Authorship Verification via Linear Correlation Methods of n-gram and Syntax Metrics

* Jared Ray Nelson
Department of Engineering
Utah Valley University
Orem, Utah, United States
Jared.Nelson@uvu.edu

Mohammad Shekaramiz, *IEEE Member*
Department of Engineering
Utah Valley University
Orem, Utah, United States
mshekaramiz@uvu.edu

***Abstract*—This research evaluates the accuracy of two methods of authorship prediction: syntactical analysis and n-gram, and explores its potential usage. The proposed algorithm measures n-gram, and counts adjectives, adverbs, verbs, nouns, punctuation, and sentence length from the training data, and normalizes each metric. The proposed algorithm compares the metrics of training samples to testing samples and predicts authorship based on the correlation they share for each metric. The severity of correlation between the testing and training data produces significant weight in the decision-making process. For example, if analysis of one metric approximates 100% positive correlation, the weight in the decision is assigned a maximum value for that metric. Conversely, a 100% negative correlation receives the minimum value. This new method of authorship validation holds promise for future innovation in fraud protection, the study of historical documents, and maintaining integrity within academia.**

***Index Terms*—n-grams, syntactical data, authorship, machine learning, stylometry, Part of Speech (PoS).**

I. Introduction

This research aims to solve the problem of authorship recognition by identifying stylometry and including a computer algorithm that analyzes written data and predicts whether two documents have the same author. Stylometry is the study of the statistical differences in two author's writing [1]. This topic is applicable in areas of cybersecurity, history, and academia. Lately, data security has become a high priority as more companies hold sensitive information such as social security numbers, emails, phone numbers, account numbers, etc. Thus, hackers have implemented new ways to lure employees through phishing tactics so that they can breach a secured system in seconds. A method of attack can be through email or a text message trying to imposter a boss or colleague. Unfortunately, this type of attack has been effective, and many companies have lost precious resources. Despite many efforts of companies to prevent hackers, employees continue to fall for cyber attacks. Thus, there is a growing demand for effective and fast software solutions to tackle these issues. As the volume of phishing attacks grows, the demand for fraud detection in writing increases. Researchers have studied Stylometry to identify a means of authorship verification. In this case, machine learning (ML) algorithms can provide more robust systems to train a computer to recognize an author's style and compare it to others.

Stylometry is too complex for humans to solve promptly. Conversely, computers have enough memory and processing power to collect data, identify patterns and differences, and represent the data faster than humans. In this regard, ML algorithms can predict authorship. ML algorithms can identify whether famous authors plagiarize or change their writing styles throughout their careers. Numerous literature has unidentified authors, and ML is a tool to help researchers identify authorship and gain new insight into eras of history. Problems in cybersecurity, history, and academia persist with identifying authorship and do not have a solution. Sifting through mass amounts of data and performing complex algorithms demands more computation resources. Therefore, this research uses an algorithm to calculate stylometry by measuring syntax and n-gram, and adds its performance with other algorithms to explore its potential uses in strengthening the aforementioned weaknesses. Below, we describe stylometry, a need for authorship recognition, the history of stylometry from Mendenhall's work, defining n-gram, distinguishing between supervised and unsupervised learning, and the goals of this research.

A. Stylometry

Stylometry involves collecting the components of a person's writing to predict authorship. Stylometry's history dates to the English Renaissance era. Scholars attempted to identify poets' styles and compared each poet's works [2]. Recently, with more computing capabilities, researchers have applied various methods to determine the authorship. Many algorithms have taken old documents to identify the context and possible correlation. This paper proposes a new algorithm that utilizes n-gram [3] and syntactical analysis. Counting and forming averages in Mendenhall's characteristic curve (MCCC) [4] and k-nearest-neighbor [5] have inspired this approach.

