# CHAPTER THREE

# RESEARCH METHODOLOGY

## 3.1 Introduction

The research method used to explain the sentiment in emails will be explained in this chapter. The research design and the principles driving the study, the approaches taken, and the resources utilized. Additionally, the chapter will discuss the sampling strategy employed to collect the emails, the ethical considerations that were considered, and the limitations of the research design. Furthermore, the chapter will provide a detailed description of the data analysis process and the interpretation of the results obtained.

## 3.2 Research Philosophy

Research philosophy is the assumptions and beliefs that underpin the advancement of knowledge, it decides the overall approach and structure of the study (Saunder et al.2019) and captures the relationship between research and theory. Tuli (2010) describes two orientations, epistemology, and ontology orientations. Ontology is the study of the nature of existence and aims to provide an answer to a research question by identifying existing types of knowledge. While epistemology is concerned with how a researcher acquires knowledge to uncover reality.

Saunders et al (2019) recognized interpretivism, positivism, critical realism, pragmatism, and postmodernism as five ideologies that support management research. Positivism and interpretivism are the two major divisions made by Creswell & Creswell (2017) in research philosophy. According to positivism, gathering quantitative data and applying scientific research techniques to the study of social phenomena are very important. Strijker, Bosworth, and Bouter (2020) suggests positive thinkers think that scientific investigation can reveal laws that can be used to explain social phenomena.

Contrarily, interpretivist emphasizes the significance of comprehending social processes from the perspective of the individuals who experience them. According to interpretivists, social events are produced by human interaction, and researchers should try to understand the interpretations that individuals make of their experiences (Snyder 2019). They contend that qualitative techniques like observation and interviewing are the most effective means of comprehending social processes (Alharahsheh et al. 2020). Several crucial aspects of interpretivism are pertinent to this study. First, interpretivists contend that social phenomena are shaped by human interaction and that researchers should work to comprehend the meanings that individuals give to their experiences. This is particularly important in this study because it aims to analyze emails' sentiments to spot fraud.

This study's research technique is based on both positivist and interpretivist schools of thought. The study adopts both lexicon-based and machine learning-based approaches. The use of these two approaches, lexicon-based and machine learning-based approaches in research do not inherently align with either positivism or interpretivism, as these are epistemological paradigms dealing with how knowledge is generated and interpreted, rather than specific research methods. However, it is worth noting that the use of machine learning-based approaches may be more closely aligned with positivism, which tends to prioritize quantitative data analysis and the use of objective, empirical methods, while the use of lexicon-based approaches may be more closely aligned with interpretivism, which tends to prioritize qualitative data analysis and the use of subjective, interpretive methods. Ultimately, the epistemological orientation of the researcher will influence how they interpret and make sense of the results of their research, regardless of the specific methods they use (Denzin & Lincoln, 2011).

## 3.3 Research Design

A research design serves as a roadmap that outlines the techniques, strategies, and procedures to be employed in a research investigation. This aspect of the research is of great importance as it significantly impacts the study's feasibility, validity, and reliability. The research design determines the major variables and associations to be examined, as well as the procedures for selecting a sample and collecting data (Hilmola 2018). The selection of the appropriate research design depends on the research topic, the type of data, and the available resources. Some of the research designs available include exploratory, descriptive, correlational, experimental, and mixed methods research designs (Saunders et al. 2019).

To better comprehend the sentiment analysis of business emails, a mixed method research design combining both descriptive and mixed methods will be used in this study to classify emails using sentiment analysis (Snyder 2019). The descriptive methods will be utilized to identify any data quality issues such as missing data and outliers that may require resolution before further analysis to ensure the accuracy of the results.

A mixed methods design involves utilizing both qualitative and quantitative research methods in a single study to obtain a more comprehensive understanding of a research question or problem (Taherdoost 2022). In the context of sentiment analysis in email classification, a mixed methods design may entail employing both lexicon-based and machine learning-based approaches to analyze the sentiment of emails, as well as collecting qualitative data through interviews or surveys to gain a deeper understanding of the context in which the emails were written and the attitudes and perceptions of the individuals who wrote them.

The quantitative component of the study may include using machine learning algorithms to automatically classify emails as either fraudulent or legitimate based on their sentiment scores. The qualitative component may involve conducting interviews with individuals who have encountered email fraud previously to gain insight into their perceptions of the problem and their strategies for avoiding it (Morgan 2022).

By combining both quantitative and qualitative research methods, a mixed methods design can provide a more nuanced understanding of the sentiment analysis of emails and identify new insights and trends that may not be apparent through quantitative analysis alone.

## 3.4 Research Approach

The research approach is the overall strategy or procedure utilized in carrying out a research study. The research approach picked will determine how data is gathered and processed. The two main types of research methods are quantitative and qualitative (Saunders et al.2018).

The natural and social sciences typically apply the quantitative research approach, which is used if a researcher wishes to use data that is numeric to test a certain hypothesis or idea. The methodology used in this approach is deductive, based on statistical analysis, and starts with a theory before being supported by empirical data compared to an inductive approach that is theorydriven and creates research hypotheses that are intended to fill the existing knowledge gap (Stockermer 2019).

The qualitative research approach, on the other hand, is used when the researcher wants to comprehend the perspectives, perceptions, and experiences of individuals or groups. It is often used in social sciences and humanities and places emphasize on gathering and analyzing nonnumerical data such as interviews, written materials and observations. This strategy applies an inductive research process, creating a theory out of the data gathered (Saunders et al. 2018).

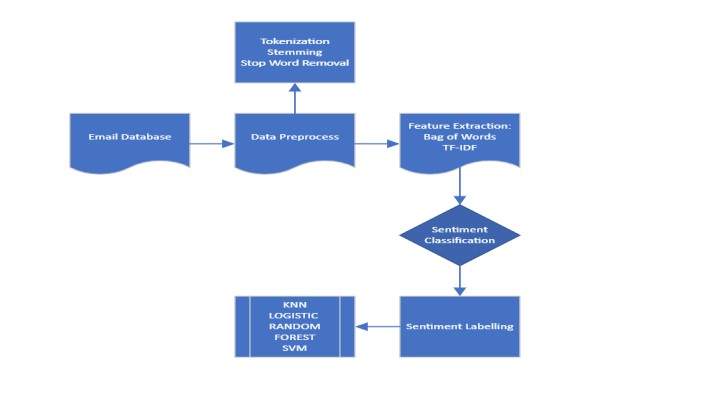
This study will utilize both quantitative and qualitative techniques to analyze the sentiment of emails in this study. To preprocess the text data (emails), a qualitative research strategy will be used, this is because at this stage it is text data. Once the qualitative data has been transformed into numerical data, to test the hypothesis and theory a quantitative research approach will be used.

Mixed-methods strategy is the name for this fusion of the two methods (Saunders et al. 2019).

## 3.5 Data Description and Preprocessing

The dataset, which can be accessed at (https://www.cs.cmu.edu/enron/), was gathered and produced by the CALO Project and contains financial data and text features that were extracted from emails composed of 146 users with 21 features apiece. The information is a pre-processed list of email texts taken from the dataset from the Enron Corporation. One of the top energy marketers in North America, Europe, and the rest of the world, the corporation sells liquids, electricity, natural gas, and crude oil. The second set of information that will be used in this study will be taken from the Kaggle website (https://www.kaggle.com/datasets/llabhishekll/fraud-emaildataset?resource=download).

The emails collected are preprocessed for sentiment analysis (Pradha et al. 2019) which involves several steps to prepare the data for analysis as shown below.

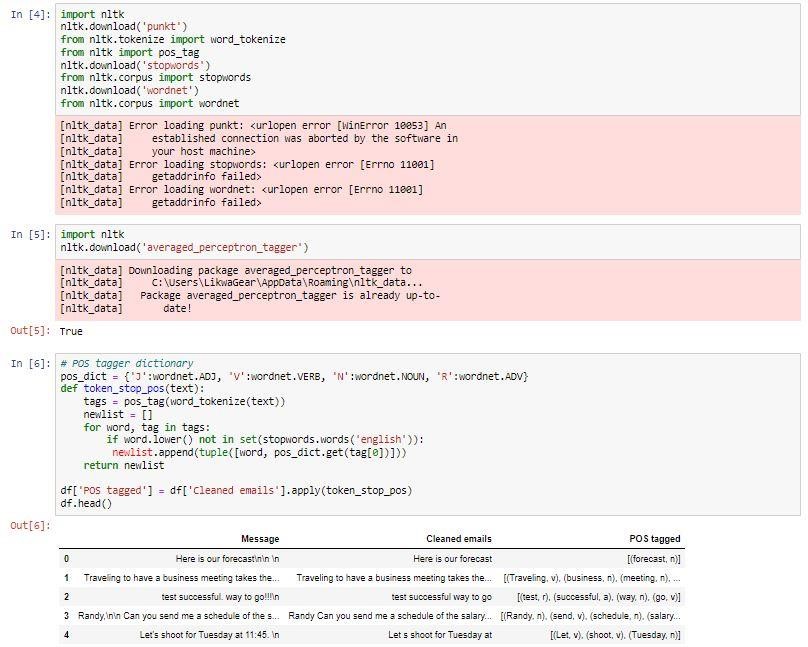


***Figure 3-1: The sentiment analysis method applied to the unlabeled email dataset in this study (source: Author, 2023)***

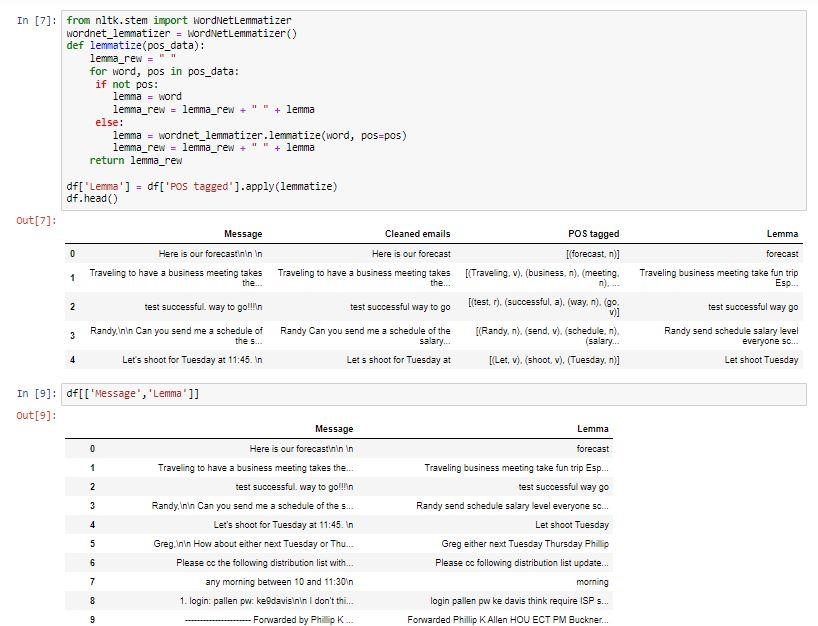
* Data cleaning: In this process, the dataset is cleaned up from unnecessary missing or duplicate data.



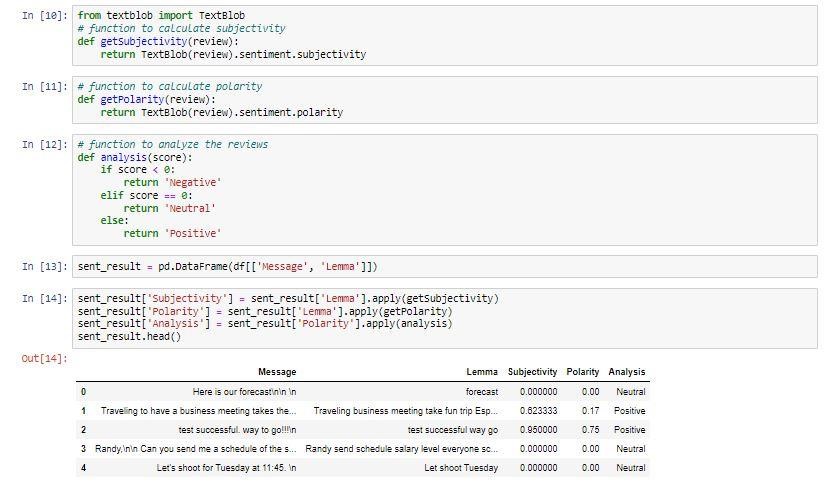
* Data normalization: It involves transforming the data into a common format like lowercase. This is to ensure consistency throughout the dataset.
* Data tokenization: This step involves separating the data into tokens which are individual words or phrases to analyze the data.
* Stop word removal: This step involves removing words that are often used but do not significantly contribute to the analysis are removed.



* Stemming or Lemmatization: This process entails breaking down words into their root forms to increase analysis’s accuracy.
* Part-of-Speech tagging: This stage entails determining the token’s part of speech such as their noun or verb. This increases the analysis’s accuracy.



* Sentiment labeling: This step classifies the data as either positive, negative, or neutral based on the sentiment expressed.



These pretreatment activities are essential to ensure that the data is clean, consistent, and ready for analysis. Moreover, reducing the number of dimensions in the data, stopping words, stemming or lemmatization, and part-of-speech labelling help sentiment analysis be more accurate (Rustam et al. 2023).

## 3.6 Data Analysis

Data manipulation is a step in the analysis process that aims to provide insights, reach judgments, and enhance decision-making (Pradha et al.2019). Various processes, including data preparation, sentiment analysis, and machine learning modeling, might be a part of this.

To make the data ready for analysis, it must be cleaned, transformed, and normalized (Rustam et al. 2023). Stop words, punctuation, special characters, and changing the text's case, among other things, can be eliminated. Understanding the opinion or feeling that is being expressed in a text is known as sentiment classification (Srinivasarao et al.2021). One well-known tool for this is TextBlob, which classifies text as positive, negative, or neutral using a trained model. TextBlob uses the word2vec technique for feature extraction and the sentiment function to classify the sentiment of the text, generating a polarity score that ranges from -1 (negative) to 1 (positive) and a subjectivity score that runs from 0 (objective) to 1 (subjective) (Srinivasarao et al.2021).

The Bag of Words and TF-IDF feature extraction methods were both used in this work along with a machine learning strategy (Singh, 2020). With the training set being used to train the model and the testing set is used to validate it, the data was divided into two sets: training and testing. The input text is represented by the feature variable, but the sentiment or result is represented by the label variable.

The preprocessed and categorized data are evaluated using several machine learning method, including the K-Nearest Neighbors Algorithm (KNN), Logistic Regression, Random Forest, and

Support Vector Machine (SVM).

