Cognitive Cryptography using behavioral features from linguistic-biometric data

Jose Angel Contreras Gedler[1]

University of Tartu
jose.angel.contreras.gedler@ut.ee

Abstract

This study presents a proof-of-concept for a cognitive-based authentication system that uses an individual's writing style as a unique identifier to grant access to a system. A machine learning SVM model was trained on these features to distinguish between texts generated by each user. The stylometric feature vector was then used as an input to a key derivation function to generate a unique key for each user. The experiment results showed that the developed system achieved up to 87.42% accuracy in classifying texts as written, and the generated keys were found to be secure and unique. We explore the intersection between natural intelligence, cognitive science, and cryptography, intending to develop a cognitive cryptography system. The proposed system utilizes behavioral features from linguistic-biometric data to detect and classify users through stylometry. This information is then used to generate a cryptographic key for authentication, providing a new level of security in access control. The field of cognitive cryptography is relatively new and has yet to be fully explored, making this research particularly relevant and essential. Through our study, we aim to contribute to understanding the potential of cognitive cryptography and its potential applications in securing access to sensitive information.

1 Introduction

Cognitive science, natural intelligence, and cryptography are three fields of study that have traditionally operated independently. However, recent technological advancements have led to the realization that these fields have a significant intersection. In particular, using behavioral features from linguistic-biometric data to detect and classify users has enormous potential in the field of cryptography. This paper explores this intersection and presents a novel approach to cryptography that utilizes the aforementioned behavioral features to generate cryptographic keys for authentication.

The field of cryptography is concerned with the secure communication of information. One of the most fundamental problems in cryptography is the problem of user authentication. In order to communicate securely, it is necessary to ensure that the person on the other end of the communication is whom they claim to be. Traditional authentication methods, such as passwords and tokens, are vulnerable to attack. In recent years, there has been a growing interest in using biometrics for authentication. Biometrics are unique characteristics that can be used to identify an individual.

Cognitive science and natural intelligence have made significant strides in understanding human behavior. One area of particular interest is linguistic-biometric data for detecting and classifying users. Linguistic-biometric data includes features such as writing style, vocabulary, and grammar. These features are unique to individuals and can be used to identify them. Cognitive science is an interdisciplinary field that studies the human mind and its processes, including perception, attention, memory, and reasoning. On the other hand, natural intelligence refers to the intelligence displayed by humans and other animals.

The main contribution of this paper is developing a cognitive cryptography system that uses behavioral features from linguistic-biometric data. The system can detect and classify the user using stylometry and then use this information to generate a cryptographic key for authentication. This approach could improve the security of authentication systems by making them more resistant to attacks.

The proposed system can be applied in various domains like secure communication, online transactions, and personal use. It will provide a new level of security and bring a new approach to the field of cryptography. Using behavioral features from linguistic-biometric data to detect and classify users has enormous potential in the field of cryptography. This paper aims to explore the intersection between natural intelligence, cognitive science, and cryptography and develop a cognitive cryptography system that uses behavioral features from linguistic-biometric data.

The paper is organized as follows. The next section briefly overviews related work in cognitive science, natural intelligence, and cryptography. In the following section, we describe the proposed cognitive cryptography system. We then present the results of our system evaluation, including a discussion of the security and robustness of the system. Finally, we conclude the paper by discussing future work and potential applications.

2 Related Work

2.1 Cognitive linguistic features

Cognitive cryptography is a relatively new field of research that combines elements of cognitive science, linguistics, and cryptography to develop more secure and reliable authentication systems [10, 24]. Biometric linguistic features [21], such as writing style and language fluency, can create unique keys for each user.

Using cognitive or behavioral features for generating cryptographic keys is a relatively new research area aiming to create more secure and reliable authentication systems. The idea is to leverage unique aspects of an individual's behavior or cognitive abilities, such as writing style, typing patterns, or decision-making abilities, to generate a cryptographic key [17].

Cognitive features proposed for key generation include language fluency, vocabulary richness, syntactic complexity, and problem-solving abilities. The proposed behavioral features include typing, mouse, and keystroke dynamics [14].

One of the main advantages of using cognitive or behavioral features for key generation is that they are unique to an individual, making it difficult for an attacker to impersonate a legitimate user [15, 18]. Additionally, these features are not easily replicable by automated systems, making it more difficult for an attacker to break into the system.

