Spam Email Text Analysis and Classifier

Jyun-Hao Chen
University of Wisconsin, Madison, USA

1 Introduction

Spam emails are pervasive, with billions of messages containing harmful content sent daily. These messages cause inconvenience and frustration for recipients and may lead to phishing scams, malware, and viruses. As a result, accurately classifying spam emails and distinguishing them from legitimate ones is essential to safeguard user's personal information and ensure their protection.

As someone who has personally experienced the effects of spam emails, including scam phone calls, text messages, and phishing emails, I understand the importance of filtering out fraudulent information. In today's world, where people are inundated with vast information, distinguishing relevant information from noise can be challenging. Therefore, developing effective strategies to classify spam emails and protect users' personal information is critical.

2 Background

- Information Overload Theory [1]: Spam emails result from information overload, and it serves as a method for companies to get their message across to potential customers in a sea of information.
- Social Influence Theory [2] : Social factors, such as peer pressure or social norms, influence clicking on spam emails, making people more vulnerable to phishing scams.

3 Methods

- Data Import and Preprocessing:
  Combined and changed two sets columns to ‘Label’ and ‘text’.
  Concatenated and dropped duplicate rows to (10448, 2).
- Text Cleaning and Visualization [3]:
  Used nltk (PorterStemmer, word_tokenize, stopwords) to clean texts.
  Generated the spam and ham s’ WordCloud diagram.

Before: LIFE has never been this much fun and great until you came in. You made it truly special for me. I won't forget you! enjoy @ one gbp/sms
After: life never much fun great came made truli special wo forget enjoy one

- Train-Test Split and Naive Bayes Classifier [4]:
  Compared three Naive Bayes classifiers (ComplementNB, BernoulliNB, and MultinomialNB) and visualized.