B. Need for Author Recognition

There are several methods for author identification including biometrics, passwords, dual-authentication, and multi-authentication [6]. Biometrics are more secure than passwords because they make it more difficult for a hacker to penetrate a system. Examples such as fingerprints are unique to an individual, and they prevent a hacker from generating a random fingerprint to access a user's account [6]. Since biometrics

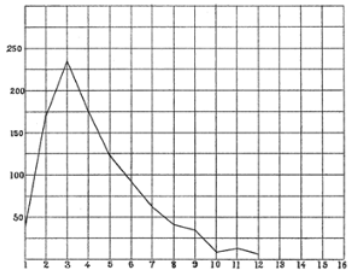

Fig. 1. B. Mendenhall's Characteristic Curve of Composition [4].

are gaining popularity, these are used as the main source of verification on many systems. Thus, hackers resort to penetrating trusted systems that have been previous unlocked through biometrics. Passwords are convenient, but hackers easily hack them. Authenticators such as dual/multi-authentication provide more initial protection, but these typically allow a window of time without re-authenticating. Despite these efforts, hackers send phishing attacks that lure people to click a link that downloads malware on their computers. Hackers know that penetrating a trusted computer gives them access to sensitive information. The two strategies of n-gram and syntactical analysis are used in this research to collect data on the author's stylometry, and we propose a new method that combines these metrics linearly to predict authorship.

Phishing emails and other methods demonstrate an imposter creating a message pretending to take on the original author's voice. When successful, the company's data is compromised, and the company often pays money to receive the data. Companies invest in software to prevent the effects of phishing, but this method doesn't completely resolve the issue, but text analyzing software could be the solution. Researchers use algorithms to identify authorship by measuring data from a writer. The number of adjectives, apostrophes, adverbs, nouns, verbs, dashes, and types of punctuation are measured to form a stylometry profile. The slight variation in these values gives greater insight when comparing different authors. The successful implementation of this approach provides a cost-effective approach to preventing phishing, unlocking greater insight into historical literature, and recognizing plagiarism.

C. B. Mendenhall's Characteristic Curve of Composition [4]

Mendenhall's Characteristic Curve of Composition (MCCC) is one of the first attempts that anyone has tried to tie the word length and frequency of use into data analysis. Fig. 1 demonstrates the length of the word as the x-axis and occurrence as the y-axis. The MCCC takes in the average word length ranging 1-n (n representing the arbitrary number). Mendenhall took text lengths of 1000 words and created a graph that would contain a person's writing style. He noted that there were significant changes in the frequency of use and the length of the words [4].

Mendenhall's research focused on giving each person an identity through the data collected from their writing compared to others. Mendenhall's research sparked interest in the study of stylometry as a means of verifying authorship. He achieved some success by counting the lengths of words and their repetition. Fig. 1 demonstrates that words ranging 2-4 letters in length had the most occurrence. Mendenhall's research was one of the first attempts using syntactical data to predict authorship.

D. n-gram

An n-gram is an adjacent sequence of n items from a text selection. The term n represents the number of words analyzed per iteration. The word unigram means analysis of 1 word, bigram is 2 words, and trigram is 3 [5]. An n-gram demonstrates how a writer uses words to create a sentence. For example, "I like to run vigorously when I'm trying to make it under time." This sentence analyzed with bi-gram would look like: *I like, like to, to run, run vigorously, etc*. The repetition of the words clumped together provides insight into the author's writing style. This approach of verifying authorship assumes writers will often use the same combination of words throughout their writing. It further uses the assumption that a writer will use these clusters of words in other written documents. When comparing two documents, the likelihood that an author uses a similar cluster of adjacent words decreases as the size of n increases. Each n-gram is summed by its repetition throughout the document and normalized to identify the n-grams repetition *vs* other n-gram. The correlation between two texts of the same author using 2-gram, 3-gram, to 4-gram decreases at a rate comparable to an exponential decay equation. This observation provides insight to better identify an author through n-gram analysis by forming an exponential curve between the correlation of the two texts. The value of correlation found in the n-gram analysis can be coupled with the understanding of the exponential decay trend to form a decision-making curve. In the work by Cavnar, they were able to take the value of repetition of each n-gram vs. the number of n-grams in a technical document. Their research suggests that the repetition of certain n-gram will be more frequent than others, and their findings point to an exponential decay function [3].

E. Machine Learning

ML entails using training data (data as sample data gathered from a source) and making predictions from the data. These algorithms improve automatically through experience. In the early ages of programming, many intelligent programs used a series of if-else statements to process the data and adjust the user input. ML algorithms limit the use of if-else instructions and create more clear solutions to problems. [7]. For example, an ML algorithm can take two different species of flowers, and based on the measurements of its petals and stem length, the model makes predictions on the flower type. A person approaches an unknown flower, measures its petals, and uses the algorithm to determine its type. The algorithm predicts flower species with high accuracy. Many ML applications use statistical models, but not all ML is statistical.