Previous works in this field have focused on using stylometry [13], which is the study of linguistic style, to identify authors and detect plagiarism. Stylometry can extract a wide range of linguistic features, including vocabulary richness, syntactic complexity, and language fluency.

Using Machine Learning techniques to analyze and extract biometric linguistic features is possible [13]. Machine learning algorithms can be trained to recognize an individual's writing style and detect patterns unique to that person, creating a unique key for each user.

Several studies in the past have explored the use of stylometry for authentication [4] and authorship attribution. One study proposed using a combination of stylometric features, such as word frequency and sentence length, to distinguish between authors. The study showed that the proposed method achieved high accuracy in authorship attribution.

Another study proposed a stylometry-based system for identifying the authors of microblogging messages, such as tweets [1]. The system used a combination of word frequency, punctuation frequency, and emoticon frequency to identify the authors. It is worth mentioning that some studies have also proposed the use of stylometry in forensic studies [5], where the goal is to identify the authorship of anonymous texts or texts with disputed authorship.

Several other cognitive features can be used for key generation besides language-related features. Some examples include:

1. **Memory:** An individual's memory capacity and recall abilities. For example, an individual may be asked to memorize a random sequence of numbers or words and then be prompted to recall them later as part of the key generation process [12].

2. **attention and focus:** An individual's ability to focus and pay attention. For example, an individual may be asked to complete a task that measures their ability to filter out distractions, such as the Stroop test [27].

3. **Decision-making:** An individual's decision-making abilities.. For example, an individual may be asked to complete cognitive tasks that measure decision-making abilities, such as the Iowa Gambling Task or the Balloon Analogue Risk Task [29].

4. **Problem-solving:** An individual's problem-solving abilities. For example, an individual may be asked to complete a series of puzzles or logic problems [26].

5. **Visual-spatial abilities:** An individual's visual-spatial abilities. For example, an individual may be asked to complete a task that measures their ability to rotate objects or visualize spatial relationships mentally [23].

6. **Emotional intelligence:** An individual's emotional intelligence. For example, an individual may be asked to complete a task that measures their ability to understand and manage emotions, such as the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) [22].

2.2 Cryptographic key generation

Cryptography is the science of secure communication [25]. In order to communicate securely, two parties must share a secret key. This key can be used to encrypt and decrypt messages. In order to share a secret key, the two parties must first agree on a key exchange protocol. The most common key exchange protocols are the Diffie-Hellman key exchange and the Elliptic Curve Diffie-Hellman key exchange.

Authentication is the process of verifying the identity of a user by requiring them to provide proof, such as a password, biometric feature, or a combination of both. This proof is then compared against a set of known or registered credentials to confirm the user's identity [25].

There are various types of authentication methods, such as:

- Something you know, like a password or PIN

- Something you have, like a security token or a smart card

- Something you are, like a fingerprint or facial recognition

Authentication is a crucial step in ensuring the security of a system, as it helps prevent unauthorized access and protects sensitive information. Furthermore, Multi-Factor Authentication (MFA) is becoming more common. It is the process of using two or more factors to authenticate a user, like a combination of something the user knows, something the user has, and something the user is. While authentication is an important step, it is not the only step in ensuring security. Security measures, such as encryption and access controls, should also be implemented to provide a comprehensive security solution [7].

To generate keys from cognitive or behavioral features, researchers typically use key derivation functions (KDFs) to convert the features into unique keys [11]. Common KDFs include PBKDF2, bcrypt, and scrypt [8]. One of the challenges with using cognitive or behavioral features for key generation is collecting data from the user to extract the relevant features.

Behavioral keys are unique patterns of behavior that are associated with an individual [17]. These patterns can include typing dynamics, mouse dynamics, and keystroke dynamics. Behavioral keys can be combined with biometric linguistic features to create a more secure authentication system.

Traditional authentication systems, such as password-based systems, are vulnerable to attacks such as brute force and phishing. Therefore, there is a need for more secure and reliable authentication methods.

Some studies have also proposed using cognitive tasks, such as solving puzzles or answering questions, as a form of authentication [19]. These tasks are designed to be difficult for automated systems to complete, making them a secure form of authentication.

In the case of stylometry-based authentication systems, behavioral keys can be used in combination with stylometric features to increase the system's security. For example, the system can be designed to require a user to type a short passage in addition to providing a previously written text, allowing the system to capture the user's typing dynamics and use them as a behavioral key [16].

