Scientific Research Publishing

Journal of Software Engineering and Applications, 2024, 17, 463-473
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

# User Profile & Attitude Analysis Based on Unstructured Social Media and Online Activity

**Yuting Tan, Vijay K. Madisetti**

School of Cybersecurity and Privacy, Georgia Institute of Technology, Atlanta, GA, USA
Email: ytan328@gatech.edu, madisetti.vijay@gmail.com

**How to cite this paper:** Tan, Y.T. and Madisetti, V.K. (2024) User Profile & Attitude Analysis Based on Unstructured Social Media and Online Activity. *Journal of Software Engineering and Applications*, **17**, 463-473.
https://doi.org/10.4236/jsea.2024.176025

**Received:** April 24, 2024
**Accepted:** June 4, 2024
**Published:** June 7, 2024

## Abstract

As social media and online activity continue to pervade all age groups, it serves as a crucial platform for sharing personal experiences and opinions as well as information about attitudes and preferences for certain interests or purchases. This generates a wealth of behavioral data, which, while invaluable to businesses, researchers, policymakers, and the cybersecurity sector, presents significant challenges due to its unstructured nature. Existing tools for analyzing this data often lack the capability to effectively retrieve and process it comprehensively. This paper addresses the need for an advanced analytical tool that ethically and legally collects and analyzes social media data and online activity logs, constructing detailed and structured user profiles. It reviews current solutions, highlights their limitations, and introduces a new approach, the Advanced Social Analyzer (ASAN), that bridges these gaps. The proposed solution's technical aspects, implementation, and evaluation are discussed, with results compared to existing methodologies. The paper concludes by suggesting future research directions to further enhance the utility and effectiveness of social media data analysis.

## Keywords

Social Media, User Behavior Analysis, Sentiment Analysis, Data Mining, Machine Learning, User Profiling, Cybersecurity, Behavioral Insights, Personality Prediction

## 1. Introduction

The widespread use of social media and other types of online activity, such as shopping behavior, across various age groups makes it a great platform for individuals to share their lives, opinions, and preferences. This behavioral data is invaluable to businesses, researchers, policymakers, and the cybersecurity sector, offering insights into user personalities that can enhance products, understand social trends, make informed decisions, and identify potential cyber risks. However, the data's unstructured nature, stemming from the diverse functionalities and freedom afforded by social media platforms, poses significant challenges in extracting relevant information. Most current tools and solutions offer limited features in effectively retrieving and analyzing such data. Therefore, there is an urgent need to develop an advanced tool that not only collects and analyzes user data ethically and legally but also constructs structured and comprehensive social media user profiles. This paper begins by reviewing several existing solutions and their limitations. It then introduces a proposed solution that addresses these gaps and meets social needs. The paper delves into the technical aspects of the project's implementation and evaluation, analyzes the results, and compares the proposed solution with existing ones, highlighting the improvements made. Finally, it concludes with a summary of the project and paper and suggests potential directions for future research based on the current solution's limitations.

## 2. Existing Solutions and Limitations

Several existing solutions, including open-source tools and commercial software, serve as fundamental tools for collecting and analyzing social media profiles.

### 2.1. Open-Source Tools

***Qeeqbox Social Analyzer***

- This tool is developed with a focus on enhancing cybersecurity measures by analyzing and identifying potentially fake or malicious social media profiles across more than 1000 social media platforms. Its primary goal is to safeguard users from online threats by providing detailed profile assessments [1].
- **Limitation—Comprehensive Feature Shortcomings:** Predominantly focused on the identification of security threats, this tool lacks the capability to perform broader analyses on content or to deduce psychological insights such as user personality traits, thus constraining its applicability to security assessments.

***Sherlock Project***

- The Sherlock Project is an open-source utility designed to facilitate the aggregation of social media profiles by querying usernames across more than 300 platforms. It aims to construct a comprehensive digital footprint of individuals, which is essential for investigations and digital identity tracking [2].
- **Limitation—Analytical Depth Deficiencies:** The Sherlock Project, while efficient in aggregating user account data across platforms, does not extend its functionalities to the analysis of content or the extraction of behavioral insights, thus limiting its utility to only presence and location of accounts without deeper behavioral analysis such as user-generated data identification and sentiment analysis.

