Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

DEHRADUN CAMPUS

MINI PROJECT

Bitcoin Price Prediction

B-Tech. C.S.E.

SEM-VI

2023-24

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GRAPHIC ERA HILL UNIVERSITY, DEHRADUN

SUBMITTED TO:

Mr. Samir Rana

Asst. Professor

Department of Computer

Science & Engineering

SUBMITTED BY:

Prachet Sahoo

Roll No.- 40

Section- B

Student ID- 210111178

Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

CERTIFICATE

This is to certify that Mr. Prachet Sahoo has satisfactorily completed the **Spam Email Classification** project in this college. The project in partial fulfilment of the requirement in **VIth Semester** of **B-Tech. C.S.E.** degree course prescribed by **Graphic Era Hill University, Dehradun** during the year **2023-24.**

Concerned Faculty Head of the Department

Name of the Examiner:

Signature of the Examiner:

Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)

Acknowledgement

I would like to particularly thank my Machine Learning Faculty **Mr. Amit Gupta** and my class coordinator **Mr. Samir Rana** for their patience, support and encouragement throughout the completion of this Mini Project. At last, but not the least I am greatly indebted to all the other persons who directly or indirectly helped me during the course.

SMS Spam Classifier Using Machine Learning

***Abstract*—** Short Message Service (SMS) spam is a growing nuisance, posing threats like phishing and financial scams. This paper investigates the development of an SMS spam classifier using machine learning (ML) techniques. We explore the effectiveness of various ML algorithms in identifying spam messages. The project involves data pre-processing techniques to clean and transform SMS text into a format suitable for ML models. Feature engineering techniques are employed to extract relevant characteristics that distinguish spam from legitimate messages. We evaluate the performance of different ML classifiers, including Naive Bayes, Support Vector Machines (SVM), Random Forest, and potentially deep learning models like Recurrent Neural Networks (RNNs). The paper compares the accuracy, precision, recall, and F1-score of these models to determine the most effective approach for SMS spam classification. This research contributes to the ongoing effort to combat SMS spam and improve mobile communication security.

Keywords—SMS,Email, Machine Learning, Naïve Bayes

I. INTRODUCTION

The ubiquitous nature of Short Message Service (SMS) has transformed communication, enabling instant messaging to permeate nearly every facet of our lives. SMS allows us to send quick updates, coordinate plans, and share information with friends, family, and colleagues across vast distances. This constant connection fosters a sense of immediacy and convenience that has become an indispensable part of modern communication.

However, this convenience has also become a target for malicious actors who exploit SMS for spam campaigns. These unsolicited messages often promote dubious products, spread misinformation, or attempt phishing scams to steal personal information or financial data.

Curbing SMS spam is crucial for safeguarding mobile communication and protecting users. This paper presents a machine learning (ML) based approach to develop a robust SMS spam classifier. ML algorithms excel at identifying patterns in data, making them well-suited for analyzing SMS content and distinguishing spam from legitimate messages.

Our research explores various ML algorithms, including established methods like Naive Bayes and Support Vector Machines (SVM) alongside potentially more powerful deep learning models like Recurrent Neural Networks (RNNs). Through data pre-processing, feature engineering, and model training, we aim to develop a classifier that can accurately categorize incoming SMS messages as spam or legitimate ("ham").

This paper delves into the details of the SMS spam classifier's development, including the chosen ML algorithms, feature extraction techniques, and evaluation metrics. We compare the performance of different models based on accuracy, precision, recall, and F1-score to identify the most effective approach for SMS spam classification.

By contributing to the fight against SMS spam, this research has the potential to enhance mobile security and provide users with a safer communication experience.

II. NLP

Unveiling the Mystery: A Deep Dive into
NLP

Natural Language Processing (NLP) is a fascinating field
at the intersection of computer science, artificial intelligence (AI), and linguistics. Its core mission is to bridge the gap between human language and machines, enabling computers to understand, interpret, and even generate human language. This journey into NLP is akin to deciphering a complex code, where we break down human language into its building blocks and teach computers to process and derive meaning from it.