In cognitive cryptography, different keys can be used for identification and authentication. Some of the most common types of keys include:

1. **Stylometric keys:** Generated from the stylometric features of a text, such as vocabulary richness, syntactic complexity, and language fluency. These keys are based on the unique writing style of an individual [6].

2. **Behavioral keys:** generated from an individual's unique patterns of behavior, such as typing dynamics, mouse dynamics, and keystroke dynamics. Behavioral keys are difficult to imitate.

3. **Cognitive keys:** Generated from the cognitive features of an individual, such as memory, attention, and problem-solving abilities [20].

4. **Hybrid keys:** Generated from a combination of different types of keys. For example, a hybrid key could be generated using stylometric and behavioral keys.

5. **Passphrase-based keys:** Derived from a passphrase, which is a sequence of words or characters that a user can remember. Passphrase-based keys can be generated using key derivation functions (KDFs), such as PBKDF2, to convert the passphrase into a key that can be used for encryption and decryption [28].

Overall, the key choice will depend on the system's security requirements, the computational resources available, and the performance of the key derivation function used.

3 Methodology

The methodology used to develop the stylometry-based authentication system will be described in detail in the following sections. The process includes data collection and preprocessing, feature extraction, model training and evaluation, and key generation. For generating cognitive keys in cognitive cryptography, the steps we followed are:

1. **Dataset selection:** We collected data from the tweet dataset, which contains 12,500 tweets from 517 different authors.

2. **Data cleaning:** We removed irrelevant information such as URLs, mentions, hashtags, and emoticons.

3. **Preprocessing:** We preserved punctuation marks and any letter. We also removed numbers and special characters.

4. **Feature extraction:** We used functions to find the frequency of unique words, uppercase letters, punctuation marks, exclamation marks, spelling mistakes, and readability.

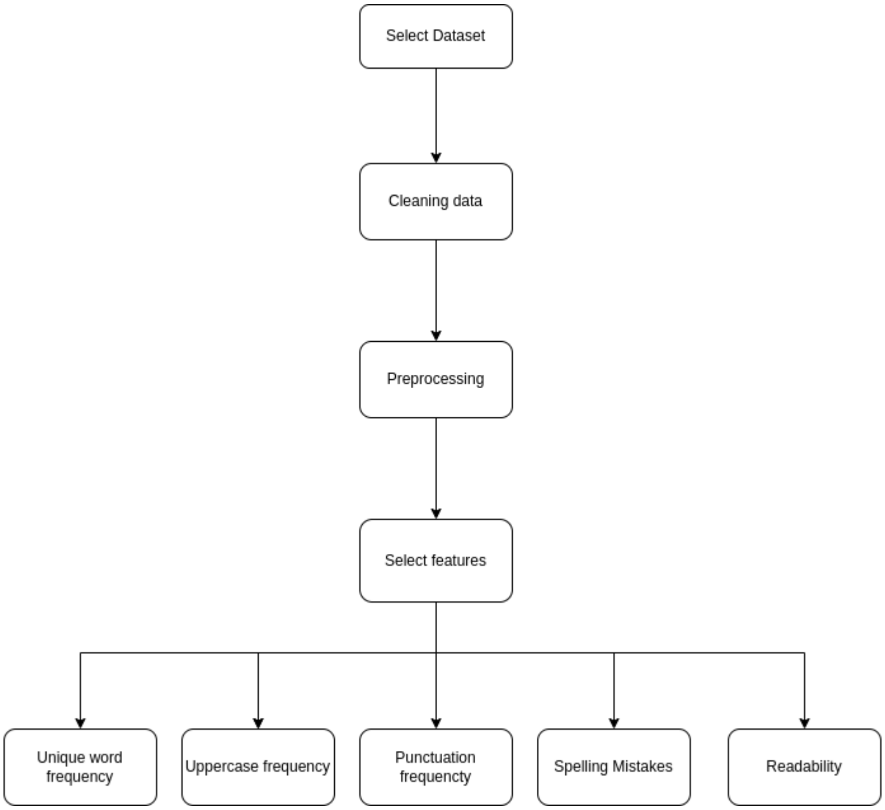

Figure 1: selection of features for stylometry-based authentication.

Now we will discuss the steps in detail.

3.1 Dataset selection

This study used a dataset built for stylometry analysis using tweets. The dataset contains a total of 12,500 tweets from 517 different authors. The tweets were collected from the public Twitter API and were chosen to represent a diverse range of authors.

