**Email Spam Classification Using BERT and Naive Bayes: A Semantic Deep Learning Approach**

**Authors:** Kalam Uddin, Abhigya Malla, Ronik Karki, Nischal Subedi

# ***Abstract*- Email Spam poses a significant hassle by clogging up inboxes, wasting people’s time and often trying to scam people with malware and fraud. So, Email Spam detection is an important exercise in maintaining secure and effective communication. This study gives a comparative analysis between the traditional approaches and BERT (Bidirectional Encoder Representations from Transformers), a modern and powerful technique. The traditional method relied on manual features such as character count, word count and sentence count. Spam messages are usually longer- about 137 characters on average compared to 71 for regular messages. Even though this approach helped on spotting basic patterns, it often missed the bigger picture- for example, when same words such as “FREE” comes up on different context to the message. On the other hand, BERT offered a smarter way to read and understand the message. It processes the text in both directions and analyzes how words are used and fits in with the sentences. By truly understanding the context of the message, it makes more accurate predictions.**

**Keywords- *Spam Detection, Email Classification, BERT, Transformers, Semantic Analysis, Natural Language Processing, Naïve Bayes***

# **Introduction**

Natural Language Processing has witnessed significant advancement in recent years, particularly in the domain of semantic understanding. The ability to capture the meaning and relationships between words remains a fundamental challenge in computational. Traditional Machine Learning approaches have long been employed for text classification and semantic analysis, but they usually struggle to capture the relationships between words that human language encompasses. Contextualized word embeddings have emerged as powerful tools that can better represent the semantic meaning of words by considering their surrounding context. These models not only aim to encode lexical information but also semantic roles which are crucial for deeper language understanding [1]

The Naive Bayes algorithms have been widely used for text classification due to their simplicity and computational efficiency. However, its effectiveness is limited by a strong reliance on the “bag-of-words” or “TF-IDF” approach that disregard word order and context [2].

This fundamental limitation prevents Naïve Bayes from capturing semantic nuances as it treats each word as an independent entity without considering how meaning changes in different contexts. While semantic enhancements to Naïve Bayes have been proposed to mitigate these issues, such as incorporating lexicon-based methods and semantic orientation calculators, these approaches still fall short when dealing with complex semantic relationships [2].  
BERT(Bidirectional Encoder Representation from Transformers) represents a paradigm shift in how machines understand language, offering a more sophisticated approach to semantic understanding. Unlike Naïve Bayes, BERT pre-trains deep bidirectional representations by jointly conditioning on both left and right context in all layers [3].

This bidirectional architecture enables BERT to develop a deeper understanding of semantic relationships between words. Study on this have shown that BERT encodes information about entity types, relations, and semantic roles showing its capacity to capture semantic knowledge that statistical model cannot [1].

Furthermore, BERT can be fine-tuned with just one additional output layer to create state-of-the-art models for various semantic tasks without substantial task-specific architecture modifications [[3].](https://aclanthology.org/N19-1423.pdf)This research paper explored how BERT’s approach to semantic understanding outperforms Naïve Bayes algorithms, analyzing the underlying mechanisms that enable this superior performance and demonstrating its practical implications across multiple natural language understanding tasks.

# **Related Works**

This study has explored various research for email spam detection. This research compares different machine learning approaches for spam detection. Various types of approach can be taken for email spam detection such as- content based, case based, rule based. These approaches are still considered a traditional method as compared to advanced machines and deep learning approach to solve the thriving issue of spam emails. With continuous development of technology with time, advanced machine learning methods have been proved effective to solve email spam detection across multiple studies [4]. Bayesian Classification (BC), Support Vector Machine (SVM), Random Tree (RT) are considered reliable algorithms for machine learning approaches [5]. Upon testing, among these algorithms, RT classifier surpasses these other classifiers [5].