**General Limitation:** Deployment and operational management of open-source tools frequently demand substantial technical expertise. The absence of extensive documentation and user support further exacerbates the challenge, rendering these tools less accessible to individuals lacking technical proficiency.

## 2.2. Commercial Software

***Hootsuite***

- Hootsuite is a commercial platform that integrates management of multiple social media accounts, enabling post scheduling and real-time keyword monitoring. It addresses the need for coordinated social media strategies and analytics by providing tools that streamline posting schedules and track engagement metrics [3].
- **Limitation—Prohibitive Cost Factors:** Full access to Hootsuite's capabilities necessitates a premium subscription, which may represent a financial barrier to some entities. Additionally, the platform's primary focus on management rather than deep analytical tasks limits its effectiveness in comprehensive user-generated content analysis.

***Sprout Social***

- Sprout Social offers a suite of management tools that facilitate deep engagement with social media content, coupled with analytics to measure audience interaction. The platform seeks to optimize user engagement through data-driven insights and tailored content delivery [4].
- **Limitation—Functionality Focused Limitations:** While Sprout Social excels in monitoring user engagement and sentiment, it lacks the necessary features to conduct in-depth analyses such as personality assessments or detailed examinations of user-generated content.

***Dripify***

- Dripify specializes in automating LinkedIn activities, targeting professionals and businesses seeking to enhance their online engagement without manual intervention. It focuses on streamlining lead generation and sales processes through automated sequences, thus maximizing efficiency on LinkedIn [5].
- **Limitation—Platform-Specific Applicability Constraints:** Dripify's exclusive focus on LinkedIn restricts its functionality to a single platform, thus failing to meet the analytical requirements across diverse social media platforms or to provide insights beyond professional engagement.

**General Limitation:** Commercial software often incorporates complex features that require a steep learning curve and are predominantly accessible only through higher-cost subscriptions, thereby limiting their utility to entities capable of sustaining such financial commitments.

# 3. Proposed Solution—ASAN

This paper develops a proposal that aims to enhance the existing solutions by developing a comprehensive tool, the Advanced Social Analyzer or ASAN, extending on the Qeeqbox Social Analyzer that incorporates advanced data mining capabilities, along with sentiment and behavioral analysis techniques. This enriched and enhanced tool will collect not only metadata and page contents but will also identify user posts, or online activity (shopping, travel, or other traces of "digital exhaust"), and social or business activities to construct structured, detailed profile reports. These profiles will summarize user personalities, attitudes, and opinions on key topics, based on the personality models of social trends. **Figure 1** illustrates the architecture and workflow of the entire program.

During the operation of the program, the web scraping engine will collect data from social media platforms and other online sources. This data will be filtered to isolate crucial behavioral components and attitude analyses for personality prediction. Following a dynamic analysis of these personality prediction features, the program will generate profiles for specific users, classifying their personalities according to various models.

We selected two commonly-used personality models to generate user profiles within our analysis framework: the Big 5 (also known as the OCEAN model) and the Myers-Briggs Type Indicator (MBTI model). These models are widely recognized for their comprehensive approach to personality assessment, providing a robust framework for analyzing and interpreting user-generated content on social media platforms.

***Big* 5 *Personality Model***

The Big Five personality model breaks down human personality into five broad dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). This model is widely used in psychology for research and assessment, reflecting how people interact with the world and each other [6].

***MBTI Model***

The Myers-Briggs Type Indicator (MBTI) is a personality framework that categorizes people into 16 different personality types based on four dichotomies: Introversion vs. Extraversion, Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving. It is commonly used in career counseling, team building, and personal development to understand work styles and interpersonal dynamics [7].

ASAN significantly enhances existing solutions by substantially reducing both tangible and intangible costs associated with the acquisition, implementation, and operation of social media analysis tools without corruption to functionality or usability. Tangible costs, such as subscription fees for premium features, are eliminated, making the tool financially accessible. Simultaneously, intangible costs—such as the labor and time expended by teams to deploy, maintain, and integrate various tools—are also minimized. Furthermore, the tool simplifies the data collection and analysis process, making it easily operable by individuals without technical background. For technical users, it provides an integrated solution that addresses the common challenges of tool interoperability, thereby preventing potential conflicts and redundancies during tool integration. This dual approach not only democratizes data analysis, making it accessible to a broader audience but also enhances efficiency for users with varying levels of expertise.

