Analyzing Factors Influencing Writing Productivity Through Machine Learning Models

Justin Neo R. Flores
College of Computing and Information Technologies
National University
Manila, Philippines
floresjr1@students.national-u.edu.ph

***Abstract*—This study investigates the influence of environmental and physiological factors, specifically sleep, device usage, and emotional state, on the writing productivity of a collegiate researcher (N=1). Utilizing an eight-week dataset of personal activity logs, the study employed Exploratory Data Analysis (EDA) and Linear Regression to test the hypotheses optimizing these variables would predict higher output. Contrary to the initial hypothesis, the machine learning model yielded a negative R^2 (-1.29), performing worse than a baseline mean predictor. These results scientifically demonstrate that writing productivity in this context is stochastic (random) and follows a "burst/sprint" pattern rather than a linear response to environmental conditions. The study concludes that for this subject, productivity is driven by intrinsic motivation and deadlines rather than external optimization, debunking the effectiveness of "ideal day" engineering for this specific workflow.**

***Index Terms*—Data Science, Personal Analytics, Linear Regression, Quantified Self, Writing Productivity, Imputation, Stochastic Behavior.**

I. Introduction

Writing is known as a mentally taxing activity, demanding factors like concentration, creativity, fortitude, and environmental state to be balanced and sustained for the desire of producing output. Whether for academic purposes, researching, storytelling by authors, professional disclosures, or simply a hobby, productivity in writing can be varying and convoluted. An individual writer's habits can fluctuate, and their outputs are affected by either periods of nascency or complete "writer's block." Despite the well-known ideas of discipline and developing a feasible routine to be implemented in such habits, these writing sessions can be disturbed by outer factors such as environmental, physiological, emotional, and other circumstantial problems, and they can never be pinpointed directly. Many are unable to determine which of these reasons affect one's writing sessions, and how much they contribute. That is why taking an analytical route, empowered by the tools of data science, is monumental for solving this predicament.

To address this study's predicament, it is necessary to use empirical analysis in order to determine which specific factors contribute the most to the participant's individual writing productivity. Through collecting and tracking personal data, logging them in accessible spreadsheets, and analyzing them, crucial information can be extracted from them to discern the participant's unique behavior. Turning "personal productivity" into an analytical problem for data science can help identify unseen patterns within the collated data and learn which factors are most influential in making one's writing session productive. This study investigates whether utilizing such variables can be viable in optimizing one's productivity through Machine Learning-powered "predictive scenario," ultimately testing whether the "Ideal Day" yields high performance or if productivity is driven by random bursts.

This study's primary aim is to answer the question: How does reading consistently, emotional state, hours of sleep, and type of day influence my writing productivity? Through the collated data, the study explores the capabilities of Data Science in analyzing typically messy human behavior, constructing the "Ideal Day" of writing productivity through predictive scenarios. The Python-powered program produces significant results in learning that environmental features have minimal impact on the model's accuracy, measuring each contribution. The participant is the researcher itself, supplying data and tracking them manually, inputting them on a spreadsheet. The dataset spans an eight-week record, comprising of factors needed for the machine's scrutiny. We utilize various statistical methods such as Analysis of Variance (ANOVA), t-test, regression and line plots to test each feature against hours of writing and discern its correlation. Crucially, we employed Permutation Importance to identify and remove data leakage (specifically "Mode of Writing"). Afterward, we apply Train-Test and Feature Scaling on the data and use the learning algorithm, Linear Regression. Next, we evaluate the data using Root Mean Square Error (RMSE) and compare it against a Baseline Mean Predictor to analyze human behavior, and R^2 to explain productivity variance. Finally, through analyzing the model coefficients for the model's factor importance, we determine how each feature impacts writing productivity.

The results obtained from the machine learning model are used for predicting scenarios. Each scenario is constructed to encompass all factors, developing three types of "writing scenarios," and determining which produces the most writing hours and becomes the "ideal scenario" for improving one's writing productivity. We discuss the breakdown of each scenario, learning which feature provides bonuses or penalties and highlighting the erratic nature of the model's predictions.