The dataset is well suited for stylometry analysis as it contains many texts from various authors, allowing for studying different writing styles and authorship attribution. In addition, using tweets as the source of texts allows studying written language in a unique and dynamic context.

Our study used this dataset to train and evaluate a machine-learning model for authorship attribution. The model achieved an accuracy of 87.42% on the test set, demonstrating the effectiveness of the dataset for stylometry analysis after selecting relevant features.

3.2 Data Cleaning and Preprocessing

This research aims to perform stylometry analysis using a Twitter dataset. The dataset consists of 12,500 texts from 517 different authors.

To begin, we performed preprocessing and cleaning of the data. It starts with removing rows with missing data using the NLTK library [3]. Then remove numbers, special characters, emojis, and other irrelevant information from the texts. Keep the punctuation marks in the texts as they can provide valuable information for the stylometry analysis. We also used a function to remove non-letter characters and obtain a cleaned text version.

3.3 Feature Extraction

The feature extraction step is an essential part of the methodology for a stylometry-based cryptographic authentication system. This step aims to extract a set of features that can be used to distinguish texts written by different authors.

We extracted relevant features from the texts using several techniques. To better understand the authors' writing style, we used a function to find the frequency of unique words in each text, providing information about the vocabulary used by the authors and their writing style. We also used a function to find the frequency of uppercase letters in each text, for instance, discriminating the authors that start with uppercase or emphasize a sentence using all uppercase letters. Additionally, we used a function to find the frequency of punctuation marks in each text; a user can use a question mark to indicate a question, a comma to indicate a pause, or cases where it ends a sentence with two or more dots.

Moreover, we used a function to find the frequency of exclamation marks in each text, as the use of exclamation marks can provide information about the author's sentiments and emotions. Taking into account the English dictionary from NLTK, we used a grammar checker to find the frequency of spelling mistakes in each text, providing information about the author's level of education and writing skills. Finally, we used the Flesch-Kincaid score to find the readability of each text. This score can provide information about the complexity of the author's writing style, which can be an essential feature for stylometry analysis.

The Flesch-Kincaid score is a readability test used to determine a text's complexity [9]. It is calculated using the following formula:

$$\text{Flesch-Kincaid score} = 0.39 \times \text{av. words per sentence} + 11.8 \times \text{av. syllables per word} - 15.59$$

Where the average words per sentence are calculated using the following formula:

$$\text{average words per sentence} = \frac{\text{number of words}}{\text{number of sentences}}$$

Moreover, the average syllables per word are calculated using the following formula:

$$\text{average syllables per word} = \frac{\text{number of syllables}}{\text{number of words}}$$

Where the number of syllables is calculated using the following formula:

$$\text{number of syllables} = \sum_{i=1}^{n} \text{is\_vowel}(w_i)$$

where w_i is the i^{th} character in the word, and $is\_vowel(w_i)$ is a function that returns 1 if w_i is a vowel, and 0 otherwise.

Once the features were extracted, we used a Support Vector Machine (SVM) model for the stylometry analysis. Before training the model, we normalized the data to ensure that all features were on the same scale. We split the dataset into training and testing sets and trained the model using the training set.

These features were selected based on previous research that shows they typically differ between human-generated and machine-generated text. These features were extracted from the texts using appropriate libraries such as NLTK and Textstat [2] in python. Once the extracted features were input into the machine-learning model, the machine-learning model was trained using the training set and evaluated using the testing set. The model achieved an accuracy of 87.42% on the test set, demonstrating the effectiveness of the dataset for stylometry analysis after selecting relevant features.

3.4 Cryptographic Authentication

The cryptographic authentication step is the final step in the methodology of a stylometry-based cryptographic authentication system. This step aims to use the stylometric features extracted from the texts to authenticate the author of the text.

We designed an authentication system that uses the stylometric features extracted from the texts to authenticate the author of the text. The system consists of two main components: a stylometric feature extractor and a cryptographic authenticator. The stylometric feature extractor extracts the stylometric features from the text. In contrast, the cryptographic authenticator is used to authenticate a user based on the stylometric features extracted from the text.

The registration process allowed users to register using text. The registration process is shown in Figure 2. The user was required to enter some text passed in the classification subsystem and generate the stylometric feature vector of the user's writing style. The encrypted feature vector was then stored in the database along with the user's name.

