Machine learning for text classification in classical cryptography

Floe Foxon
University of Leeds
United Kingdom
floefoxon@protonmail.com

Abstract

This study furthers previous work on text classification to distinguish between ciphertext and gibberish. The statistical/linguistic properties of four text types were studied: meaningful English text, and three gibberish types ($n = 1,250$ each; total $N = 5,000$). Dimension reduction techniques (PCA, t-SNE, and UMAP) were used to reduce the statistical/linguistic feature space of the texts to two dimensions, revealing distinct regions of (lower dimensional) feature space occupied by each text, with some overlap. Machine learning models including random forests, neural networks (NNs), and support vector machines (SVMs) were used to classify the four text types based on their statistical/linguistic properties. Nested cross-validation revealed better generalization performance for the NNs and SVMs, classifying texts with $> 90\%$ accuracy. Applied to the Dorabella cryptogram, the models suggest that this text resembles meaningful English text more closely than gibberish types, which comports with the Dorabella cryptogram as a monoalphabetic substitution cipher, but this classification should be interpreted with caution. Features that better separate meaningful English from English-like gibberish are needed, and other encryption schemes/cryptograms should be explored with these methods.

1 Introduction

Previous work (Foxon, 2024) explored the the possibility of distinguishing between English text and 'gibberish' text generated at random from a uniform probability distribution in the context of classical cryptography, namely for the identification of 'hoax' cryptograms. The statistical and linguistic characteristics of those two datasets (English and uniform gibberish) were calculated, and a simple neural network (no hidden layers) was trained to distinguish between the two. The previous model had a high cross-validated mean accuracy (99.8%), and classified the Dorabella cryptogram as most likely representing English text (probability 0.9999) as opposed to uniform gibberish (probability 0.0001).

While taking a necessary first step toward successful meaningful/gibberish classification for putative cryptograms, the previous work had significant limitations in its scope. First, uniform gibberish is not the only alternative to meaningful English text that may be represented by a putative cryptogram. Second, the actual differences between the English and uniform gibberish texts were obscured by the high dimensionality of the data; no attempt was made to visualize the feature space. Third, the model in the previous work was selected arbitrarily with no optimization.

This work expands upon the previous work in three significant ways. First, the previous binary dataset is expanded to include two additional categories of text besides English and uniform random to potentially improve classification. Second, these texts are visualised in lower dimensional feature space using dimension reduction techniques to assess their separability. Third, a comprehensive approach to model selection is taken with nested cross-validation, comparing different types of machine learning models across different model hyperparameters.

Proceedings of the 8th International Conference on Historical Cryptology, pages 60-64
Poznań, Poland, 16-18 June, 2025

2 Methods

The methods used in this study are described in detail in the online supplemental materials, which contain a fully documented Jupyter notebook with statistical analysis code and text description along with the data:
`https://doi.org/10.17605/OSF.IO/ZYMTD`.
These methods are summarised only succinctly in the following. Readers are directed to the supplement for elaboration.

2.1 Data

As in the previous work, a training and testing set of meaningful English-language plaintexts was generated from Wikipedia articles (Davies, 2015), and a training and testing set of uniform gibberish was generated from a uniform probability distribution in which each character (lowercase English alphabet only) had equal probability $\left(\frac{1}{26}\right)$ of being selected next. These two types of text are referred to as 'English Corpus' and 'Uniform Random' hereafter.

Two new types of text were defined for the present study as follows.

- 'English-Like Random' text was generated by taking the frequency distribution of the English Corpus text and generating gibberish in which the letter 'e' had the highest probability of being selected next ($p \approx 0.116$), and 'q' the lowest ($p \approx 0.001$), as in the English Corpus.