From Text to Understanding: The NLP Pipeline

NLP tasks can be broadly categorized into two main areas: Natural Language Understanding (NLU) and Natural Language Generation (NLG). NLU focuses on extracting meaning from text and speech, while NLG involves creating human-like text or speech based on a given intent. Here's a breakdown of the typical NLP pipeline:

1. **Data Preprocessing:** Raw text data is often messy and unstructured. NLP begins by cleaning and preparing the data for further processing. This may involve tasks like removing punctuation, converting text to lowercase, and stemming/lemmatization (reducing words to their root form).
2. **Tokenization:** Sentences are broken down into smaller units like words or phrases, called tokens. This segmentation allows NLP models to analyze individual components of a sentence.
3. **Part-of-Speech (POS) Tagging:** Each token is assigned its grammatical function (e.g., noun, verb, adjective) to understand the sentence structure and relationships between words.
4. **Named Entity Recognition (NER):** NLP systems can identify and classify named entities within a text, such as people, organizations, locations, dates, or monetary values. This is crucial for tasks like information extraction and question answering.
5. **Text Representation:** Extracted features from the text need to be converted into a numerical format that computers can understand. Techniques like word embeddings, where words are mapped to highdimensional vectors, capture semantic relationships between words.
6. **Machine Learning Models:** Here's where the magic happens! NLP leverages various machine learning algorithms trained on massive amounts of labeled text data. These algorithms learn to identify patterns and relationships within the data, enabling them to perform specific tasks like sentiment analysis, topic modeling, machine translation, or spam filtering.

Unveiling the Secrets of Language: Core NLP Techniques

Let's delve deeper into some of the fundamental techniques employed in NLP:

- **Rule-based NLP:** This traditional approach relies on hand-crafted rules that define language structure and grammar. While effective for specific tasks, it can be inflexible and requires significant human effort to adapt to new languages or variations.
- **Statistical NLP:** This method utilizes statistical models to analyze language patterns. Techniques like n-grams (sequences of n words) and language models capture the co-occurrence of words and predict the likelihood of a word appearing in a specific context.
- **Machine Learning (ML) for NLP:** Supervised learning algorithms are trained on labeled datasets where text is paired with corresponding labels (e.g., sentiment, topic). The model learns to map text features to these labels, enabling it to perform tasks like sentiment analysis or spam classification on unseen data.
- **Deep Learning for NLP:** Deep learning architectures like Recurrent Neural Networks (RNNs) and their variants (LSTMs, GRUs) excel at capturing long-range dependencies in language. These models are particularly adept at handling complex tasks like machine translation, text summarization, and question answering that require understanding context across longer sequences of words.

Challenges and the Road Ahead

Despite significant progress, NLP still faces challenges.
Language is inherently ambiguous, with words having multiple meanings depending on context. Sarcasm, humor, and cultural nuances can further complicate NLP tasks. Additionally, the vast amount of data required to train NLP models can be a barrier, especially for lowresource languages.

However, the future of NLP holds immense promise. As research continues, we can expect advancements in areas like:

- **Explainable NLP:** Developing models that can explain their reasoning behind decisions will be crucial for building trust and transparency in NLP applications.
- **Multilingual NLP:** Breaking down language barriers through robust multilingual translation and understanding is essential for a truly globalized world.
- **Conversational AI:** Creating natural and engaging chatbots that can understand complex user queries and respond in a human-like manner will be a gamechanger for human-computer interaction.

NLP is revolutionizing how we interact with machines and transforming industries like customer service, healthcare, and education. By unlocking the power of human language for computers, NLP is paving the way for a more seamless and intelligent future.

III. DATA SET

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged acording being ham (legitimate) or spam.

v1 class	v2 sms		
ham 87% spam 13%	5169 unique values	[null] 99% bt not his girlfrnd..... 0% Other (47) 1%	[null] 100% MK17 92H. 450Pp... 0% Other (10) 0%
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a...		
ham	Ok lar... Joking wif u oni...		
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...		
ham	U dun say so early hor... U c already then say...		
ham	Nah I don't think he goes to usf, he lives around here though		
spam	FreeMsg Hey there		

IV. METHODOLOGY

In this research paper we will use machine learning algorithms for the prediction of price of bitcoin.