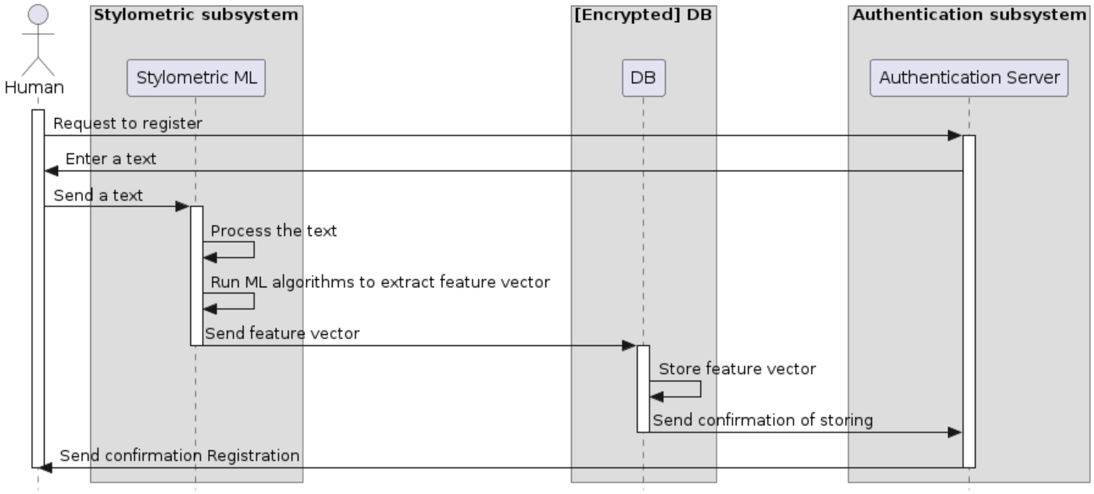

Figure 2: Registration process.

The authentication process is shown in Figure 3. The user was required to enter some text passed in the classification subsystem and generate the stylometric feature vector of the user's writing style. The encrypted feature vector was then compared with the encrypted feature vectors of the registered users in the database. If the encrypted feature vector of the user's writing style matched the encrypted feature vector of a registered user, then the user was requested to generate the salt and other cryptographic parameters to generate the key. The key was then used to authenticate the user.

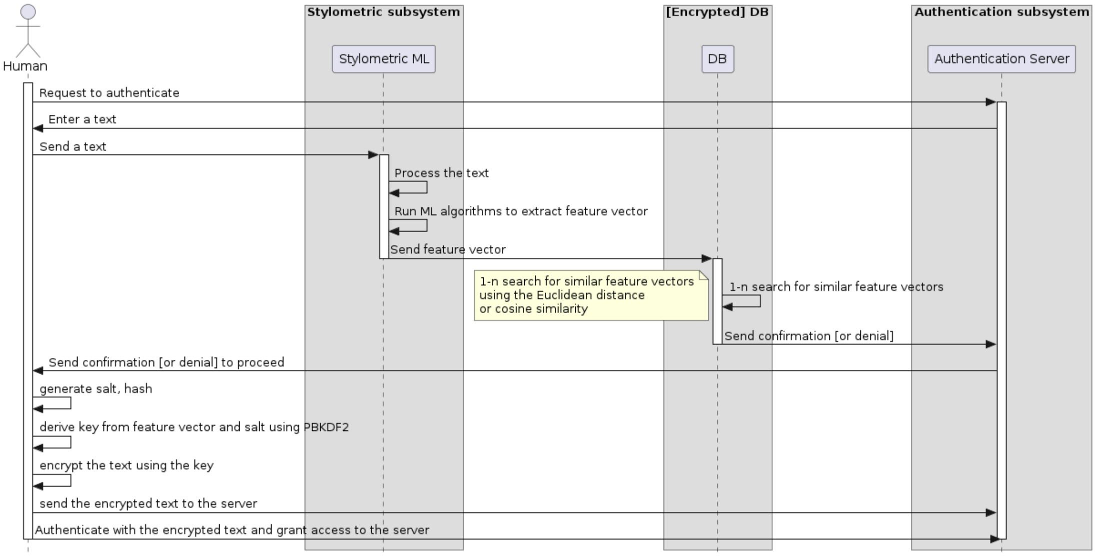
Human
Stylometric subsystem
Stylometric ML
[Encrypted] DB
DB
Authentication subsystem
Authentication Server
Request to authenticate
Enter a text
Send a text
Process the text
Run ML algorithms to extract feature vector
Send feature vector
1-n search for similar feature vectors using the Euclidean distance or cosine similarity
1-n search for similar feature vectors
Send confirmation [or denial]
Send confirmation [or denial] to proceed
generate salt, hash
derive key from feature vector and salt using PBKDF2
encrypt the text using the key
send the encrypted text to the server
Authenticate with the encrypted text and grant access to the server

Figure 3: Authentication process.