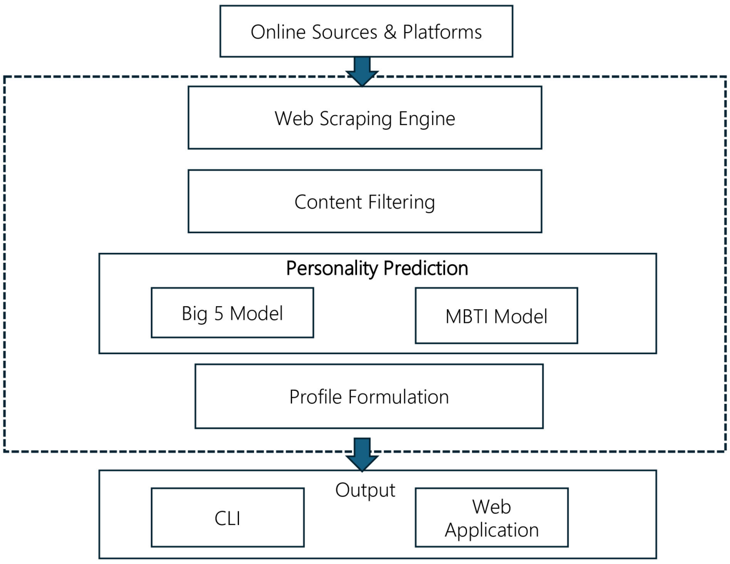

**Figure 1.** General program flow.

## 4. Implementation Details

Figure 2 offers an advanced overview of our software system design of ASAN. It retains all functionalities of the Qeeqbox Social-Analyzer, each configured within separate code files. The process controller manages the overall workflow, handling user input, invoking the appropriate functionality based on user-defined settings, and generating the desired reports.

### 4.1. User Generated Data Filtering

To facilitate the identification of user behavioral data, static analysis techniques were integrated into the process controller, emphasizing simplicity and efficiency. Through comprehensive research into the structures of prominent social media platforms and other online activity, our methods were meticulously tailored to leverage web scraping and content filtering. Utilizing the BeautifulSoup Python library [8], these methods specifically target commonly utilized HTML tags associated with user-generated content, optimizing data extraction accuracy and speed.

The output handler was enhanced with versatile flag options, enabling the selective activation of the user-generated data extraction feature. This enhancement includes additional fields in the output reports that display user posts, enriching the data presentation. These output configurations, standardized across the project's various implementations, ensure uniformity and coherence in report generation.

### 4.2. Big 5 Personality Prediction

According to the architecture of the tool selected for Big 5 personality prediction, the Big 5 prediction feature is separated from the main process controller to enhance compatibility with the ASAN. This distinct configuration ensures

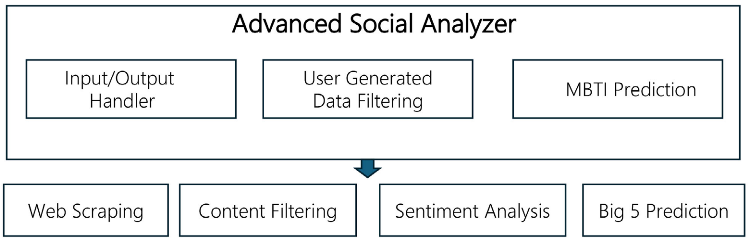

**Figure 2.** Software architecture of ASAN.

seamless interaction between the core functionalities of the social-analyzer and the isolated Big 5 prediction module, which leverages open-source machine learning algorithms from the sklearn Python library [9]. The dynamic analysis capability of this module is further refined through training on datasets sourced from the myPersonality project, initially associated with Facebook user data.

The implementation of the Big 5 prediction feature initiated with the elimination of redundant web scraping functionalities that were exclusively focused on retrieving data from Facebook. This modification primarily aims at streamlining the programming processes, thereby simplifying the system's architecture, and reducing its computational load.

In the subsequent phase of tool design, which previously relied solely on pre-processed databases for input, the capability to accept customized user inputs is incorporated. This enhancement is carefully overlaid onto the existing data collection framework without disrupting the core functionalities of the tool, thereby preserving its integrity and robustness.