- 'Typewriter-Like Random' text was generated, in theory, by a human typing 'at random' on a typewriter. The text used to represent Typewriter-Like Random text was the lowercase English alphabet characters of the 'Rilke cryptogram', for which there is some statistical evidence that the 'cryptogram' is a random sequence created with a typewriter (Foxon, 2022). While this identification is not certain, it was hypothesised that the Rilke text effectively represents a distinct category from the other three text categories in this study (later confirmed in the results). Readers may think of this category as 'Unknown' if preferred.

For each of the four text categories, $n = 1,250$ strings were generated for a well-balanced total of $N = 5,000$ strings, each 87-characters in length to match the Dorabella cryptogram.

As before, the values for 15 statistical and linguistic properties, including alphabet length, frequency, distance, entropy, index of coincidence, Zipf's exponent, and mean associated contact count were calculated for each string in each of the four text types.

2.2 Dimension Reduction and Data Visualisation

Since it is not possible to visualize the differences between each text in 15-dimensional space, dimension reduction techniques were used to reduce the number of dimensions of the feature space while retaining much of the variance. Three dimension reduction techniques were applied to the four text types in this study, namely Principal Component Analysis (PCA; Hotelling (1933); implemented with scikit-learn 1.3.0 (Pedregosa et al., 2011)), t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton (2008); implemented with scikit-learn) and Uniform Manifold Approximation and Projection (UMAP; implemented with the UMAP Python package; see McInnes and Healy (2018)). By reducing the original 15-dimensional feature space to two effective dimensions, it was possible to visualize the separability of the data and provide an initial prediction for the classification of the Dorabella cryptogram prior to rigorous modelling. Figures were generated with matplotlib 3.7.2 (Hunter, 2007).

2.3 Machine Learning Models

Multi-class classification of the four-text dataset was achieved with three machine learning approaches. Namely, random forests (RFs; Breiman (2001)), neural networks (NNs; Rosenblatt (1958)), and support-vector machines (SVMs; Cortes & Vapnik (1995)), all implemented with sci-kit learn. To estimate the generalization performance of these models, 3×3 nested cross validation (NCV) was implemented. Grid searching was used over the hyperparameter space of each model, including a range of maximum tree depths, impurity criteria, and number of estimators for the RFs; solvers, activation functions, learning rates, and hidden layer sizes for the NNs; and kernel types, polynomial degrees, and regularization parameter values for the SVMs, described in the supplemental materials. These models were applied to the full 15-dimensional dataset.

Finally, the most accurate models were applied to the Dorabella cryptogram to obtain a possible

identification of its text type.

Analyses were conducted in Python version 3.8.16 with the additional packages Numpy 1.24.3, Pandas 2.0.3, Scipy 1.10.1, and Natural Language Toolkit 3.9.1.

3 Results

Figure 1 shows the letter/unigram frequency distributions for each type of text. Notably, the English-Like Random and English Corpus types have identical distributions (by design), the Uniform Random distribution is flat, and the Typewriter-Like Random distribution lies somewhere in-between.

Figure 2 shows the results of the dimension reduction techniques, plotting each string in two-dimensional feature space, where each of the reduced dimensions is a function of the original 15 dimensions of the data. These plots suggest that each type of text occupies a relatively distinct area of the reduced feature space (or 'text space'), with some overlap, particularly between the English Corpus and English-Like Random sets. The Dorabella cryptogram lies somewhere in the boundary between the clusters of English Corpus and English-Like Random text, and is quite distinct in its properties from Uniform Random and Typewriter-Like Random text. Note however that this is only a projection of the feature space onto two dimensions; there may be significant difference between English Corpus and English-Like Random text obscured by the loss of variance in dimension reduction. These plots at least suggest that classification of the four text types is feasible.

Table 1 shows the testing and training accuracies for the machine learning models. NNs and SVMs had $\sim 94\%$ NCV testing accuracy, with little difference between testing and training accuracies, indicating little-to-no overfitting. RFs performed somewhat worse with $\sim 90\%$ NCV testing accuracy and much higher training accuracy, suggesting that the RF models were overfitted. NNs and SVMs were carried forward for Dorabella classification.