**Random Forest (RF):** The random forest machine learning algorithm machine builds multiple decision and aggregates their prediction to make a final prediction based on the majority vote among the decision trees. Random Forest was used because it can handle noisy and unstructured data as that used for this study. Also, just like other machine learning models, it requires enough high-quality data to train on, and the performance of the model will depend on the quality of the dataset and the preprocessing step taken (Shah et al.2020).

**K-Nearest Neighbors (KNN):** KNN was used because the data for this study is texted data associated with sentiments. KNN works by computing the similarity between each pair of data points in the training set using a distance metric and then finding the k-nearest neighbors for a new data point and classifying it based on the majority vote of their labels. KNN is effective but can be computationally expensive for large datasets and requires the careful turning of the value k (Shah et al.2020).

**Support Vector Machine (SVM):** SVM was used because it can handle high dimensional data and is robust to noise in the data. SVM works by transforming the text data into higher dimensional feature space and finding the hyperplane defined by support vectors that maximally separate the positive and negative sentiment labels. New data points are then classified based on their position relative to the hyperplane. SVM is effective but sensitive to the choice of kernel function and regularization parameter requiring careful tuning to optimize performance (Boateng et al.2020)

**Logistic Regression:** Thisworks by estimating the probability that a given text data belongs to a particular sentiment class. The input features are mapped to a probability score using a logistic function. The model then predicts the sentiment of the text data based on the class with the highest probability score. Logistic regression was used in this study because it can handle the non-linear relationship between input features and output labels. (Bisong 2019).

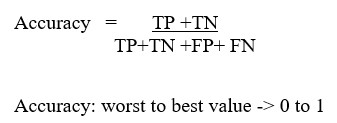
The issue of imbalance classification is investigated using imbalance resampling techniques,

SMOTE+ENN, and SMOTE+Tomek techniques.

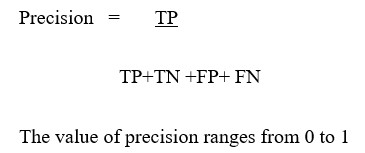
## 3.7 Model Evaluation

Model evaluation is a crucial stage in data analysis where the effectiveness of a machine learning model is evaluated. It aids in determining a model's benefits and drawbacks as well as its applicability to a certain task. Several metrics, including accuracy, precision, recall, and F1-score, can be used to assess a machine learning model. These metrics are employed to evaluate how well the models are performing on a test dataset and to contrast the performance of several models (Elmrabit et al.2020).

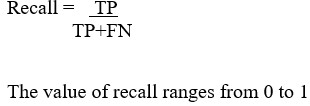
**Accuracy**: The proportion of test dataset cases that were correctly classified. A high accuracy value means that most of the cases in the test dataset can be correctly identified by the model.



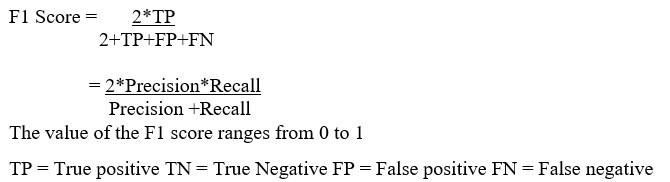
**Precision**: This is the proportion of real positive instances among all positive instances predicted by the model. High accuracy indicates that the model has a low false positive rate.

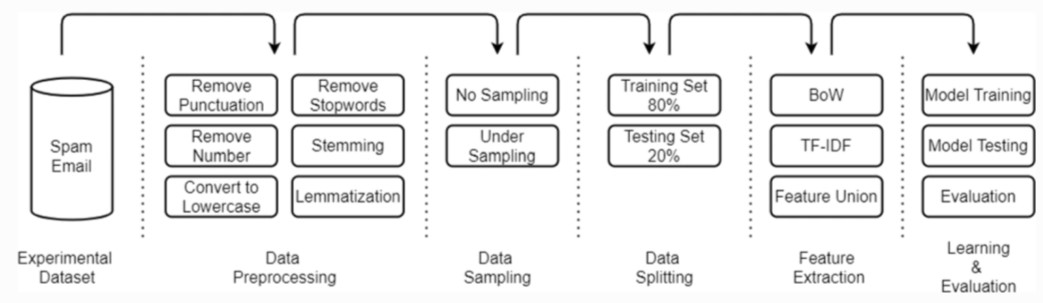


**Recall**: The percentage of positive cases in the test dataset that are correctly identified as genuine positive examples is called recall. A low false negative rate for the model is indicated by a high recall.



**F1-score**: A balance between the two metrics is represented by the harmonic means of recall and precision. The precision-to-recall ratio of a model with a high F1 score is superior.





**Figure 3-2: Workflow methodology used in this study (Rustam et al. 2023)**

**3.8 Limitation of Study**

The study has limitations that should be acknowledged. The data used in the study is limited to emails from Enron Corporation and the Kaggle website, which may not be representative of all types of emails. The study only used basic feature engineering techniques, and only four machine learning models were evaluated, which may not be comprehensive enough to determine the best approach. Additionally, the study only evaluated the using models precision, recall, F1-score, and accuracy, which may not be sufficient for a comprehensive evaluation of the models' performance.

## 3.9 Ethical Consideration

The dataset used for this study is publicly available and was ethically sourced. The University of Bradford’s ethical committee assessed and approved the study before it started to ensure all ethical requirements are followed. All data was kept private and secured against unwanted access. The study was transparent about any possible conflicts of interest. Data accuracy and integrity were guaranteed when gathering, storing, and analyzing data, research findings are accurately and honestly reported, and the study limitations are made explicit.

## 3.10 Conclusion

The importance of data analysis in the discovery of email fraud has been covered in chapter three on fraudulent email detection using sentiment analysis. The many steps involved in data preparation, sentiment analysis, and machine learning modeling have been described.

In the next chapter, the research methodology discussed in this chapter will be applied and the result presented and interpreted the model's evaluation and the study's findings will be discussed in more detail.

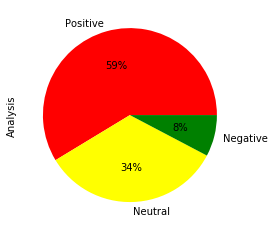
# CHAPTER FOUR

# DATA ANALYSIS, RESULTS AND FINDINGS

## 4.1 Introduction

The data used for this study was collated from two online sources with 2500 emails and 11972 emails respectively, this resulted in a total of 14472 emails. The 11972 emails were extracted from Enron Corporation while the 2500 emails were extracted from the Kaggle website. The extracted data will be cleaned and sentimental analysis carried out on the data using the hybrid-based approach. The hybrid-based approach is the utilization of both the lexicon-based and Machine Learning (ML) classification approach. This approach implements the classification ability of the two approaches to optimizing accuracy. The ML classifiers to adopt for this study are the KNearest Neighbors Algorithm (KNN), Logistic regression, Random Forest, and Support Vector Machine (SVM). The analysis is carried out on both the dataset and the combined dataset and ML classification is carried out using a balanced class by applying techniques to handle the imbalanced class in the classification.

## 4.2 Analysis using the Enron Corporation Dataset



**Figure 4-1: Illustration of result on sentiments of the emails on Enron datasets in a pie chart**

The sentiment analysis result derived from Enron data set using the lexicon-based approach is displayed. The figure shows that 8% of the 11972 emails are classified as negative or fraudulent emails, 34% of the total emails are classified as neutral emails and 59% of the total emails are classified as positive emails.

**Table 4-1: Predictive and Classification performances of the selected machine learning models using the training set of Enron**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| **Random**  **Forest** | Precision | 1 | 1 | 1 | 1 |
| Recall | 0.99 | 1 | 1 | 1 |
| F1-score | 1 | 1 | 1 | 1 |
| Accuracy |  |  |  | 0.9980 |
| KNN | Precision | 0.77 | 0.52 | 0.94 | 0.74 |
| Recall | 0.18 | 0.99 | 0.54 | 0.57 |
| F1-score | 0.29 | 0.68 | 0.69 | 0.55 |
| Accuracy |  |  |  | 0.6660 |
| SVM | Precision | 1 | 0.97 | 0.96 | 0.98 |
| Recall | 0.63 | 0.99 | 0.99 | 0.87 |
| F1-score | 0.77 | 0.98 | 0.98 | 0.91 |
| Accuracy |  |  |  | 0.9657 |
| Logistic | Precision | 0.97 | 0.90 | 0.91 | 0.93 |
| Recall | 0.23 | 0.97 | 0.96 | 0.72 |
| F1-score | 0.37 | 0.93 | 0.94 | 0.75 |
| Accuracy |  |  |  | 0.9109 |

Table 4.1 suggests that the random forest model outperformed the other selected models since its precision (100%), recall (99%), F1-score (100%), and accuracy value of 99.80% are the highest compared to the other selected models. The model accuracy of the random forest suggested that 99.80% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 99.80% accurate. The recall of the model is 100%. This means the model properly detects 100% of all positive emails. Also, the model suggests a precision score of 100%, this implies that there is 100% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

**Table 4-2: Predictive and Classification performances of the selected machine learning models using the testing set of Enron.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 0.92 | 0.77 | 0.88 | 0.86 |
| Recall | 0.17 | 0.94 | 0.89 | 0.66 |
| F1-score | 0.28 | 0.85 | 0.89 | 0.67 |
| Accuracy |  |  |  | 0.8398 |
| KNN | Precision | 0.78 | 0.50 | 0.93 | 0.73 |
| Recall | 0.10 | 0.99 | 0.52 | 0.54 |
| F1-score | 0.18 | 0.66 | 0.66 | 0.50 |
| Accuracy |  |  |  | 0.6382 |
| **SVM** | Precision | 1.00 | 0.80 | 0.89 | 0.90 |
| Recall | 0.15 | 0.94 | 0.91 | 0.67 |
| F1-score | 0.26 | 0.87 | 0.90 | 0.86 |
| Accuracy |  |  |  | 0.8557 |
| Logistic | Precision | 0.88 | 0.81 | 0.88 | 0.86 |
| Recall | 0.11 | 0.93 | 0.92 | 0.65 |
| F1-score | 0.19 | 0.86 | 0.90 | 0.65 |
| Accuracy |  |  |  | 0.8511 |

Table 4.2 suggests that the support vector machine outperformed the other selected models since its precision (90%), recall (67%), F1-score (86%), and accuracy value of 85.57% are the highest compared to the other selected models. The model accuracy of the SVM suggested that 85.57% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 85.57% accurate. The recall of the SVM model is 67%. This means the model properly detects 67% of all positive emails. Also, the model suggests a precision score of 90%, this implies that there is 90% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

### 4.2.1 Email Classification using Balanced Dataset of the Enron Corporation

**Table 4-3: Predictive and Classification performances of the selected machine learning models using SMOTETOMEK Techniques of the Enron Corporation Emails**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 0.97 | 0.72 | 0.74 | 0.81 |
| Recall | 1 | 1 | 1 | **1** | 0.50 | 0.97 | 0.84 | 0.77 |
| F1-score | 1 | 1 | 1 | **1** | 0.66 | 0.83 | 0.79 | 0.76 |
| Accuracy |  |  |  | **0.9985** |  |  |  | 0.7707 |
| KNN | Precision | 0.64 | 0.83 | 1.00 | 0.82 | 0.55 | 0.71 | 0.98 | 0.75 |
| Recall | 0.99 | 0.93 | 0.34 | 0.75 | 0.83 | 0.87 | 0.27 | 0.66 |
| F1-score | 0.78 | 0.88 | 0.51 | 0.72 | 0.67 | 0.79 | 0.42 | 0.62 |
| Accuracy |  |  |  | 0.7549 |  |  |  | 0.6584 |
| SVM | Precision | 0.99 | 0.96 | 1 | 0.98 | 0.99 | 0.78 | 0.81 | **0.86** |
| Recall | 0.97 | 1 | 0.99 | 0.98 | 0.65 | 0.96 | 0.90 | **0.84** |
| F1-score | 0.98 | 0.98 | 0.99 | 0.98 | 0.78 | 0.86 | 0.85 | **0.83** |
| Accuracy |  |  |  | 0.9835 |  |  |  | **0.8373** |
| Logistic | Precision | 0.96 | 0.90 | 0.98 | 0.95 | 0.91 | 0.81 | 0.79 | 0.83 |
| Recall | 0.94 | 0.98 | 0.92 | 0.95 | 0.66 | 0.96 | 0.86 | 0.83 |
| F1-score | 0.95 | 0.94 | 0.95 | 0.95 | 0.76 | 0.88 | 0.82 | 0.82 |
| Accuracy |  |  |  | 0.9459 |  |  |  | 0.8255 |

Table 4.3 suggested that random forest outperformed the other selected models with the training set while SVM performed best with the training set while using the SMOTETOMEK techniques.

**Table 4-4: Predictive and Classification performances of the selected machine learning models for the training set using SMOTEENN Techniques of the Enron Corporation Emails**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 0.97 | 0.80 | 0.87 | 0.88 |
| Recall | 1 | 1 | 1 | **1** | 0.74 | 0.99 | 0.89 | 0.88 |
| F1-score | 1 | 1 | 1 | **1** | 0.84 | 0.89 | 0.88 | 0.87 |
| Accuracy |  |  |  | **1.000** |  |  |  | 0.8685 |
| KNN | Precision | 0.93 | 0.99 | 1.00 | 0.97 | 0.76 | 0.86 | 0.97 | 0.86 |
| Recall | 1.00 | 0.93 | 0.94 | 0.96 | 0.88 | 0.76 | 0.77 | 0.80 |
| F1-score | 0.96 | 0.96 | 0.97 | 0.96 | 0.81 | 0.80 | 0.86 | 0.83 |
| Accuracy |  |  |  | 0.9626 |  |  |  | 0.8130 |
| SVM | Precision | 1 | 0.98 | 1 | 0.99 | 0.97 | 0.85 | 0.77 | 0.86 |
| Recall | 0.99 | 1 | 1 | 0.99 | 0.78 | 0.98 | 0.93 | 0.90 |
| F1-score | 0.99 | 0.99 | 1 | 0.99 | 0.86 | 0.91 | 0.84 | 0.87 |
| Accuracy |  |  |  | 0.9929 |  |  |  | 0.8829 |
| Logistic | Precision | 0.98 | 0.96 | 1 | 0.98 | 0.93 | 0.89 | 0.87 | **0.90** |
| Recall | 0.97 | 0.99 | 0.94 | 0.97 | 0.86 | 0.95 | 0.87 | **0.89** |
| F1-score | 0.98 | 0.98 | 0.97 | 0.97 | 0.89 | 0.92 | 0.87 | **0.89** |
| Accuracy |  |  |  | 0.9756 |  |  |  | **0.9032** |

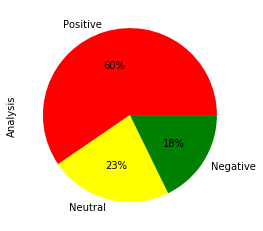
Table 4.4 suggested that random forest outperformed the other selected models with the training set and the Logistic regression model performed best among the selected classifiers for the testing set using the SMOTE+ENN techniques.