While majority of the existing research studies still focuses on traditional machine learning approach to solve email spam, BERT-Based model can be effective- by using well known datasets, it achieved excellent accuracy to classify emails by context aware spam [5]. Data preprocessing is a vital step for machine learning where the data are formatted for models to improve overall performance. Tokenization, a data preprocessing method where emails are sliced into token is most effective in this case. After comprehensively going through various projects and research paper, it can be concluded that BERT methodology has been proven to be an outperforming approach for email spam detection [6].

# **Dataset and Preprocessing**

We utilized the Email Spam Classification Dataset, publicly available on Kaggle, which comprises 5,172 email samples, each structured as a row in a CSV file. The dataset includes 3,002 columns: the first column represents an anonymized identifier, the last column indicates the binary target label (1 for spam, 0 for non-spam), and the remaining 3,000 columns correspond to the most common words across the dataset, with each cell reflecting the frequency of that word in the given email.

The dataset is numerically structured and formatted in a way that is particularly well-suited for traditional machine learning models. Since the word counts are already extracted, the classical models such as Naive Bayes were able to consume this data without requiring additional tokenization or textual feature extraction.

For transformer-based models like BERT, however, the raw text form of emails is essential. To support semantic modeling, we used an alternate version of the dataset containing raw email content. This version underwent tokenization using the BERT tokenizer, which converts the text into input IDs and attention masks that are compatible with the BERT architecture.

We also applied standard preprocessing steps, including:

* Feature normalization for numerical columns (used with classical models)
* Label encoding (binary: spam = 1, non-spam = 0)
* An 80/20 train-test split to ensure reliable evaluation and generalization assessment across all models.

These preprocessing steps ensured that the dataset was structured and formatted optimally for both statistical and transformer-based approaches, enabling robust binary classification performance analysis.

# **Methodology**

This section discusses the implementation of both traditional and modern approaches for text classification, specifically addressing the task of semantic understanding and classification. The methodology is divided into two distinct approaches: a classical model using TF-IDF with Naïve Bayes, and a transformer-based model using BERT.

## Classical Model: TF-IDF and Naïve Bayes

Feature Extraction with TF-IDF

The term Frequency-Inverse Document Frequency (TF-IDF) approach was implemented as our feature extraction method for the classical model. This statistical evaluates the importance of words within documents relative to a corpus by:

* Calculating Term Frequency (TF): Measures how frequently a term appears in a document.
* Computing Inverse Document Frequency (IDF): Reduces the weight of terms that appear in many documents.
* Combining these metrics (TF -IDF): Produces a composite weight for each term in each document.

This transformation converts text data into numerical feature vectors that captures the relative importance of words, mitigating the impact of commonly occurring terms. The implementation uses scikit-learn’s TfidfVectorizer, which allows for customization of parameters such as n-gram range, maximum features and minimum document frequency.

**Naïve Bayes Classification**

Two variants of the Naïve Bayes algorithm were implemented for the text classification task:

* **MultinomialNB:** This variant is well-suited for text classification problems where input features are discrete, such as term frequencies or TF-IDF values. It models the likelihood of features using multinomial distribution, which is appropriate for datasets where features represent the count or frequency of word occurrences in text documents. MultinomialNB is a standard and effective baseline for spam detection and document categorization tasks due to its simplicity and strong performance on high-dimensional data [7].
* **GaussianNB:** In contrast, Gaussian Naïve Bayes assumes that the features follow a normal (Gaussian) distribution. While this variant is generally used for continuous data, we included it in our experiment to observe how it performs on TF-IDF-transformed text features. As expected, the assumption of continuous Gaussian-distributed features made it less effective than MultinomialNB for our specific application, where feature values represent sparse, non-negative frequencies [8].

Both implementations leverage the conditional independence assumption between features given there is the class label, allowing for proper parameter estimation and prediction.

**Baseline Advantages**

The TF-IDF with naïve Bayes approach is an excellent baseline for many reasons:

* Computational Efficiency: Training and inference are extremely fast, with O(n) complexity, making it a good choice for resource constrained environments.
* Minimal Hardware Requirements: Can be implemented on standard Cpu’s without specialized hardware requirement.
* Effectiveness with Limited Data: Performs relatively well even with small training datasets.
* Simplicity: Requires minimal hyperparameter tuning compared to deep learning approaches.

