained using the Filipino dataset and transformed by the TF-IDF. The model allowed more flexibility to fit the training data with the use of C parameter to prevent overfitting. The model limits large feature weights separating fake and real news through the learned weights that adapt to the Filipino language uniqueness, providing interpretable results for local news content.

3) **Support Vector Machine Algorithm**
Support Vector Machine roles is to study the dataset's word patterns and phrases using TF-IDF to transform those phrases into numerical algorithms. To find the optimal line (hyperplane) of fake and real news using LinearSVC that handles large datasets for clear interpretations of language structure and overlaps this is executed to assess class imbalance.

4) **Random Forest Algorithm**
Random forest algorithm uses its tree to train the random feature of the Filipino n-grams. These trees then take a majority vote on whether the data is fake or real by using the parameters like n_estimators and max_depth that help the algorithm to train the data accurately while avoiding overfitting. This provides a stable model and avoids growing of trees uncontrollably to distinguish the differences between linguistics

IV. Results And Discussions

As mentioned in the methodology section, models went through three phases. Models were evaluated based on Accuracy, Precision, Recall, and F1-score, which were selected to provide a balanced view of each model's ability to correctly classify both phony and real news. Confusion matrices were also analyzed to visualize the true and false classification for each model.

A. Phase 1: Trained Model with English-language Dataset Evaluation on English Dataset

1) Models using Default Parameter: .

TABLE II: Performance of Models on English Datasets (Default Parameters)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Support Vector Machine	0.9975	0.9979	0.9975	0.9977
Random Forest	0.9935	0.9961	0.9916	0.9939
Logistic Regression	0.9917	0.9936	0.9907	0.9922
Naive Bayes	0.9503	0.9595	0.9462	0.9528

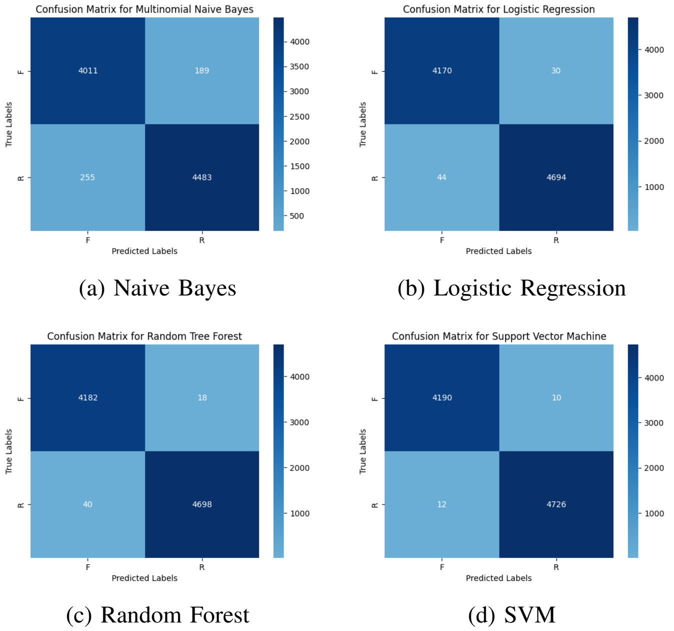
To further validate consistency, 5-fold cross-validation was conducted. The mean F1-scores and standard deviations are summarized.

TABLE III: Cross-Validation Mean F1-Scores (Default Parameters)

Model	**Mean F1-Score**	**Standard Deviation**
Support Vector Machine	0.9946	±0.0013
Random Forest	0.9871	±0.0047
Logistic Regression	0.9846	±0.0036
Naive Bayes	0.9149	±0.0238

The results indicate that the Support Vector Machine (SVM) algorithm achieved the highest overall performance. It attained an accuracy of 99.75% and an F1-score of 0.9977. Furthermore, SVM achieved the highest mean F1-score of 0.9946 with the lowest standard deviation of ±0.0013–indicating both high accuracy and consistency across folds. Random Forest and Logistic Regression followed closely. Random Forest achieved an accuracy score of 99.54%. a mean F1-score of 0.9871, a standard deviation of ±0.0047, while Logistic Regression attained an accuracy score of 99.17%, a mean F1-score of 0.9846, and score ±0.0036 on standard deviation. In contrast, Naive Bayes model recorded the lowest performance among 4 models, with an accuracy score of 95.05%, a mean F1-score of 0.9149 and standard deviation of ±0.0238.

