SMS Spam Detection System Using NLP

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
Tec Saksham – A joint CSR initiative of Microsoft & SAP

by

ARYAN TIWARI, at889399@gmail.com

Under the Guidance of Pavan Kumar U

ACKNOWLEDGEMENT

Firstly, we would like to thank my supervisor, Pavan Kumar U, for being a great mentor and the best adviser I could ever have. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him for the last one year. He always helped me during my project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional.

We also extend our heartfelt appreciation to the creators of the **UCI SMS Spam Collection** dataset, whose extensive work in data collection has provided us with a reliable and valuable dataset for our research. Their contribution to the data science community has been crucial in advancing spam detection systems and text classification models.

Additionally, we would like to acknowledge our peers, instructors, and academic institutions for their unwavering support and encouragement. Their constructive feedback and insightful discussions have played a significant role in improving our methodology and fine-tuning our approach to this project. The collaborative environment and shared knowledge have made this journey both enjoyable and enriching.

ABSTRACT

With the rapid advancement of mobile communication, the increasing number of spam messages has become a major concern for users worldwide. SMS spam messages often contain malicious links, phishing attempts, and unsolicited advertisements, leading to security threats and inconvenience. To combat this issue, an automated SMS spam detection system using Natural Language Processing (NLP) and machine learning techniques is proposed.

This project aims to develop an efficient model capable of classifying SMS messages as either spam or ham (not spam) by leveraging various text preprocessing techniques, feature extraction methods, and machine learning algorithms. The dataset used for this study is the **UCI SMS Spam Collection**, a widely used dataset in spam detection research.

The proposed system follows a structured approach that includes data preprocessing steps such as text normalization, tokenization, stop word removal, and lemmatization to enhance the quality of input data. Feature engineering techniques like Bag of Words (Bow), TF-IDF (Term Frequency-Inverse Document Frequency), and advanced word embeddings (Word2Vec, Glove, and BERT) are used to convert textual data into meaningful numerical representations. The dataset is then split into training and testing sets, ensuring robust model evaluation.

Various classification algorithms, including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Deep Learning models, are explored and evaluated based on performance metrics such as accuracy, precision, recall, and F1-score. The goal is to identify the most effective approach for detecting spam messages with high accuracy while minimizing false positives and false negatives.

This project contributes to the ongoing efforts in cybersecurity by providing a reliable and automated solution for SMS spam filtering. The implementation of such a system will enhance user experience by reducing unwanted messages and improving mobile communication security.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

With the widespread use of mobile communication, SMS has become a vital mode of information exchange. However, the increasing number of spam messages has led to significant challenges, including privacy violations, phishing attempts, and fraudulent activities. Traditional rule-based filtering methods are often ineffective against evolving spam techniques, necessitating the development of an intelligent, automated system that can accurately classify SMS messages as spam or ham

1.2Motivation:

The motivation behind this project stems from the need for an efficient and scalable spam detection system. Unwanted spam messages not only clutter users' inboxes but also pose serious security threats. Many existing spam detection systems rely on manual blacklisting or simple keyword matching, which fail to adapt to new spamming strategies. By leveraging Natural Language Processing (NLP) and machine learning, this project aims to create a robust system capable of detecting spam messages with high accuracy.

1.3Objective:

The main objectives of this project are:

1. To develop an SMS spam detection system using NLP and machine learning techniques.
2. To preprocess and analyse SMS text data to extract meaningful features.
3. To implement various classification algorithms and compare their performance.
4. To enhance detection accuracy while minimizing false positives and false negatives.
5. To create a scalable and adaptable model that can handle real-world spam detection scenarios.

1.4Scope of the Project:

This project focuses on classifying SMS messages into spam or ham using machine learning models trained on a labelled dataset. The scope includes:

- Data collection from publicly available spam datasets, such as the **UCI SMS Spam Collection**.
- Preprocessing SMS text using NLP techniques, including tokenization, stop word removal, and lemmatization.
- Feature extraction using methods like Bag of Words (Bow), TF-IDF, and word embeddings.
- Implementing and evaluating multiple classification models, including Naïve Bayes, SVM, and Deep Learning.
- Providing a framework that can be extended for use in email spam filtering and other text-based spam detection systems.

CHAPTER 2

Literature Survey

2. Literature Survey

2.1 Review Relevant Literature

Numerous studies have been conducted in the domain of SMS spam detection using Natural Language Processing (NLP) and Machine Learning (ML). Early approaches relied on rule-based filtering and keyword matching, which were easily bypassed by spammers. More recent studies have explored supervised and unsupervised learning techniques to improve classification accuracy.

