## Explain the pros and cons about your model, including limitation (can be both quantitative and qualitative), including how can you improve your model in future.

### Logistic Regression (Without TFIDF)

**Pros:**

- High accuracy (98.13%), simple to implement and interpret.
- Works well when dealing with smaller, less complex datasets.

**Cons:**

- Without TFIDF, the model may not capture the relative importance of certain words.
- Not effective for large datasets with complex structures.
- May overfit, especially on unbalanced data, as was the case with the provided dataset.

### Logistic Regression (With TFIDF)

**Pros:**

- TFIDF adds weight to important words, improving feature representation.
- More robust than the raw word count in capturing the significance of words.

**Cons:**

- Slightly lower accuracy (95.41%) compared to the non-TFIDF model, indicating possible overfitting or information loss with sparse matrices.
- Still not suited for highly contextual or large-scale datasets.

### BERT

**Pros:**

- Highest accuracy (99.07%) due to its deep understanding of language context.
- Handles more nuanced features, like word meanings and sentence structure, which is essential for spam detection.

**Cons:**

- Computationally expensive and slow to train compared to logistic regression.
- Requires a lot of labelled data for fine-tuning and may overfit on small datasets.

### Improvements:

- Fine-tuning BERT with more training data.
- Using ensemble methods to combine models.
- Experimenting with other NLP models like GPT, LSTM, or hybrid models.
- Using a more balanced dataset to prevent overfitting of models

## Explain why it is important to know text analytics.

Text analytics is important because a vast amount of business information is stored in unstructured text, such as emails, customer feedback, and social media posts. Understanding and analysing this data allows businesses to extract meaningful insights, improving decision-making processes. In the case of spam email classification, text analytics helps in detecting patterns and anomalies within email content that differentiate between spam and legitimate emails. Moreover, it enhances security by identifying phishing and malware attempts embedded within emails. As businesses increasingly rely on digital communication, having a system that can automatically analyse and sort through text data is vital for improving operational efficiency, customer satisfaction, and safeguarding sensitive information.

## How can classification of spam email help your company.

Classifying spam emails enhances the overall productivity of a company by preventing unnecessary distractions caused by unwanted emails. As mentioned previously, it also reduces the risk of exposure to phishing attempts, malware, and ransomware attacks that can compromise sensitive data. Efficient spam detection systems protect both employees and the organization’s network from potential threats, thereby maintaining the integrity of internal and external communications.

## What do you need to get a better result.

To improve spam classification results, more data is essential. A larger, more diverse dataset provides the model with better exposure to varying types of spam, which helps in generalization. Feature engineering, including adding features like email metadata (e.g., sender information, timestamps, and email headers), could improve model performance. Using ensemble models—combinations of different algorithms such as BERT with TFIDF-based models—can also provide a more robust classification. Finally, continuous model training with updated datasets can help the model adapt to evolving spam patterns.

## What are the most common characteristics that identify an email as spam, and how can email systems distinguish between spam and legitimate email?

Spam emails usually contain suspicious characteristics like irrelevant subject lines, unsolicited offers, and suspicious links or attachments. They often promote deals, such as free gifts or huge discounts, and may contain poor grammar and spelling, indicating low-effort, automated content generation. Email systems distinguish between spam and legitimate emails by using techniques like blacklisting known spammers, analysing email content for suspicious keywords or phrases, and utilizing natural language processing (NLP) to understand context. AI models are also used to recognize patterns in both the content and metadata, such as the sender’s IP address or the structure of the email, which further helps in identifying spam.

## How can individuals and organizations protect themselves from the potential harm caused by spam emails, such as phishing attempts or malware?

Individuals and organizations can protect themselves from the potential harm caused by spam emails by using spam filters, which automatically detect and block unwanted messages. Educating employees on recognizing phishing attempts and suspicious emails is critical to preventing security breaches. Antivirus software, kept up to date, provides an additional layer of protection against malware distributed via email. Organizations should also use two-factor authentication (2FA) to protect sensitive accounts and limit the impact of compromised credentials. These strategies collectively reduce the risks posed by malicious spam emails.

## What are the legal implications of sending spam emails, and how do laws like the CAN-SPAM Act in the United States enforce regulations against spammers?

The **CAN-SPAM Act** in the United States regulates commercial emails, requiring senders to provide accurate subject lines, headers, and opt-out mechanisms. The act aims to reduce deceptive practices in email marketing. Violations can lead to fines up to $43,792 per email, which can accumulate quickly for businesses that engage in large-scale spamming. Globally, other regulations like the European Union’s **General Data Protection Regulation (GDPR)** impose strict guidelines on unsolicited communications. These laws serve as deterrents against spam and ensure businesses adhere to ethical practices. Non-compliance not only results in hefty fines but can also damage a company’s reputation and trustworthiness.

## How do spammers collect email addresses, and what strategies can people use to keep their email addresses private and reduce the amount of spam they receive?

Spammers collect email addresses through methods like web scraping, where they extract publicly available email information from websites. They also purchase email lists from unethical sources or exploit data breaches to access large numbers of addresses. Sometimes, they use phishing schemes to trick individuals into providing their email addresses voluntarily. To protect themselves, individuals and businesses should avoid posting emails in public forums or websites. Disposable email services, such as Temp Mail and Burner Mail, can also be useful for signing up for websites or services that may not be trusted, reducing the risk of spam.

## What role does artificial intelligence play in the fight against spam emails, and how effective are modern AI-powered spam filters in screening unwanted email?

Artificial intelligence plays a crucial role in combating spam by using machine learning models to analyse patterns within email content and metadata. Modern AI-powered spam filters leverage advanced models like deep learning, which can detect even subtle differences between legitimate and spam emails. AI improves the ability to handle large volumes of emails, quickly identifying phishing attempts, malware, and other malicious activities. Over time, these models learn from new data, continuously adapting to new types of spam. AI-powered systems are more efficient and accurate than traditional rule-based methods, providing dynamic protection against evolving spam tactics and significantly reducing false positives.