Figure 4 revealed that misclassifications rates were minimal across all models. For instance, SVM misclassified a total of 22 samples (10 false positives and 12 false negatives), highlighting its superior discriminative ability in distinguishing



(a) Naive Bayes

(b) Logistic Regression

(c) Random Forest

(d) SVM

Fig. 4: Confusion Matrices of Machine Learning Models (Default Parameters)

between fake and real news. Logistic Regression and Random Forest resulted in slightly higher but minimal misclassifications. Furthermore, Naive Bayes showed reliable accuracy but with a slightly higher number of false negatives.

2) Models After Hyperparameter Tuning: .

After establishing the results of baseline models, GridSearchCV was applied to optimize each model's parameters, such as regularization for Logistic Regression and SVM. Whereas for Random Forest, parameters such as number of estimators (n_estimators), and tree depth (max_depth), and smoothing parameter (alpha) for Naive Bayes. The tuned models were then retrained and re-evaluated using the same English-language dataset.

TABLE IV: Performance of Models on English Datasets (After Hyperparameter Tuning)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Support Vector Machine	0.9974	0.9977	0.9975	0.9976
Random Forest	0.9943	0.9956	0.9937	0.9946
Logistic Regression	0.9961	0.9962	0.9964	0.9963
Naive Bayes	0.9578	0.9605	0.9599	0.9602

After tuning, all models maintained or slightly improved their overall performance which is presented in Table IV. Naive Bayes, particularly, achieved an F1-score increase from 0.9528 to .09602, and Logistic Regression, which reached a near-perfect accuracy score of 99.6%.

TABLE V: Cross-Validation Mean F1-Scores (After Hyperparameter Tuning)

Model	**Mean F1-Score**	**Standard Deviation**
Support Vector Machine	0.9890	±0.0053
Random Forest	0.9947	±0.0011
Logistic Regression	0.9922	±0.0012
Naive Bayes	0.9211	±0.0255

The cross-validation results presented in Table V, confirms the consistency performance improvements across folds, with slightly reduced variance for tuned models. The improvement of Naive Bayes and Logistic Regression shows that hyperparameter tuning positively affected generalization and slightly reduced performance variance. SVM maintained impressive results both and after tuning, showing that it was already near optimal in its default configuration.

B. Phase 2: Cross-Linguistic Evaluation – English-trained Models on Filipino Dataset

In this phase, the researchers trained the models exclusively on English-language fake and real news data, and then evaluated with Filipino-language articles to assess cross-linguistic generalization performance. This phase demonstrated the study's main objective—determine whether the drop in accuracy across datasets is attributed to language mismatch rather than model inefficiency.

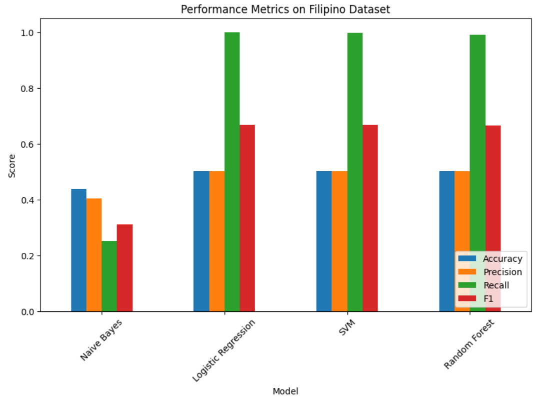


Fig. 5: Performance Metrics on Filipino Datasets

As shown in Figure 5, the accuracy, precision, recall, and F1-scores of the models tested on the Filipino dataset, all classifiers showed a significant drop when it comes to performance compared to their results on the English dataset. Among all models, Naive Bayes exhibited the lowest F1-score with 0.31, which shows the model's limitation when capturing contextual and morphological variations between English and Filipino text. In contrast, Logistic Regression, SVM, and Random Forest achieved an impressive high F1-scores, ranging 0.66–0.67, but maintained accuracy scores close to 0.50, shows that these models primarily predicted one dominant class.

