AN ARTIFICIAL INTELLIGENCE PROJECT PROPOSAL

on

SMS SPAM SHIELD: MULTI-CATEGORY XAI SPAM DETECTOR

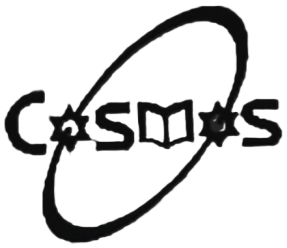
Submitted By

Alok Kumar Jha (230302)
Bibek Kumar Jha (230310)
Kushal Prasad Joshi (230345)

Submitted To

Er. Ranjan Raj Aryal

in partial fulfilment of requirement for the practicals of
Artificial Intelligence (CMP 346) course.



Cosmos College of Management & Technology
(Affiliated to Pokhara University)
Sitapaila, Kathmandu, Nepal

January 18, 2026

Cosmos College of Management & Technology
(Affiliated to Pokhara University)
Sitapaila, Kathmandu, Nepal

APPROVAL

This is to certify that the project proposal titled:

SMS Spam Shield:
Multi-Category XAI Spam Detector

has been reviewed and approved by the project assigner Er. Ranjan Raj Aryal for the further working on project in partial fulfilment of requirement for the practicals of Artificial Intelligence (CMP 346) course.

Project group memebers of Bacholer of Engineering in Computer Engineering named as Alok Kumar Jha (230302), Bibek Kumar Jha (230310) and Kushal Prasad Joshi (230345) can work on the project titled SMS Spam Shield: Multi-Category XAI Spam Detector and submit the final report to fulfill the requirement for the practicals of Artificial Intelligence (CMP 346) course by Pokhara University.

_____

Er. Ranjan Raj Aryal
Course Lecturer

Date of approval: _____

ABSTRACT

This project proposes **SMS Spam Shield: Multi-Category XAI Spam Detector**, an intelligent and explainable system for classifying SMS messages into multiple actionable categories such as *spam, phishing, promotional, transactional, and legitimate* messages. Unlike conventional binary spam filters, the proposed system aims to provide fine-grained classification while offering transparent, human-interpretable explanations for each prediction.

The system is designed to combine classical machine learning models, including Logistic Regression, Naive Bayes, and SVM, with a deep learning-based recurrent neural network (RNN/LSTM). An ensemble-based aggregation strategy is employed to improve robustness and generalization across diverse SMS patterns. To address the black-box nature of automated text classifiers, explainable artificial intelligence techniques such as LIME and SHAP are incorporated to generate token-level explanations and confidence measures for classification decisions.

The project focuses on English-language SMS messages and utilizes offline-trained models evaluated using standard multi-class performance metrics, including precision, recall, F1-score, and confusion matrices. By integrating ensemble learning with explainable AI (XAI), the proposed system aims to enhance both the accuracy and transparency of SMS spam detection, benefiting end users, system administrators, and researchers seeking interpretable and trustworthy text classification solutions.

Keywords: SMS spam detection, multi-category classification, explainable AI (XAI), ensemble learning, LSTM, LIME, SHAP.

PREFACE

This document is submitted in partial fulfillment of the requirements for the Bachelor of Engineering degree in Department of Information Communication and Technology (ICT). The proposed project, *SMS Spam Shield: Multi-Category XAI Spam Detector*, aims to address the increasing variety and sophistication of unsolicited SMS messages by developing an accurate and interpretable SMS classification system. The motivation for this work arises from the growing societal and economic impact of SMS-based spam, phishing, and fraudulent communication, as well as the increasing demand for transparency in automated decision-making systems used in security and communication domains.

Through this project, we seek to explore practical applications of artificial intelligence in cybersecurity, design a robust and user-friendly SMS spam detection framework, and contribute towards improving message safety and user trust through explainable classification mechanisms.

