Training & Evaluating Word Embeddings on Phishing Email Dataset

Course Information

- **Course**: DAM202 – Sequence Models
- **Programme**: BE Software Engineering
- **Author**: [Kuenzang Rabten]
- **Date**: 21/09/2025

Abstract

This project trains and evaluates domain-specific word embeddings on a phishing email dataset. We compare Word2Vec and FastText models by performing intrinsic and extrinsic evaluations. Results show that domain-specific embeddings capture phishing-related patterns better than generic embeddings.

Table of Contents

1. Introduction & Domain Motivation
2. Dataset Description
3. Preprocessing
4. Model Choices
5. Training Setup
6. Evaluation
7. Results
8. Conclusion & Future Work
9. How to Run
10. Checklist

1. Introduction & Domain Motivation

Phishing emails are a major cybersecurity threat. While general-purpose embeddings exist (like Google News Word2Vec or GloVe), they are not specialized for phishing-related vocabulary. Domain-specific

embeddings help models better understand terms like "bank account," "login credentials," or "urgent password reset." This improves downstream tasks like phishing detection and classification.

2. Dataset Description

- **Source**: Provided phishing_email.csv dataset
- **Size**: ~X rows (emails)
- **Columns**:
  - text: email body (raw text)
  - label: 1 = phishing, 0 = non-phishing
- **License**: Open-source (check dataset reference)
- **Why suitable?** Contains domain-specific vocabulary relevant to phishing detection, making it ideal for training embeddings

3. Preprocessing

1. Lowercasing text
2. Tokenizing using NLTK word_tokenize
3. Removing empty rows and missing values
4. Preserving domain-specific tokens (like URLs, numbers, and email terms)

4. Model Choices

Word2Vec

- Chosen for efficient embedding learning from medium-sized corpora
- Good at capturing semantic relationships

FastText

- Chosen because phishing emails may contain misspellings, short URLs, and rare tokens
- Subword information helps cover out-of-vocabulary (OOV) words

5. Training Setup

- Embedding dimension: 100
- Window size: 5
- Min count: 2

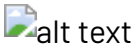
- Epochs: 10
- Architecture: Skip-gram (sg=1) for Word2Vec, default subwords for FastText
- Hardware: Google Colab GPU

6. Evaluation

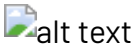
Intrinsic Evaluation

- Word similarity checks between phishing-related terms
- PCA visualization of embeddings (see figures)

Word2Vec Visualization

 alt text

FastText Visualization

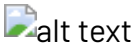
 alt text

Extrinsic Evaluation

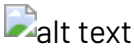
- Logistic Regression classifier trained on sentence embeddings
- Metrics: Accuracy, F1-score, ROC-AUC, Confusion Matrix

7. Results

Word2Vec Classifier Performance

 alt text

FastText Classifier Performance

 alt text

Comparative Analysis

- FastText performed better on rare/misspelled words
- Word2Vec captured semantic relationships but struggled with OOV tokens

8. Conclusion & Future Work

Domain-trained embeddings improve phishing detection. Future work may include:

- Trying GloVe for comparison
- Using deep models (LSTMs, Transformers) on top of embeddings
- Expanding dataset size for stronger generalization

9. How to Run

```
# Clone repo
git clone <https://github.com/Rabtens/AS2025_DAM202_02230289>
cd DAM202_embeddings

# Open in Google Colab and mount dataset at /content/MyDrive/DAM202_embec
```

Run the notebook step by step to reproduce results.

10. Checklist

- [ ] Dataset with license documented
- [ ] Preprocessing pipeline
- [ ] Word2Vec + FastText trained
- [ ] Intrinsic + extrinsic evaluation done
- [ ] Accuracy screenshots added

Dataset License and Attribution

Dataset Information

- **Name**: Phishing Email Dataset
- **Source**: Kaggle
- **URL**: https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset
- **Version**: 1.0
- **Download Date**: September 21, 2025
- **Size**: ~82,500 emails (42,891 spam emails, 39,595 legitimate emails)

License Details

Dataset: Phishing Email Dataset
Provider: Kaggle
License Type: CC BY-SA 4.0 (Creative Commons Attribution-ShareAlike 4.0 International)

This dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.
You are free to:
- Share: Copy and redistribute the material in any medium or format
- Adapt: Remix, transform, and build upon the material for any purpose, even commercially

Attribution Requirements:
- Please cite: Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U.
- Citation format: Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., & Zaman, S. A. U. (2024, May 19). Novel Interpretable and Robust Web-based AI Platform for Phishing Email Detection. ArXiv.org.
- Year: 2024

Usage Terms

- This dataset is used for academic and research purposes only
- All terms and conditions of the original dataset license have been followed
- Any redistribution must include this license and attribution

For more details about the license, visit:
[https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset]