The detailed classification metrics report are summarized on table VI

TABLE VI: English-trained Models Performance Tested on Filipino Datasets

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Machine	0.5018	0.5020	0.9980	0.6680
Random Forest	0.5012	0.5017	0.9914	0.6662
Logistic Regression	0.5022	0.5022	1.0000	0.6686
Naive Bayes	0.4379	0.4045	0.2525	0.3109

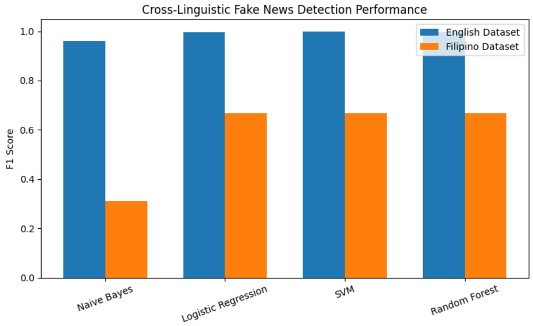


Fig. 6: Cross-Linguistic Fake News Detection Performance

A comparative visualization of the F1-scores between English and Filipino datasets are shown on Figure 6—showed a drastic drop in model performance when evaluated on cross-linguistic data. The average decline in F1-score across models was approximately 33-65%, which indicates that even high-performing English-trained machine learning models fail to generalize effectively to Filipino text.

C. Phase 3: Performance of Models Trained and Tested on Filipino Dataset

For this final phase, researchers evaluated the performance of models that were trained and tested entirely on the Filipino-language fake news dataset. The purpose of this phase was to determine whether retraining the models on native-language data would restore the model's classification performance that have been drop during Phase 2 which is cross-linguistic testing.

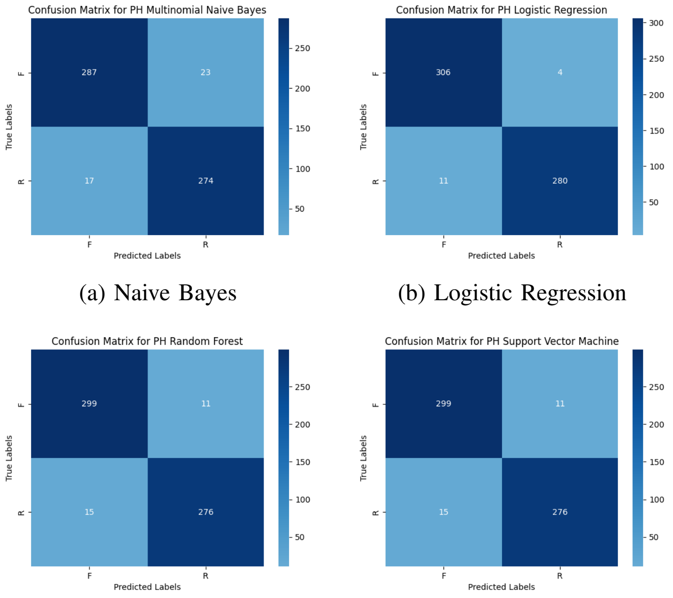
TABLE VII: Filipino-trained Models Performance Tested on Filipino Datasets

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Machine	0.9750	0.9859	0.9622	0.9739
Random Forest	0.9567	0.9617	0.9485	0.9550
Logistic Regression	0.9750	0.9859	0.9622	0.9739
Naive Bayes	0.9334	0.9226	0.9416	0.9320

TABLE VIII: Cross-Validation Mean F1-Scores (Filipino-trained model on Filipino Dataset)

Model	Mean F1-Score	Standard Deviation
Support Vector Machine	0.9890	±0.0053
Random Forest	0.9947	±0.0011
Logistic Regression	0.9922	±0.0012
Naive Bayes	0.9211	±0.0255

Table VII shows the overall evaluation metrics for all classifiers after retraining on Filipino fake news dataset. As observed, all models achieved a significant improvements compared to cross-linguistic counterpart from Phase 2. The Support Vector Machine (SVM) and Logistic Regression both achieved the highest accuracy score 97.50% and an F1-scores of 0.9739. Furthermore, Random Forest and Naive Bayes demonstrated a strong performance as well, with an accuracy scores of 95.67% and 93.34%, respectively.