This project is intended to be carried out under the supervision of Er. Ranjan Raj Aryal, whose expertise and guidance are expected to be invaluable throughout the development process. With this proposal, we formally seek approval to proceed with the proposed work and look forward to the opportunity to contribute to academic learning and applied research in artificial intelligence.

We, Alok Kumar Jha (230302), Bibek Kumar Jha (230310) and Kushal Prasad Joshi (230345), hope that this proposal clearly communicates the objectives and planned approach of the proposed SMS Spam Shield: Multi-Category XAI Spam Detector, and serves as a strong foundation for its successful execution under the guidance of Er. Ranjan Raj Aryal.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to our respected supervisor, Er. Ranjan Raj Aryal, for his continuous support, encouragement, and expert guidance throughout the process of preparing this project proposal. His valuable feedback and insights have been instrumental in shaping the direction of our work.

We are also thankful to the Department of Information Communication and Technology (ICT) and all the faculty members of Cosmos College of Management & Technology (Affilated to Pokhara University), Sitapaila, Kathamandu, Nepal for their continuous support and for providing us with the opportunity and resources to carry out this proposed project.

We would also like to express our kind regards to the people around us who have directly or indirectly contributed to the successful completion of this proposal. Also we will thank our college friends who gave us valuable suggestions and feedback during the preparation of this project proposal.

Parts of this proposal were drafted and refined with the assistance of AI-powered language models, including ChatGPT [1]. The AI tools were used solely to help with structuring, phrasing, and clarity of the text. All research, analysis, design, and conclusions presented in this proposal are entirely the author's own work.

Finally, we extend our sincere thanks to our family and friends for their unwavering support and encouragement during this endeavour.

We are grateful to all of you.

Alok Kumar Jha (230302), Bibek Kumar Jha (230310) and Kushal Prasad Joshi (230345)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
LSTM	Long Short-Term Memory
NB	Naive Bayes
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SMS	Short Message Service
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
XAI	Explainable Artificial Intelligence

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Short Message Service (SMS) remains one of the most widely used communication mediums due to its simplicity, low cost, and universal availability across mobile devices. However, this widespread adoption has also made SMS an attractive channel for unsolicited and malicious messages, including spam, phishing attempts, and fraudulent promotions. Traditional SMS filtering solutions often focus on binary classification: labeling messages as either spam or legitimate; which is increasingly insufficient in modern threat landscapes [2].

Recent advances in machine learning have enabled more accurate text classification techniques, yet most deployed systems operate as black boxes, offering limited insight into why a particular message was flagged. This lack of transparency raises concerns regarding trust, accountability, and regulatory compliance, especially when automated systems influence user communication [3]. These challenges motivate the development of an intelligent, transparent, and multi-category SMS spam detection system.

1.2 PROBLEM STATEMENT

Existing SMS spam detection systems suffer from three primary limitations:

1. **Binary classification constraint:** Most systems classify SMS messages only as spam or non-spam, failing to distinguish between different spam categories such as phishing, promotional, or scam messages.
2. **Limited explainability:** Users and administrators are rarely provided with understandable explanations for classification decisions, reducing trust in automated filtering systems.
3. **Model rigidity:** Single-model approaches struggle to generalize across diverse message structures and evolving spam patterns [4].

Therefore, there is a need for a robust SMS filtering system that supports multi-category classification while providing interpretable and explainable outputs.

1.3 PROJECT OBJECTIVES

The primary objective of this project is to design and implement **SMS Spam Shield: Multi-Category XAI Spam Detector**, an explainable and extensible SMS classification system.

The specific objectives are:

- To collect and preprocess SMS data suitable for multi-category classification.
- To extract meaningful textual features using statistical and sequential representations.
- To train and evaluate multiple machine learning and deep learning models, including Logistic Regression, Naive Bayes, Support Vector Machines, and Recurrent Neural Networks.
- To design an ensemble-based result aggregation mechanism for improved prediction robustness [4].
- To integrate XAI techniques that provide human-interpretable explanations for each prediction [5, 6].