F. Objectives of This Paper and the Overall View

This paper predicts authorship by comparing 1-gram, 2-gram, and 3-gram and stylometric metrics such as adjectives, verbs, adverbs, commas, dashes, apostrophes, sentence length from two scripts. The first script will be the training set data, containing 1-2 chapters of a book, and the test data includes anything from a paragraph up to a chapter's length. The algorithm normalizes the n-gram values for the testing and training data and finds its correlation. The correlation rate in the n-gram decreases exponentially as the value n increases. Thus, the algorithm uses an exponential decay function as a basis to calculate the correlation of the n-gram. This system gathers the syntactical data and compares the normalized occurrences of each metric. This is a form of statistical classification. The greater the correlation, the greater the influence in authorship prediction. This process gives insight into whether stylometric and n-gram metrics can provide accurate results in authorship prediction. As the main application of this work, we focus on historical literature by comparing authors of the same era, but in future iterations, the research will focus on cybersecurity and creating identification matrices for individual writing.

II. Related Work

A. Overview

A study from 1994 by Cavnar and Trenkle used n-gram as a "highly effective way for classifying documents" [3]. They implemented frequency statistics with the n-gram, where they took the number of occurrences of a group size of n-words and normalized the data set. Cheng in [8] introduced various metrics and used machine learning to predict male and female writers. Cheng, Chaski, and Iqbal in [8] [9] [10], struggled to obtain similar results when more writers and fewer data were introduced into the model. Authorship characterization and verification are vital elements of Stylometry. Author verification represents training and testing data that are compared and demonstrates whether the correct model is used. Author characterization takes the written text and determines whether the algorithm gathers accurate information.

B. Authorship Verification

Authorship prediction introduces challenges when the sample size increases dramatically. Iqbal et al. demonstrated that the prediction results drops from 90% to 80% accuracy when the author size increases beyond 3 [9]. They used an identification method named "AuthorMiner" and used syntactical, structural, content-specific, and lexical patterns. When the sample size of authors went from 6 to 10, the accuracy dropped from 80.5% to 77%. Chaski received a more accurate results of 95.70% with 10 authors but noticed a decrease in accuracy as the population size increased [10]. Authorship prediction presents challenges when predicting authors of novels because of the tone of voice. Thus, authorship prediction is currently limited in its scalability.

C. Authorship Characterization

Cheng in [8] analyzed many elements in the author's syntactical writing. These elements include tabs, newline characters, digital characters, a-z characters, and special characters. Cheng identified the differences between male and female writing by using some of these filtering methods, among others, but also encountered a scalability issue. This research identifies a valid way to predict authorship by calculating the syntactical structure of the author's sentences. The process identifies occurrences to conclude the author's patterns. Fig. 2 displays a methodology that formulates a class containing information about the authors' apostrophe, comma, dash, sentence length, n-gram, adjective, adverb, noun, and verb usage.

III. Methodology

We propose a process that contains two significant steps. The first step entails mining and formatting the data, and the second step includes the decision-making algorithm. Fig. 2 illustrates the diagram that receives a text file as input, stores and categorizes its values, normalizes each metric, and creates an object that gives accessible variables for each measured metric. The algorithm uses this process by reading the text file as a list containing each line as an index in the list in Python. It parses the list by characters, words, and punctuation to identify each metric. Lists and dictionaries are the results of the different parsing styles. The process normalizes the dictionary's values and finalizes the object containing all the needed information to verify authorship.

A. Books

The books used for this project are excerpts from *Moby Dick*, *Typee*, *The Confidence Man* by Herman Melville, *Treasure Island* and *Kidnapped* by Robert Louis Stevenson, and *Alice in Wonderland* by Lewis Carroll [11]. Gutenburg.org has an adventure section. This section contained Moby Dick, Kidnapped, and Alice in Wonderland and was written in the years 1851, 1886 1865, respectively. Melville and Stevenson wrote Moby Dick and Kidnapped in first-person voice, and Carroll wrote Alice in Wonderland in third-person.