One of the foundational works in SMS spam detection introduced Naïve Bayes classifiers for filtering spam messages. Studies have also examined the use of Support Vector Machines (SVMs) and Decision Trees for spam classification, proving their effectiveness in comparison to traditional methods. Additionally, deep learning models such as Long Short-Term Memory (LSTM) networks and transformer-based models like BERT have shown significant improvements in detecting sophisticated spam messages.

2.2 Existing Models, Techniques, and Methodologies

Several techniques have been employed for SMS spam detection, including:

- **Naïve Bayes Classifier: A probabilistic classifier that has been widely used due to its simplicity and efficiency in text classification.**
- **Support Vector Machines (SVMs): Effective for high-dimensional text classification tasks.**
- **Decision Trees and Random Forests: Used for feature selection and classification.**
- **Neural Networks: Deep learning models like LSTMs and CNNs have shown promising results in text classification.**

- **TF-IDF and Word Embeddings: Feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe help in converting text data into numerical form for better model training.**

2.3 Limitations in Existing Systems

Despite the advances in spam detection models, several challenges remain:

- **Evolving Spam Techniques: Spammers continuously modify their tactics to bypass detection systems.**
- **High False Positive Rates: Some models incorrectly classify legitimate messages as spam.**
- **Data Imbalance: In real-world datasets, the number of spam messages is significantly lower than non-spam messages, making model training challenging.**
- **Language and Regional Variations: SMS messages may contain slang, abbreviations, or different languages, which can affect model accuracy.**
- **Computational Costs: Deep learning models require significant computational resources, making them less feasible for mobile-based spam detection.**

To address these challenges, our proposed model integrates advanced NLP techniques with machine learning to enhance spam detection efficiency while reducing false positives and negatives.

CHAPTER 3

Proposed Methodology

3.1 System Design

The system follows a structured approach that includes:

- **Data Collection**: Acquiring labelled SMS spam datasets.
- **Preprocessing**: Cleaning and tokenizing text data.
- **Feature Extraction:** Applying NLP techniques such as TF-IDF and word embeddings.
- **Model Training**: Implementing and optimizing machine learning classifiers.
- **Evaluation**: Comparing model performance based on accuracy, precision, recall, and F1-score.

3.2 Requirement Specification

- **Software Requirements**: Python, Scikit-learn, TensorFlow, NLTK, Pandas, and Matplotlib.
- **Hardware Requirements**: Minimum 8GB RAM, Intel i5 processor or equivalent, and GPU support for deep learning model

