Artificial neural network for hoax cryptogram identification

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

Abstract

Numerous putative cryptograms remain unsolved. Some, including the Dorabella cryptogram, have been suggested as hoaxes, i.e., some sort of gibberish with no meaningful underlying plaintext. The statistical properties of a putative cryptogram may be modelled to determine whether the cryptogram groups more closely with real or with randomly generated plaintext. Ten thousand plaintexts from an English-language corpus, and ten thousand (pseudo-)randomly generated English-alphabet gibberish texts were studied through their statistical properties, including the alphabet length; the frequency, separation, and entropy of n-grams; the index of coincidence; Zipf's law, and mean associated contact counts. An artificial neural network (deep learning) model was fitted to these data, with a cross-validated mean accuracy of 99.8% (standard deviation: 0.1%). This model correctly predicted that arbitrary, out-of-sample simple substitution ciphers represented meaningful English plaintext (as opposed to gibberish) with probabilities close to 1; correctly predicted that arbitrary, out-of-sample gibberish texts were gibberish (as opposed to simple substitution ciphers) with probabilities close to 1; and assigned a probability of meaningful English plaintext of 0.9996 to the Dorabella cryptogram.

1 Introduction

Lists of dozens of putative unsolved cryptograms have been published, such as the 'Top 50 Unsolved Encrypted Messages' by Klaus Schmeh (2023) and 'Famous Unsolved Codes and Ciphers' by Elonka Dunin (2023), and even these are not exhaustive. Some of these putative cryptograms have remained unsolved for many decades, such as the first and third Beale ciphers, published in 1885; the Voynich manuscript, which drew modern attention in 1912; and the Dorabella cryptogram, published in 1937.

It seems probable (though the author will not prove) that the longer a putative cryptogram goes undeciphered, the more likely it is to be identified as a hoax; that there is no cipher to be solved at all, and that the 'cryptogram' was only a fake, perhaps designed to attract media attention or sell merchandise. Indeed, the three putative cryptograms mentioned above have been described as 'bamboozlement' (Kruh, 1982), 'gibberish' (Gaskell and Bowern, 2022), and a 'full-fat hoax' (Pelling, 2019).

However, proving a negative is challenging (and often times practically impossible). Cryptanalysts must be wary of the example of Z340, a cryptogram that went unsolved for 51 years and was considered to be a possible pseudo-cipher (Juzek, 2019) before being solved completely in 2020 by Blake et al. (2021).

Thus, there is a need in classical cryptology to develop more sophisticated ways of distinguishing between real but unsolved cryptograms and actual hoaxes. In data science and statistics, a popular and effective way of categorizing data is with machine learning classification algorithms. In essence, this involves taking records (which may represent anything from animals to cryptograms) and applying an algorithm to these data which identifies, for each record, which category the record most likely belongs to. For example, the properties of vocalisations made by particular species of bird may be used to identify the species of an unknown bird (Qian et al., 2015).

These methods may be extended to cryptology by studing the linguistic and other statistical prop-

erties of putative cryptograms. With an appropriately trained model, a putative cryptograms of unknown status (real or hoax) may be identified rigorously and accurately (at least in theory).

The Dorabella cryptogram is of particular interest to the application of machine learning methods in this study due to its simplicity (having only a short alphabet) and brevity (having only 87 characters). The Dorabella cryptogram has been comprehensively described in other works (e.g. Bauer (2017)). Briefly, it is apparently a simple or monoalphabetic substitution cipher (MASC) prepared by Edward Elgar in an 1897 letter to an acquaintance named Dora Penny. It is assumed to be a MASC because the same symbols appear in definite MASCs in other cryptologic writings by Elgar. Despite its apparently simple encryption method, no solution has been generally accepted by the cryptologic community, hence its possible identification as a hoax (Elgar was allegedly known to be cruel), or as a different type of cipher. Cryptanalytic methods designed for MASCs have yet to yield a solution to the Dorabella cryptogram, but have identified interesting properties such as possible vowels (Schmeh, 2018).

The aim of the present study is to build an accurate machine learning model using statistical properties of cryptograms designed specifically for the Dorabella cryptogram. The model developed is used to determine whether the Dorabella cryptogram is statistically more likely to be real, or more likely to represent a hoax.

2 Methods

2.1 Data

To create a training and testing set of real (i.e., meaningful) English-language plaintexts, a collection of Wikipedia articles totalling 1.8 million English words were used as an English-language corpus (Davies, 2015). Ten thousand successive blocks of text were taken from the corpus, each 87 characters in length (the same length as the Dorabella cryptogram; all lowercase). This provided ten thousand real English plaintexts. An example of one of the 10,000 87-character texts from the English-language corpus used in the study is as follows:
turnwasinvestedwiththeduchyforhimselfand-
hisheirsalbertsruleinprussiawasfairlyprosperous
(from the Wikipedia article for Albert, Duke of Prussia).