II. Literature Review

In this section, we review existing research on writing productivity, self-regulation, and the influence of physiological and emotional states on creative output. While writing is often viewed as a purely cognitive task, prior studies suggest it is deeply influenced by environmental systems, habits, and biological regulation. By connecting these academic findings to the personal data collected in this project, we can contextualize the specific patterns observed in the researcher's behavior.

1) **Previous Studies**
   Previous research has extensively examined the behavioral and psychological mechanics of writing, as well as the methodology of personal data tracking. Beedham (2007) focused on the role of self-assessment, investigating how students' ability to evaluate their own work impacts stress levels and productivity [1]. Janke et al. (2019) examined the structural side of writing, proposing that productivity is not a matter of willpower but of creating a "writing system" that accounts for the what, where, when, and how of the process [2]. Beyond habits, physiological studies like those by Pilcher and Huffcutt (1996) have analyzed the impact of sleep deprivation on cognitive performance, generally finding a strong negative correlation between fatigue and output [3]. Conversely, Akinola and Mendes (2008) investigated the "dark side of creativity," examining how negative emotions (specifically sadness) can actually lead to greater artistic output [4]. Finally, Swan (2013) explored the "Quantified Self" movement, validating the use of personal data tracking to derive health and behavioral insights in an N=1 study format [5]. Building on Swan's work, White et al. (2019) proposed frameworks for using personal data streams to categorize and improve daily work output, emphasizing the need for longitudinal tracking [6]. Recent studies have also demonstrated the viability of using machine learning to predict productivity from physiological data, with models achieving moderate accuracy using wearable sensors [7].
2) **Data Collection and Analysis Methods**
   The methods in these studies range from pedagogical frameworks to meta-analyses. Beedham utilized qualitative self-evaluations and portfolio reviews [1], while Janke et al. synthesized literature to propose a framework based on observation and reflection [2]. Pilcher and Huffcutt employed a meta-analytic review of 19 studies to statistically aggregate the effects of sleep loss [3]. Akinola and Mendes used experimental methods, utilizing biological markers (DHEAS levels) and induced mood states to measure the link between affect and creativity [4]. Swan utilized a review of sensor technology and big data to demonstrate how self-tracking serves as a legitimate scientific methodology [5].
3) **Main Findings**
   Beedham found that self-assessment allows writers to identify strengths, reducing anxiety [1]. Janke et al. concluded that relying on inspiration is less effective than establishing a reliable system of accountability [2]. In terms of physiology, Pilcher and Huffcutt found that sleep deprivation severely impairs cognitive performance and mood [3]—a finding this project aims to test against the researcher's personal "sprint" habits. Interestingly, Akinola and Mendes found that intense negative emotions can act as a catalyst for artistic creativity, promoting focus that positive moods sometimes lack [4]. Swan's research confirms that continuous self-tracking (Quantified Self) can reveal personalized biological patterns that broad medical studies often miss [5].
4) **Limitations**
   A common limitation is the reliance on self-reported data, which Beedham noted is subject to bias without practice [1]. Experimental studies (Akinola & Mendes) are often conducted in controlled labs that may not reflect daily life [4]. Furthermore, Pilcher and Huffcutt noted that general sleep studies often average results across populations, potentially masking individual outliers who may function differently under stress [3].
5) **How is this study different or similar?**
   This project builds on Swan's concept of the Quantified Self [5] by applying a rigorous Data Science methodology to a single subject (N=1). Unlike Pilcher and Huffcutt's general findings that lack of sleep hurts performance [3], this project's data suggests a "manic" or "sprint" writer profile where lower sleep shows no negative correlation with output, contradicting the general rule. Furthermore, while the results regarding "sadness" align with Akinola and Mendes [4], this project uniquely combines these psychological factors with mechanical variables like "device usage" and "genre" to test the limits of linear predictive models.