Predictions for the classification of the Dorabella cryptogram were similar for the final NN and SVM (Table 2), with both assigning the most probable classification to English Corpus text (probability ≥ 0.75), and the least probable classification to the English-Like Random text (probability ≤ 0.01).

Statistic	**RF**	**NN**	**SVM**
NCV Testing Accuracy % (SD)	90.4 (3.6)	93.8 (1.8)	94.0 (1.5)
NCV Training Accuracy % (SD)	99.1 (1.2)	95.0 (0.4)	95.3 (0.3)
Difference (pp)	8.7	1.2	1.3

Table 1: Nested cross-validation results for machine learning models.

Type	**NN**	**SVM**
English Corpus	0.75	0.81
Typewriter-Like Random	0.24	0.07
Uniform Random	0.01	0.11
English-Like Random	0.00	0.01

Table 2: Predicted probabilities for the Dorabella cryptogram by machine learning models.

4 Discussion

Comporting with the previous study, the present study reifies the use of machine learning models to classify text in historical cryptography, and found that the Dorabella cryptogram resembles meaningful English text more closely than gibberish types of text.

However, this result should be interpreted with caution. First, despite being more comprehensive than the previous binary dataset, the four text categories explored in this study do not capture all possible explanations, for example, other encryption schemes. Second, of the four text types in the present study, the final NN and SVM models were least accurate in classifying English Corpus text (though still with accuracies $> 90\%$).

This study also explored the concept of a 'textspace', i.e. a two-dimensional feature space for representing the characteristics of types of text related to historical cryptography. Results suggest that the textspace is helpful for determining the separability of the space, though the location of the Dorabella cryptogram so close to the boundary between English Corpus and English-Like Random - despite the machine learning models assigning very low probabilities to an English-Like Random classification - is not intuitive. This highlights the limitations of dimension reduction techniques.

Future work should find features that better separate meaningful English from English-like gibberish, apply these methods to other putative cryptograms, and explore other encryption schemes.

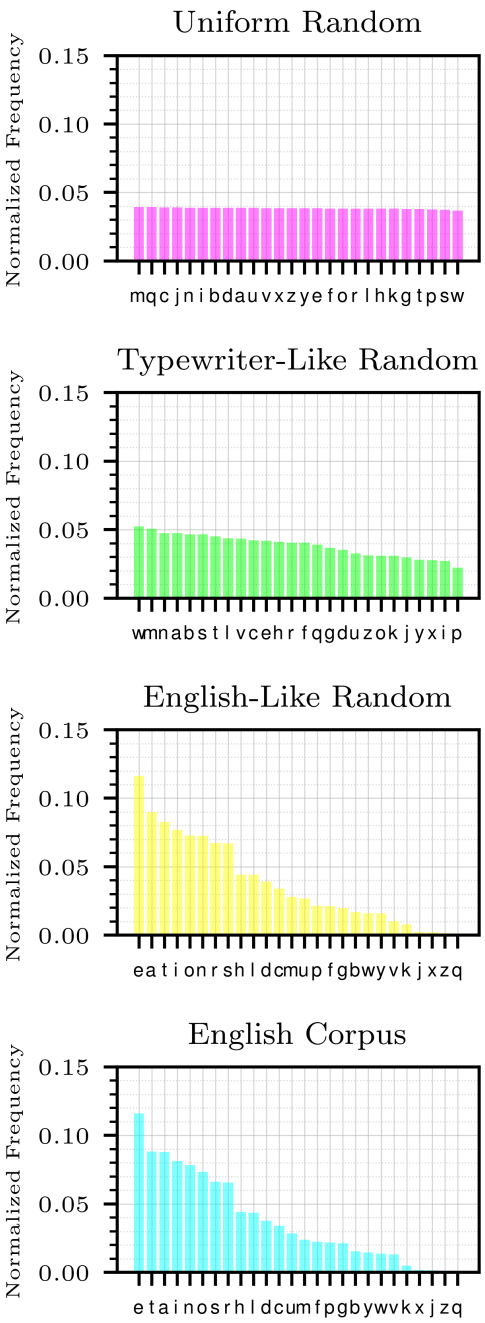

Figure 1: Letter/unigram frequency distributions for each text type.