B. Mining the Data

We use the excerpts from each book and placed their contents in a text file. The algorithm reads the text files and creates a class on values mentioned in Fig. 2. The objects with the information from the text files contain dictionaries, lists, average values for each metric, and strings. These objects store information about the training set data and the testing set data. The algorithm calculates 2-gram, 3-gram, and 4-gram. 1-gram is not included as it includes too many commonly used words. Similarly, 5-gram is not considered as it demonstrates too few cases of correlation to predict outcomes. The observed level of correlation between two text documents from the same author showed an exponential decay curve when observing 2, 3, and 4-gram. The text documents from the same author averaged 15-20% correlation for 2-gram, 1-3% for 3-gram, and roughly 0.1-0.5% for 4-gram. The proposed algorithm uses the NLTK

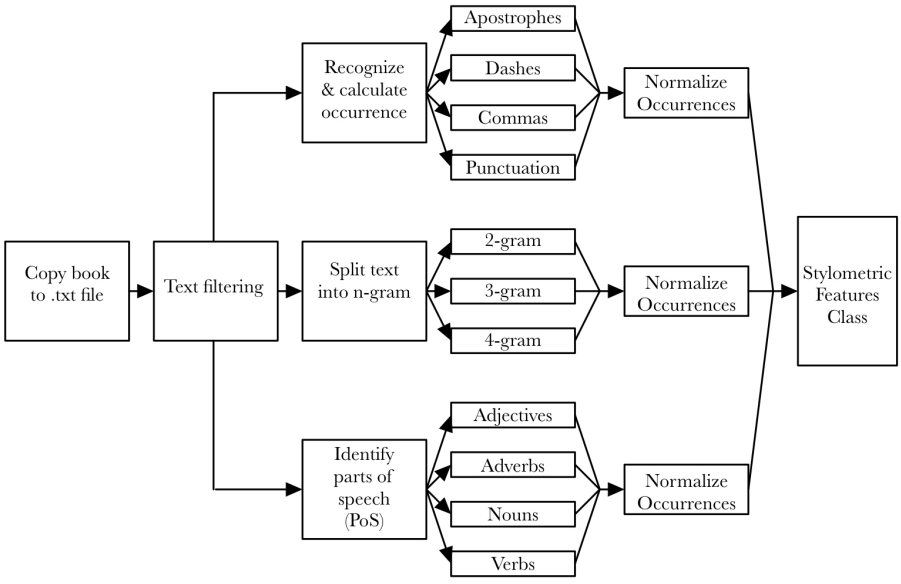

Fig. 2. Class Formation Made from Text File.

{'to go': 0.1523, 'I am' : 0.12...}

{'to go': 0.1432, 'I like': 0.023..}

Fig. 3. Dictionary Example.

module to identify PoS. For each PoS metric, the algorithm summed the number of adverbs, verbs, adjectives, and nouns. Then, it divided each metric by the number of sentences to identify occurrence per sentence.

C. Analysis of Data

The algorithm calculates each metric and represents the data either as a normalized value or occurrence of PoS per sentence. The process saves these values in an object. The Python objects contain the training data and the testing data. The algorithm calculates the percent error between these objects and stores these values for final assessment. The smaller the percent error, the metric has a greater influence verifying both files come from the same author.

$$\% \text{ error} = |\frac{\#\text{experimental} - \#\text{actual}}{\text{actual}}| \quad (1)$$

The algorithm incorporates two linear equations to determine the final prediction. The first equation has a greater priority on the authorship prediction and it uses the correlation exists in the n-gram, and the second equation uses the percent error for the PoS and punctuation. There are maximum weights each metric has in the prediction because no metric should have an overwhelming influence on the final prediction.

D. Exponential Function Explained

Through repeated trials in determining n-gram, we observed that the correlation values between the training data and test

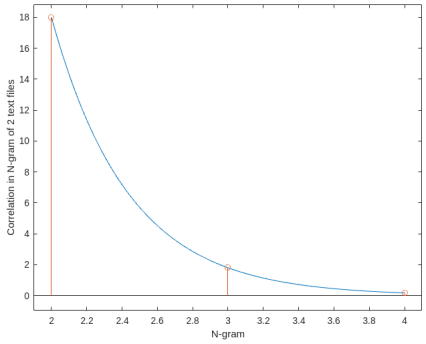

Fig. 4. Exponential Function Example.