(a) Naive Bayes

(b) Logistic Regression

(c) Random Forest

(d) SVM

Fig. 7: Confusion Matrices of Filipino-trained models

The classification reports shown at Table VII and Table VIII further confirms the consistency of these results across both classes (”Fake” and ”Real”). Notably, Figure 7 revealed balanced prediction with minimal misclassifications, suggesting that the models effectively learned the linguistic and contextual nuances of Filipino fake news content.

The findings across the three experimental phases clearly reveal how linguistic context and dataset alignment strongly influence the performance of machine learning models for fake news detection. While the four algorithms, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Naive Bayes—exhibited consistently high accuracy when trained and tested on English-language data, their performance declined significantly when applied to Filipino-language data without retraining. This observation demonstrates that language mismatch, rather than algorithmic inefficiency, serves as the primary source of performance degradation.

In Phase 1, all models performed exceptionally well when the models were trained and tested with English dataset. SVM got an almost perfect result with a 99.75% accuracy score and an F1-score of 0.9977. Logistic Regression and Random Forest maintained a high score for precision and recall. The minimal difference between pre-tuned and post-tuned performances indicates that these algorithms were already optimized for the linguistic and statistical patterns of English data. Furthermore, cross-validation results showed a low variance between folds, which confirms the model's stability robustness in a monolingual context.

However, during Phase 2 where the researchers performed a cross-linguistic experiment which exposed a contrast results. When English-trained models were tested on the Filipino-language fake news dataset, all classifiers experienced a severe accuracy and F1-score drops. For instance, Naive Bayes's F1-score dropped from 0.95 to 0.31, while even the top-performing models such as SVM and Logistic Regression only managed to have a 0.66-0.67 F1-score, and accuracy scores hovering near 0.50. These results suggest that the models failed to recognize linguistic cues and token patterns present in Filipino text—effectively defaulting to predict one dominant class. The degradation pattern was consistent across all four algorithms; underscoring was not due to the model's inefficiency, but to semantic and morphological differences between training and testing corpora.

Lastly, Phase 3 showed that retraining the models with a Filipino-language dataset restored high performance scores across all models. SVM and Logistic Regression both achieved an accuracy score 97.50% and an F1-score of 0.9739, effectively matching their English-language benchmarks. Naive Bayes, who scored the lowest with cross-linguistic generalization recovered to an F1-score of 0.9320 after being trained on Filipino text. These confirms the study's main hypothesis strong: performance decline in cross-linguistic fake news detection is caused by language mismatch, not by model inefficiency. The same architectures, once exposed to native-language data will regain their predictive strength.

Overall, the findings give light that machine learning models are not inherently weak at detecting misinformation but are highly sensitive to language and domain alignment. This study highlights a critical consideration for multilingual societies like the Philippines—relying solely on English-trained fake news classifiers will likely produce a misleading outcomes. Developing or fine-tuning models on language-specific datasets is therefore important for achieving trustworthy and contextually releveant fake news detection.

V. Conclusion

This study successfully achieved its objectives of detecting fake news from real news by implementing the desired machine learning algorithms namely Naive Bayes, Logistic Regression, Support Vector Machine, and Random forest to train and test US-based and Filipino-based datasets from Kaggle and Huggingface. Dividing the analysis of each datasets is the best move in giving each algorithm to perform efficiently and accurately. This experiment demonstrated that while all the models achieved the expected accuracy on all of the phases, the Support Vector Machine and Logistic Regression were consistently higher in metric evaluation capturing linguistic generalization with an approximately percentage of 95% to 99.97%. The observation second motioned that model performance was not attributed to the algorithms since it demonstrated consistent effectiveness under similar conditions rather than linguistic characteristics and structural complexity of the dataset is the huge influence.