**Table 4-5: Comparison of the predictive performance of the selected machine learning model across the sampling techniques of the Enron Corporation Emails**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Training** |  |  | **Testing** |  |
| **Models** | **Criteria** | **Original data** | **Smote+Tomek** | **Smote+ENN** | **Original data** | **Smote+Tomek** | **Smote+ENN** |
| **Random**  **Forest** | **Precision** | 1 | 1 | **1** | 0.86 | 0.81 | **0.88** |
| **Recall** | 1 | 1 | **1** | 0.66 | 0.77 | **0.88** |
| **F1-score** | 1 | 1 | **1** | 0.67 | 0.76 | **0.87** |
| **Accuracy** | 0.9980 | 0.9985 | **1.000** | 0.8398 | 0.7707 | **0.8685** |
| **KNN** | **Precision** | 0.74 | 0.82 | **0.97** | 0.73 | 0.75 | **0.86** |
| **Recall** | 0.57 | 0.75 | **0.96** | 0.54 | 0.66 | **0.80** |
| **F1-score** | 0.55 | 0.72 | **0.96** | 0.50 | 0.62 | **0.83** |
| **Accuracy** | 0.6660 | 0.7549 | **0.9626** | 0.6382 | 0.6584 | **0.8130** |
| **SVM** | **Precision** | 0.98 | 0.98 | **0.99** | 0.90 | 0.86 | **0.86** |
| **Recall** | 0.87 | 0.98 | **0.99** | 0.67 | 0.84 | **0.90** |
| **F1-score** | 0.91 | 0.98 | **0.99** | 0.86 | 0.83 | **0.87** |
| **Accuracy** | 0.9657 | 0.9835 | **0.9929** | 0.8557 | 0.8373 | **0.8829** |
| **LOGISTIC** | **Precision** | 0.93 | 0.95 | **0.98** | 0.86 | 0.83 | **0.90** |
| **Recall** | 0.72 | 0.95 | **0.97** | 0.65 | 0.83 | **0.89** |
| **F1-score** | 0.75 | 0.95 | **0.97** | 0.65 | 0.82 | **0.89** |
| **Accuracy** | 0.9109 | 0.9459 | **0.9756** | 0.8511 | 0.8255 | **0.9032** |

Table 4.5 shows the comparison of the classification performance of the classifiers across different sampling techniques. The table suggests that the selected classifiers (RF, KNN, SVM, Logistic) performed better when using the Smote+ENN techniques compared to the Smote+Tomek and original dataset since the precision, recall, F1-score, and accuracy values are the highest for the training and testing set.

## 4.3 Analysis using the Kaggle Dataset



**Figure 4-2: Illustration of result on sentiments of the emails on Kaggle datasets in a pie chart**

Figure 4.2 displays the result derived from sentiment analysis using the lexicon-based approach, the figure shows that 18% of the 2500 emails are classified as negative or fraudulent emails, 23% of the total emails are classified as neutral emails and 60% of the total emails are classified as positive emails.

**Table 4-6: Predictive and Classification performances of the selected machine learning models using the training set of the Kaggle Emails**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| **Random**  **Forest** | Precision | 1 | 1 | 1 | 1 |
| Recall | 1 | 1 | 1 | 1 |
| F1-score | 1 | 1 | 1 | 1 |
| Accuracy |  |  |  | 1.000 |
| KNN | Precision | 0.95 | 0.36 | 0.97 | 0.76 |
| Recall | 0.60 | 0.98 | 0.45 | 0.68 |
| F1-score | 0.74 | 0.53 | 0.61 | 0.63 |
| Accuracy |  |  |  | 0.4979 |
| SVM | Precision | 1 | 1 | 0.99 | 1 |
| Recall | 0.98 | 1 | 1 | 0.99 |
| F1-score | 0.99 | 1 | 1 | 0.99 |
| Accuracy |  |  |  | 0.9953 |
| Logistic | Precision | 1.00 | 0.97 | 0.94 | 0.97 |
| Recall | 0.84 | 0.94 | 0.99 | 0.92 |
| F1-score | 0.91 | 0.96 | 0.96 | 0.94 |
| Accuracy |  |  |  | 0.9539 |

Table 4.6 suggests that the random forest model outperformed the other selected models since its precision (100%), recall (100%), F1-score (100%), and accuracy value of 100% are the highest compared to the other selected models. The model accuracy of the random forest suggested that 99.83% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 99.83% accurate. The recall of the model is 100%. This means the model properly detects 100% of all positive emails. Also, the model suggests a precision score of 100%, this implies that there is 100% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

**Table 4-7: Predictive and Classification performances of the selected models using the testing set of the Kaggle Emails**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1.00 | 0.98 | 0.96 | **0.98** |
| Recall | 0.93 | 0.95 | 0.99 | **0.96** |
| F1-score | 0.96 | 0.97 | 0.98 | **0.97** |
| Accuracy |  |  |  | **0.9728** |
| KNN | Precision | 0.97 | 0.31 | 0.96 | 0.75 |
| Recall | 0.43 | 0.99 | 0.33 | 0.59 |
| F1-score | 0.60 | 0.47 | 0.49 | 0.52 |
| Accuracy |  |  |  | 0.4979 |
| **SVM** | Precision | 1.00 | 0.98 | 0.96 | 0.98 |
| Recall | 0.90 | 0.95 | 0.99 | 0.95 |
| F1-score | 0.95 | 0.97 | 0.97 | 0.96 |
| Accuracy |  |  |  | 0.9686 |
| Logistic | Precision | 1.00 | 0.94 | 0.88 | 0.94 |
| Recall | 0.72 | 0.84 | 0.98 | 0.85 |
| F1-score | 0.84 | 0.89 | 0.92 | 0.88 |
| Accuracy |  |  |  | 0.9038 |

Table 4.7 suggests that the RF outperformed the other selected models since its precision (98%), recall (96%), F1-score (97%), and accuracy value of 97.28% are the highest compared to the other selected models. The model accuracy of the RF suggested that 97.28% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 97.28% accurate. The recall of the RF model is 96%. This means the model properly detects 96% of all positive emails. Also, the model suggests a precision score of 98%, this implies that there is 98% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

### 4.3.1 Email Classification using Balanced Dataset of the Kaggle Website

**Table 4-8: Predictive and Classification performances of the selected machine learning models using SMOTETOMEK**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 1.00 | 0.99 | 0.90 | 0.96 |
| Recall | 1 | 1 | 1 | **1** | 0.93 | 0.96 | 0.99 | 0.96 |
| F1-score | 1 | 1 | 1 | **1** | 0.96 | 0.97 | 0.94 | 0.96 |
| Accuracy |  |  |  | **1** |  |  |  | 0.9594 |
| KNN | Precision | 0.96 | 0.64 | 1.00 | 0.87 | 0.94 | 0.58 | 1.00 | 0.84 |
| Recall | 0.92 | 1.00 | 0.28 | 0.73 | 0.92 | 1.00 | 0.28 | 0.73 |
| F1-score | 0.97 | 0.78 | 0.59 | 0.78 | 0.93 | 0.73 | 0.44 | 0.70 |
| Accuracy |  |  |  | 0.8010 |  |  |  | 0.7343 |
| SVM | Precision | 1 | 1 | 1 | 1 | 1.00 | 0.99 | 0.93 | **0.97** |
| Recall | 1 | 1 | 1 | 1 | 0.99 | 0.97 | 0.98 | **0.97** |
| F1-score | 1 | 1 | 1 | 0.99 | 0.98 | 0.98 | 0.96 | **0.97** |
| Accuracy |  |  |  | 0.9991 |  |  |  | **0.9722** |
| Logistic | Precision | 0.99 | 0.98 | 0.98 | 0.98 | 0.99 | 0.94 | 0.88 | 0.94 |
| Recall | 0.98 | 1.00 | 0.97 | 0.98 | 0.92 | 0.93 | 0.95 | 0.94 |
| F1-score | 0.99 | 0.99 | 0.98 | 0.98 | 0.95 | 0.94 | 0.91 | 0.94 |
| Accuracy |  |  |  | 0.9844 |  |  |  | 0.9350 |

Table 4.8 suggested that RF outperformed the other selected models with the training set and the

SVM classifier performed best among the selected classifiers for the testing set using the SMOTE+TOMEK techniques.

**Table 4-9: Predictive and Classification performances of the selected machine learning models for the training set using SMOTEENN**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 0.98 | 0.96 | 0.88 | 0.94 |
| Recall | 1 | 1 | 1 | **1** | 0.96 | 1.00 | 0.98 | 0.92 |
| F1-score | 1 | 1 | 1 | **1** | 0.88 | 0.74 | 0.80 | 0.92 |
| Accuracy |  |  |  | **1.000** |  |  |  | 0.9648 |
| KNN | Precision | 0.99 | 0.99 | 1.00 | 1.00 | 0.98 | 0.91 | 1.00 | 0.97 |
| Recall | 1.00 | 1.00 | 0.97 | 0.99 | 0.93 | 1.00 | 0.68 | 0.87 |
| F1-score | 0.99 | 1.00 | 0.99 | 0.99 | 0.96 | 0.95 | 0.81 | 0.91 |
| Accuracy |  |  |  | 0.9941 |  |  |  | 0.9479 |
| SVM | Precision | 1 | 1 | 1 | 1 | 1.00 | 0.97 | 0.84 | **0.94** |
| Recall | 1 | 1 | 1 | 1 | 0.94 | 1.00 | 0.97 | **0.97** |
| F1-score | 1 | 1 | 1 | 1 | 0.97 | 0.98 | 0.90 | **0.95** |
| Accuracy |  |  |  | 1.00 |  |  |  | **0.9698** |
| Logistic | Precision | 0.99 | 1 | 1 | 1.00 | 0.97 | 0.96 | 0.85 | 0.93 |
| Recall | 1.00 | 1.00 | 0.98 | 0.99 | 0.94 | 1.00 | 0.76 | 0.90 |
| F1-score | 0.99 | 1.00 | 0.99 | 0.99 | 0.95 | 0.98 | 0.81 | 0.91 |
| Accuracy |  |  |  | 0.9948 |  |  |  | 0.9581 |

Table 4.9 suggested that random forest outperformed the other selected models with the training set and the SVM classifier performed best among the selected classifiers for the testing set using the SMOTE+ENN techniques.

**Table 4-10: Comparison of the predictive performance of the selected machine learning model across the sampling techniques**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | Training |  |  | Testing |  |
| Models | Criteria | Original data | Smote+Tomek | Smote+ENN | Original data | Smote+Tomek | Smote+ENN |
| Random  Forest | Precision | **1** | **1** | **1** | 0.98 | 0.96 | 0.94 |
| Recall | **1** | **1** | **1** | 0.96 | 0.96 | 0.92 |
| F1-score | **1** | **1** | **1** | 0.97 | 0.96 | 0.92 |
| Accuracy | **1** | **1** | **1** | 0.9728 | 0.9594 | 0.9648 |
| KNN | Precision | 0.76 | 0.87 | **1** | 0.75 | 0.84 | **0.97** |
| Recall | 0.63 | 0.73 | **0.99** | 0.59 | 0.73 | **0.87** |
| F1-score | 0.62 | 0.78 | **0.99** | 0.52 | 0.70 | **0.91** |
| Accuracy | 0.6709 | 0.8010 | **0.9941** | 0.4979 | 0.7343 | **0.9479** |
| SVM | Precision | 0.98 | 1 | **1** | 0.98 | **0.97** | 0.93 |
| Recall | 0.92 | 1 | **1** | 0.95 | **0.97** | 0.97 |
| F1-score | 0.95 | 0.99 | **1** | 0.96 | **0.97** | 0.95 |
| Accuracy | 0.9731 | 0.9991 | **1.00** | 0.9686 | **0.9722** | 0.9698 |
| LOGISTIC | Precision | 0.94 | 0.98 | **1.00** | 0.94 | 0.94 | **0.93** |
| Recall | 0.79 | 0.98 | **1.00** | 0.85 | 0.94 | **0.90** |
| F1-score | 0.82 | 0.98 | **0.99** | 0.88 | 0.94 | **0.91** |
| Accuracy | 0.9163 | 0.9844 | **0.9948** | 0.9038 | 0.9350 | **0.9581** |

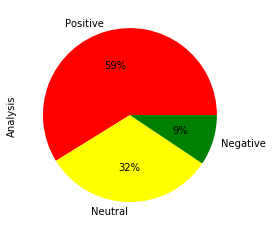
Table 4.10 shows the comparison of the classification performance of the classifiers across different sampling techniques using the Kaggle dataset. The table suggests that the RF classifier performed optimally with the original data and using SmoteTomek and SmoteENN sampling techniques using the training set with 100% accuracy while the RF model performed better using the original data with 97.28% accuracy compared to smoke-tomek and smote-enn sampling

techniques using the testing set.

Also, the table suggested that the KNN and logistic classifier performed optimally when using the Smote+ENN techniques compared to the Smote+Tomek and original dataset since the precision, recall, F1-score, and accuracy values are the highest for the training and testing set respectively.

The SVM classifier performed optimally with the Smote+ENN techniques compared to the Smote+Tomek and original dataset using the training set with 100% accuracy but performed better using the Smote+Tomek with 97.22% accuracy.

## 4.4 Analysis using the Combined Dataset



**Figure 4-3: Illustration of result on sentiments of the emails on combined datasets in a pie chart**

Figure 4.3 displays the result derived from sentiment analysis on the combined datasets using the lexicon-based approach, the figure shows that 9% of the 14472 emails are classified as negative or fraudulent emails, 32% of the total emails are classified as neutral emails and 59% of the total emails are classified as positive emails.