Despite its limitation in capturing word order and contextual relationships talked about in the introduction, this classical approach provides a strong benchmark against which more complex models can be evaluated.

## Transformer-Based Model: BERT

**Architecture Overview**

For our newer approach, we implemented the BERT-base-uncased model, a transformer-based architecture that has revolutionized NLP tasks through contextual word representations. The specific implementation uses:

* BERT-base-uncased: A **12-layer,768-hidden,12 head,110M** parameter model trained on lower-cased English text.
* Bidirectional Context: Unlike traditional models that process text in one direction, BERT processes words with all other words in a sentence.
* Pre-training + Fine-tuning: Leveraging transfer learning by using a model pre-trained on a large corpus and fine-tuning it on our specific classification task.

**Tokenization Process**

The tokenization process converts raw text into a format suitable for BERT processing:

A screenshot of a computer

AI-generated content may be incorrect.

Figure : Implementation of Tokenization before processing into BERT Model

This process involves:

* Word Piece Tokenization: Breaking words into sub word units to handle out-of-vocabulary words.
* Special Token Addition: Inserting [CLS] at the beginning and [SEP] at the end of sequences.
* Input ID Generation: Converting tokens to their corresponding numerical IDs.
* Attention Mask Creation: Differentiating actual tokens from padding tokens (1s for real tokens,0s for padding).

**Fine-tuning Setup**

The fine-tuning process was carefully configured to optimize model performance:

* Training Duration: 3 Epochs
* Batch size: Implemented with consideration for available computational resources
* Optimization:
* Optimizer: Adam W with weight decay to prevent overfitting
* Learning rate: Small learning rate (typically 2e-5 to 5e-5) to preserve pre-trained knowledge
* Learning Rate Scheduler: Linear decay with linear warmup
* Loss Function: Cross-entropy loss for classification tasks
* Evaluation Metrics: Accuracy, Precision, recall and F1-score

The training process achieved a remarkably low training loss of 0.073, indicating strong model performance on the classification task.

**Implementation Advantages:**

The BERT-based approach offers several advantages over classical methods:

* Contextual Understanding: Captures semantic relationships between words based on their context
* Transfer learning: leverages knowledge from pre-training
* Reduced Feature Engineering: Eliminates the need for manual feature extraction

The implementation we discussed shows how Bert’s bidirectional architecture enables deeper semantic understanding compared to the traditional models. BERT not only gives advantages over traditional model but performs better in accuracy and interpretability of the corpus.

# **Experiments and Results**

For evaluating the spam detection model effectiveness, comparison of three approaches (BERT, Multinomial Naïve Bayes, and Gaussian Naïve Bayes) was done. Models were tested on SMS spam dataset and evaluation was done on basis of accuracy, precision, recall, F1-score and false positive rate (FPR).

The results are summarized in the table and ROC curve below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **FPR** |
| **BERT** | 99.00% | 100.00% | 94.44% | 97.14% | 0.00% |
| **MultinomialNB** | 96.77% | 100.00% | 77.22% | 87.14% | 0.00% |
| **GaussianNB** | 88.79% | 56.73% | 87.97% | 68.98% | 11.08% |

Table : Performance comparison table BERT vs Naive Bayes

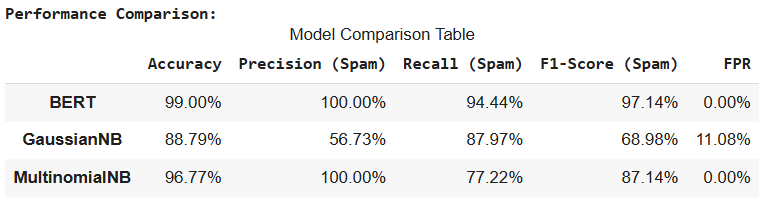


Figure : Model Comparison table

A graph of a graph showing different types of data

AI-generated content may be incorrect.