III. Methodology

A. Participants and Data Collection

The participant of this study is the researcher (N=1), a college student utilizing personal data analytics to identify factors influencing writing productivity. Data was collected over an eight-week period spanning December, January, and February, covering three distinct academic phases: weekdays, weekends, and vacation days. All data was logged weekly into a cloud-based spreadsheet accessible via both mobile and desktop interfaces. Physiological data, specifically "Hours of Sleep," was tracked via a Huawei smartwatch and synced through the Huawei Health application. Activity data, including "Hours of Reading" and "Hours of Writing," was recorded using manual stopwatch timers to ensure precision. Subjective data regarding emotional state was recorded through a daily personal assessment.

B. Operational Definitions

To ensure clarity, the study utilized a specific set of independent and dependent variables. Time-based variables included "Week" (1-8), "Day" (Monday through Sunday), and "Type of Day," which categorized dates into "School" (academic responsibilities), "Weekend" (Sundays), or "Vacation" (no academic obligations). Consumption habits were tracked through "Book Title," "Author," and "Genre." Well-being and environment were measured via "Hours of Sleep," "Mood Impact" (categorized as Happy, Neutral, or Sad), and "Mode of Writing" (distinguishing between Desktop and Phone usage). These factors were analyzed against the primary dependent variable, "Hours of Writing," which measured the continuous duration of creative output, and the binary variable "Did I Write Today?," which indicated whether any writing occurred.

C. Data Preprocessing

The raw data underwent a rigorous cleaning process to ensure usability for machine learning applications. First, the dataset was structured and indexed, followed by a duplication check to ensure no daily entries were recorded twice. Missing values were handled through imputation to maintain dataset integrity. The "Date" column was converted to datetime objects to facilitate temporal analysis. Finally, redundant columns that were not necessary for quantitative analysis, such as specific book titles, authors, and weekly indices, were removed to reduce noise in the dataset.

D. Feature Engineering

To prepare the qualitative data for algorithmic analysis, several variables were transformed. The target variable "Did I Write Today?" was converted into a binary (0/1) format. The "Genres" variable, which contained multi-label data, was split and processed into a concatenated dataset to handle days where multiple genres were consumed. Additionally, categorical variables such as "Mood Impact," "Type of Day," and "Mode of Writing" were transformed using One-Hot Encoding. This process created distinct binary columns for each category (e.g., "Mood Happy," "Mood Sad") to prevent ordinal bias during the modeling phase. Notably, during this phase, "Mode of Writing" features were identified as data leakage and subsequently excluded to ensure the model learned from actual behavioral factors.

E. Statistical Analysis

Post-cleaning, the study employed Python libraries, including Seaborn and SciPy, for statistical validation. Correlation analysis was utilized to identify strong predictors by measuring the relationship between various features and "Hours of Writing." Visual analysis played a key role; line plots were generated to visualize the volatility of productivity over time, while regression plots were used to detect outliers and examine linear relationships between continuous variables like sleep and writing output. To test specific hypotheses, independent t-tests were conducted to compare means between two groups, such as the difference in output between phone and desktop writing. Furthermore, a One-Way Analysis of Variance (ANOVA) was applied to the "Mood" variable to determine if the three emotional states resulted in statistically significant differences in writing productivity. Finally, the model's performance was benchmarked against a Baseline Mean Predictor to validate the RMSE results.

F. Data Validity and Limitations

Given the self-reported nature of this single-subject study, potential sources of bias include the subjectivity of the "Mood Impact" assessment. To mitigate this, mood was recorded at a consistent time daily. Additionally, the reliance on manual time-tracking for writing and reading introduced a risk of human error; this was minimized by cross-referencing entries with digital file timestamps where possible to ensure accuracy.