4 Results

Word Cloud

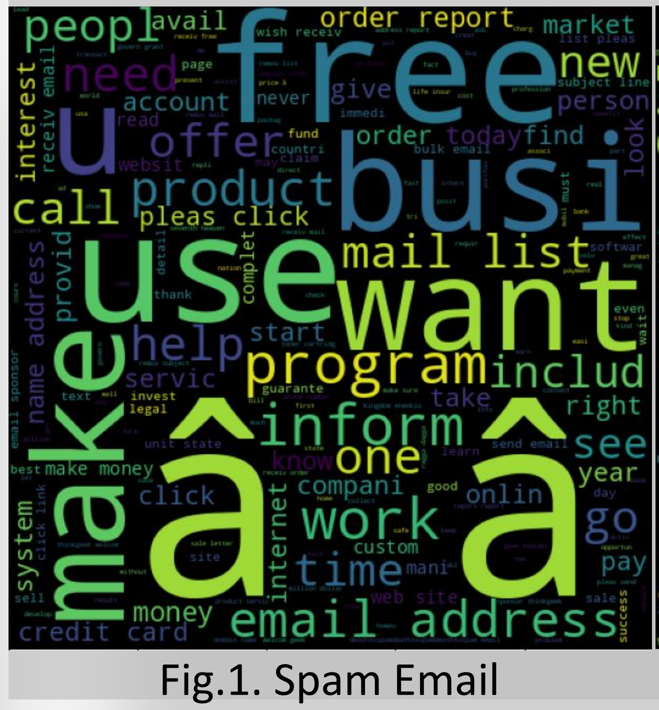


Fig.1. Spam Email

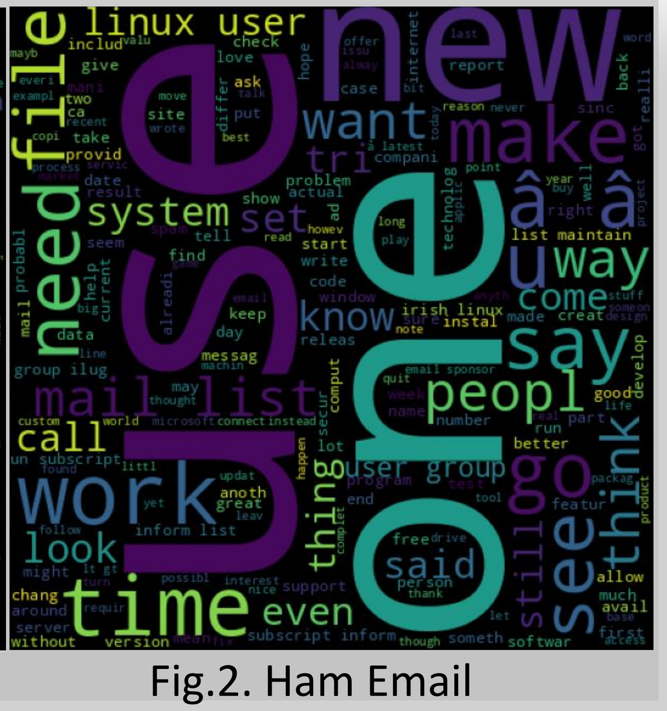
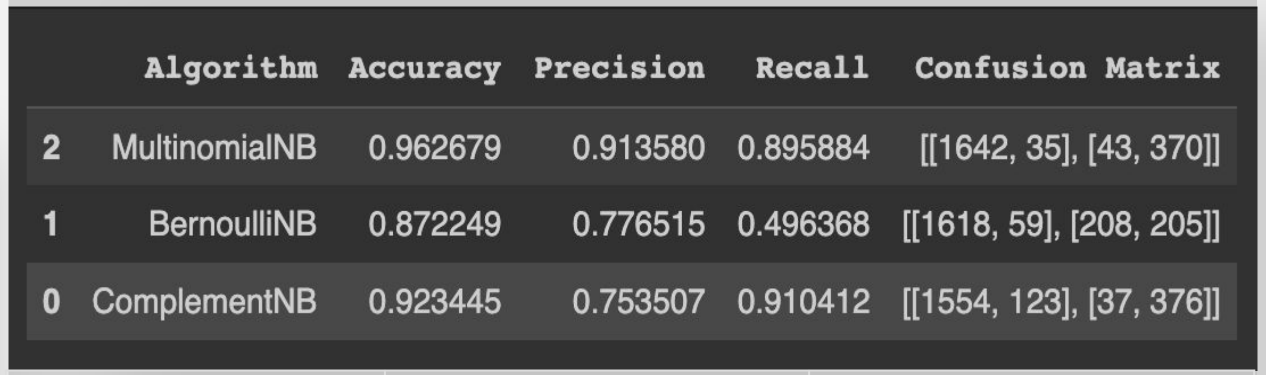


Fig.2. Ham Email

Naive Bayes Classifier [5]



	Algorithm	Accuracy	Precision	Recall	Confusion Matrix
2	MultinomialNB	0.962679	0.913580	0.895884	[[1642, 35], [43, 370]]
1	BernoulliNB	0.872249	0.776515	0.496368	[[1618, 59], [208, 205]]
0	ComplementNB	0.923445	0.753507	0.910412	[[1554, 123], [37, 376]]

Confusion Matrix	Predicted: Real Email	Predicted: Spam Email
Actual: Real Email	True Negatives (TN)	False Positives (FP)
Actual: Spam Email	False Negatives (FN)	True Positives (TP)

5 Conclusion

- Spam emails use words like "free," "receive," and "money" to entice readers.
- Similar words like "get," "order," "use," and "click" are used in both spam and legitimate emails.
- MultinomialNB has the best predict results.
- It can be challenging to distinguish spam from legitimate emails based on text alone.
- Other factors like sender, title, typesetting, foreign characters, and suspicious links should be considered to identify fraudulent emails effectively.

6 Future

- Create a suite of Chrome plug-ins for detecting harmful email content.
- Customized kit tailored to individual usage habits can provide comprehensive protection.
- Incorporate new data sets, algorithms, and multimedia scanning to enhance the capabilities of the mail filter.

References
[1] Roman Soucek, Klaus Moser(2010, Nov.). Coping with information overload in email communication: Evaluation of a training intervention, Computers in Human Behavior(Vol 26, Iss. 6, pp. 1458-1466). https://doi.org/10.1016/j.chb.2010.04.024
[2] Kathryn Parsons, Marcus Butavicius, Paul Delfabbro, Meredith Lillie(2019, Aug.) Predicting susceptibility to social influence in phishing emails, International Journal of Human-Computer Studies(Vol. 128, pp. 17-26). https://doi.org/10.1016/j.ijhcs.2019.02.007
[3] SHEKHAR BANERJEE(2023, Jan.). Spam classifier on SMS/Email dataset. https://www.kaggle.com/code/shekharbanerjee/spam-classifier-on-sms-email-dataset
[4] DEEP PATEL(2023. Mar.) Email_spam_detection_using_naive_bayes_classifier. https://www.kaggle.com/code/deeppatel9095/email-spam-detection-using-naive-bayes-classifier
[5] Hugo Ferreira(2018, Apr.) Confusion matrix and other metrics in machine learning. https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning-894688cb1c0a