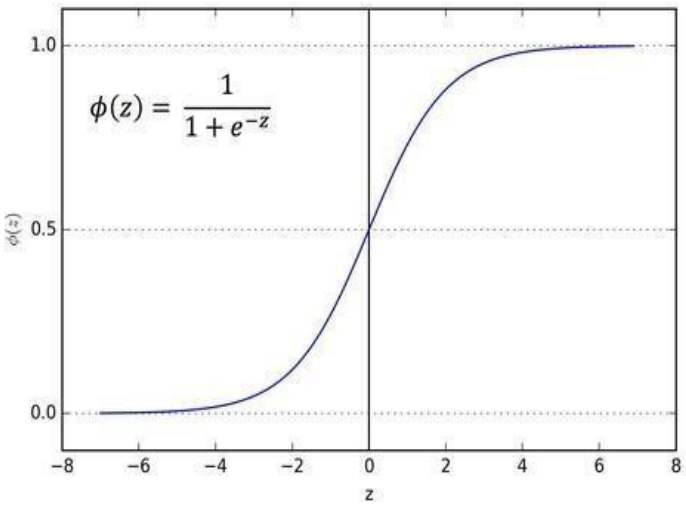
1) Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for classification tasks where we want to predict the probability of an instance belonging to a class or not. It is a statistical algorithm as it analyzes the relationship between dependent and independent variable

.

The output we get is categorical dependent. It can either be YES or NO, 1 or 0 or between 1 and 0, true or false etc.

In this the fitting line is formed in an S shape signifying 0 or 1.



Sigmoid curve

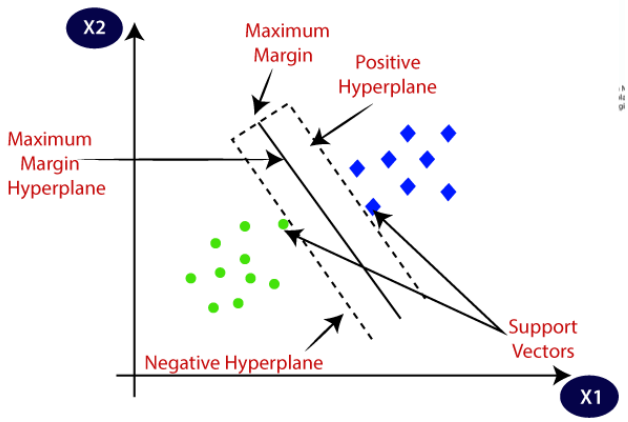
2) Support Vector Machine:-

Support vector machine is powerful machine learning algorithm used for linear and non-linear classification, regression, and outlier detection.

Some of its applications include image classification spam detection, handwriting identification, gene expression analysis etc.

The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries to ensure that the margin between the closest points of different classes should be as maximum as possible.

The dimensionality of a hyperplane is directly linked to the number of features in the dataset. When dealing with two input features, the hyperplane manifests as a simple line. With three input features, the hyperplane transforms into a twodimensional plane. This pattern persists as the number of features increases, with each additional feature contributing another dimension to the hyperplane.



3) XGBClassifier :-

XGBoost is an implementation of Gradient Boosted decision trees.

Gradient boosting is an ensemble learning method that combines multiple weak learners to create a strong learner. In gradient boosting, each weak learner is trained to minimize the loss function of the previous weak learner. This process is repeated until the desired level of accuracy is achieved.

Flow of code :-

1) First, we will perform Exploratory Data Analysis to discover patterns and trends.

2) We will check for null value in our data set.

