Spam Classification

1. Problem Definition & Motivation

Spam emails have transformed from simple mass-advertising to **high-risk social engineering attacks**. These messages are crafted to mimic real communication, using friendly tone, personalized details, and even AI-generated text, making it harder to distinguish from legitimate emails.

Incorrect classification leads to serious consequences:

- **False negative (missed spam/phishing):**
  user receives harmful links, malware, or fraudulent financial requests
- **False positive (blocked legitimate email):**
  important emails never reach the user

Traditional filters that search for suspicious words (like *"free money"*) are no longer enough. Modern spam requires a classifier that learns **linguistic patterns**, not just keywords.

Research question:
Can classical NLP (TF-IDF + Multinomial Naïve Bayes) still classify spam effectively in 2025, and how much does text preprocessing improve performance?

2. Used Datasets

To understand spam across different contexts, three datasets were selected intentionally each representing a different era and communication medium.

Ling-Spam Corpus (primary dataset)
Source: http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz
Contains emails from a linguistics mailing list mixed with real spam.
This dataset uniquely comes in four versions (raw, stop-words removed, lemmatized, and lemmatized + stopwords), allowing us to test how text preprocessing affects performance.
Its 10-fold partition design makes it perfect for controlled experimentation.

Spam Email Classification (modern dataset)
Source: https://www.kaggle.com/datasets/ashfakyeafi/spam-email-classification
Contains 5158 real-world spam and phishing emails written with modern persuasive techniques.
Unlike Ling-Spam, these messages reflect today's attackers who intentionally write professional-sounding emails.

SMS Spam Collection
Source: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
A widely used dataset of 5,574 mobile text messages labeled as spam or ham.
This dataset reveals additional challenges, shorter messages, abbreviations, informal tone, and shows how spam classification behaves on compact text.

Together, these datasets let us study how spam **evolved over time**:
mailing list spam → modern phishing email → SMS/mobile spam.

3. Methodology

The project follows a standard NLP-based machine learning workflow.
First, the dataset is loaded, and each email is labeled as either **spam (1)** or **ham (0)**. After labeling, the text is cleaned so the model focuses only on meaningful language.

Preprocessing steps include:

- converting all text to lowercase
- removing punctuation and unnecessary characters
- removing common stop-words (e.g., *"the," "and," "of"*)
- **lemmatization**, which reduces words to their base form (e.g., *"running"* → *"run"*)

Once the text is clean, the emails are transformed into numeric features using **TF-IDF**, a technique that highlights important terms (like *"bank account"*) and reduces weight for common words.

Next, the processed data is used to train a **Multinomial Naïve Bayes classifier**, selected because prior research shows it is highly effective for text classification tasks.

Finally, model performance is evaluated using **10-fold cross-validation**, ensuring that every part of the dataset is used for both training and testing, so the model does not overfit.

Evaluation metrics:

- **Accuracy** – how often predictions are correct
- **Precision** – avoiding false spam flags (blocking real emails)
- **Recall** – avoiding missed spam (letting dangerous emails through)
- **F1-score** – balance between precision and recall

The main objective is to measure how different preprocessing steps influence the accuracy of Naïve Bayes in detecting spam.

4. Prior Research (Literature Review)

1. Androutsopoulos et al. (2000) introduced the Ling-Spam corpus and demonstrated that Naïve Bayes performs well for email classification, especially when lemmatization and stop-word removal are used. This paper sets the foundation of our experiment.
2. Klimt & Yang (2004) released the Enron Email Corpus, a large dataset of real corporate emails. They showed that many models trained on "clean academic datasets" fail when applied to real inbox behavior, demonstrating the importance of realistic data.

3. Metsis et al. (2006) compared different types of Naïve Bayes models and proved that **Multinomial Naïve Bayes performs best for text classification**, providing direct justification for our model choice.
4. Cormack (2007) introduced **TREC Spam Track**, where spam was evaluated chronologically rather than randomly. This revealed that spam evolves over time and filtering systems should be tested under realistic time-based conditions.
5. Almeida et al. (2011) worked on SMS spam classification and discovered that as dataset size increases, models like SVM can outperform Naïve Bayes. This shows that classical models are efficient but may hit performance limits on large datasets.
6. "Email Spam Detection using Deep Learning and Novel Feature Selection" (2023) explored deep learning methods and demonstrated that better feature selection significantly boosts deep learning accuracy, even outperforming classical ML.
7. "Real-Time Phishing Detection Using DistilBERT" (2024) applied transformers to email phishing classification and achieved extremely high accuracy (>99%) by understanding context, not just individual words.

In summary:
Classical Naïve Bayes is simple, explainable, and effective but deep learning now dominates when dealing with modern, realistic phishing emails.

5. Expected Outcome

At the end of Project-1, we expect to:

- build a working spam classifier
- measure and compare model performance across preprocessing variations
- demonstrate whether classical ML can still compete in 2025

Expected accuracy: **~95 - 97%**, based on prior work.

6. References (APA)

1. Androutsopoulos et al., 2000 – Ling-Spam / Naive Bayes foundation
Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). *An evaluation of Naive Bayesian anti-spam filtering*. In Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (pp. 9–17), Barcelona, Spain.

2. Klimt & Yang, 2004 – Enron corpus dataset
Klimt, B., & Yang, Y. (2004). *The Enron corpus: A new dataset for email classification research*. In European Conference on Machine Learning (ECML) (pp. 217–226). Springer.

3. Metsis, Androutsopoulos, & Paliouras, 2006 – Which Naive Bayes performs best
Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). *Spam filtering with Naive Bayes: Which Naive Bayes?* In 3rd Conference on Email and Anti-Spam (CEAS).

4. Cormack, 2007 – TREC Spam Track (real-world evaluation)
Cormack, G. V. (2007). *TREC 2007 Spam Track overview*. University of Waterloo.

5. Almeida et al., 2011 – SMS Spam Collection dataset
Almeida, T. A., Gómez-Hidalgo, J. M., & Yamakami, A. (2011). *Contributions to the study of SMS spam filtering: New collection and results*. In Proceedings of the ACM Symposium on Document Engineering (DocEng).

6. Nasreen et al., 2024 – Deep learning + BERT hybrid spam detection
Nasreen, G., Khan, M. M., Younus, M., Zafar, B., & Hanif, M. K. (2024). *Email spam detection by deep learning models using novel feature selection technique and BERT*. Egyptian Informatics Journal.

7. Damatie et al., 2024 – Real-time phishing using DistilBERT
Damatie, E. M., Eleyan, A., & Bejaoui, T. (2024). *Real-time email phishing detection using a custom DistilBERT model*. In 2024 International Symposium on Networks, Computers and Communications (ISNCC). IEEE.