Despite positive results, there are limitations that can be identified. First, the study primarily concentrated on developing machine learning and did not go overboard on adding deep learning layers to learn patterns automatically. Second, the models show limited support for cross-linguistic adaptability that affects the quality and balance of datasets. Followed by, the lack of TF-IDF to capture the semantic meaning or contextual relationships between the words in feature extraction that affects the performance outcomes especially in Filipino-based dataset. Lastly, since it's experimental, the study did not focus on including statistical significance testing that will determine if the results are actually real or random guesses.

For future work, it is highly recommended to incorporate and integrate deep learning architectures so it will suit more for multilingual fake news detection and enhance cross-lingual analysis. Additionally, application of feature engineering to better understand the semantic meaning of the datasets contents and refinement of pipelines and hyperparameters. In conclusion, the model created is a strong foundation of developing fake news detection especially in the Philippines news outlets and consumers. This will be valuable insights for the media, and policymakers for awareness.

References

[1] K. Jeevan and K. V. Kanimozhi, "Improved accuracy for fake news in social media using logistic regression comparing naive bayes classifier," in *Advances in Parallel Computing: Algorithms, Tools and Paradigms*, ser. Advances in Parallel Computing. IOS Press, 2022, vol. 41, pp. 481–486.

[2] J. C. Ong and R. Tapsell, "Demystifying disinformation shadow economies: Fake news work models in indonesia and the philippines," vol. 32, no. 3, pp. 251–267, 2022. [Online]. Available: https://doi.org/10.1080/01292986.2021.1971270

[3] I. B. Deinla, G. A. S. Mendoza, K. J. Ballar, and J. K. Yap, "The link between fake news susceptibility and political polarization of the youth in the philippines," vol. 30, no. 2, pp. 160–181, 2022. [Online]. Available: https://doi.org/10.1080/02185377.2022.2117713

[4] J. C. B. B. Cruz and Charibeth Cheng, "Evaluating language model finetuning techniques for low-resource languages," 2019. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.23028.40322

[5] Y. Zhu, Q. Sheng, J. Cao, S. Li, D. Wang, and F. Zhuang, "Generalizing to the future: Mitigating entity bias in fake news detection," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. ACM, Jul. 2022, p. 2120–2125. [Online]. Available: http://dx.doi.org/10.1145/3477495.3531816

[6] T. Murayama, S. Wakamiya, and E. Aramaki, "Mitigation of diachronic bias in fake news detection dataset," 2021. [Online]. Available: https://arxiv.org/abs/2108.12601

[7] J. Kudari, V. Varsha, R. Archana, and B. G. Monica, "Fake news detection using passive aggressive and tf-idf vectorizer," *International Journal of Research in Engineering and Technology*, vol. 7, pp. 2395–0072, 09 2020.

[8] A. Tanvir, E. Mahir, S. Akhter, and M. R. Huq, "Detecting fake news using machine learning and deep learning algorithms," 06 2019, pp. 1–5.

[9] Figueira and L. Oliveira, "The current state of fake news: challenges and opportunities," *Procedia Computer Science*, vol. 121, pp. 817–825, 12 2017.

[10] S. Vosoughi, M. ohsenvand, and D. Roy, "Rumor gauge: Predicting the veracity of rumors on twitter," *ACM Transactions on Knowledge Discovery from Data*, vol. 11, no. 4, pp. 1–36, 2017.

[11] A. Patel, A. K. Tiwari, and S. S. Ahmad, "Fake news detection using support vector machine," in *Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering (ICACSE 2021)*. SciTePress, 2021, pp. 34–38.

[12] C. Bisaillon, "Fake and real news dataset," Kaggle, 2020, [Online; accessed Aug. 10, 2024]. [Online]. Available: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

[13] J. C. B. Cruz, J. A. Tan, and C. Cheng, "Localization of fake news detection via multitask transfer learning," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 2596–2604.

[14] CodeSignal Learn. (2023) Tokenization: The gateway to text classification. CodeSignal. Accessed: 2024-01-15. [Online]. Available: https://codesignal.com/learn/courses/feature-engineering-for-text-classification/lessons/tokenization-the-gateway-to-text-classification

[15] GeeksforGeeks. (2025, 8) Understanding TF-IDF (term frequency-inverse document frequency). GeeksforGeeks. Accessed: 2024-12-19. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/