To create a training and testing set of fake (i.e., gibberish) English-alphabet plaintexts, characters from the English alphabet (all lowercase) were pseudo-randomly selected (with replacement) to create a string of gibberish. Ten thousand such gibberish texts were generated, each 87 characters in length as above. An example of one of the 10,000 87-character gibberish texts used in this study is as follows:
idqbnjbnalbclvdfqypkzbwhddivepqjobbfrliplhusg-
onwshzdktdmbrtowispplvymrbsqzvhkramedbtdgk.

Since there were ten thousand texts in each of the real and fake sets, the data were evenly balanced, enabling a fair learning phase in the model.

The values of linguistic and other statistical properties were calculated for each text. These properties were as follows.

- Alphabet length: The unigram alphabet length is the number of unique single characters (unigrams) in the text. E.g., the text 'aabc' has a unigram alphabet length of 3 (the alphabet is the set { 'a', 'b', 'c' }). The bigram alphabet length is the same as above but for unique pairs of characters (bigrams). E.g., the text 'aabc' has a bigram alphabet length of 3 ({ 'aa', 'ab', 'bc' }). Finally, the trigram alphabet length is the same as above but for unique trios of characters (trigrams). E.g., the text 'aabc' has a trigram alphabet length of 2 ({ 'aab', 'abc' }).

- Average frequency: The average unigram frequency is the average number of occurrences of each unigram in the text. E.g., the text 'aabc' has an average unigram frequency of $1.\dot{3}$ from $\frac{2+1+1}{3}$. The average bigram/trigram frequency is the same as above but for bigrams/trigrams.

- Average distance: The average unigram distance is the average number of 'steps' between repeated occurrences of unigrams in the text. E.g., the text 'abba' has an average unigram distance of 2 (from $\frac{3+1}{2}$). The average bigram/trigram distance is the same as above but for bigrams/trigrams.

- Entropy: the first-order Shannon character entropy or unigram entropy is given by $H_1 = -\sum_{i=1}^{n} p_i \log_2 p_i$, where p_i is the probability of occurrence of each unigram (i.e.,

the number of occurrences of the unigram divided by the total number of occurrences of all unigrams). The bigram/trigram entropies are the same as above but for bigrams/trigrams.

- Index of coincidence: The index of coincidence (IC) measures the evenness of the distribution of characters in the text (greater IC means greater unevenness) and is given by $IC = \frac{n}{N(N-1)} \sum_{i=1}^{n} n_i(n_i - 1)$, where N is the text length, n is the unigram alphabet length, and n_i is the number of occurrences of the i^{th} character in the unigram alphabet. E.g., the text 'abbba' has $IC = 0.8$ from $\frac{2}{5(5-1)}[2(2-1)+3(3-1)]$.
- Zipf's exponent: The exponent α in the equation $f = \frac{K}{r^{\alpha}}$ is obtained by regressing the unigrams' frequencies of occurrence f on their ranks r, where the most frequently-occuring unigram has rank $r = 1$, the second most frequently-occuring unigram has $r = 2$, etc. Zipf's exponent measures the gradient or slope of $\log(f)$ against $\log(r)$. Natural languages have $\alpha \approx 1$.
- Average mean associated contact counts: Burleson (1989) defines the variety of contact count (VCC) for a given unigram as the number of unique unigrams adjacent to (i.e., immediately next to or contacting) the root unigram. E.g., the unigram 'a' in the text 'cab' has $VCC = 2$, while 'c' and 'b' both have $VCC = 1$. Burleson (1989) then defines the mean associated contact count (MACC) for a given unigram as the sum of the VCC values for each adjacent unigram divided by the VCC of the root unigram. E.g., the unigram 'a' in the text 'cab' has $MACC = 1$ from $\frac{1+1}{2}$, while 'c' and 'b' both have $MACC = 2$. To provide a single statistic for the entire text, the author defines the average MACC (AMACC) as the sum of the MACC values for each unigram in the alphabet of the text divided by the alphabet length. E.g., the text 'cab' has $AMACC = 1.\dot{6}$ from $\frac{1+2+2}{3}$.