**Table 4-11: Predictive and Classification performances of the selected machine learning models using the training set for the combined emails.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| **Random**  **Forest** | Precision | 1 | 1 | 1 | 1 |
| Recall | 1 | 1 | 1 | 1 |
| F1-score | 1 | 1 | 1 | 1 |
| Accuracy |  |  |  | 0.9983 |
| KNN | Precision | 0.82 | 0.51 | 0.94 | 0.76 |
| Recall | 0.34 | 0.99 | 0.55 | 0.63 |
| F1-score | 0.48 | 0.67 | 0.70 | 0.62 |
| Accuracy |  |  |  | 0.6709 |
| SVM | Precision | 1 | 0.98 | 0.97 | 0.98 |
| Recall | 0.78 | 0.99 | 0.99 | 0.92 |
| F1-score | 0.88 | 0.98 | 0.98 | 0.95 |
| Accuracy |  |  |  | 0.9731 |
| Logistic | Precision | 0.98 | 0.90 | 0.92 | 0.94 |
| Recall | 0.43 | 0.97 | 0.97 | 0.79 |
| F1-score | 0.60 | 0.93 | 0.94 | 0.82 |
| Accuracy |  |  |  | 0.9163 |

Table 4.11 suggests that the random forest model outperformed the other selected models since its precision (99%), recall (100%), F1-score (100%), and accuracy value of 99.83% are the highest compared to the other selected models. The model accuracy of the random forest suggested that 99.83% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 99.83% accurate. The recall of the model is 100%. This means the model properly detects 100% of all positive emails. Also, the model suggests a precision score of 100%, this implies that there is 100% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

**Table 4-12: Predictive and Classification performances of the selected machine learning models using the testing set for the combined emails.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Criteria | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 0.99 | 0.81 | 0.90 | 0.90 |
| Recall | 0.40 | 0.94 | 0.90 | 0.75 |
| F1-score | 0.57 | 0.87 | 0.90 | 0.78 |
| Accuracy |  |  |  | 0.8672 |
| KNN | Precision | 0.76 | 0.48 | 0.92 | 0.72 |
| Recall | 0.21 | 0.99 | 0.48 | 0.56 |
| F1-score | 0.33 | 0.64 | 0.63 | 0.54 |
| Accuracy |  |  |  | 0.6190 |
| **SVM** | Precision | 0.99 | 0.84 | 0.90 | 0.91 |
| Recall | 0.39 | 0.95 | 0.93 | 0.76 |
| F1-score | 0.56 | 0.89 | 0.92 | 0.79 |
| Accuracy |  |  |  | 0.8829 |
| Logistic | Precision | 0.94 | 0.84 | 0.88 | 0.89 |
| Recall | 0.29 | 0.93 | 0.93 | 0.71 |
| F1-score | 0.44 | 0.88 | 0.91 | 0.74 |
| Accuracy |  |  |  | 0.8669 |

Table 4.12 shows the predictive performance of the selected machine learning models on the testing set. The result shows that the support vector machine outperformed the other selected machine learning models since its precision (91%), recall (76%), F1-score (79%), and accuracy value of 88.29% are the highest compared to the other selected models. The model accuracy of the support vector machine suggested that 88.29% of the emails are classified accurately. When the model predicts the occurrence of negative or fraudulent emails, it is 88.29% accurate. The recall of the model is 76%. This means the model properly detects 76% of all positive emails. Also, the model suggests a precision score of 91%, this implies that there is 91% certainty that the model will efficiently predict positive emails without predicting any false positive emails.

### 4.4.1 Email Classification using Balanced Dataset of the Combined Emails

**Table 4-13: Predictive and Classification performances of the selected models using the SMOTETOMEK technique for the combined emails.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 0.99 | 0.78 | 0.75 | 0.84 |
| Recall | 1 | 1 | 1 | **1** | 0.61 | 0.98 | 0.86 | 0.81 |
| F1-score | 1 | 1 | 1 | **1** | 0.75 | 0.87 | 0.80 | 0.81 |
| Accuracy |  |  |  | **0.9989** |  |  |  | 0.8121 |
| KNN | Precision | 0.83 | 0.70 | 1.00 | 0.84 | 0.74 | 0.55 | 0.94 | 0.74 |
| Recall | 0.99 | 0.99 | 0.38 | 0.79 | 0.65 | 0.99 | 0.30 | 0.64 |
| F1-score | 0.91 | 0.82 | 0.55 | 0.76 | 0.69 | 0.70 | 0.45 | 0.62 |
| Accuracy |  |  |  | 0.7883 |  |  |  | 0.6448 |
| SVM | Precision | 1 | 0.99 | 1.00 | 0.99 | 0.99 | 0.81 | 0.83 | **0.88** |
| Recall | 1 | 1.00 | 0.99 | 0.99 | 0.69 | 0.97 | 0.92 | **0.86** |
| F1-score | 1 | 1.00 | 0.99 | 0.99 | 0.81 | 0.88 | 0.87 | **0.86** |
| Accuracy |  |  |  | 0.9961 |  |  |  | **0.8589** |
| Logistic | Precision | 0.97 | 0.95 | 0.98 | 0.97 | 0.93 | 0.80 | 0.78 | 0.84 |
| Recall | 0.99 | 0.98 | 0.93 | 0.96 | 0.64 | 0.96 | 0.87 | 0.82 |
| F1-score | 0.98 | 0.96 | 0.95 | 0.96 | 0.76 | 0.87 | 0.82 | 0.82 |
| Accuracy |  |  |  | 0.9649 |  |  |  | 0.8221 |

Table 4.13 suggested that random forest outperformed the other selected models with the training set while SVM performed best with the training set while using the SMOTETOMEK techniques.

**Table 4-14: Predictive and Classification performances of the selected machine learning models for the training set using the SMOTEENN technique for the combined emails.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | Criteria |  | Training | |  |  | Testing | |  |
| Negative | Neutral | Positive | Overall | Negative | Neutral | Positive | Overall |
| Random  Forest | Precision | 1 | 1 | 1 | **1** | 1.00 | 0.83 | 0.84 | **0.89** |
| Recall | 1 | 1 | 1 | **1** | 0.83 | 1.00 | 0.93 | **0.92** |
| F1-score | 1 | 1 | 1 | **1** | 0.90 | 0.91 | 0.88 | **0.90** |
| Accuracy |  |  |  | **1.000** |  |  |  | **0.9034** |
| KNN | Precision | 0.98 | 1.00 | 1.00 | 0.99 | 0.97 | 0.70 | 0.94 | 0.87 |
| Recall | 1.00 | 1.00 | 0.95 | 0.98 | 0.68 | 0.99 | 0.81 | 0.83 |
| F1-score | 0.99 | 1.00 | 0.97 | 0.99 | 0.80 | 0.82 | 0.87 | 0.83 |
| Accuracy |  |  |  | 0.9910 |  |  |  | 0.8165 |
| SVM | Precision | 1 | 0.99 | 1 | 0.99 | 1.00 | 0.84 | 0.68 | 0.84 |
| Recall | 1 | 1 | 1 | 1 | 0.79 | 1.00 | 0.95 | 0.91 |
| F1-score | 1 | 1 | 1 | 1 | 0.88 | 0.91 | 0.79 | 0.86 |
| Accuracy |  |  |  | 0.9997 |  |  |  | 0.8862 |
| Logistic | Precision | 0.97 | 0.99 | 1 | 0.99 | 0.98 | 0.85 | 0.69 | 0.84 |
| Recall | 0.99 | 0.99 | 0.93 | 0.97 | 0.82 | 0.99 | 0.90 | 0.90 |
| F1-score | 0.98 | 0.99 | 0.96 | 0.98 | 0.89 | 0.92 | 0.86 | 0.86 |
| Accuracy |  |  |  | 0.9841 |  |  |  | 0.8948 |

Table 4.14 suggested that the random forest outperformed the other selected models with the training set and testing set using the SMOTE+ENN techniques.

**Table 4-15: Comparison of the predictive performance of the selected machine learning model across the sampling techniques using the training set**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Criteria | Original data | Smote+Tomek | Smote+ENN |
| Random Forest | Precision | 1 | 1 | 1 |
| Recall | 1 | 1 | 1 |
| F1-score | 1 | 1 | 1 |
| Accuracy | 0.9983 | 0.9989 | 1 |
| KNN | Precision | 0.76 | 0.84 | 0.99 |
| Recall | 0.63 | 0.79 | 0.98 |
| F1-score | 0.62 | 0.76 | 0.99 |
| Accuracy | 0.6709 | 0.7883 | 0.991 |
| SVC | Precision | 0.98 | 0.99 | 0.99 |
| Recall | 0.92 | 0.99 | 1 |
| F1-score | 0.95 | 0.99 | 1 |
| Accuracy | 0.9731 | 0.9961 | 0.9997 |
| LOGISTIC | Precision | 0.94 | 0.97 | 0.99 |
| Recall | 0.79 | 0.96 | 0.97 |
| F1-score | 0.82 | 0.96 | 0.98 |
| Accuracy | 0.9163 | 0.9649 | 0.9841 |

Table 4.15 suggests that there is an improvement in the classification and predictive performances across the sampling techniques. The SMOTE+ENN sampling techniques produce the highest predictive and classification accuracy for the selected models.

**Table 4-16: Comparison of the predictive performance of the selected machine learning model across the sampling techniques using the testing set**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Criteria | Original data | Smote+Tomek | Smote+ENN |
| Random Forest | Precision | 0.9 | 0.84 | 0.89 |
| Recall | 0.75 | 0.81 | 0.92 |
| F1-score | 0.78 | 0.81 | 0.9 |
| Accuracy | 0.8672 | 0.8121 | 0.9034 |
| KNN | Precision | 0.72 | 0.74 | 0.87 |
| Recall | 0.56 | 0.64 | 0.83 |
| F1-score | 0.54 | 0.62 | 0.83 |
| Accuracy | 0.619 | 0.6448 | 0.8165 |
| SVC | Precision | 0.91 | 0.88 | 0.84 |
| Recall | 0.76 | 0.86 | 0.91 |
| F1-score | 0.79 | 0.86 | 0.86 |
| Accuracy | 0.8829 | 0.8589 | 0.8862 |
| LOGISTIC | Precision | 0.89 | 0.84 | 0.84 |
| Recall | 0.71 | 0.82 | 0.9 |
| F1-score | 0.74 | 0.82 | 0.86 |
| Accuracy | 0.8669 | 0.8221 | 0.8948 |

Table 4.16 suggests that there is an improvement in the classification and predictive performances across the sampling techniques. The SMOTE+ENN sampling techniques produce the highest predictive and classification accuracy for the selected models.

## 4.5 Summary of the Major Findings

Sentiment analysis was conducted on a dataset of emails using a lexicon-based approach. The results suggest that the majority of the emails were classified as positive, with a small percentage being negative or fraudulent. Different models were evaluated for their classification performance, and it was found that the SVM model outperformed other models with the Enron dataset, while the Random Forest (RF) model outperformed other models with the Kaggle dataset. When using the combined dataset, the SVM model also outperformed other models. The SMOTE+ENN technique produced the highest accuracy for classification and prediction among the selected models, with RF having the highest predictive performance by improving the precision (0.90 to 0.89), recall (0.75 to 0.92), f1-score (0.78 to 0.90), and accuracy (0.8672 to 0.9034) metrics.

Overall, the findings suggest that both the SVM model and the SMOTE+ENN technique are effective in classifying emails based on sentiment.

In the next chapter, the model's evaluation and the study's findings will be discussed in more detail. This will provide a discussion of how the various machine learning models employed in the study performed as well as an analysis of the outcomes.

# CHAPTER FIVE

# SUMMARY, CONCLUSIONS, DISCUSSIONS, and RECOMMENDATIONS

## 5.1 Introduction

Sentiment analysis has emerged as a vital tool for communication in various industries, including finance and oil and gas. It allows organizations to analyze feedback and gain a better understanding of the emotions, thoughts, opinions, ideas, and motives expressed in emails. With the increasing use of machine learning algorithms in this context, this chapter focuses on evaluating and comparing the performance of RF, SVM, KNN, and Logistic regression algorithms for sentiment analysis of emails in the banking sector. The main objective of the study was to develop a model that could accurately classify customer emails into three categories: positive, negative, or neutral. Through a detailed discussion of the study's summary, findings, conclusions, and recommendations, this chapter provides valuable insights into the performance of various machine learning algorithms in the task of sentiment analysis.

## 5.2 Summary

This study was aimed at developing a model that could effectively classify emails into positive, negative, or neutral categories. The total number of emails considered in this study was 14472 emails which were extracted from two different sources: Enron Corporation and the Kaggle website. The emails were preprocessed using several techniques, including removing stop words, stemming, tokenization, and converting to lowercase, to make the text data more manageable for analysis. The sentiment analysis was performed using both lexicon-based and ML approaches.

The lexicon-based approach classified the emails into negative sentiments (8%, 18%, and 9%), neutral sentiments (34%, 23%, and 32%), and positive sentiments (59%, 60%, and 59%) using the Enron, Kaggle and combined dataset respectively which suggested an imbalance classification.

The study then compared the performance of Random Forest, Support Vector Machine (SVM), KNearest Neighbors (KNN) and Logistic Regression machine learning models. The models were evaluated using various metrics, including precision, recall, F1-score, and accuracy.

The results showed that the Support Vector Machine model had the best performance overall, achieving high, F1-score, recall, precision and accuracy on the overall datasets. The performance of the models using different sampling techniques, including SMOTE, Tomek links, and ENN were compared. The results showed that the SMOTEENN technique generally produced the best results, with all models achieving higher precision, recall, F1-score, and accuracy.

In conclusion, the study found that the Random Forest model, in combination with the SMOTEENN technique, was the most effective model for sentiment analysis of emails.

## 5.3 Discussions

The primary goal of the study was to categorize emails into positive, negative, and neutral groups.

The results of the lexicon-based approach indicate that only a small fraction of emails in both the Enron and Kaggle datasets were classified as negative or fraudulent. This implies that text analysis alone may not be sufficient for detecting fraudulent activities since they may not be mentioned explicitly in emails. The significant proportion of emails identified as positive or neutral suggests that most emails are authentic and do not contain any fraudulent activities. However, it is important to keep in mind that the lexicon-based approach has limitations and may not accurately capture the complexities of language and context in emails. Thus, incorporating machine learning algorithms with this approach may improve the accuracy of detecting fraudulent emails. Nonetheless, the lexicon-based approach is a crucial starting point for identifying potential fraudulent emails and can be used in combination with other techniques.

The study aimed to identify the most effective machine learning algorithm for classifying fraudulent email sentiment outcomes across three datasets. The results indicated that the SVM algorithm outperformed other models in the Enron dataset, with the highest precision (90%), recall (67%), F1-score (86%), and accuracy values (85.57%). Similarly, the RF algorithm performed best in the Kaggle dataset, with the highest precision (98%), recall (96%), f1 score (97%), and accuracy values (97.28%). SVM performed best with the highest precision (88%), recall (86%), F1-score (82%), and accuracy value (85.89%) while using the combined dataset. The outcome of this study was in line with the study by Hag Ali and El Gayar (2019), Helmini (2019), and Ruz et al. (2020). Overall, the results suggest that SVM and RF algorithms are effective in detecting fraudulent emails, but it is important to consider the dataset's specific characteristics such as data quality, data quantity, data diversity, data imbalance, and data noise (L’heureux et al. 2017). The study highlights the significance of machine learning algorithms in detecting fraudulent emails and how organizations can use these findings to prevent fraudulent activities and protect themselves from financial losses.