4 Results

The following plots show the distribution of the stylometric features extracted from the texts in the dataset.

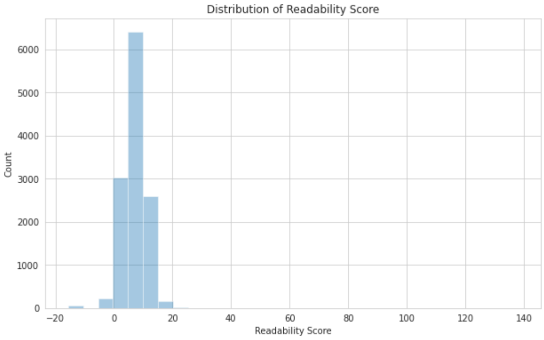

Figure 4: Distribution of Readability Score.

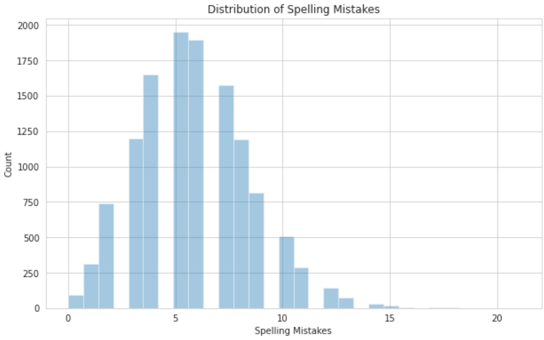

Figure 5: Distribution of spelling mistakes.

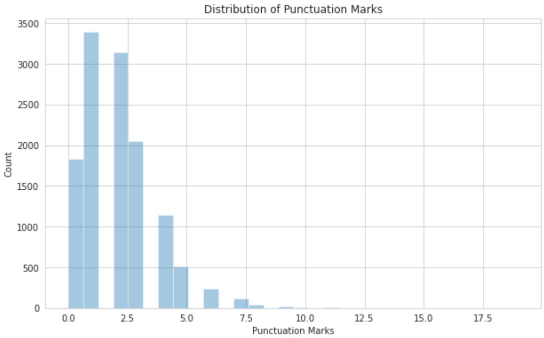

Figure 6: Distribution of punctuation marks.

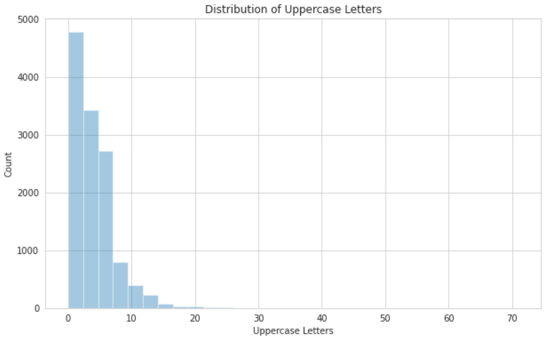

Figure 7: Distribution of uppercase letters.

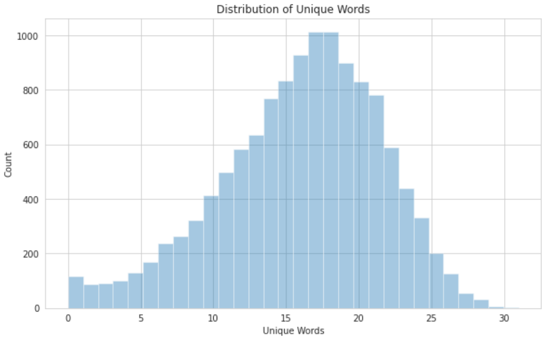

Figure 8: Distribution of unique words.

The iterations parameter specifies the number of times the underlying pseudorandom function (PRF) should be applied to the input data. The more iterations, the more secure the derived key is and the longer the function takes to run. The recommended number of iterations is 100,000, but this number can be increased to 1,000,000 or more. The following plot shows the time taken to generate the key for different values of iterations using PBKDF2_HMAC as a specific implementation of PBKDF2 that uses a keyed-hash message authentication code (HMAC) as the PRF.