2.2 Model

To classify texts, an artifical neural network was implemented in Python with the TensorFlow machine learning software library and Keras deep learning API. In this model, the target variable was the binary category of text (real or fake); the fit data were the statistical properties of text (i.e., unigram, bigram, and trigram alphabet lengths, average frequencies, average distances, and entropies; as well as the IC, Zipf's exponent, and AMACC). Briefly, a simple two-layer sequential model was implemented. The input layer contained 15 nodes (one for each input feature) and used the rectified linear unit activation function. The output layer used the sigmoid activation function. Binary cross entropy was used as the loss function for binary (real/fake) classification. The Adam algorithm was used as the optimizer for efficiency.

$N = 20{,}000$ Dorabella-like texts ($10{,}000$ real and $10{,}000$ fake, as described above) were used in training and testing the model. 5-fold cross-validation was used to evaluate the model.

No general, formal equation exists to estimate the sample size required for accurate and precise classification with neural networks. In the context of quantitative linguistics, Kubáček (1994) suggests that a sample size in the thousands may be necessary for a representative count of linguistic entities. In this study, the sample size (number of texts) was in the tens of thousands; three orders of magnitude greater than the number of fit variables. Thus, sample size is unlikely to present an issue (classification models may still be accurate with few data as long as the data are high quality).

All analyses were conducted in Python version 3.8.16 with the packages Numpy version 1.21.5, Pandas version 1.5.2, Scipy version 1.7.3, Uncertainties version 3.1.6, Natural Language Toolkit (NLTK) version 3.7, Scikit-learn version 1.0.2, and TensorFlow version 2.10.0.

3 Results

Linguistic and other statistics for the real texts, fake texts, and Dorabella cryptogram are shown in Table 1. Values for the Dorabella cryptogram are closer to the real texts except in the trigram statistics and AMACC.

A satisfactory model was obtained. Across the five folds, the average model accuracy and standard deviation were 99.8% (0.1%).

To further test the out-of-sample performance of the model, a real 87-character MASC was created from the plaintext of Ellie's essay in the theatrical play *The Whale* and the JavaScript 'Simple Substitution Cipher' generator available from Practical

Cryptography (Lyons, 2023). The model correctly predicted that this real MASC represented real English text with a probability of 0.99999 (giving a probability of fake text of just 0.00001). Likewise, the model correctly predicted that an out-of-sample 87-character random string generated with the website `random.org` was random text with a probability of 0.9999998 (giving a probability of real text of just 0.0000002).

Finally, the model was applied to the Dorabella cryptogram. The model classified the Dorabella cryptogram as real English text with a probability of 0.9996.

Statistic	**Corpus**	**DB**	**Random**
Alphabet length			
Unigram	19.7 (1.4)	20	25.1 (0.9)
Bigram	64.9 (5.5)	69	80.9 (2.1)
Trigram	78.5 (5.6)	83	84.8 (0.5)
Avg. frequency			
Unigram	4.4 (0.3)	4.4	3.5 (0.1)
Bigram	1.3 (0.1)	1.2	1.1 (0.0)
Trigram	1.1 (0.1)	1.0	1.0 (0.0)
Avg. distance			
Unigram	14.4 (2.9)	14.8	17.8 (2.6)
Bigram	7.1 (2.7)	4.7	1.8 (1.0)
Trigram	2.5 (2.6)	0.4	0.1 (0.2)
Entropy			
Unigram	4.0 (0.1)	4.0	4.5 (0.1)
Bigram	5.9 (0.2)	6.0	6.3 (0.1)
Trigram	6.2 (0.1)	6.4	6.4 (0.0)
IC	1.3 (0.1)	1.2	1.0 (0.1)
Zipf's exp.	0.6 (0.1)	0.5	0.4 (0.1)
AMACC	7.7 (0.7)	6.9	7.0 (0.4)

Table 1: Statistics for 10,000 87-character samples from an English-language corpus (Corpus), the Dorabella cryptogram (DB), and 10,000 randomly-generated 87-character English strings (Random). For the Corpus and Random results, means are presented with standard deviations.

4 Discussion

This study demonstrates the use of deep learning methods to classify putative cryptograms probabilistically. The results of this study preliminarily suggest that the Dorabella cryptogram is perhaps more likely to represent real underlying English plaintext than it is to represent purely random gibberish. This does not mean that the Dorabella cryptogram is necessarily a real cryptogram, only that it is less likely to be random gibberish. Of course, other possibilities exist, including the Dorabella cryptogram as non-random gibberish (i.e., gibberish that is designed to look more like real text than purely random text), as a non-MASC cipher, or as encrypted shorthand. Future studies can explore these possibilities by expanding the binary classification model of the present study to multi-class models with data representing other possible classifications.