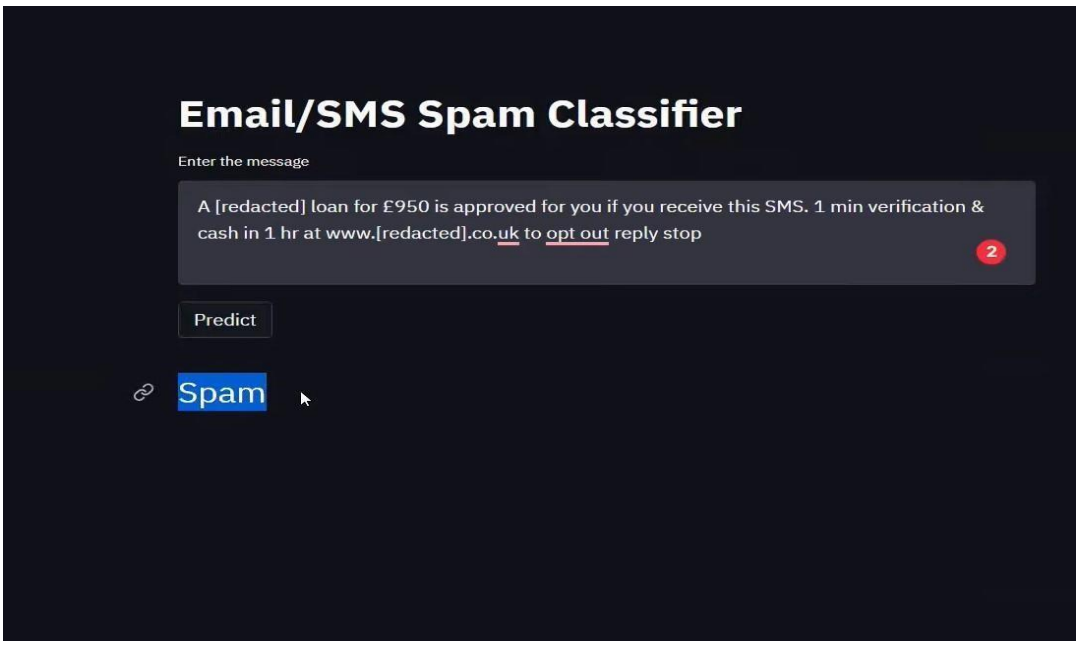
3)Now we will develop and evaluate our model with the following machine learning algorithms:-

i) Logistic Regression

ii) Support Vector Machine

iii) XGBClassifier

V. RESULTS



Out[365]:

	Algorithm	variable	value
0	ETC	Accuracy	0.977756
1	SVC	Accuracy	0.972921
2	xgb	Accuracy	0.971954
3	RF	Accuracy	0.970019
4	AdaBoost	Accuracy	0.962282
5	NB	Accuracy	0.959381
6	BgC	Accuracy	0.957447
7	LR	Accuracy	0.951644
8	GBDT	Accuracy	0.951644
9	DT	Accuracy	0.935203
10	KN	Accuracy	0.900387
11	ETC	Precision	0.991453
12	SVC	Precision	0.974138
13	xgb	Precision	0.950413
14	RF	Precision	0.990826
15	AdaBoost	Precision	0.954128
16	NB	Precision	1.000000
17	BgC	Precision	0.861538
18	LR	Precision	0.940000
19	GBDT	Precision	0.931373
20	DT	Precision	0.838095
21	KN	Precision	1.000000

VI. CONCLUSION

The realm of Natural Language Processing (NLP) offers a captivating glimpse into the intricate dance between human language and machine intelligence. By delving into the NLP pipeline, we've witnessed the transformation of raw text data into a form computers can comprehend. We explored core techniques like rule-based, statistical, and machine learning approaches, culminating in the power of deep learning architectures for complex language tasks.

While challenges like ambiguity, context dependence, and data scarcity persist, the future of NLP is brimming with possibilities. As research advances, explainable models, robust multilingual capabilities, and sophisticated conversational AI systems hold the potential to revolutionize human-computer interaction. NLP is not just about enabling machines to understand us; it's about fostering a future where communication transcends language barriers and empowers us to interact with machines in a more natural and intuitive way. The journey of NLP is far from over, and the potential for innovation and progress in this captivating field is truly limitless.

VII. ACKNOWLEDGEMENT

This work was done as the college assignment regarding Machine Learning subject and is supported by our subject teacher Dr. Amit Gupta.