data fell within these ranges, 2-gram is 15-20%, 3-gram is 1-3%, and 4-gram is 0.1-0.5%. We created an exponential decay function to determine authorship based on the correlation of the two files' metrics. Fig. 4 illustrates a decision function as an exponential decay function and the n-gram values of 2, 3, and 4 as a stem function. If the correlation value from the training and testing data is higher than the decision curve, the algorithm determines that the n-gram metric sways the decision such that both files have the same author. Anything below the curve concludes that they are not from the same author. The algorithm compares the correlation percentage to the decision curve, and the distance between the two indicates the level of certainty in the prediction. This percentage influences the weight given to n-gram analysis for final prediction.

E. Final Decision

The algorithm assesses each metric of syntax and n-gram and assigns a positive or negative value representing the

TABLE I
ACCURACY OF EACH MEASURED METRIC.

	Metrics										
	Apostrophe	Adjective	Adverb	Verb	Word Count	Comma	Dash	2-gram	3-gram	4-gram	Final Prediction
Moby Dick Train & Moby Dick Test	Correct	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct	Correct	Incorrect	Correct
Moby Dick Train & Alice in Wonderland Test	Correct	Correct	Incorrect	Incorrect	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct
Moby Dick Train & Kidnapped Test	Correct	Correct	Correct	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Incorrect	Incorrect
Moby Dick Train & Treasure Island Test	Incorrect	Incorrect	Incorrect	Correct	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect
Moby Dick Train & Typee Test	Incorrect	Correct	Correct	Correct	Correct	Correct	Correct	Incorrect	Incorrect	Correct	Correct
Moby Dick Train & The Conf. Man Test	Incorrect	Incorrect	Incorrect	Correct	Correct	Incorrect	Correct	Incorrect	Incorrect	Correct	Correct
Kidnapped Train & Moby Dick Test	Incorrect	Correct	Correct	Incorrect	Correct	Incorrect	Correct	Correct	Correct	Correct	Correct
Kidnapped Train & Alice in Wonderland Test	Correct	Correct	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Correct	Correct	Correct	Correct
Kidnapped Train & Kidnapped Test	Correct	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct	Correct	Correct	Correct
Kidnapped Train & Treasure Island Test	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Correct	Correct	Incorrect
Kidnapped Train & Typee Test	Incorrect	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct
Kidnapped Train & The Conf. Man Test	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct
Al in Wnd. Train & Moby Dick Test	Correct	Correct	Correct	Correct	Incorrect	Incorrect	Correct	Correct	Correct	Incorrect	Correct
Al in Wnd. Train & Al. in W. Test	Correct	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Incorrect	Correct	Correct	Correct	Correct
Al in Wnd. Train & Kidnapped Test	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct
Al in Wnd. Train & Treasure Island Test	Correct	Incorrect	Correct	Incorrect	Incorrect	Incorrect	Correct	Correct	Correct	Correct	Correct
Al in Wnd. Train & Typee Test	Correct	Correct	Correct	Correct	Incorrect	Incorrect	Correct	Correct	Correct	Incorrect	Correct
Al in Wnd. Train & The Conf. Man Test	Correct	Incorrect	Incorrect	Correct	Incorrect	Correct	Correct	Correct	Correct	Correct	Correct
Prediction Accuracy (%)	66.7	66.7	61.1	61.1	50	44.4	66.7	77.8	83.3	55.6	83.3

metric's correlation between the two files. This process sums each metric (as shown in Tab. I) to conclude authorship. The research indicates that n-gram is statistically more valid at determining authorship than syntax [3]. Thus, we gave a higher weight to n-gram in the final prediction.

IV. RESULTS

Tab. II demonstrates whether the algorithm's correlation matches the actual result. If the test file is from the author and a positive correlation existed between the two files within a metric, it has a value of 'Correct'. It's true when two files are not from the same author, and the algorithm observes no correlation within a metric, it has the value of 'Correct'. It's clear that the more reliable metrics were the 2-gram, 3-gram, and 4-gram, but correlation exists with apostrophes, adjectives, and dashes.