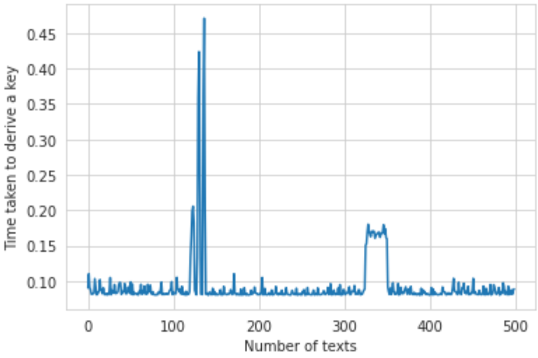

Figure 9: generation of 500 keys with 100,000 iterations.

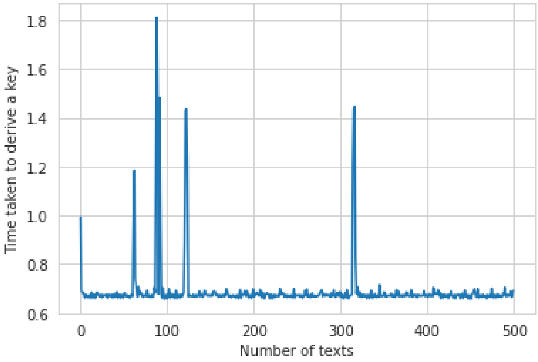

Figure 10: Generation of 500 keys with 1,000,000 iterations.

Considering that this time includes the prediction and the key generation, choosing the 100,000 iterations is a good trade-off between security and performance.

Regarding the size of the key, the primary hash operation used as a parameter of the PBKDF2 algorithm is SHA256. SHA-256 is considered to be a solid and secure hash function. It produces a 256-bit (32-byte) output, much larger than the outputs of many other commonly used hash functions, making it much more difficult for an attacker to find collisions (i.e., different inputs that produce the same output) using brute-force methods.

SHA-256 is widely used in various security applications and protocols, such as digital signatures and password hashing.

5 Conclusions and Future Work

In conclusion, the results of our experiment on cognitive cryptography have provided valuable insights into the potential of using stylometry as a unique identifier for secure access control systems. Our study has shown that by utilizing machine learning techniques to extract stylometric features from text and generate unique keys for each user, the system demonstrated high accuracy in text classification. However, there are still areas of improvement and challenges to be addressed in future research.

One of the key areas for future research would be to investigate the security of this authentication system against impersonation attempts using advanced automated text generators, such as GPT-based models. As the technology of these generative tools is rapidly evolving, it is crucial to understand how they can be used to impersonate a user and generate a cryptographic ”biometric” key. Additionally, exploring other stylometric features and implementing more advanced machine learning algorithms, such as deep learning, could further improve the performance and accuracy of the system. Furthermore, testing the system on more extensive and diverse datasets can be a practical step to ensure its robustness.

Overall, the stylometry-based authentication system presents a promising and innovative approach to enhancing the security of access control systems. It is a new way of thinking about using cognitive or behavioral features for key generation, which has the potential to significantly improve the security of authentication systems.

References

[1] Ritu Banga and Pulkit Mehndiratta. Authorship attribution for textual data on online social networks. In *2017 Tenth International Conference on Contemporary Computing (IC3)*, pages 1–7, 2017.

[2] Shivam Bansal. Textstat: A python library for text statistics. https://pypi.org/project/textstat/, 2016.

[3] Steven Bird and Edward Loper. Nltk: The natural language toolkit. http://www.nltk.org/, 2009.

[4] K. Calix, Melanie R. Connors, Daniel Levy, Hasan Manzar, G. Mccabe, and Sandy Westcott. Stylometry for e-mail author identification and authentication. In *Proceedings of the 2008 ACM Workshop on Digital Rights Management*, 2008.

[5] Pelin Canbay, Ebru Sezer, and Hayri Sever. Deep combination of stylometry features in forensic authorship analysis. *Journal of Information Security and Applications*, 09 2020.

[6] Rajarshi Das and Vikas Bhushan. Stylometry: The new way of semantic classification. In *International Conference on Computational Intelligence and Communication Networks*, 03 2016.