1.4 SCOPE OF THE PROJECT

The scope of this project includes:

- SMS messages written in the English language.
- Offline model training and evaluation using publicly available datasets.
- Explainability at the word or token level for classification decisions.

The project does not address multilingual SMS detection, real-time telecom network deployment, or encrypted message platforms.

1.5 SIGNIFICANCE OF THE PROJECT

By combining ensemble learning with explainable AI techniques, this project aims to improve both the accuracy and transparency of SMS spam detection systems. The proposed solution benefits:

- **End users**, by providing understandable reasons for message blocking.
- **System administrators**, by enabling debugging and model auditing.
- **Researchers**, by offering a modular framework for experimenting with hybrid models and XAI techniques.

CHAPTER 2: LITERATURE REVIEW

"SMS spam detection has evolved from rule-based systems to data-driven and explainable machine learning approaches."

2.1 OVERVIEW OF SMS SPAM DETECTION

SMS spam detection has been an active research area due to the increasing misuse of mobile communication channels for unsolicited and fraudulent activities. Early approaches relied heavily on rule-based systems and manually crafted keyword filters. Although effective for simple spam patterns, such systems lacked adaptability and failed to generalize against evolving spam strategies [2].

With the growth of labeled SMS datasets, machine learning-based text classification techniques became the dominant approach. These systems leverage statistical properties of text to learn discriminative patterns between legitimate and malicious messages.

2.2 TRADITIONAL MACHINE LEARNING APPROACHES

2.2.1 Naive Bayes Classifier

The Naive Bayes (NB) classifier is one of the most widely used algorithms for text classification due to its simplicity and computational efficiency. It assumes conditional independence between words given the class label. Despite this strong assumption, Naive Bayes has shown competitive performance in SMS spam filtering tasks, particularly when combined with bag-of-words or TF-IDF representations [7].

However, NB models are limited in capturing contextual relationships between words, which restricts their effectiveness in detecting sophisticated spam messages such as phishing attempts.

2.2.2 Logistic Regression

Logistic Regression (LR) is a discriminative linear model commonly applied in binary and multi-class text classification. It estimates class probabilities directly and is less sensitive to irrelevant features when regularization is applied. LR-based spam filters have demonstrated stable and interpretable performance in SMS classification tasks [2].

The linear nature of Logistic Regression limits its ability to model non-linear patterns

inherent in complex spam messages.

2.2.3 Support Vector Machines

Support Vector Machines (SVMs) are margin-based classifiers that aim to maximize the separation between classes. SVMs have been extensively used for spam detection due to their robustness in high-dimensional feature spaces [8]. When combined with TF-IDF features, SVMs often outperform simpler probabilistic models.

Despite their effectiveness, SVMs suffer from high computational cost during training and lack inherent probabilistic interpretability, which poses challenges for explainability.

2.3 DEEP LEARNING APPROACHES FOR SMS CLASSIFICATION

2.3.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are designed to model sequential data by maintaining temporal dependencies across inputs. In the context of SMS classification, RNNs capture word order and contextual information that traditional bag-of-words models ignore.

Long Short-Term Memory (LSTM) networks address the vanishing gradient problem in standard RNNs and have demonstrated improved performance in text classification tasks [9]. However, deep learning models require larger datasets and are often criticized for their black-box behavior.

2.4 ENSEMBLE LEARNING TECHNIQUES

Ensemble learning combines multiple models to improve generalization and robustness. Techniques such as voting, averaging, and stacking leverage the strengths of individual classifiers while mitigating their weaknesses [4]. In spam detection, ensemble models have been shown to outperform single-model approaches, particularly when datasets contain diverse message patterns.

This project adopts an ensemble-inspired strategy by aggregating predictions from classical and deep learning models.