The majority of the percentages of each metric are below the overall assessment. This observation demonstrates a need for a proper balance of n-gram and syntactical values that help the algorithm verify authorship. Too heavy reliance on either metric would produce poorer results. It is not recommended to place equal weights on each indicator because it's clear that n-gram has a higher percentage of prediction success, and other metrics like word count don't produce good results. When an ML algorithm analyzes the text, different context demonstrates different results. For example, an ML algorithm could identify whether characters in a book are conversing, and then it could learn that conversation in a book is different from the narration.

As illustrated in Tab. II, the overall accuracy of 3 training data and 6 testing data per training data results in 83.3% accuracy. Moby Dick, Kidnapped, and Treasure Island were the three books that the algorithm incorrectly predicted. These books are written in the first-person. The n-gram portion of the algorithm had more commonalities between these three books because of the first-person perspective, but it predicted the correct author when Typee was tested. The two files which contained a third-person narrative were The Confidence Man and Alice in Wonderland, and the algorithm didn't err when guessing between other third-person narratives. The Alice in Wonderland training file succeeded with 100% accuracy, and this illustrates the algorithm performed well compared to writings from different authors from the same era.

V. CONCLUSION

This research proposed a simple approach to data computation to find an effective way to use syntactical metrics and n-gram to predict authorship. Each metric had a weight assigned to the correlation of the training set file and testing set file. The algorithm summed these metrics to get an overall 83.3% accuracy. The results show that the correlation of the n-grams favored the prediction results, while the 4-gram was significantly less accurate. We can change the decision curve's coefficients to yield better results for the n-gram. Furthermore, the results demonstrate that each metric alone doesn't accurately predict authorship (except 3-gram). According to the results of the test data, the metrics of adjectives, dashes, and apostrophes most accurately correlated with the algorithm's authorship prediction. Thus, the usage rate of these metrics should receive more influence in the algorithm's prediction than word count or commas. The results indicate the validation of the proposed approach for the authorship verification problem. It further demonstrates a bright future for using this software in other applications such as phishing detection.

REFERENCES

[1] "Stylometry," accessed: 2021-04-01. [Online]. Available: https://www.lexico.com/en/definition/stylometry/

[2] S. Schoenbaum, *Internal Evidence and Elizabethan Dramatic Authorship*. Northwestern University Press, 2018.

TABLE II
PREDICTION RESULTS FOR EACH BOOK.

		Testing Data Set						
		Moby Dick	Alice in Wonderland	Kidnapped	Treasure Island	Typee	Confidence Man	% Accuracy
Training Data Set	Moby Dick	Correct	Correct	Incorrect	Incorrect	Correct	Correct	66.7
	Kidnapped	Correct	Correct	Correct	Incorrect	Correct	Correct	83.3
	Alice in Wonderland	Correct	Correct	Correct	Correct	Correct	Correct	100
Total % Accuracy								**83.3**

[3] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175, 1994.

[4] "The characteristic curves of composition," *JSTOR*, vol. 9, no. 214, 1887. [Online]. Available: http://www.jstor.org/stable/1764604

[5] S. U. Hassan, M. Imran, T. Iftikhar, I. Safder, and M. Shabbir, "Deep stylometry and lexical syntactic features based author attribution on plos digital repository," *Digital Libraries: Data, Information, and Knowledge for Digital Lives*, 2021, [Online; accessed 05-August-2021].

[6] B. Epifantsev, P. Lozhnikov, and A. Sulavko, "Alternative authorization scenarios for identifying users by the dynamics of subconscious movements," *Information Security Issues, All-Russian Research Institute of Inter-Branch Information-Federal Information and Analytical Center for the Defense Industry (FGUP "VIMI")*, no. 2, pp. 28–35, 2013.

[7] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python—A Guide for Data Scientists*, 4th ed., D. Schanafelt, Ed. O'Reilly Media Inc, 2018, vol. 1.

[8] N. Cheng, R. Chandramouli, and K. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78–88, 2011.

[9] F. Iqbal, H. Binsalleeh, B. C. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," *digital investigation*, vol. 7, no. 1-2, pp. 56–64, 2010.

[10] C. E. Chaski, "Who's at the keyboard? authorship attribution in digital evidence investigations," *International journal of digital evidence*, vol. 4, no. 1, pp. 1–13, 2005.

[11] "Gutenburg," https://www.gutenberg.org/, accessed: 2021-04-01.