The study examined the effectiveness of various resampling techniques in improving the classification accuracy of email sentiment outcomes. The results suggest that using the SMOTE+ENN techniques, the Logistic regression model performed best among the selected classifiers when using the Enron dataset. The RF, KNN, SVM, and Logistic classifiers performed better when using the SMOTE+ENN techniques compared to the SMOTE+Tomek and original dataset. On the other hand, using the Kaggle dataset, the RF model performed better using the original data, and KNN and logistic classifiers performed optimally when using the SMOTE+ENN techniques compared to the SMOTE+Tomek and original dataset. The SVM classifier performed best with the SMOTE+Tomek techniques compared to the SMOTE+ENN and original dataset.

Finally, using the combined dataset, the RF outperformed other selected models using the SMOTE+ENN techniques. The results suggest that using the SMOTE+ENN techniques produce the highest predictive and classification accuracy for the selected models. However, it is important to note that combining different data sets can help improve the quality of the model, increase robustness and generalization, and ensure that it is trained on a more representative and diverse set of examples (Zhong et al.2021) and the performance of the models may vary depending on the dataset and the specific characteristics of the data (L’heureux et al. 2017). The study highlights the importance of utilizing resampling techniques to improve the classification accuracy of fraudulent emails (Yaseen et al. 2021) and the results can be useful for organizations in detecting and preventing fraudulent activities.

The result suggests that machine learning algorithms can be effective in detecting fraudulent emails using different techniques and datasets. It highlights the importance of selecting the appropriate algorithm in terms of accuracy, adaptability, efficiency, generalization, and interpretability (Jain et al.2016) and the sampling technique for each dataset to achieve the best results (Sridhar et al. 2021). The study also shows that using imbalance resampling techniques, such as SMOTE+ENN, can improve the performance of classifiers (Kotb et al.2021) and increase their accuracy in identifying fraudulent emails.

The findings have practical implications for organizations that rely on email communication and need to prevent fraudulent activities. By utilizing machine learning algorithms and appropriate sampling techniques, organizations can detect and prevent fraudulent emails before any financial losses occur. It also emphasizes the importance of continuous improvement and testing of algorithms to ensure that they remain effective in detecting new and evolving types of fraudulent emails.

## 5.4 Conclusions

The first objective of the study is to classify fraudulent emails into positive, neutral, and negative sentiments, this objective was achieved using the lexicon-based approach by using the polarity scores. The second objective of the study is to determine the best classification algorithm for the sentiment fraudulent emails, this is achieved by comparing the predicting performance of the RF, SVM, KNN, and logistic models. The third objective is to model and classify the sentiment outcomes by adopting balanced and imbalanced classes, this was achieved by implementing SMOTE+ENN and SMOTE+TOMEK to resample the sentiment class and the classification algorithms were adopted and their predictive performance of the resampled classification was compared with the original datasets. Finally, the fourth objective determined and compared the predictive and classification performances of the models using accuracy, precision, recall, and F1score values, and the algorithms with the highest performance metrics were chosen as the best.

Based on the objectives and findings of the study, it can be concluded that

1. The lexicon-based approach classified the emails into negative sentiments (8%, 18%, and

9%), neutral sentiments (34%, 23%, and 32%), and positive sentiments (59%, 60%, and 59%) using the Enron, Kaggle and combined dataset respectively.

1. The classification of fraudulent emails into positive, neutral, and negative sentiments was achieved with high accuracy using machine learning models.
2. The Support Vector Machine algorithm was identified as the best classification algorithm for fraudulent email sentiments due to its high precision, recall, F1 score, and accuracy.
3. The adoption of balanced and imbalanced classes in modeling and classifying the occurrence of fraudulent emails has a significant impact on the performance of machine learning models.
4. The SMOTE+ENN performed best in improving the performance and classification

accuracy of the selected classifiers.

## 5.5 Recommendations

Based on the results and implications discussed, the following recommendations can be made:

1. Further research should be conducted to improve the performance of the models in detecting fraudulent emails, especially in cases where the data is highly imbalanced.
2. The use of multiple algorithms and ensemble methods should be considered to improve the overall classification performance.
3. It is recommended to combine multiple datasets to improve the performance of the models, as demonstrated in this study.
4. Organizations should invest in implementing fraud detection systems that utilize AI and ML models to improve their ability to detect fraudulent emails and other fraudulent

activities.

1. Finally, it is important to continuously update and improve the datasets used in fraudulent email detection to keep up with evolving fraud patterns and techniques.

### 5.5.1 Future Studies

Based on the results and implications of this study, some potential areas for future research are:

1. **Investigating the impact of additional data sources:** In this study, three different datasets were used to train and test the models. Future research could explore the impact of using additional data sources, such as social media data or financial transaction data, on fraudulent email detection.
2. **Testing additional machine learning algorithms:** The current study tested four different machine learning algorithms for fraudulent email detection. Future research could explore additional algorithms, such as deep learning models, to determine if they could improve the accuracy of email classification.
3. Exploring the impact of class imbalance: The current study investigated the impact of class imbalance on fraudulent email detection. Future research could explore additional techniques for handling imbalanced data, such as ensemble methods or cost-sensitive learning.
4. **Investigating the impact of human factors:** In this study, fraudulent email detection was performed solely using machine learning models. Future research could investigate the impact of human factors, such as expert judgment or human-in-the-loop approaches, on email classification accuracy.

**REFERENCES**

Al-Haddad, R., Sahwan, F., Aboalmakarem, A., Latif, G. and Alufaisan, Y.M. (2021). *Email text analysis for fraud detection through machine learning techniques. 3rd Smart Cities Symposium (SCS 2020).*

Alharahsheh, H.H. and Pius, A. (2020). A review of key paradigms: Positivism VS interpretivism. *Global Academic Journal of Humanities and Social Sciences*, *2*(3), pp.39-

43.

Ali, M.Z., Ehsan-Ul-Haq, Rauf, S., Javed, K. and Hussain, S. (2021). *Improving Hate Speech*

*Detection of Urdu Tweets Using Sentiment Analysis. IEEE Access, 9, pp.84296–84305.*

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). *Phishing Attacks: A Recent*

*Comprehensive Study and a New Anatomy. Frontiers in Computer Science, 3.*

Al-Musib, N.S., Al-Serhani, F.M., Humayun, M. and Jhanjhi, N.Z.(2021). Business email compromise (BEC) attacks. *Materials Today: Proceedings*.

AminiMotlagh, M., Shahhoseini, H. and Fatehi, N. (2022). *A reliable sentiment analysis for classification of tweets in social networks. Social Network Analysis and Mining, 13(1).*

Apruzzese G, Colajanni M, Ferretti L, Marchetti M (2019) *Addressing adversarial attacks against security systems based on machine learning. In: 2019 11th International conference on cyber conflict (CyCon), pp 1–18.*

Atlam, H.F. and Oluwatimilehin, O. (2022). *Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review. Electronics, 12(1), p.42. doi:10.3390/electronics12010042.*

Bagui, S., Nandi, D., Bagui, S. and White, R.J.(2021). Machine learning and deep learning for phishing email classification using one-hot encoding. *Journal of Computer Science*, *17*(7), pp.610-623.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.’’ [Online].*

Bakarich, K.M. and Baranek, D.( 2020). Something phish-y is going on here: A teaching case on business email compromise. *Current Issues in Auditing*, *14*(1), pp.A1-A9.

Banu, R., M., A., Kamath C., A., S., A., Ujwala, H. S., & S.N., H. (2019). *Detecting Phishing*

*Attacks Using Natural Language Processing And Machine Learning. IEEE Xplore.*

Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaß, G., & Strobel, S. (2010). *New filtering approaches for phishing email. Journal of Computer Security, 18(1), 7–35.*

Bisong, E. and Bisong, E.(2019). Logistic regression. *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pp.243250.

Boateng, E.Y., Otoo, J. and Abaye, D.A.(2020). *Basic tenets of classification algorithms Knearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 8(4), pp.341-357.*

Creswell, J. W., and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approach. Sage publications.*

Cross, C. and Gillett, R.(2020). *Exploiting trust for financial gain: An overview of business email compromise (BEC) fraud. Journal of Financial Crime, 27(3), pp.871-884.*

Dankers, F.J., Traverso, A., Wee, L. and van Kuijk, S.M.(2019). *Prediction modeling methodology. Fundamentals of clinical data science, pp.101-120.*

Darad, S. and Krishnan, S. (2023). *Sentimental analysis of COVID-19 twitter data using deep learning and machine learning models. Ingenius, (29), pp.108–117.*

Denzin, N. K., & Lincoln, Y. S. (2011). *The SAGE handbook of qualitative research (4th ed.). Thousand Oaks, CA: Sage Publications.*

Dewis, M., & Viana, T. (2022). *Phish Responder: A Hybrid Machine Learning Approach to Detect*

*Phishing and Spam Emails. Applied System Innovation, 5(4), 73.*

Ding, X., Liu, B., Jiang, Z., Wang, Q., & Xin, L. (2021). *Spear Phishing Emails Detection Based on Machine Learning. 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD).*

Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J. and Ijaz, M.F.(2022). *A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, 2022.*

Elmrabit, N., Zhou, F., Li, F. and Zhou, H.(2020). *June. Evaluation of machine learning algorithms for anomaly detection. In 2020 international conference on cyber security and protection of digital services (cyber security) (pp. 1-8). IEEE*.

Ezpeleta, E., Velez de Mendizabal, I., Hidalgo, J.M.G. and Zurutuza, U. (2020). *Novel email spam detection method using sentiment analysis and personality recognition. Logic Journal of the IGPL, 28(1), pp.83–94.*

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). *Phishing Email Detection Using*

*Improved RCNN Model With Multilevel Vectors and Attention Mechanism. IEEE Access, 7, 56329–56340.*

Galbusera, F., Casaroli, G. and Bassani, T. (2019). *Artificial intelligence and machine learning in spine research. JOR spine, 2(1), p.e1044.*

Gangavarapu, T., Jaidhar, C.D. and Chanduka, B.( 2020). *Applicability of machine learning in spam and phishing email filtering: review and approaches. Artificial Intelligence*

*Review, 53, pp.5019-5081.*

Géron, A.(2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "

O'Reilly Media, Inc.".

Ghorbani, R. and Ghousi, R.(2020). *Comparing different resampling methods in predicting students’ performance using machine learning techniques. IEEE Access, 8, pp.6789967911.*

Glorot, X., Bordes, A., & Bengio, Y. (2011). *Domain Adaptation for Large-Scale Sentiment*

*Classification: A Deep Learning Approach. ICML. in Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011, pp. 513–520.*

Goldberg, S.B., Flemotomos, N., Martinez, V.R., Tanana, M.J., Kuo, P.B., Pace, B.T., Villatte, J.L., Georgiou, P.G., Van Epps, J., Imel, Z.E. and Narayanan, S.S.( 2020). *Machine learning and natural language processing in psychotherapy research: Alliance as example use case. Journal of counseling psychology, 67(4), p.438.*

Gutierrez, C. N., Kim, T., Corte, R. D., Avery, J., Goldwasser, D., Cinque, M., & Bagchi, S.

(2018). *Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks. IEEE Transactions on Dependable and Secure Computing, 15(6), 988–1001.*

Hag Ali, R.S. and El Gayar, N. (2019). *Sentiment Analysis using Unlabeled Email data. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE).*

Hamid, I. R. A., & Abawajy, J. (2011). *Phishing Email Feature Selection Approach. 2011IEEE*

*10th International Conference on Trust, Security and Privacy in Computing and Communications.*

Helmini, S. (2019). *Empirical Analysis on Email Classification Using the Enron Dataset. [online] Medium. Available at: https://towardsdatascience.com/empirical-analysis-on-emailclassification-using-the-enron-dataset-19054d558697 [Accessed 13 Mar. 2023].*

Hilmola, O.P. (2018). *Research Methodology and Identifying Key References Used. In: Supply*

*Chain Cases. Palgrave Pivot, Cham.*

Hina, M., Ali, M., Javed, A.R., Srivastava, G., Gadekallu, T.R. and Jalil, Z.( 2021). *October. Email classification and forensics analysis using machine learning. In 2021 IEEE SmartWorld,*

*Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable*

*Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI) (pp. 630-635). IEEE.*

Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I. and Tygar, J.D.( 2011 October). *Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58).*

Islam, M.K., Al Amin, M., Islam, M.R., Mahbub, M.N.I., Showrov, M.I.H. and Kaushal, C.(2021)*September. Spam-detection with comparative analysis and spamming words extractions. In 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1-9). IEEE.*

Islam, M.T., (2021). Multi-layer hybrid balancing technique to remove data imbalance.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E. and Alegre, E. (2022). *A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review.*

Jonathan, B., Putra, P.H. and Ruldeviyani, Y.(2020). *July. Observation imbalanced data text to predict users selling products on female daily with smote, tomek, and smote-tomek. In 2020*

*IEEE International Conference on Industry 4.0, Artificial Intelligence, and*

*Communications Technology (IAICT) (pp. 81-85). IEEE*.