### 4.3. MBTI Personality Prediction

The MBTI prediction feature is strategically integrated as an internal part of the process controller, aligning with the program's overall design principles to enhance the analytical efficiency of the tool. This feature employs a suite of diverse machine learning models from the sklearn Python library to facilitate dynamic analysis. Model training leverages a publicly available dataset from Kaggle [10], ensuring the applicability and robustness of the predictive outputs.

The initial phase of the MBTI prediction feature's implementation involves the elimination of redundant text filtering functionalities. These are previously managed by the social analyzer and are removed to streamline the process, thereby reducing unnecessary computational overhead, and simplifying the system architecture.

Furthermore, during the integration phase, additional functionalities such as dataset analysis and testing, which are irrelevant to the core project objectives, are carefully eliminated. This process is executed with precision to ensure it does not disrupt the ongoing analytical processes, maintaining the integrity and focus of the program.

All discrete components have been seamlessly integrated with the QeeqBox Social Analyzer. Comprehensive testing phases were conducted to ensure that the entire program operates without defects. This validation process includes unit tests, integration tests, and system-wide checks to confirm flawless functionality and to certify that the program is entirely bug-free.

## 5. Evaluation Details and Results

### 5.1. Evaluation Design

Given the substantial resources and efforts needed for automatic evaluation methods, manual testing and data collection are employed to assess the Advanced Social Analyzer (ASAN). For a thorough evaluation, distinct datasets have been meticulously curated for each personality model to ensure their validity and relevance.

For the Big 5 personality model, the evaluation is conducted utilizing a specialized dataset compiled by the Media Manager Club. This dataset focuses on a project that manually assesses the personality types of prominent celebrities based on Twitter posts [11]. All 141 available records are systematically included in the evaluation process. This approach allows for a targeted analysis of social media behavior and its correlation to the Big 5 personality traits, providing a relevant context for evaluating the model's predictive accuracy.

For the MBTI personality model, the assessment employs two significant sources. The first is a public dataset from Crystal, a leading analytics service that specializes in user behavior and preference analysis [12]. The second source is the Personality Database, a dynamic community platform that engages users in the analysis and discussion of both real and fictional characters' personalities based on established psychological theories [13]. This combination of sources enriches the dataset with a diverse range of inputs, enhancing the evaluation framework for the MBTI model. 150 records of real celebrities are randomly selected and systematically included in the evaluation process to ensure the validity and relevance of the testing results.

The accuracy of the prediction results for both models is quantitatively determined using the equation in **Figure 3** [14]. The accuracy of each personality trait across different models is calculated separately before computing the average value to ensure more reliable results.

This equation provides a clear metric for evaluating the effectiveness of the ASAN in accurately predicting personality traits based on user-generated content across various social media platforms, or in online applications.

### 5.2. Evaluation Results of ASAN

$$\text{Accuracy} = \frac{\text{True Negatives} + \text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

**Figure 3.** Accuracy evaluation equation.

#### 5.2.1. Big Five Personality Traits Accuracy

**Figure 4** shows the ASAN's accuracy in identifying the Big Five personality traits:

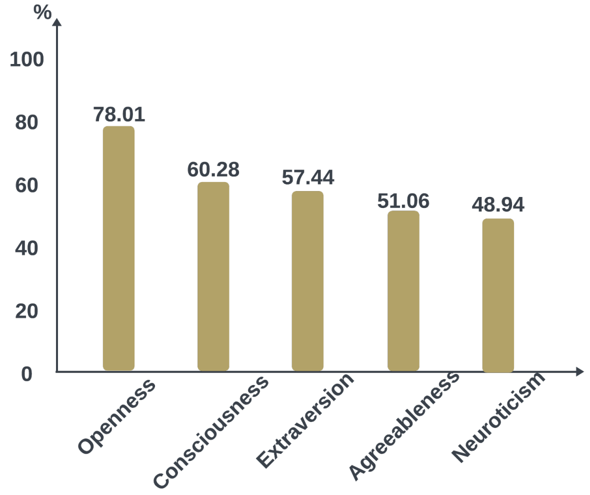

**Figure 4.** Big 5 prediction evaluation results.

- **Openness:** The tool exhibited the highest accuracy in predicting openness, with a percentage of 78.01%. This indicates a strong predictive capability regarding individuals' receptiveness to new experiences and creative thinking.
- **Conscientiousness:** With an accuracy of 60.28%, ASAN showed a good level of reliability in predicting conscientiousness, reflecting the ability to assess traits related to organization and responsibility.
- **Extraversion:** Corresponding to the MBTI extraversion dimension, ASAN's accuracy for predicting extraversion in the Big Five model was 57.44%, affirming a reasonable accuracy in identifying sociable and energetic behavior.
- **Agreeableness:** ASAN scored 51.06% in predicting agreeableness, indicating a moderate performance in detecting cooperative versus competitive dispositions among individuals.
- **Neuroticism:** The accuracy for neuroticism was 48.94%, which was the lowest among the Big Five traits but still underlines a fair predictive capability in identifying individuals' tendencies toward emotional stability.