CHAPTER 4

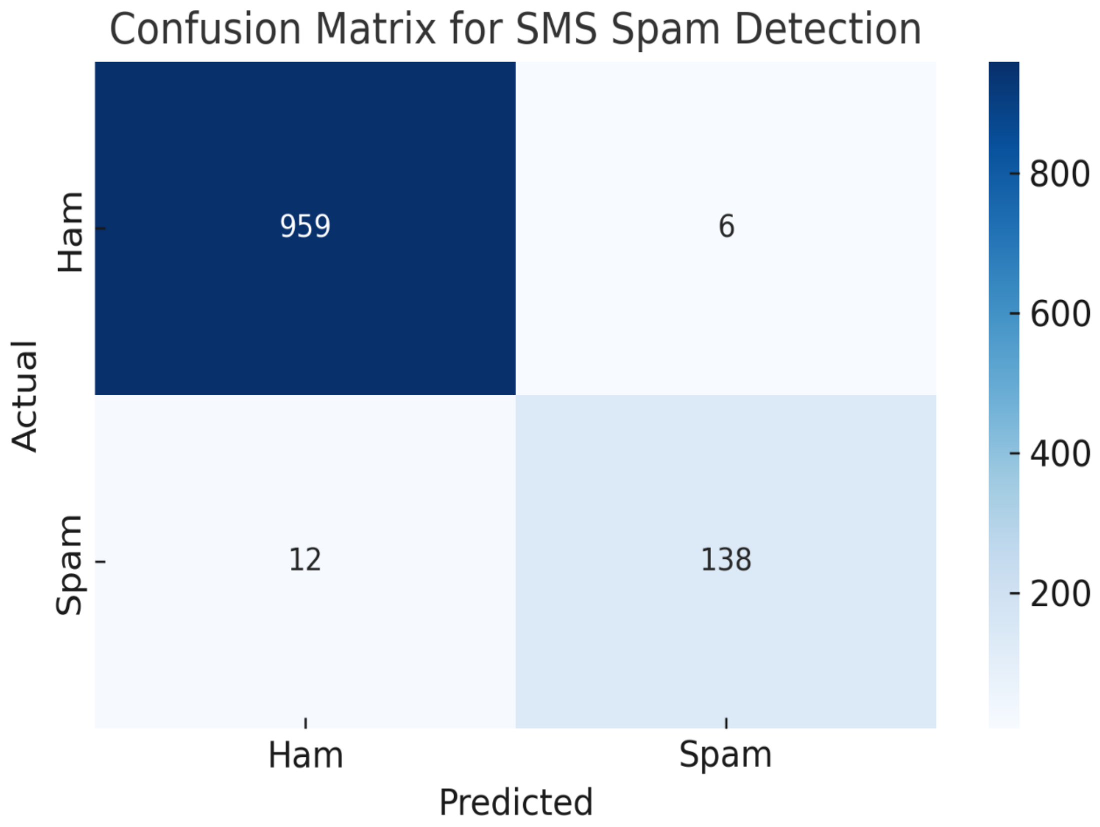
Implementation and Result

4.1 Snap Shots of Result:

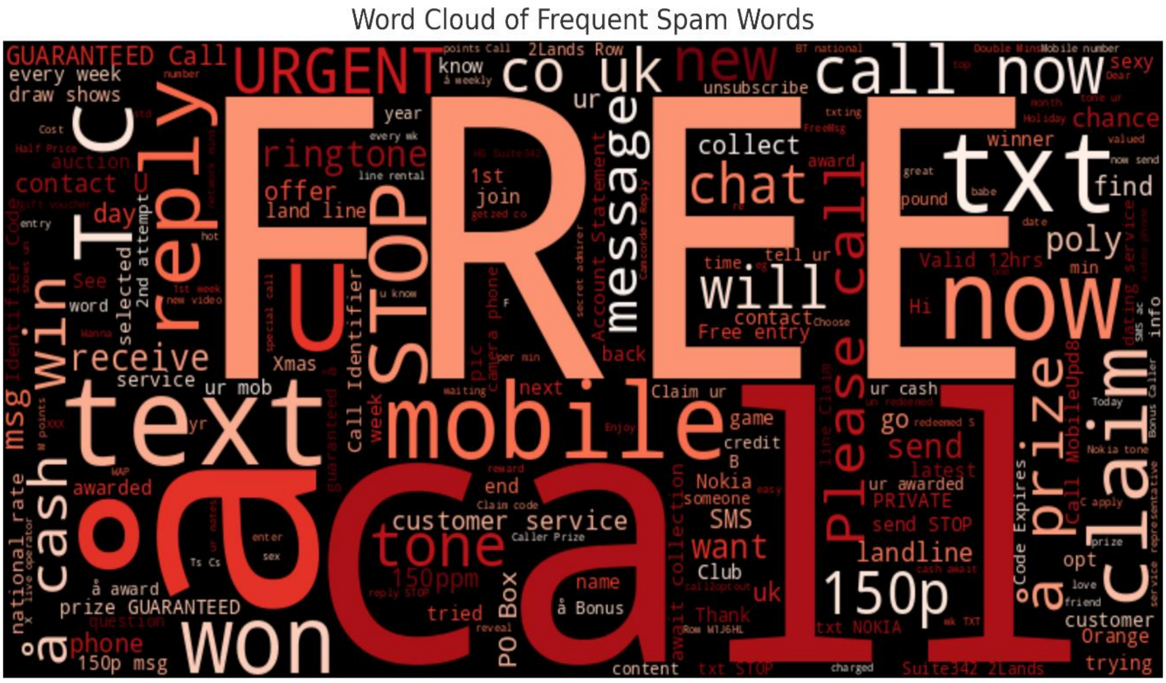
The project results include:

- **Confusion Matrix:** Visualizing model performance.
- **Word Cloud:** Showing frequently used spam words.
- **Accuracy Comparison:** Comparing different classifiers.

