**1 Exploratory Data Analysis (EDA)**

**Class balance.** The corpus contains 18,650 messages with a mild imbalance (Safe ≈ 11k; Phish ≈ 7k). *Figure 1* illustrates this distribution and motivates the later use of class weighting in the classifier.

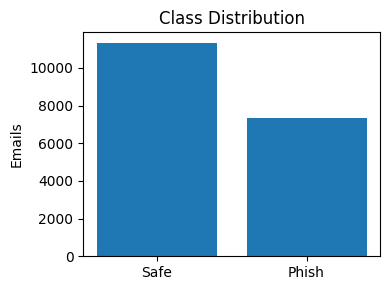


Figure Class Distribution (Safe vs Phish)

**Message length.** After trimming the longest 1 % outliers, 90 % of e-mails are shorter than 300 words (*Figure 2*). This informs the choice of a simple bag-of-words model instead of sequence architectures.

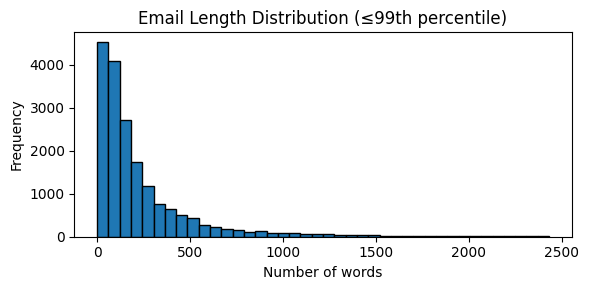


Figure E-mail Length Histogram (≤ 99th percentile)

**Lexical fingerprints.** TF-IDF averages reveal distinct vocabularies: phishing mails emphasise action and reward (*click, free, money*), whereas legitimate mails reference corporate or archival terms (*enron, 2002, thanks*) as illustrated in *Figures 3* (Phish) and 4 (Safe).

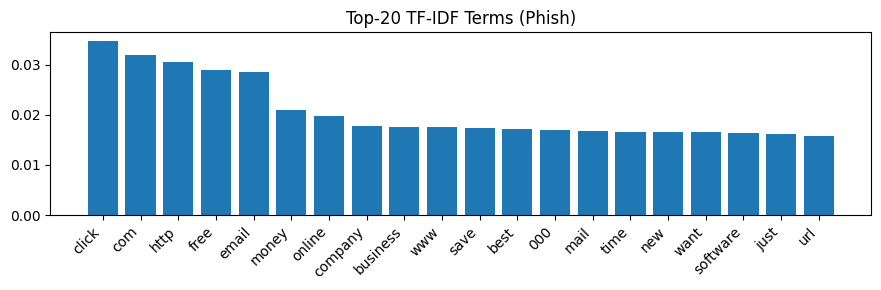


Figure Top-20 TF-IDF Terms – Phish

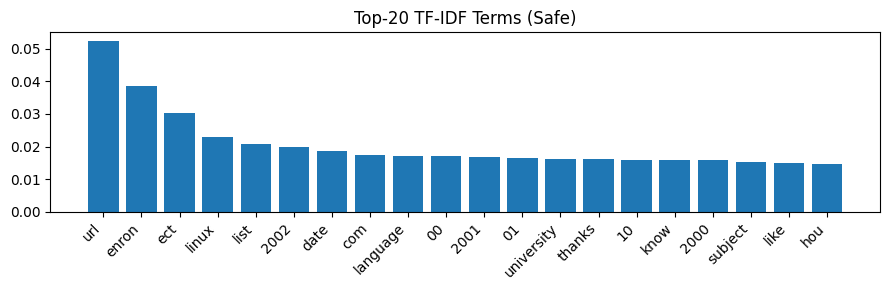


Figure Top-20 TF-IDF Terms – Safe

**2 Feature-Selection Methods and Results**

Two filter techniques were applied to the 3,000-term TF-IDF matrix:

1. **Chi-square (χ²) statistic** — ranks tokens by dependence on the class label. High-score terms such as *click, viagra, and offer* are strong phishing indicators (*Figure 5*).

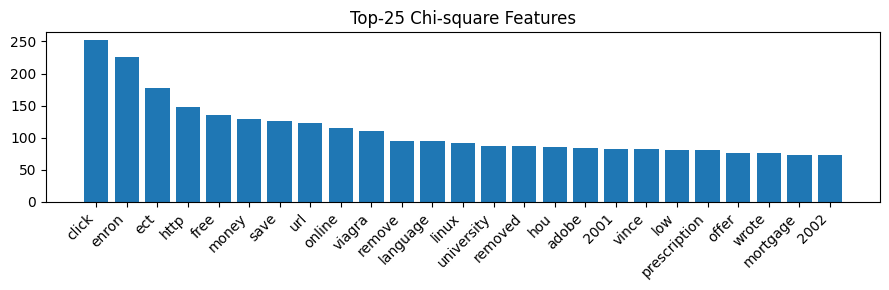


Figure Top-25 χ²-Score Features

1. **Mutual Information (MI)** — measures information gain. The top tokens (*com, time, information*) include both spam and ham cues, yielding a more balanced set (*Figure 6*).

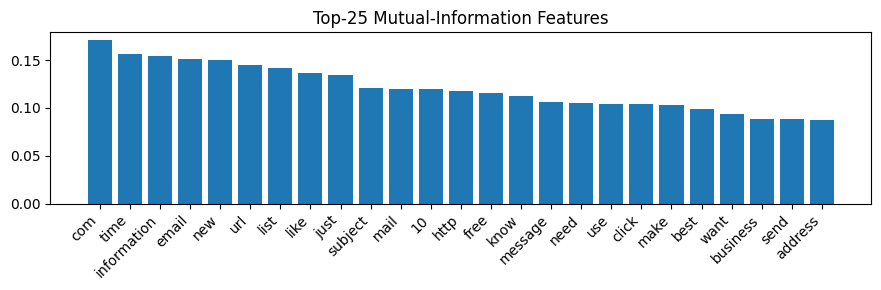


Figure Top-25 Mutual-Information Features

These ranked lists reduce dimensionality for future ablation studies while retaining semantically meaningful features.

**3 Pre-processing Pipeline and Outcomes**

|  |  |  |
| --- | --- | --- |
| **Stage** | **Action** | **Rationale** |
| Text normalization | lower-case, punctuation removal, URL → token *url* | Reduces sparsity; abstracts session-specific links. |
| Stop-word removal | built-in English list | Removes high-frequency functional words. |
| Tokenisation & TF-IDF (1-2 g) | vocab = 3,000, max\_df=0.9 | Captures phrase patterns; limits dimensionality. |
| Train–test split | 80 20, stratified | Preserves class ratio (train = 14 920; test = 3 730). |

The cleaned corpus (18,363 rows after minor null removal) was vectorised in under 10 s and is stable for reproducible experiments.