2.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

2.5.1 Need for Explainability

As machine learning systems increasingly influence user-facing decisions, explainability has become a critical requirement. Black-box models undermine user trust and complicate debugging, auditing, and regulatory compliance [3].

2.5.2 Model-Agnostic Explanation Techniques

Local Interpretable Model-agnostic Explanations (LIME) generate local surrogate models to explain individual predictions by approximating the decision boundary around a specific instance [5]. Similarly, SHAP values provide a unified framework for feature attribution based on cooperative game theory [6].

These techniques are particularly suitable for SMS classification, as they allow token-level interpretation regardless of the underlying model.

2.6 RESEARCH GAP AND JUSTIFICATION

From the reviewed literature, the following gaps are identified:

- Most existing works evaluate accuracy but do not assess interpretability quality.
- Limited focus on multi-category SMS spam classification.
- Lack of integrated explainability in ensemble-based SMS filters.
- Insufficient emphasis on user-understandable explanations in deployed systems.

The proposed **SMS Spam Shield: Multi-Category XAI Spam Detector** addresses these gaps by combining multi-model classification with explainable AI techniques in a unified framework.

GLOSSARY

Short Message Service (SMS): A text messaging service component of mobile communication systems used for exchanging short text messages between mobile devices [2].

Spam: Unsolicited or unwanted messages sent in bulk, often for advertising, fraudulent, or malicious purposes [2].

Phishing: A form of cyberattack in which deceptive messages attempt to obtain sensitive information such as passwords, banking details, or personal data [10].

Machine Learning (ML): A field of artificial intelligence that enables systems to learn patterns from data and make predictions without explicit programming [10].

Binary Classification: A classification task in which input data is assigned to one of two possible classes, such as spam or non-spam.

Multi-Category Classification: A classification task in which input data is assigned to one of several predefined categories rather than a binary decision [11].

Explainable Artificial Intelligence (XAI): Methods and techniques that make the predictions and internal behavior of artificial intelligence models understandable to humans [5].

Black-Box Model: A machine learning model whose internal decision-making process is not directly interpretable or transparent to human users.

Logistic Regression: A supervised machine learning algorithm used for classification that estimates class probabilities using a logistic function [11].

Naive Bayes Classifier: A probabilistic machine learning algorithm based on Bayes' theorem with an assumption of conditional independence among features [7].

Support Vector Machine (SVM): A margin-based supervised learning algorithm that separates data points using an optimal hyperplane in a high-dimensional feature space [8].

Recurrent Neural Network (RNN): A neural network architecture designed for sequential data processing by maintaining internal memory states across time steps [10].

Long Short-Term Memory (LSTM): A specialized type of recurrent neural network capable of learning long-range dependencies by mitigating the vanishing gradient problem [9].

Ensemble Learning: A machine learning technique that combines predictions from multiple models to improve accuracy, robustness, and generalization [4].

Local Interpretable Model-Agnostic Explanations (LIME): An explainability technique that explains individual predictions by approximating the model locally using an interpretable surrogate model [5].

SHapley Additive exPlanations (SHAP): A unified framework for model interpretation based on cooperative game theory that assigns contribution values to individual features [6].

REFERENCES

[1] OpenAI, "Chatgpt (gpt-5.1)," aI assistant used for drafting and analysis. [Online]. Available: https://chat.openai.com

[2] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: new collection and results," *arXiv preprint arXiv:1103.4678*, 2011, https://arxiv.org/abs/1103.4678.

[3] I. Sommerville, *Software Engineering*, 9th ed. Addison-Wesley, 2011.

[4] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple Classifier Systems*, pp. 1–15, 2000.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017.

[7] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[8] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*. Kluwer Academic Publishers, 1995.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Dorling Kindersley (India) Pvt. Ltd., 2014.

[11] E. Rich and K. Knight, *Artificial Intelligence*, 2nd ed. Tata McGraw-Hill Publishing Company Limited, 2006.