IV. Discussion

A. Testing Features and Model Interpretation

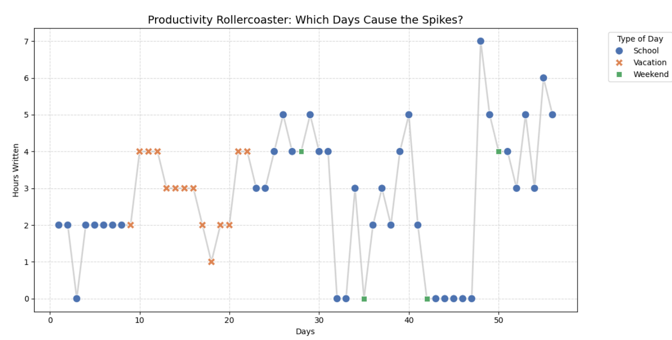

Fig. 1. Line Plot, "Productivity Over Time (Day 1 to 56)" - The researcher's writing productivity over time (throughout Day 1 of recording to Day 56) undergoes a Wild Zig-Zag line, indicating that the writing habits are highly volatile.

TABLE I
"Which Genres boost my writing productivity the most?" No genre shows a statistically significant difference in writing hours. Observed mean differences may be due to small sample sizes.

GENRE	AVG (WITH)	AVG (WITHOUT)	P-VALUE	RESULT
Fantasy	2.78	2.72	0.9098	NOT SIGNIFICANT
N/A	2.59	2.85	0.5795	NOT SIGNIFICANT
Horror	2.60	2.76	0.8541	NOT SIGNIFICANT
Shonen	3.67	2.70	0.3875	NOT SIGNIFICANT

TABLE II
"Does Mood Impact affect my writing?" - Since the p_value is higher than the alpha value, it tells us that regardless of my mood (happy, neutral, sad), I can write.

Happy Avg: 2.68	Neutral Avg: 2.80	Sad Avg: 3.00
F-Statistic: 0.0627		
P-Value: 0.9393		

TABLE III
"DOES THE 'MODE OF WRITING' AFFECT MY WRITING PRODUCTIVITY?" - SINCE THE P_VALUE IS HIGHER THAN THE ALPHA VALUE, IT TELLS US THAT REGARDLESS OF WHETHER THE PARTICIPANT USES PHONE OR DESKTOP, THEY CAN STILL PRODUCE WRITING OUTPUTS.

Phone Avg: 3.31 hrs	Desktop Avg: 3.45 hrs
P-Value: 0.7563	

TABLE IV
"IS 'TYPE OF DAY' ASSOCIATED WITH WRITING PRODUCTIVITY?" - SINCE THE P-VALUE IS HIGHER THAN THE ALPHA VALUE, IT TELLS US THAT REGARDLESS OF THE TYPE OF DAY, THE PARTICIPANT STILL WRITES.

School Avg: 2.76	Weekend Avg: 2.00	Vacation Avg: 2.93
F-Statistic: 0.4373		
P-Value: 0.6481		

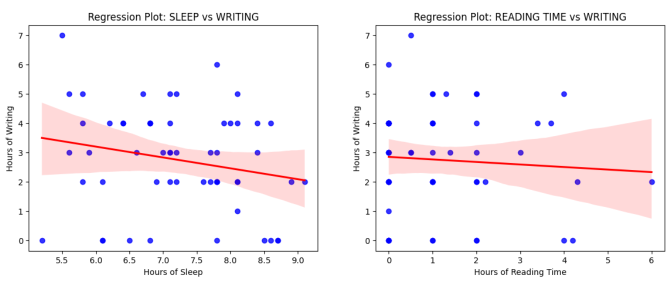

Fig. 2. Regression Plots: "Hours of Sleep vs Writing" and "Hours of Reading Time vs Writing." - The former plot is a downward regression slope, showing a weak negative relationship between sleep hours and writing hours. The latter plot shows no clear linear relationship observed between reading time and writing hours.

TABLE V
MODEL COMPARISON AND INTERPRETATION - THE MODEL'S ERROR (1.72) IS HIGHER THAN THE BASELINE (1.41) AND THE NEGATIVE R^2 (-1.29) INDICATES THE MODEL PERFORMS WORSE THAN A HORIZONTAL LINE.