### 5.2.2. Myers-Briggs Type Indicator (MBTI) Accuracy

Figure 5 illustrates the accuracy of ASAN in determining the four dichotomies of the MBTI model:

Extraversion vs. Introversion: The tool achieved 67.33% accuracy in predicting the correct orientation towards extraversion or introversion, indicating a substantial capability to discern between outward and inward energy focus.

- **Sensing vs. Intuition:** In distinguishing between a sensing and an intuition preference, the tool's accuracy was 59.33%. This suggests moderate reliability in predicting whether individuals are more likely to interpret information directly through their senses or through pattern recognition and possibilities.
- **Thinking vs. Feeling:** The tool demonstrated 47.33% accuracy in the thinking vs. feeling dimension. This nearly balanced score is the lowest among the four MBTI dimensions, indicating a challenge for the tool in reliably predicting decision-making based on logic versus values.

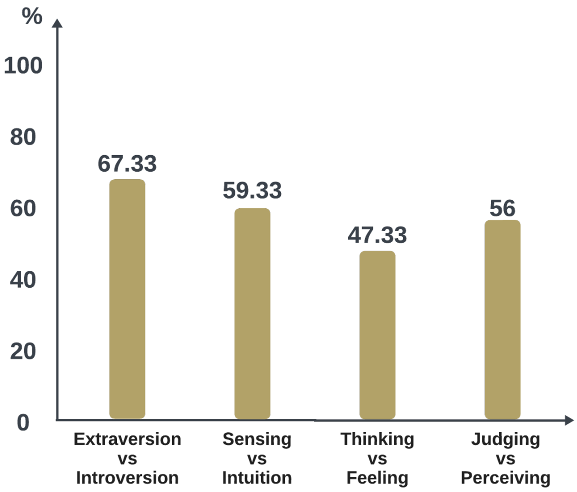

**Figure 5.** MBTI prediction evaluation results.

• **Judging vs. Perceiving:** The accuracy of predicting a preference for judging or perceiving was 56%. This result points to a moderately reliable performance of the tool in forecasting whether individuals tend to prefer structured or adaptable lifestyles.

In conclusion, ASAN demonstrated variable accuracy across different personality traits, with the highest accuracy observed in predicting openness (78.01%) and the lowest in predicting neuroticism (48.94%) and thinking vs. feeling (47.33%). These results suggest that the tool is more effective in predicting certain traits over others, which could inform further refinement of its data collection methods and predictive algorithms.

ASAN achieved average accuracy rates of 59% for the Big 5 prediction model and 58% for the MBTI prediction model. These outcomes substantiate the effectiveness of the implemented models as competent baselines for future tools aimed at predicting personality traits based on user-generated data. Consequently, the results affirm the potential of the developed models to serve as foundational benchmarks, thereby encouraging further refinement and development in the field of personality prediction analytics on user generated data.

## 6. Comparison with Prior Work

Compared to existing solutions [15] [16] [17] in social media analysis and other online activity, the proposed Advanced Social Analyzer, ASAN, excels in delivering organized and representative personality and profile classification data, proving especially valuable for researchers across various disciplines. This enhanced capability allows for a more precise understanding of user behavior through online activity, which is critical for studies in social science, marketing, psychology, and beyond.

Furthermore, ASAN is designed with user accessibility in mind. It provides comprehensive documentation that details all functionalities, along with clear instructions on how to utilize these features. Each function within the tool can be easily activated by appending the appropriate flags as specified in the user guide. This design not only facilitates ease of use but also consolidates the entire workflow for personality prediction from user-generated data into a single, streamlined process. By integrating steps such as data collection, content filtering, sentiment analysis, and personality prediction, the tool significantly reduces the complexity traditionally associated with accessing and analyzing data pertinent to personality and behavioral studies.