Kisambu, K.L., & Mjahidi, M. (2022). *Evaluation of Machines Learning Algorithms in Detection of Malware-based Phishing Attacks for Securing E-Mail Communication. Artificial Intelligence and Machine Learning.*

Kisambu, K.L., & Mjahidi, M. (2022). *Evaluation of Machines Learning Algorithms in Detection of Malware-based Phishing Attacks for Securing E-Mail Communication. Artificial Intelligence and Machine Learning.*

Kisambu, K.L., & Mjahidi, M. (2022). *Evaluation of Machines Learning Algorithms in Detection of Malware-based Phishing Attacks for Securing E-Mail Communication. Artificial Intelligence and Machine Learning.*

Kotb, M.H. and Ming, R. (2021). *Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. International Journal of Advanced Computer Science and Applications, 12(9).*

Koteluk, O., Wartecki, A., Mazurek, S., Kołodziejczak, I. and Mackiewicz, A.(2021). *How do machines learn? artificial intelligence as a new era in medicine. Journal of Personalized Medicine, 11(1), p.32.*

Koumakis, L.(2020). *Deep learning models in genomics; are we there yet?. Computational and*

*Structural Biotechnology Journal, 18, pp.1466-1473.*

Loyola-González, O., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A. and García-Borroto, M., (2016*). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing, 175, pp.935-947.*

Lu, H., Li, Y., Chen, M., Kim, H. and Serikawa, S.( 2018). *Brain intelligence: go beyond artificial intelligence. Mobile Networks and Applications, 23, pp.368-375.*

Machuca, C.R., Gallardo, C. and Toasa, R.M. (2021). *Twitter Sentiment Analysis on Coronavirus:*

*Machine Learning Approach. Journal of Physics: Conference Series, 1828(1), p.012104.*

Malhotra, P., & Malik, S. (2022). *Spam Email Detection Using Machine Learning and Deep*

*Learning Techniques. SSRN Electronic Journal.*

Malik, E.F., Khaw, K.W. and Chew, X.(2022). *New Hybrid Data Preprocessing Technique for*

*Highly Imbalanced Dataset. Computing and Informatics, 41(4), pp.981-1001.*

Millstein, F.(2020)*. Natural language processing with python: natural language processing using*

*NLTK. Frank Millstein.*

Misra, S., Li, H. and He, J.(2019). *Machine learning for subsurface characterization*. *Gulf*

*Professional Publishing.*

Mittal, A., Engels, D.D., Kommanapalli, H., Sivaraman, R. and Chowdhury, T. (2022). *Phishing Detection Using Natural Language Processing and Machine Learning. SMU Data Science*

*Review, [online] 6(2). Available at:*

*https://scholar.smu.edu/datasciencereview/vol6/iss2/14 [Accessed 16 Feb. 2023].*

Mohamed, G., Visumathi, J., Mahdal, M., Anand, J., & Elangovan, M. (2022). *An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approac)h. Processes, 10(7), 1356.*

Morgan, H.(2022). *Conducting a Qualitative Document Analysis. Qualitative report, 27(1).*

Nguyen, T. H., & Grishman, R. (2015). *Relation Extraction: Perspective from Convolutional*

*Neural Networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing.*

Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014*). Detection of fraudulent emails by employing advanced feature abundance. Egyptian Informatics Journal, 15(3), pp.169– 174.*

Park, G. and Taylor, J.M. (2015). *Using syntactic features for phishing detection. [Online].*

*Available:*

Papernot N, McDaniel P,Jha S,Fredrikson M,Celik ZB,Swami A(2015).*The limitation of deep learning in adversarial settings.*

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik, Celik ZB, Swami A(2017).*Practical blackbox attacks against machine learning.*

Pereira, R.M., Costa, Y.M. and Silla Jr, C.N.(2020). *MLTL: A multi-label approach for the Tomek*

*Link undersampling algorithm. Neurocomputing, 383, pp.95-105.*

Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E. and Loukas, G.(2019). *A taxonomy and survey of attacks against machine learning. Computer Science Review, 34, p.100199*.

Pradha, S., Halgamuge, M.N. and Vinh, N.T.Q.(2019)*. Effective text data preprocessing technique for sentiment analysis in social media data. In 2019 11th international conference on knowledge and systems engineering (KSE) (pp. 1-8). IEEE.*

Repke, T. and Krestel, R. (2018). *Bringing back structure to free text email conversations with recurrent neural networks,’’ in Proc. Eur. Conf. Inf. Retr.Cham, Switzerland: Springer, 2018, pp. 114–126.*

Rodrigues, A.P., Fernandes, R., A, A., B, A., Shetty, A., K, A., Lakshmanna, K. and Shafi, R.M. (2022). *Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques. Computational Intelligence and Neuroscience, 2022, pp.1–14.*

Rustam, F., Saher, N., Mehmood, A., Lee, E., Washington, S. and Ashraf, I.(2023). *Detecting ham and spam emails using feature union and supervised machine learning models. Multimedia Tools and Applications, pp.1-17.*

Ruz, G.A., Henríquez, P.A. and Mascareño, A. (2020). *Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer*

*Systems, [online] 106, pp.92–104.*

Salloum, S., Gaber, T., Vadera, S. and Shaalan, K. (2021). *Phishing Email Detection Using*

*Natural Language Processing Techniques: A Literature Survey. Procedia Computer Science, 189, pp.19–28.*

Salloum, S., Gaber, T., Vadera, S. and Sharan, K.(2022).*A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access.*

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., and Jinks, C. (2018). *Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & quantity, 52(4), pp.1893-1907.*

Saunders, M., Lewis, P., Thornhill, A. and Bristow, A. (2019). *Research Methods for Business Students" Chapter 4: Understanding research philosophy and approaches to theory development. 128-171.*

Shah, K., Patel, H., Sanghvi, D. and Shah, M.(2020). *A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Researc h, 5, pp.1- 16.*

Shi Y, Erpek T, Sagduyu YE, Li JH (2019*) Spectrum data poisoning with adversarial deep learning.*

Siddique, Z. B., Khan, M. A., Din, I. U., Almogren, A., Mohiuddin, I., & Nazir, S. (2021). *Machine Learning-Based Detection of Spam Emails. Scientific Programming, 2021, 1–11.*

Singh, J. (2011). *Detection of phishing e-mail in Proc. IJCST, vol. 2, no. 1, 2011, pp. 547–549.*

Smadi, S., Aslam, N., & Zhang, L. (2018). *Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. Decision Support Systems, 107, 88–102.*

Snyder, H. (2019). *Literature review as a research methodology: An overview and guidelines.*

*Journal of Business Research, 104, pp.333-339.*

Srinivasarao, U. and Sharaff, A.(2021). *Email sentiment classification using lexicon-based opinion labeling. Intelligent computing and communication systems, pp.211-218.*

Srivastava, S., Bisht, A. and Narayan, N., (2017, January). *Safety and security in smart cities using artificial intelligence—A review. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 130-133). IEEE.*

Stockemer, D.( 2019*). Quantitative Methods for the Social Sciences: A Practical Introduction with*

*Examples in SPSS and Stata (1st Ed.). Ontario: Springer International Publishing*.

Taherdoost, H.(2022). *What are different research approaches? Comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. Journal of Management Science & Engineering Research, 5(1), pp.53-63.*

Unnithan, N.A., Harikrishnan, N.B., Akarsh, S., Vinayakumar, R. and Soman, K.P., (2018). *Machine learning based phishing e-mail detection. Security-CEN@ Amrita, pp.65-69.*

Vazhayil, A.; Harikrishnan, N. B.; Vinayakumar, R. and Soman, K. P., PED-ML (2018). *Phishing email detection using classical machine learning techniques, in Proc. 1st AntiPhishing*

*Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal. (IWSPA), A. D. R. Verma, Ed. Tempe, AZ, USA, 2018, pp. 1–8.*

Veprytska, O. and Kharchenko, V., (2022, December). AI powered attacks against AI powered protection: classification, scenarios and risk analysis. In *2022 12th International*

*Conference on Dependable Systems, Services and Technologies (DESSERT)* (pp. 1-7). IEEE.

Verma, R. and Hossain, N. (2013). *Semantic feature selection for text with application to phishing email detection, in Proc. Int. Conf. Inf. Secur. Cryptol. Cham, Switzerland: Springer, 2013, pp. 455*–*468*

Verma, R., Shashidhar, S. and Hossain, R. (2012). *Detecting phishing emails the natural language way, in Proc. Eur. Symp. Res. Comput. Secur. Berlin, Germany: Springer, 2012, pp. 824–*

*841.`1*

Von Krogh, G. (2018). *Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. Academy of Management Discoveries.*

Wang, D., Al-Rubaie, A., Hirsch, B. and Pole, G.C. (2021). *National happiness index monitoring using Twitter for bilanguages. Social Network Analysis and Mining, 11(1).*

Wojcicki, N.M. (2019). *Phishing Attacks: Preying on Human Psychology to Beat the System and*

*Developing Cybersecurity Protections to Reduce the Risks. World Libraries, 23(1).*

Yacouby, R. and Axman, D. (2020, November). *Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In Proceedings of the first workshop on evaluation and comparison of NLP systems (pp. 79-91).*

Yaseen, Y.A., Qasaimeh, M., Al-Qassas, R.S. and Al-Fayoumi, M. (2019). *Email Fraud Attack Detection Using Hybrid Machine Learning Approach. Recent Patents on Computer Science, 12.*

Zhang, J. and Li, X. (2017). *Phishing detection method based on borderlinesmote deep belief network, in Proc. Int. Conf. Secur., Privacy Anonymity Comput., Commun. Storage. Cham, Switzerland: Springer, 2017,pp. 45–53.*

**Appendice A**

**Research** **Proposal**

**Course:** **Applied** **Artificial** **Intelligence** **and** **Data** **Analytics** **Capstone** **Dissertation**

**Title:** **The** **Application** **of** **Artificial** **Intelligence** **Techniques;** **Sentiment** **Analysis** **for** **Fraudulent** **Email** **Detection** **Using** **Hybrid** **Approach**

**by:** **Eghe** **Ikponmwosa-Eweka**

**UB** **Number:** **21039838**

**Date:** **26th** **September** **2022**

1

Table of Contents

**The** **rationale** **for** **thStudy**.........................................................................................3

**LiteratureRevi**ew..........................................................................................................6

**Research Questions**.....................................................................................................10

**Research Aim and** **Objectives**...................................................................................10

**Research** **Methodology** ................................................................................................11

**SamplingProcedure**......................................................................................................11

**DataCollection**..............................................................................................................11

**Data** **Analysis**............................................................................................................12

**Limitations and Ethical** **Consideration**...............................................................13

**Conclusion** ....................................................................................................................13

**References………………………………………………………………………..**14

**Appendices**……………………………….……………………………………………21 **The** **Rationale** **for** **the** **Study**

Artificial Intelligence techniques are enhancing solutions in recent times, data is the new fuel

that powers the evolution of Artificial Intelligence in decision-making. Because it is high in

volume, value, variety, variability, and velocity, organizations are shifting from using

traditional methods to becoming data-driven using this advanced technology to extract high-

value insights from data using big data Artificial Intelligence powered analytics (Li et al. 2022).

Hence the use of artificial Intelligence techniques in fraudulent email detection is an active area

of research.

The number of unsolicited emails has increased as a result of the increased usage of social

media globally. The Internet has also become a more integral part of daily life for many people,

who find it comfortable to utilize it for their convenience. Internet users view emails as a

dependable mode of communication (Kisambu & Mjahidi, 2022).

Over time, email services have become a powerful tool for communicating a variety of

information. Due to the rise in e-mail use, there is an increase in spam attacks. Spams can be

sent from anywhere on the planet by anyone who has access to the Internet and nefarious

intentions. Spams are unwanted emails that are sent to people who do not need them. These

spam emails contain fake materials frequently linked to phishing scams, spoofing, etc, and are

frequently distributed to several recipients. They are made to steal the private information of

users and use it without their will to make money from it. These emails might contain malicious

content or links to risky information. These emails are also known as phishing emails (Siddique

et al., 2021).

Hackers in the United States used email messages in 2006 to lure customers into giving up their

passwords and login information for their American Online accounts. Phishing tactics have

developed, and this makes it more challenging to recognize fake emails. Verizon's 2016 data

breach report stated that around 636,000 phishing emails were sent, yet just 3% of the recipients

who were targeted notified management of a potential phishing email (Guptal et al.2018).

In May 2017, Google was the subject of a large phishing attempt that acquired access to the

email history of millions of Gmail users (Mashtalyar et al. 2021). With the use of these

credentials, the hackers were able to send emails to recipients asking them to open an

attachment. When consumers clicked the link to the malicious file, they were prompted to grant

bogus software access to control their email accounts. The risk of losing sensitive information

to fraudsters has grown along with the continuous expansion of email use and technology.

Phishing assaults have increased significantly since the start of 2020. Data provided by the

Anti-Phishing Working Group (APWG) in the third quarter of 2021 shows the greatest monthly

total in the history of APWG reporting, 260,642 attacks were logged in July 2021. According

to APWG, business email compromise (BEC) attacks saw an average wire transfer request rise

from $48,000 in Q3 to $75,000 in Q4 of 2020. Webmail services and software as a service

were the most often used targets in the fourth quarter of 2021, accounting for 29.1% of all

attacks. (Alkhalil et al., 2021).

These figures demonstrate the widespread usage of the internet and the huge number of internet

users who fall prey to phishing scams in both the United States and the rest of the world. In

more recent years, research like those in (Banu et al., 2019) and (Alkhalil et al., 2021) have

shown an upsurge in phishing assaults during the Coronavirus pandemic (COVID-19), and phishers utilize COVID-19 to trick their targets and users, particularly those in healthcare

facilities. Attackers generated a lot of spam and scam mail with Coronavirus schemes, taking

advantage of people's concern about catching COVID-19.

This research focuses on the application of Artificial Intelligence techniques (machine

learning) to phishing, fraudulent, and spam email detection even if there is a human component

to the process. To lessen the risk of spam and phishing, several existing solutions employ

machine learning (ML) or natural language processing (NLP). For instance, Ding et al. (2021)

have suggested an email detection method for spear phishing that makes use of ML algorithms

like Decision Trees and Random Forest. Also, Banu et al. (2019) determine whether an email

is fraudulent by using both NLP and ML. But there is a need to further investigate and

effectively identify spam, phishing, and fraudulent emails, this research suggests a detection

method that combines deep learning, machine learning, and natural language processing.

Due to the enormous amount of labor involved in detecting fake emails, no set of attributes has

been identified as the most effective at detecting phishing. The underlying categorization

technique also uses the same non-deterministic scenario. The amazing thing is that phishing

attempts continue to evolve every year, despite the existence of email filters that employ

machine learning (ML) techniques and several studies about phishing attack detection and

mitigation. Additionally, phishers and malware are growing through obfuscation to become

smarter. Accordingly, it was noted that the conflict between security experts and malware

creators is ongoing, with malware’s convolution changing as swiftly as the rate of

transformation has increased (Alkhalil et al. 2021).

Due to the varied character of the vulnerabilities, there is no one solution to the phishing

problem, so it is necessary to continue investigating and improving the precision and

accuracy of the detection approaches (Alkhalil et al. 2021). To secure email communication,

it is important to evaluate NLP, ML, and DL algorithms for the identification of malware-based

phishing attempts.

The results of this study will be helpful to corporations, financial institutions, parastatals run

by governments, and non-governmental organizations. Because phishing assaults hurt

enterprises, the study will assist in identifying bogus emails in real-time. Businesses frequently

attempt to conceal the fact that they have experienced phishing assaults because of the

reputational harm associated with it as customers frequently support brands, they consider

reliable and trustworthy. The disclosure of a breach will damage the brand's reputation as well

as undermine the established level of trust. Gaining back a customer's trust is no simple task,

and a brand’s customer base has a direct impact on its value.