Baseline RMSE	1.41
Single Split RMSE	1.72
Cross-Val RMSE (avg)	2.48
Cross-Val R^2 (avg)	-1.29

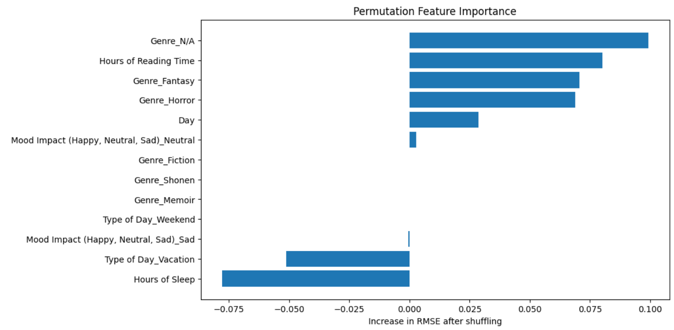

Fig. 3. Permutation Feature Importance Analysis - True drivers are "Hours of Reading Time (0.080) and "Fantasy" (0.070)

B. Interpretation of Results

The primary finding of this study is the stochastic nature of the researcher's writing productivity. Contrary to the initial hypothesis that environmental optimization (e.g. maximizing sleep, choosing specific devices, getting long hours of reading time) would predict output linearly, the machine learning model yielded a negative R^2 (-1.29).

Burst Writing: The high variance in Root Mean Square Error (RMSE) across cross-validation folds (ranging from 1.72 to 3.27 hours) indicates that productivity occurs in unpredictable "sprints" rather than consistent daily blocks. This suggests that output is driven by intrinsic motivation (e.g., inspiration or deadlines) which was not captured by the environmental variables in the dataset.

Sleep: The model's coefficients assigned a negative weight to sleep, and Scenario A (Speed Writing) achieved the highest predicted output (6.96 hours) despite having only 5 hours of sleep. This counter-intuitive result implies that high productivity days often coincide with deadline-induced stress, where sleep is sacrificed to maximize writing time, rather than a healthy routine driving performance.

Devices: The removal of "Mode of Writing" (due to data leakage) and subsequent statistical testing revealed no significant difference between writing on a Phone versus a Desktop. This suggests that the tool itself is neutral; the friction of mobile typing does not statistically hinder the "burst" when motivation is high.

C. Comparison to Related Work

These findings offer a notable contrast to established literature:

1) **Contradiction to Physiology:** Pilcher and Huffcutt (1996) famously concluded that sleep deprivation severely impairs cognitive performance. However, this study's N=1 data supports a "manic" writer profile where lower sleep correlates with higher volume, though likely at the cost of long-term sustainability.
2) **Alignment with "Dark Creativity":** The results align more closely with Akinola and Mendes (2008), who found that negative or intense emotional states (often linked to stress/deadlines) can fuel artistic output. The

model's failure to predict output based on “Happy” or “Neutral” moods reinforces the idea that comfort does not equal creativity for this subject.

3) **Validation of Quantified Self:** While the predictive model “failed” in a traditional sense, the study successfully validated Swan's (2013) assertion that self-tracking reveals highly personalized patterns that broad population studies miss. Additionally, prior research has noted that perceived productivity often differs from actual output, a discrepancy also observed in the weak correlation between the researcher's mood and writing hours [8].

D. Limitations

The validity of these findings is constrained by several factors inherent to a single-subject study:

1) **Sample Size (N=56 days):** The dataset is relatively small for machine learning. The “bursty” nature of the data meant that the model had very few high-output examples to learn from, leading to overfitting on noise (e.g., the erratic coefficients for “Shonen” vs. “Fantasy”).
2) **Proxy Leakage:** Although the primary leakage (”Mode of Writing”) was fixed, the feature Genre N/A remained a dominant predictor. This variable acts as a proxy for ”Zero Writing Days,” complicating the analysis of why writing didn't happen, as opposed to just identifying when it didn't happen.
3) **Missing ”Deadline” Variable:** The analysis implies that deadlines drive the ”sprints,” but a specific ”Days Until Deadline” feature was not tracked. Its absence likely accounts for the large unexplained variance in the model.