Figure : ROC curve Comparison- BERT vs Naive Bayes

Furthermore, the ROC curve demonstration shows that BERT has achieved the highest AUC(Area Under the Curve) of 1.00. MultinomialNB has AUC of 0.98 and GaussianNB scored 0.88. Based on the results above, conclusion can be drawn that Transformer-based models outperforms tradional statistical methods in classification, confidence and generalization.

# **Discussion**

Both BERT and Naïve Bayes on curated set of 10 tricky messages was tested. These included phishing-style prompts and contextually misleading phrases—typical of real-world spam scenarios. As shown in the figure below, BERT achieved an accuracy of 90%, while Naive Bayes only reached 70%:

A screenshot of a computer error

AI-generated content may be incorrect.

Figure 4: Classification comparaison on Trickey sentences

One of the main drawbacks of traditional models like Naive Bayes is that they focus only on surface-level features, such as how often certain keywords appear. They treat text as if it’s just a collection of words, completely overlooking the order, context, and intent behind the message. This limitation makes them prone to misclassifying genuine messages that just happen to include spam-like keywords.

Take, for example, how Naive Bayes misclassified these sentences.

"Did you bring your textbook for the class?"

"Project deadline has been moved to next week."

Clearly, these are legitimate messages, but Naive Bayes flagged them as spam simply because they contained words like "textbook" or "deadline," which often show up in spam training data.

On the other hand, BERT takes a different approach. It’s pre-trained on a vast amount of natural language using techniques like masked language modeling and next sentence prediction. This allows it to grasp the semantic meaning of words in context. For instance, "click here" in a phishing email carries a very different connotation than "click here to view your class notes."

Thanks to this contextual understanding, BERT can pick up on the subtle intent behind sentences, accurately identifying legitimate academic, conversational, and transactional messages as genuine, even when they might superficially look like spam.

# **Conclusion**

Our findings show that BERT’s deep contextual modeling brings some serious advantages over traditional spam detection methods. Unlike Naive Bayes, which just counts word frequency and misses the bigger picture, BERT actually understands meaning through sentence structure, intent, and context. This ability to grasp semantics is crucial in today’s world of clever spam and phishing attacks, where scammers carefully design messages to dodge common spam words while still trying to trick users. That’s why transformer-based models like BERT aren’t just accurate; they’re also tough against the ever-changing tricks of language. This makes them a much better fit for critical applications like email servers, messaging platforms, and enterprise security systems. Looking ahead, we might want to dive into smaller versions like DistilBERT or TinyBERT, which could deliver similar semantic performance without the heavy computational load — making real-time deployment on edge devices a reality.

# **References**

|  |  |
| --- | --- |
| [1] | O. K. A. R. Anna Rogers, "A Primer in BERTology: What We Know About How BERT Works," *Transactions of the Association for Computational Linguistics,* vol. 8, p. 842–866, 2020. |
| [2] | M. J. M. B. R. C. S. J. M. V. A. Juvi C. Tesoro, "A Semantic Approach of the Naïve Bayes Classification," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 9, no. 3, p. 3483–3487, 2020. |
| [3] | M.-W. C. K. L. K. T. Jacob Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, Minnesota, 2019. |
| [4] | M. A. I. M. A. R. A. H. A. M. U. Ekramul Haque Tusher, "IEE Xplore," 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10693429. |
| [5] | D. M. M. Thaer Sahmoud, "ResearchGate," June 2022. [Online]. Available: https://www.researchgate.net/publication/361134881\_Spam\_Detection\_Using\_BERT. [Accessed April 2025]. |
| [6] | M. M. K. M. Y. Z. M. K. H. Ghazala Nasreen, "ScienceDirect," June 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110866524000367. [Accessed April 2025]. |
| [7] | K. N. Andrew McCallum, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI, 1998. |
| [8] | L. S. J. T. D. R. K. Jason D. M. Rennie, "Tackling the poor assumptions of naive Bayes text classifiers.," in *ICML'03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 2003. |