Additionally, it will disrupt companies that offer services related to transportation, technology,

waste management, and other crucial infrastructure, which might severely harm the economy.

Financial losses and harm to the company's reputation will result from this. Financial losses

are brought on by data breaches, privacy violations, and regulatory body fines for

noncompliance.

Artificial intelligence (AI) techniques will be used in this project to detect email fraud. This

study will employ secondary data that will be sourced online, and the representative sample or

observation will be huge enough to suit the methodologies used for this study. AI and ML

function best and are most successful in the presence of vast volumes of data.

**Literature** **Review**

Artificial Intelligence is the ability of machines to think and act like humans, various sectors

are significantly impacted by AI (Javaid et al. 2022) One of the most intriguing recent

advancements in artificial intelligence is machine learning, which is crucial because it allows

machines to develop intelligence comparable to humans without explicit programming. Areas

such as artificial intelligence, data mining, detection of phishing emails, phishing emails, and

other security breaches, etc have found machine learning to be quite beneficial. Types of

Machine Learning include Supervised Learning, Unsupervised Learning, Reinforced Learning,

and Recommender Systems (Das et al.2015).

The majority of AI models are predictive and prescriptive. Predictive models include

forecasting and projection, whereas prescriptive models are used to inform normative

recommendations. Machine learning, deep learning, natural language processing, cognitive

computing, and neural networks are just a few of the subfields of artificial intelligence that

have emerged over time. Sentiment Analysis is a form of text analytics that uses NLP and ML,it

determines if a given text contains positive, negative or neural emotions(Glorot et al.2011)

With the advent of email, the ease of communication has given rise to the issue of a lot of spam,

particularly phishing attacks via email. To combat phishing assaults, many anti-phishing

technologies have been put forth. Phishing email detection has transitioned into the era of

ML with the advancement of AI. In particular, phishing email detection has greatly benefited

from the integration of NLP and machine learning. In the past, contextual features (Verma et

al., 2012), syntactic features (Park and Taylor, 2015), and semantic features (Verma and

Hossain, 2013) have all been applied in this field.

Starting with the most fundamental machine learning techniques, Vazhayil et al. (2018)

employed supervised classification in conjunction with SVM, logistic regression, decision

trees, and random forests, to identify phishing emails. Using both content and behavior, Hamid,

Abawajy, and Hamid (2011) suggested a hybrid feature selection method. The detection

approach for phishing emails using machine learning primarily requires tagged phishing emails

and authentic emails to train the classification algorithm in the algorithm and develop the

classifier model for email classification. Three sets of features were proposed by Bergholz et

al. (2010) and are categorized as basic features (Bergholz et al., 2008), latent topic model

features (Singh, 2011), and dynamic Markov chain features (Gu and Wang, 2009).

Without additional processing, the fundamental features of an email can be recovered. Potential

features, which are not visible in emails, are called topic model features. Particularly, it

generally consists of a few words that are connected and may appear together. Dynamic

Markov chain features are text features that are based on the bag-of-words; in other words, the

purpose of modeling each type of message content is to achieve the objective of recording the

likelihood of an email belonging to a particular category.

NLP based on machine learning has limitations in detecting phishing emails since it relies on

surface-level text rather than deep semantics. Therefore, machine learning-based NLP finds

other variations such as the usage of synonyms and varied phrase structures to be difficult

(Gutierrez et al., 2018). Additionally, the machine learning approach primarily uses feature

engineering to create features that represent emails and carry out activities through these

features. Blacklisting and feature engineering must both be done manually, which necessitates

a lot of effort and specialists with the necessary subject knowledge and reduces the effectiveness of detection. The research that follows concentrates on deep learning

methodologies to address the issues with the first two methods.

NLP tasks like text categorization (Glorot et al. 2011), information extraction (Nguyen, &

Grishman, 2015), and machine translation, have been well-represented by deep learning

(Bahdanau et al., 2014). As an alternative to manually extracting email properties, it might also

automatically create useful features from emails to recognize phishing emails. Deep learning

for phishing email detection gives a comprehensive description of the email text information.

With the use of deep learning and word embedding, Repke and Krestel (2018) gave free text

email discussions some structure again. Using deep learning and word embedding to analyze

emails is informative even though the goal of this work is not to identify fraudulent emails.

Convolutional neural network (CNN) technology and Keras word embedding were suggested

by Fang et al. (2019) as a way to create a model for phishing email detection. The Deep Belief

Network (DBN) and the Recurrent Neural Network (RNN) are just two of the various deep

learning algorithms in use today (Zhang and Li, 2017; Smadi et al., 2018). These deep learning

techniques for detecting phishing emails currently ignore the distinctions between phishing

emails and other targets and merely apply natural language processing (NLP) technology to

the problem. To some extent, context is disregarded. The development of phishing email

detection, of them, has led to constraints.

To detect spam emails, Siddique et al. (2021) employed both machine learning and deep

learning techniques. Their research makes use of Naive Bayes, CNN, SVM, and LSTM as well

as other known machine learning techniques to recognize and classify email content.

Their results show that the LSTM model outperforms other models, with a maximum accuracy

score of 98.4%.

An automated and deep learning-based detection algorithm was proposed by Malhotra and

Malik in 2022. Using the Spam Email Dataset, they also used NLP to identify spam and fake

news. LSTM, Bi-LSTM, and Dense Classifier Sequential Neural networks were used to

compare accuracy and results. Recall, accuracy, and F1-score are a few metrics that are used

to evaluate the dataset's performance. The study showed that applying Bi-LSTM classification

improves the dataset's overall accuracy.

Dewis and Viana (2022) used a hybrid ML approach to identify spam and phishing emails.

According to the paper, the efficiency of the chosen model was attained by having the Phish

Responder participate in an experiment where it obtained an average accuracy of 99% using

the LSTM model for text-based datasets. Additionally, Phish Responder demonstrated an

average MLP model accuracy of 94% for datasets that were based on numbers.

Mohamed et al. (2022) used machine learning to detect phishing attacks. The classification

accuracy for phishing detection using three different classifiers was determined to be 95.18%,

85.45%, and 78.89% for NN, SVM, and RF, respectively. The findings imply that the best

method for phishing detection is machine learning.

**Research** **Questions**

This study tends to carry out sentimental analysis for fraudulent email detection. This study

tends to answer the following research questions.

1. How to choose the ideal combination of attributes for fraudulent email detection?
2. How to choose the most effective classification algorithm for fraudulent email

detection?

1. How to improve the performance of the best classifiers and features that were chosen?

1. How to assess the integration of multiple classification algorithms for the detection of

fraudulent emails?

**Research** **Aim** **and** **Objectives**

This research aims to carry out a sentiment analysis of fraudulent emails using a hybrid-based

approach. The specific objectives are:

1. To determine and evaluate the best feature extraction techniques that best classify

Emails into positive, neutral, and negative (fraudulent) emails.

1. To determine the best classification algorithm for fraudulent email detection.

1. To model and classify the occurrence of fraudulent emails.

1. To determine and compare the classification and predictive performance of the AI and

ML models using some selected performance metrics.

**Research** **Methodology**

This study adopts a quantitative research methodology because numerical data will be utilized

and analyzed (Apuke 2017). This will be done using artificial intelligence techniques to answer

the research questions, making it a deductive research approach as well, also a bit inductive

reasoning in finding meaning of the data (Hong 2022). The positivist and constructivist

philosophies underpin this study, data used for this study is secondary data sourced online which means it comes from known social experiences as such, the method of analysis followed

will be the positive position of natural science. According to Bryman (2012) two philosophies,

epistemology and ontology inform relationships between research and theory. Epistemology

philosophy is underpinned in this study as it is according to the natural science procedures

(Liu 2022).

**Sampling** **Procedure**

The sample will be collected from Kaggle (it consists of 11959 rows) and from Enron

Corporation (composed of 146 users with 21 features a piece). Probability sample method will

be used to draw the data which means that every data in the set has an equal chance of being

included. Although probability sampling is the least biased, it could be the most time- and

resource-consuming sample for a particular level of sampling error. (Taherdoost 2016).

**Data** **Collection**

Secondary data collection will be used as it is much more appropriate based on the chosen area

of research, however, there are two methods of collecting data, primary and secondary. Primary

data are new, original, and obtained for the first time while secondary data is that that has been

previously collected and utilized for another purpose and has undergone the statistical process.

(Mahar et al. 2021).

The original dataset, which can be accessed at (https://www.cs.cmu.edu/enron/), was gathered

and produced by the CALO Project and contains financial data and text features that were

extracted from emails composed of 146 users with 21 features apiece. The information is a pre-

processed list of email texts taken from the dataset from the Enron Corporation. One of the top

energy marketers in North America, Europe, and the rest of the world, the corporation sells liquids, electricity, natural gas, and crude oil. The second set of information that will be used

in this study will be taken from the Kaggle website

(https://www.kaggle.com/datasets/llabhishekll/fraud-email-dataset?resource=download).

Using a hybrid technique, the retrieved data will be cleansed before being subjected to

sentiment analysis.

**Data** **Analysis**

This research study will adopt supervised (Vazhayil et al., 2018) and unsupervised learning

algorithms. The unsupervised learning that will be utilized is the NLP and the supervised

learning is the ML algorithms. The hybrid-based approach is the utilization of both the lexicon-

based and Machine Learning (ML) classification approach (Yaseen et al., 2019; Ding et al.,

2021). This approach implements the classification ability of the two approaches to optimizing

accuracy. The ML classifier to adopt for this study are the K-Nearest Neighbors Algorithm

(KNN), Support Vector Machine (SVM), Logistic regression, Decision trees, Neural Network,

Multilayer perceptron (MLP), and Random Forest, and the data undergo text document

preprocessing steps which involve tokenization, stop word removal, lowercase conversion, and

stemming (Yaseen et al., 2019).

For statistical algorithms to work with text, the text will be converted to numbers. Three main

approaches can be used to achieve this, i.e. Bag of Words, TF-IDF and Word2Vec. This study

will adopt the three approaches in which the lexicon-based sentiment analysis utilized the

Word2Vec approach, and the machine learning approach adopted both the Bag of Words and

TF-IDF (Srinivasarao and Sharaff, 2021).

The lexicon-based approach will be carried out by computing the subjectivity and polarity score

(Srinivasarao and Sharaff, 2021). A score less than zero will be classified as negative emails,

a score equal to zero will be classified as neutral emails and scores greater than zero will be

classified as positive emails. The data will be divided into an 80% training set and a 20% testing

set. The training set will be used for modeling or training the model while the testing set will

be used for model validation. The label variable is the sentiment outcome, and the feature

variable is the emails (text). The analysis will be carried out on each of the datasets, then both

datasets will be combined. The classification will also be carried out on both balanced and

unbalanced classes.

**Limitations** **and** **Ethical** **Consideration**

This study's drawback is that not all AI algorithms or methodologies will be used in it. The

source of the data collected from Kaggle is unknown, and data from Enron Corporation dates

to 2015. Recent data sets are needed to address this issue to keep up with the latest methods

used by spammers to bypass systems. Data is publicly available and does not contain personal

information that can be linked to individuals. All ethical approvals required for the data sets

that will be used must be secured before the research may move forward to preserve ethical

standards (Locke et al. 2013).

**Conclusion**

At the end of this study, it is expected that the best model that best classifies, predicts, or detects

fraudulent emails would have been identified and discussed to help the financial institution and

business organization formulate strategies for mitigating phishing, spam, or email fraud.

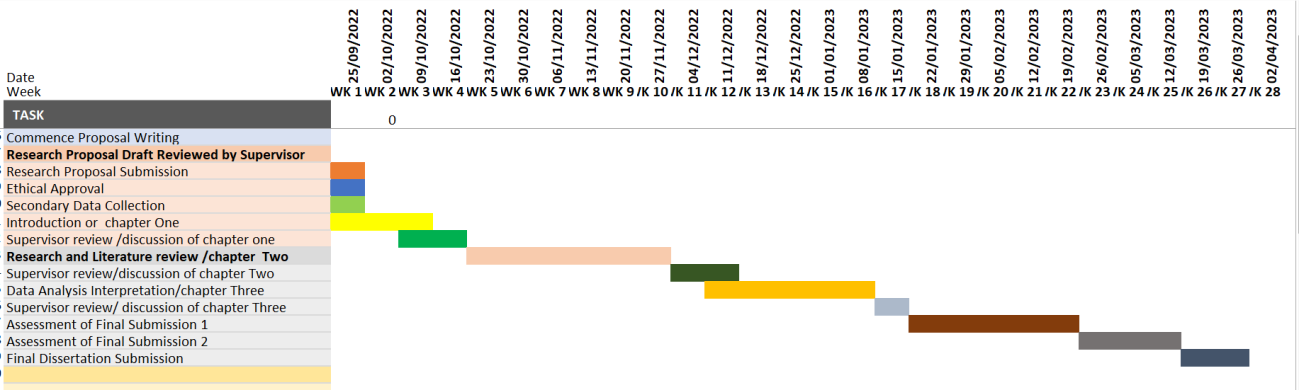
**Appendix** **A**

**Gantt** **Chart** **for** **the** **Proposed** **Research**

A workflow schedule with the proposed chapters is presented below using a Gantt Chart. Below

Gantt chart gives a visual representation of tasks displayed against time, as such, planning,

scheduling, and managing this research is made easy (Agrawal et al.2021).