Moreover, ASAN's architecture, built entirely on open-source components, ensures that all elements are thoroughly vetted through rigorous testing and analysis. This approach not only eliminates the direct costs associated with purchasing software but also substantially lowers the intangible costs related to configuring and integrating disparate components of behavioral analysis. As a result, researchers can deploy the ASAN with minimal financial and operational overhead, further democratizing access to advanced analytical capabilities.

## 7. Conclusions

As social media continues to serve as a pivotal source of valuable behavioral data, including user sentiment and attitudes, for a diverse array of stakeholders including businesses, researchers, policymakers, and cybersecurity experts, the necessity for a comprehensive analytical tool that can systematically organize and extract meaningful insights from this unstructured data becomes increasingly critical. The present project lays a formidable foundation for future endeavors in this arena by merging features for user-generated data extraction and personality prediction with existing methodologies, thereby addressing a significant gap in meeting contemporary social and analytical needs.

Looking forward, to enhance both the accuracy and efficiency of the tool, future research should concentrate on the incorporation of dynamic data filtering techniques utilizing advancements in machine learning and natural language processing. These methods have the potential to significantly diminish the rate of false positives in the data identification process. Additionally, optimizing the management of static training data can streamline the training phases of the tools, making them more effective and easier to use. Such improvements will not only refine the functionality of ASAN but also extend its applicability and effectiveness in real-world settings.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Qeeqbox (n.d.) Qeeqbox/Social-Analyzer: API, CLI, and Web App for Analyzing and Finding a Person's Profile in 1000 Social Media\Websites. GitHub. https://github.com/qeeqbox/social-analyzer

[2] Felipe, M. (n.d.) Sherlock-Project/Sherlock: Hunt Down Social Media Accounts by

Username across Social Networks. GitHub. https://github.com/sherlock-project/sherlock

[3] Holmes, R. (2024) Social Media Marketing & Management Dashboard. Hootsuite. https://www.hootsuite.com/

[4] Howard, J., Rankin, A., Lara, G. and Soung, P. (2024) Social Media Management Solutions. Sprout Social. https://sproutsocial.com/

[5] Mosley, C. (2024) #1 Linkedin Automation and Prospecting Tool. Dripify, https://dripify.io/

[6] Lim, A.G.Y. (2023) Big 5 Personality Traits: The 5-Factor Model of Personality. Simply Psychology. https://www.simplypsychology.org/big-five-personality.html#:~:text=The%20Big%20Five%20Personality%20Traits,Extraversion%2C%20Agreeableness%2C

[7] Rai, N. (n.d.) Naivedya-Rai/Personality-Prediction: Personality Prediction Using Machine Learning—MBTI Personality Predictor. GitHub, https://github.com/Naivedya-Rai/Personality-Prediction

[8] Beautiful Soup Documentation. https://beautiful-soup-4.readthedocs.io/en/latest/

[9] Scikit-Learn. https://scikit-learn.org/stable/

[10] (MBTI) Myers-Briggs Personality Type Dataset. Kaggle, https://www.kaggle.com/datasets/datasnaek/mbti-type

[11] Jason (n.d.) JCL132/Personality-Prediction-from-Text: Predict Big 5 Personality Traits from Text GitHub. https://github.com/jcl132/personality-prediction-from-text

[12] Public Figures by Personality Type. https://www.crystalknows.com/famous-people

[13] Pdb (n.d.) Personality Database. https://www.personality-database.com/en-US/profile?pid=1

[14] Erika Dauria (2019) Accuracy, Recall & Precision. Medium. https://medium.com/@erika.dauria/accuracy-recall-precision-80a5b6cbd28d

[15] Natalia, Dettmers, T., *et al.* (2014) Personality Types of World's Most Powerful Celebrities by Analyzing Their Twitter Stream Content. Media Managers Club. http://mediamanagersclub.org/personality-types-world%E2%80%99s-most-powerful-celebrities-analyzing-their-twitter-stream-content

[16] Cherry, K. (2023) Myers-Briggs Type Indicator: The 16 Personality Types. Verywell Mind. https://www.verywellmind.com/the-myers-briggs-type-indicator-2795583

[17] Stillwell, D.J. and Kosinski, M. (2015) myPersonality Project Website. https://sites.google.com/michalkosinski.com/mypersonality