The present study is not the first to apply classification or clustering methods to classical cryptography. For example, the Neural Cipher Identifier (NCID) by Leierzopf et al. (2021) uses an ensemble neural network classifier for 55 standardized classical cipher types. More relevantly, Juzek (2019) clustered true ciphers and pseudo-ciphers using support-vector machines on entropy. The present study expands upon the NCID by including random or gibberish text as a possible classification (which the NCID does not), and expands upon the Juzek analysis by including other statistical properties besides entropy. Yet more statistical properties may be included in future works, as well as other classification types.

The results of the present study comport with other recent analyses of the Dorabella cryptogram. Schmeh (2018) concluded that "frequency count and the contact counts of the Dorabella Cryptogram are consistent with the English language," and Hauer et al. (2021) reported "evidence for English as the language of the cipher" from n-gram language models, and that "the occurrence of several pairs of mirrored symbols is unlikely to be due to chance, suggesting that Dorabella is not a hoax." Still, no convincing MASC solution to the Dorabella cryptogram has yet been found, even using modern, state-of-the-art algorithms with high decipherment rates (Wase, 2023).

It is not certain whether these findings are accurate, and the author emphasises the need for larger multi-class models to better understand the nature of the Dorabella cryptogram and other putative cryptograms. One limitation in applying deep learning methods for classification is the necessity for the training dataset to reflect the underlying plaintext of the putative cryptogram; if the plaintext is written in French or Old English but the training dataset consists of Modern English text, accurate classification is not possible. Thus, the corpus must be appropriate for the use case.

References

Craig Bauer. 2017. *Unsolved! The history and mystery of the world's greatest ciphers from Ancient Egypt to online secret societies*. Princeton University Press.

Sam Blake. 2021. The solution of the zodiac killer's 340-character cipher. https://blog.wolfram.com/2021/03/24/the-solution-of-the-zodiac-killers-340-character-cipher/.

Donald R. Burleson. 1989. The "macc"–a statistical technique for vowel isolation. *The Cryptogram*, March–April:4–6.

Mark Davies. 2015. The wikipedia corpus. https://www.english-corpora.org/wiki/.

Elonka Dunin. 2023. Famous unsolved codes and ciphers. https://web.archive.org/web/20231025073911/https://elonka.com/UnsolvedCodes.html.

Daniel E Gaskell and Claire L Bowern. 2022. Gibberish after all? voynichese is statistically similar to human-produced samples of meaningless text. In *Proceedings of the 1st International Conference on the Voynich Manuscript*.

Bradley Hauer, Colin Choi, Anirudh Sundar, Abram Hindle, Scott Smallwood, and Grzegorz Kondrak. 2021. Experimental analysis of the dorabella cipher with statistical language models. *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt 2021, June 20-22, 2022, University of Amsterdam, The Netherlands*, pages 70–79.

Tom S. Juzek. 2019. Using the entropy of n-grams to evaluate the authenticity of substitution ciphers and z340 in particular. *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt 2019, June 23-26, 2019, Mons, Belgium*, 158:117–125.

Louis Kruh. 1982. A basic probe of the beale cipher as a bamboozlement. *Cryptologia*, 6(4):378–382.

Lubomír Kubáček. 1994. Confidence limits for proportions of linguistic entities. *Journal of Quantitative Linguistics*, 1(1):56–61.

E. Leierzopf, N. Kopal, B. Esslinger, H. Lampesberger, and E. Hermann. 2021. A massive machine-learning approach for classical cipher type detection using feature engineering. *Proceedings of the 4th International Conference on Historical Cryptology HistoCrypt 2021*.

James Lyons. 2023. Simple substitution cipher. http://practicalcryptography.com/ciphers/simple-substitution-cipher/.

Nick Pelling. 2019. Dorabella cipher: timeline, texts, and keith massey... https://web.archive.org/web/20230528045150/https://ciphermysteries.com/2019/11/15/dorabella-cipher-timeline-texts-and-keith-massey.

Kun Qian, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2015. Bird sounds classification by large scale acoustic features and extreme learning machine. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1317–1321. IEEE.

Klaus Schmeh. 2018. Examining the dorabella cipher with three lesser-known cryptanalysis methods. In *Proceedings of the 1st International Conference on Historical Cryptology, HistoCrypt*, pages 145–152.

Klaus Schmeh. 2023. The top 50 unsolved encrypted messages. https://web.archive.org/web/20230601011324/https://scienceblogs.de/klausis-krypto-kolumne/the-top-50-unsolved-encrypted-messages/.

Viktor Wase. 2023. Dorabella unmasced – the dorabella cipher is not an english or latin monoalphabetical substitution cipher. *Cryptologia*, pages 1–10.