**Proposed** **Chapters**

The proposed research will be written in proposed chapters as stated below(Youssef 2022)

Chapter 0ne-Introduction

Chapter Two-Literature Review

Chapter Three-Research Methodology

Chapter Four-Results Presentation and Discussion

Chapter Five-Conclusion and Recommendation

**Acronyms**

**Abbreviations:** **Definition** **of** **Terms**

NLP-Natural Language Processing

DT-Decision Tree

RF-Random Forest

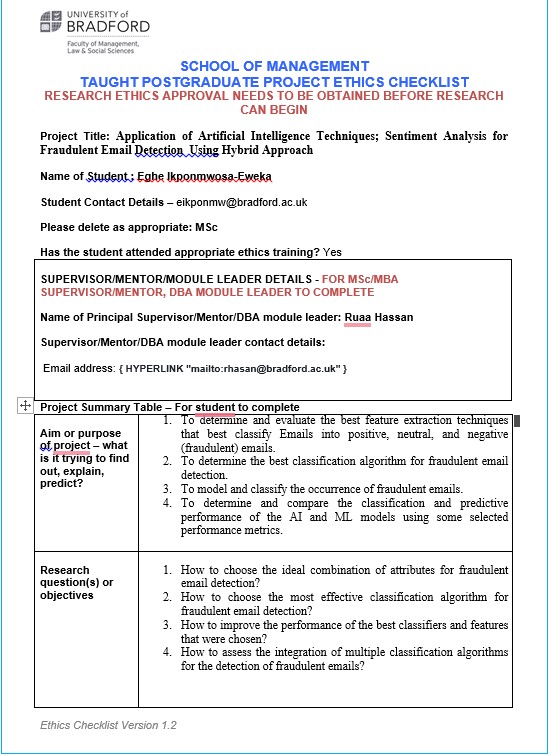
NN-Neural Networks

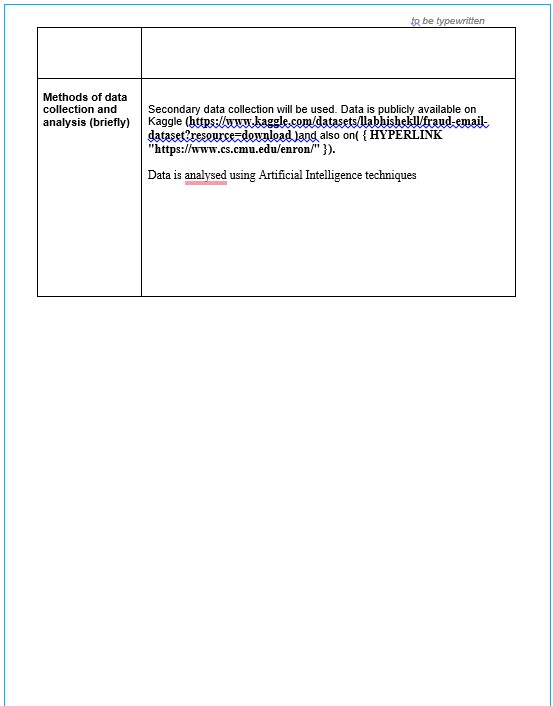
SVM-Support Vector Machine

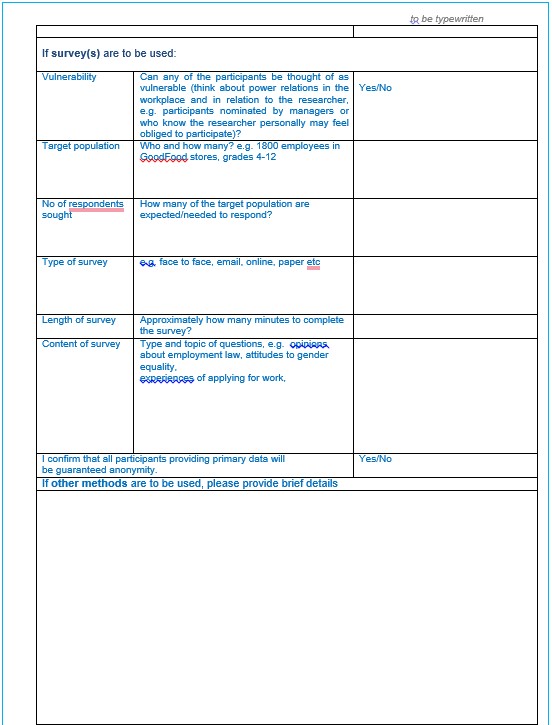
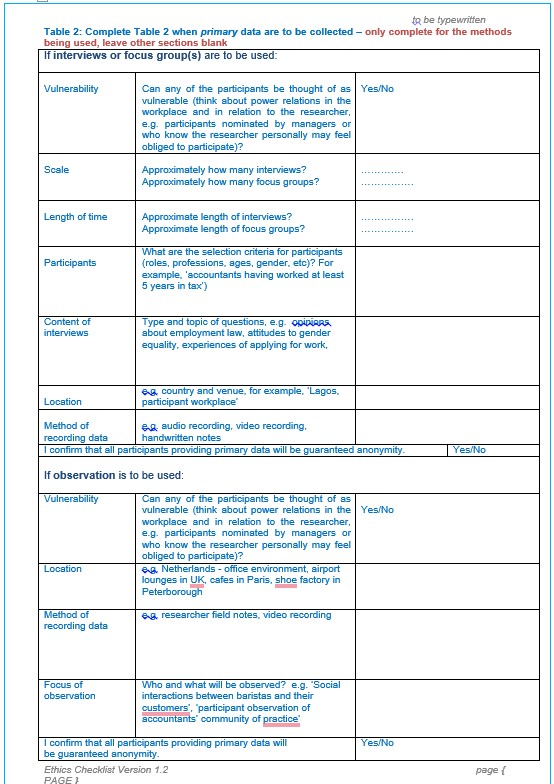
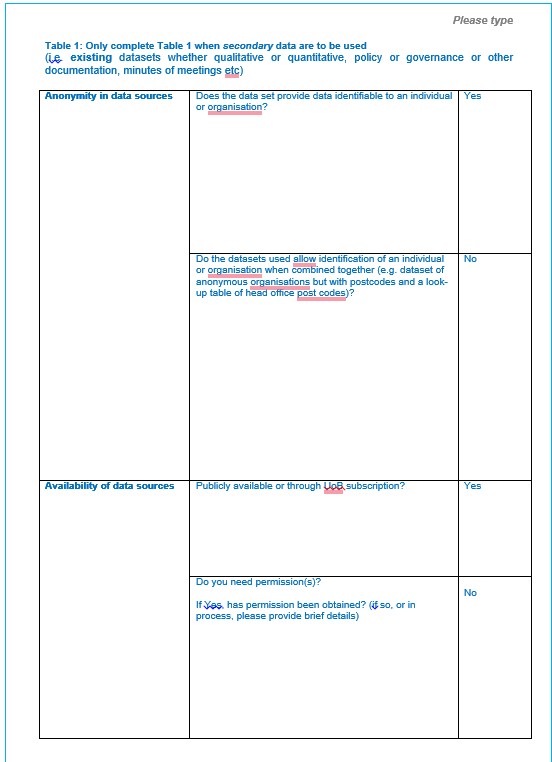
LSTM-Long Short-Term Memory Network

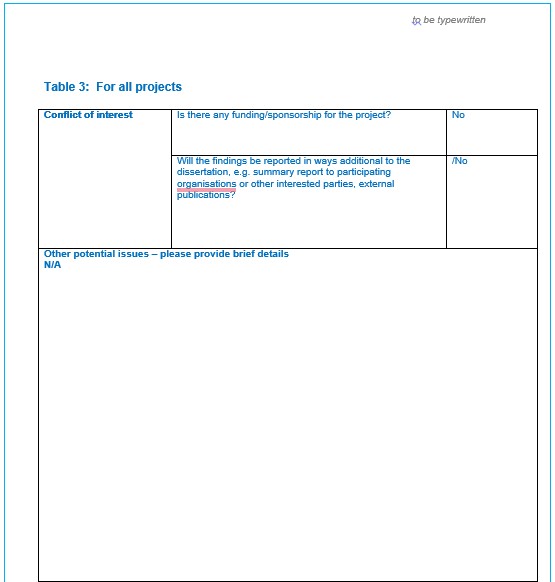
MLP-Multilayer Perceptron

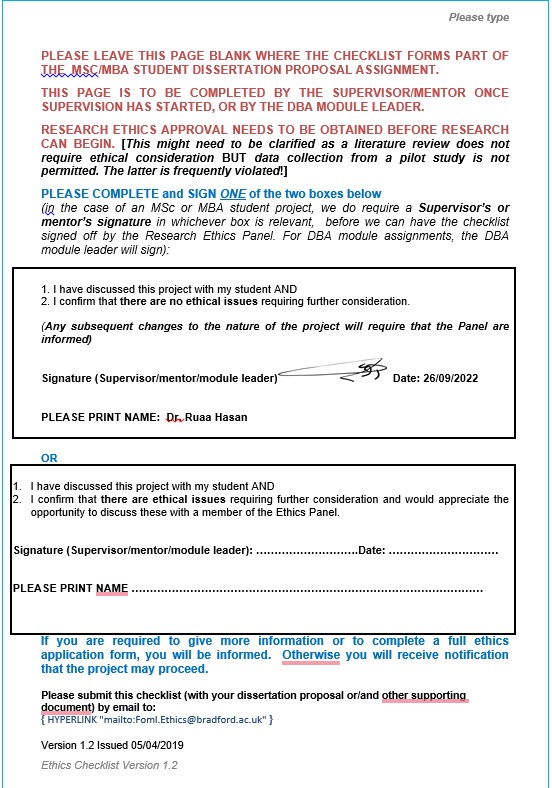
TF-IDF-Term Frequency-Inverse Document Frequency **Appendix B**

 **Ethics Checklist**



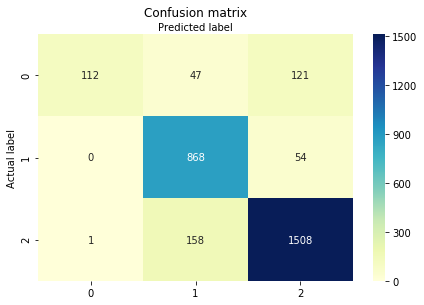






**Appendice B**

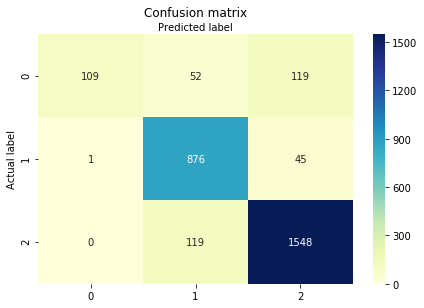
**Confusion Matrix for the Combined Dataset**



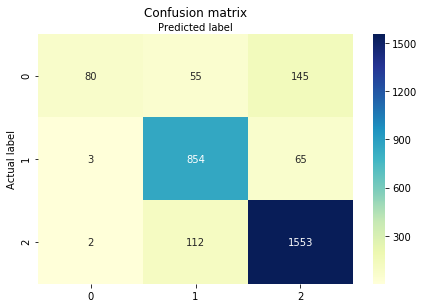
**Figure 1: Confusion Matrix of the Random Forest classifier using the testing set**



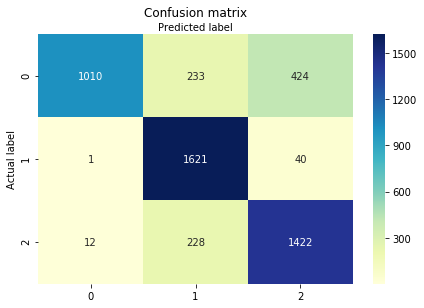
***Figure 2: Confusion Matrix of the KNN classifier using the testing set***



***Figure 3: Confusion Matrix of the SVM classifier using the testing set***

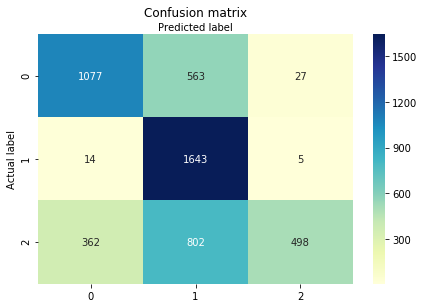


***Figure 4: Confusion Matrix of the Logistic classifier using the testing set***

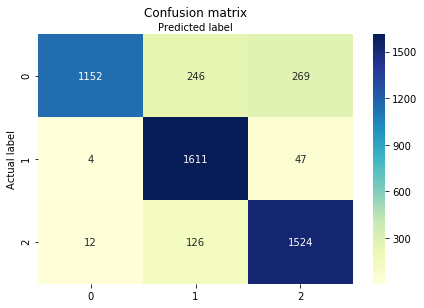


***Figure 5: Confusion Matrix of the Random Forest classifier for the testing set using***

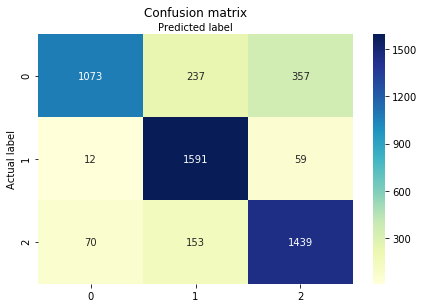
***SMOTE+TOMEK techniques***



***Figure 6: Confusion Matrix of the KNN classifier for the testing set using SMOTE+TOMEK techniques***

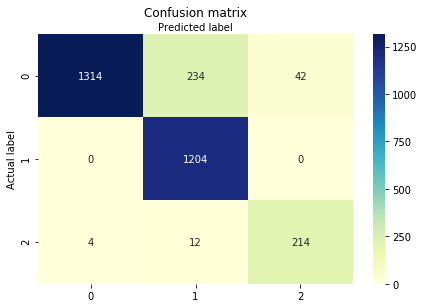


***Figure 7: Confusion Matrix of the SVM classifier for the testing set using SMOTE+TOMEK techniques***

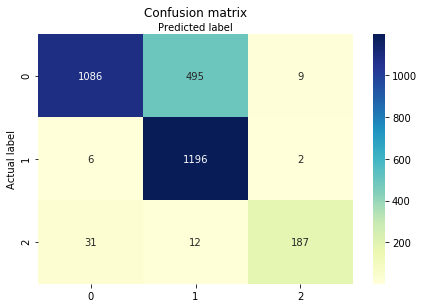


***Figure 8: Confusion Matrix of the Logistic classifier for the testing set using***

***SMOTE+TOMEK techniques***

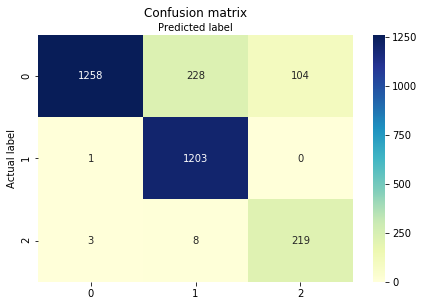


***Figure 9: Confusion Matrix of the random forest classifier for the testing set using SMOTE+ENN techniques***

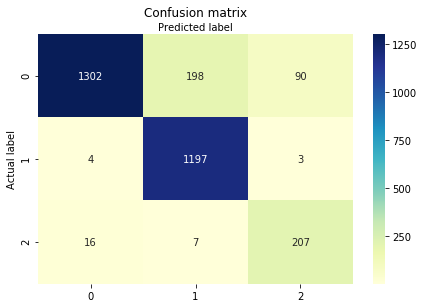


***Figure 10: Confusion Matrix of the KNN classifier for the testing set using SMOTE+ENN***

***techniques***



***Figure 11: Confusion Matrix of the SVM classifier for the testing set using SMOTE+ENN techniques***



***Figure 12: Confusion Matrix of the logistic regression classifier for the testing set using***

***SMOTE+ENN techniques***