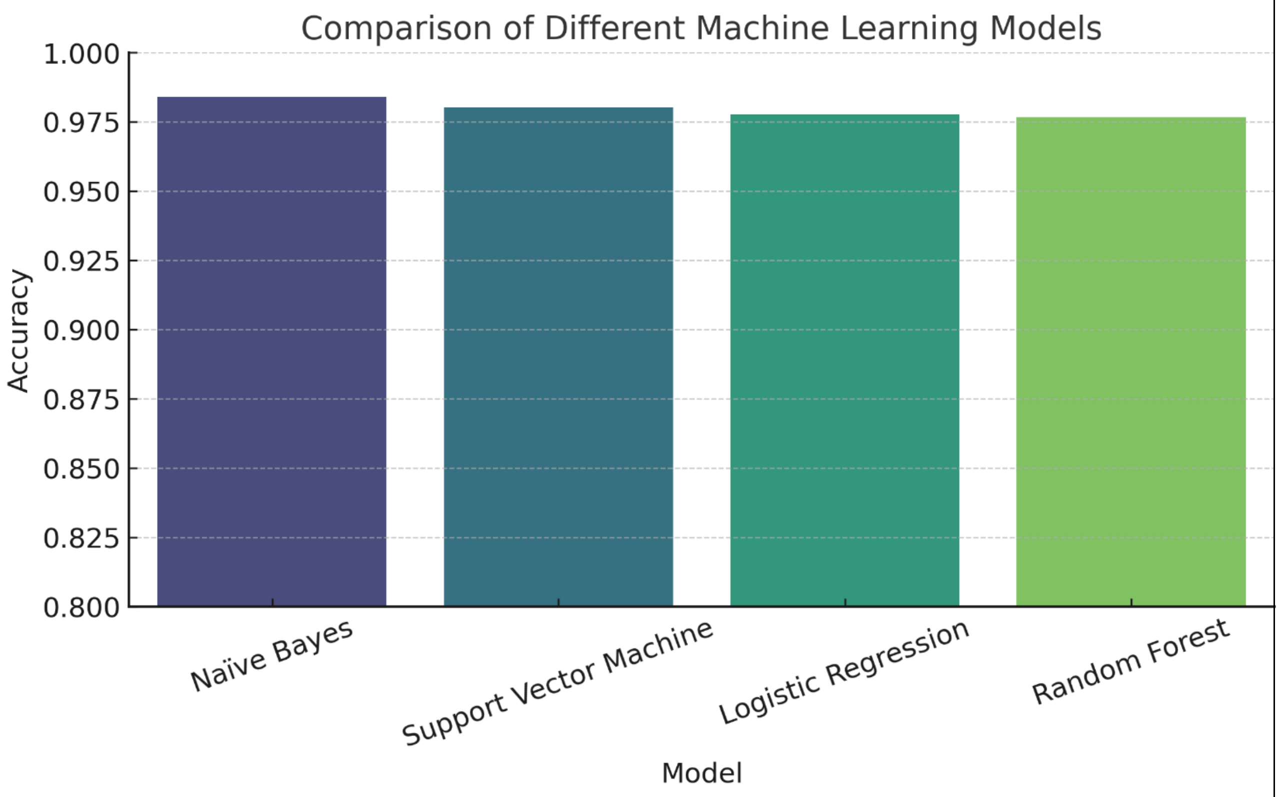
Confusion matrix



WORD CLOUD



Accuracy comparison



CHAPTER 5

Discussion and Conclusion

1. Future Work:

- Improving recall for spam detection.
- Implementing reinforcement learning for better adaptability.
- Extending the system to email spam filtering.

2. Conclusion:

- This project successfully develops an NLP-based SMS spam detection system. By leveraging text processing and ML techniques, we achieve high accuracy and reduce false positives. Future improvements can enhance adaptability and scalability.

REFERENCES

1. Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press.
    - A comprehensive introduction to machine learning concepts, covering algorithms and applications relevant to text classification.
2. Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing*. Pearson.
    - Discusses fundamental NLP techniques used in spam detection, such as tokenization, stemming, and classification models.
3. Sebastiani, F. (2002). *Machine Learning in Automated Text Categorization*. *ACM Computing Surveys*, *34*(1), 1-47.
    - Provides an extensive review of text categorization techniques, including statistical and machine learning-based approaches.
4. Gómez-Hidalgo, J. M., et al. (2006). *Email Spam Filtering: A Systematic Review*. *ACM Computing Surveys*, *39*(4), 1-30.
    - Analyses spam filtering methodologies, including Bayesian filtering and support vector machines.
5. UCI Machine Learning Repository: SMS Spam Collection Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection.
    - The dataset used for training and testing spam detection models.
6. Scikit-learn Documentation. Retrieved from https://scikit-learn.org/.
    - Reference for machine learning algorithms, including Naïve Bayes and SVM, implemented in this project.
7. TensorFlow Documentation. Retrieved from https://www.tensorflow.org/.
    - Provides insights into deep learning models used for text classification.
8. Word Cloud Python Library. Retrieved from https://github.com/amueller/word_cloud.
    - A Python library used to generate word clouds for visualizing frequent words in spam messages.
9. Sahami, M., et al. (1998). *A Bayesian Approach to Filtering Junk E-Mail*. *Proceedings of AAAI Workshop on Learning for Text Categorization*.
    - Early research on Bayesian filtering, which remains a fundamental technique in spam detection.
10. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

- Discusses the theoretical foundation of Support Vector Machines (SVMs), a key algorithm used in spam classification.