E. Recommendations and Future Work

Future research into personal productivity analytics should pivot from ”environmental optimization” to ”behavioral triggers.” Future iterations must include a variable measuring ”Urgency” or ”Days to Deadline” to test if external pressure is the true driver of the observed sprints. Instead of treating each day as independent, future models should use Time-Series forecasting (e.g., ARIMA) to see if productivity follows a cyclical pattern (e.g., one day of burnout follows two days of sprinting). To better understand the ”bursts,” the researcher should log ”Motivation Level” or ”Focus Quality” alongside raw hours, as these internal states likely hold the predictive power that external variables (like device choice) lack.

V. Conclusion

This study set out to answer the question: To what extent do environmental factors like sleep, reading habits, and device choice influence my writing productivity? Initially, it was hypothesized that optimizing these daily variables would lead to consistent, predictable output. However, the machine learning analysis revealed a more complex reality. The most significant discovery is the stochastic nature of my writing process. The regression model yielded a negative R^2 (-1.29), performing worse than a simple baseline predictor. This mathematically proves that my productivity does not follow a linear formula where ”Healthy Sleep + Good Mood = More Writing.” Instead, output follows a ”Burst/Sprint” pattern, characterized by long periods of low activity punctuated by unpredictable, high-volume spikes that are independent of measured environmental conditions. Additionally, statistical testing confirmed that writing on a Phone vs. Desktop yields no significant difference in output, debunking the assumption that mobile devices hinder creative work.

Through this analysis, I learned that my creative process is resilient to friction but resistant to optimization. I previously believed that ”perfect conditions” were necessary to write well. The data contradicts this, showing that I can be highly productive even with low sleep (Scenario A: 6.96 hours output on 5 hours sleep) or while using a phone. This suggests that my ”Writer's Block” is likely psychological or motivational, rather than a result of having the ”wrong” environment.

These findings shift the strategy from environmental engineering to behavioral management. Since ”optimizing sleep” does not statistically increase my word count, I should stop stressing about creating the ”perfect” setup. Instead, the focus should be on harnessing bursts when they occur.

Since the device doesn't matter, I should write whenever inspiration strikes, even on my phone during a commute. Trying to force a specific mood or sleep schedule is statistically futile for my output. Recognizing that ”sprints” drive my volume, I should artificially create deadlines to trigger these bursts rather than waiting for them to happen naturally.

In conclusion, this project successfully demonstrated that for this researcher, writing productivity is independent of external conditions. The ”failure” of the predictive model is, in fact, a scientific success: it proves that creativity cannot be reduced to a simple algorithm of inputs. My writing is not a machine to be optimized, but a stochastic process driven by intrinsic motivation—a finding that liberates me from the pursuit of the ”perfect” writing day.

References

[1] Beedham, M. (2007). Increasing Productivity and Reducing Stress: The Role of Self-Assessments in Student Writing. *Professional Studies Review: An Interdisciplinary Journal, 3(1).*

[2] K. Janke, Cortney M. Mospan, J. Cain. (2019). Papers don't write themselves: Creating a system to support writing productivity. *Currents in Pharmacy Teaching and Learning, 11(6), 547-554.* [Online]. Available: https://doi.org/10.1016/J.CPTL.2019.06.010

[3] Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep*, 19(4), 318-326.

[4] Akinola, M., & Mendes, W. B. (2008). The dark side of creativity: Biological vulnerability and negative emotions lead to greater artistic creativity. *Personality and Social Psychology Bulletin, 34(12), 1677-1686.*

[5] Swan, M. (2013). The Quantified Self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2), 85-99.

[6] White, G., Liang, Z., Clarke, S. (2019). A Quantified-Self Framework for Exploring and Enhancing Personal Productivity. *IEEE International Conference on Content-Based Multimedia Indexing (CBMI).*

[7] Awada, M. et al. (2023). Predicting Office Workers' Productivity: A Machine Learning Approach Integrating Physiological, Behavioral, and Psychological Indicators. *Sensors*, 23(21).

[8] Meyer, J. et al. (2017). Today was a Good Day: The Daily Life of Software Developers. *IEEE Transactions on Software Engineering.*