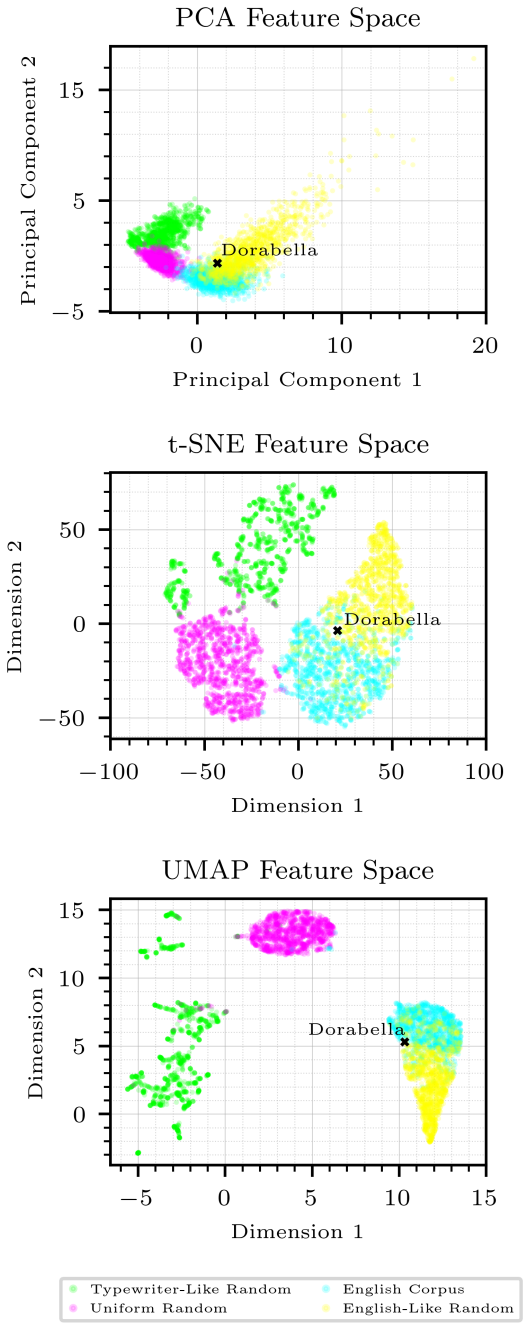

Figure 2: Dimension reduction of the original 15 features to two-dimensional feature space.

References

L. Breiman. 2001. Random Forests. *Machine Learning*, 45:5–32. https://doi.org/10.1023/A:1010933404324

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Machine Learning*, 20:273–297. https://doi.org/10.1007/BF00994018

M. Davies. 2015. The Wikipedia Corpus. https://www.english-corpora.org/wiki/

F. Foxon. 2022. A Treatise on the Rilke Cryptogram. *Cryptologia*, 47(6):493–510. https://doi.org/10.1080/01611194.2022.2092784

F. Foxon. 2024. Artificial Neural Network for Hoax Cryptogram Identification. *Proceedings of the 7th International Conference on Historical Cryptology, HistoCrypt*, pages 86–90. https://doi.org/10.58009/aere-perennius0094

H. Hotelling. 1933. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24:417–441. https://psycnet.apa.org/doi/10.1037/h0071325

J. D. Hunter. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

L. McInnes and J. Healy. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, 1802.03426. https://doi.org/10.48550/arXiv.1802.03426

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

F. Rosenblatt. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological review*, 65(6):386–408. https://psycnet.apa.org/doi/10.1037/h0042519

L. van der Maaten and G. Hinton. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html