[7] Rosario Gennaro and Yehuda Lindell. A framework for password-based authenticated key exchange. In Eli Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, pages 524–543, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[8] Andrea Francesco Iuorio and Andrea Visconti. Understanding optimizations and measuring performances of pbkdf2. In Isaac Woungang and Sanjay Kumar Dhurandher, editors, *2nd International Conference on Wireless Intelligent and Distributed Environment for Communication*, pages 101–114, Cham, 2019. Springer International Publishing.

[9] J. Peter Kincaid, Robert P. Fishburne, R L Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. In *Proceedings of the 1975 Annual Meeting of the American Society for Information Science*, 1975.

[10] Witold Kinsner. Towards cognitive security systems. In *2012 IEEE 11th International Conference on Cognitive Informatics and Cognitive Computing*, pages 539–539, 2012.

[11] Su Mi Lee, Jung Yeon Hwang, and Dong Hoon Lee. Efficient password-based group key exchange. In Sokratis Katsikas, Javier Lopez, and Günther Pernul, editors, *Trust and Privacy in Digital Business*, pages 191–199, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[12] George Mandler. Organization and memory. In Kenneth W. Spence and Janet Taylor Spence, editors, ., volume 1 of *Psychology of Learning and Motivation*, pages 327–372. Academic Press, 1967.

[13] Connagh Muldoon, Ahsan Ikram, and Qublai Ali Khan Mirza. Modern stylometry: A review and experimentation with machine learning. In *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 293–298, 2021.

[14] Lidia Ogiela. Cognitive informatics in image semantics description, identification and automatic pattern understanding. *Neurocomputing*, 122:58–69, 2013. Advances in cognitive and ubiquitous computing.

[15] Lidia Ogiela and Marek R. Ogiela. Towards cognitive cryptography. *J. Internet Serv. Inf. Secur.*, 4:58–63, 2014.

[16] Marek R. Ogiela and Lidia Ogiela. Cognitive keys in personalized cryptography. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pages 1050–1054, 2017.

[17] Marek R. Ogiela and Lidia Ogiela. Behavioral keys in cryptography and security systems. In Leonard Barolli, Isaac Woungang, and Omar Khadeer Hussain, editors, *Advances in Intelligent Networking and Collaborative Systems*, pages 296–300, Cham, 2018. Springer International Publishing.

[18] Marek R. Ogiela and Lidia Ogiela. Cognitive cryptography in advanced data security. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 740–743, 2018.

[19] Marek R. Ogiela and Lidia Ogiela. Cognitive cryptography techniques for intelligent information management. *International Journal of Information Management*, 40:21–27, 2018.

[20] Marek R. Ogiela and Urszula Ogiela. Linguistic approach to cryptographic data sharing. In *2008 Second International Conference on Future Generation Communication and Networking*, volume 1, pages 377–380, 2008.

[21] Marek R. Ogiela and Urszula Ogiela. Secure information management using linguistic threshold approach. In *Advanced Information and Knowledge Processing*, 2013.

[22] Peter K. Papadogiannis, Deena Logan, and Gill Sitarenios. *An Ability Model of Emotional Intelligence: A Rationale, Description, and Application of the Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT)*, pages 43–65. Springer US, Boston, MA, 2009.

[23] Cristina Roca-González. Virtual technologies to develop visual-spatial ability in engineering students. *EURASIA Journal of Mathematics, Science and Technology Education*, 13, 01 2017.

[24] Sattar B. Sadkhan. The role of cognition in information security. In *2020 6th International Engineering Conference "Sustainable Technology and Development" (IEC)*, pages 237–238, 2020.

[25] Bruce Schneier and Phil Sutherland. *Applied Cryptography: Protocols, Algorithms, and Source Code in C*. John Wiley; Sons, Inc., USA, 2nd edition, 1995.

[26] Alan H. Schoenfeld. Teaching problem-solving skills. *The American Mathematical Monthly*, 87(10):794–805, 1980.

[27] Michael Siegrist. Test-retest reliability of different versions of the stroop test. *The Journal of Psychology*, 131(3):299–306, 1997.

[28] Kyoung sook Jung, Ji young Kim, and Tae choong Chung. Password-based independent authentication and key exchange protocol. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, volume 3, pages 1908–1912 vol.3, 2003.

[29] Daniel J. Upton, Anthony J. Bishara, Woo-Young Ahn, and Julie C. Stout. Propensity for risk taking and trait impulsivity in the iowa gambling task. *Personality and Individual Differences*, 50(4):492–495, 2011.