





Review

Identification and Prediction of Human Behavior through Mining of Unstructured Textual Data

Mohammad Reza Davahli ^{1,*}, Waldemar Karwowski ¹, Edgar Gutierrez ^{1,2},
Krzysztof Fiok ¹, Grzegorz Wróbel ³, Redha Taiar ⁴ and Tareq Ahram ¹

¹ Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL 32816, USA; wkar@ucf.edu (W.K.); edgar.gutierrezfranco@ucf.edu (E.G.); fiok@ucf.edu (K.F.); tahram@ucf.edu (T.A.)

² Center for Latin-American Logistics Innovation, LOGyCA, Bogota 110111, Colombia

³ Department of Logistics and Process Engineering, University of Information Technology and Management in Rzeszów, 35-225 Rzeszów, Poland; gwrobel@wsiz.edu.pl

⁴ Sport Science Department, Université de Reims Champagne-Ardenne, 51100 Reims, France; redha.taiar@univ-reims.fr

* Correspondence: mohammadreza.davahli@ucf.edu

Received: 30 September 2020; Accepted: 17 November 2020; Published: 19 November 2020



Abstract: The identification of human behavior can provide useful information across multiple job spectra. Recent advances in applying data-based approaches to social sciences have increased the feasibility of modeling human behavior. In particular, studying human behavior by analyzing unstructured textual data has recently received considerable attention because of the abundance of textual data. The main objective of the present study was to discuss the primary methods for identifying and predicting human behavior through the mining of unstructured textual data. Of the 823 articles analyzed, 87 met the predefined inclusion criteria and were included in the literature review. Our results show that the included articles could be symmetrically classified into two groups. The first group of articles attempted to identify the leading indicators of human behavior in unstructured textual data. In this group, the data-based approaches had three main components: (1) collecting self-reported survey data, (2) collecting data from social media and extracting data features, and (3) applying correlation analysis to evaluate the relationship between two sets of data. In contrast, the second group focused on the accuracy of data-based approaches for predicting human behavior. In this group, the data-based approaches could be categorized into (1) approaches based on labeled unstructured textual data and (2) approaches based on unlabeled unstructured textual data. The review provides a comprehensive insight into unstructured textual data mining to identify and predict human behavior and personality traits.

Keywords: human behavior; personality traits; data-based approaches; machine learning; unstructured textual data

1. Introduction

Human behavior is a complex phenomenon [1–5]. However, one of the basic assumptions in human behavior is that each person can be described by a set of characteristics that is stable and does not change over time [6]. This set of stable elements has been conceptualized primarily as a personality, and people's differences in social behaviors have been conceptualized as personality traits [5]. However, the concept of personality has been reported to not fully explain human behavior, and people tend to behave unstably in different situations [5].

Sticha et al. [7] have indicated that two main elements of human behavior are (1) personal characteristics and qualities, and (2) the demands of a situation. However, the relationships among

personal characteristics, situational demands, and human behavior are complex, and particular behaviors, such as suicidal behavior, are difficult to attribute to personal or situational causes [7]. Mining unstructured textual data, mainly from social media, provides an excellent opportunity to reduce this complexity and improve the assessment of human behavior for two reasons. First, extracted textual features from social media have been reported to be significantly correlated with individuals' characteristics [8]. Second, individuals post textual data over long time frames, and specific situational causes or states of mind do not play an important role [9].

Studying human behavior by mining unstructured textual data has recently received substantial attention because of the abundance of textual data, including posts in social networks, different types of written reports, online databases, sensor-based data, and other records [4,10–13]. Different characteristics and qualities of individuals have been studied by researchers in different disciplines, such as the social and psychological sciences, public health, business marketing, and computer science [10].

Personality is a unique identifier for every person because it influences mental processes and human behavior. Personality is generally referred to as a group of elements that explain behavioral characteristics [14]. In some articles, personality is defined as an individual characteristic pattern of thought, emotion, and behavior, together with the psychological mechanisms—hidden or not—behind those patterns [15]. Many psychologists believe that individual personality influences different aspects of behavior, including job performance, success, and professional behavior [16].

The primary assumption regarding personality is that well-balanced individual characteristics become well-balanced behavioral patterns among individuals; consequently, human behavior can be predicted and determined based on measurable and stable individual characteristics [17]. Personality and behavior are closely related because personality can affect individuals' behavior, and characteristics of behavior can represent the personality [18]. Finding this connection through social networks, mobile phones, and textual data can enable the classification of people into different personality types [18]. Personality can be measured through standard psychometric tests, such as the big five personality inventory [15].

In many studies, personality structure is modeled based on factors or traits, and individual personality is explained by trait value [19]. Different models exist for personality traits, including the Big Five model of personality, the Myers-Briggs Type Indicator (MBTI), dark triad personality traits, the Jackson Personality Inventory (JPI), the Eysenck three-factor model, and the Big Two model of personality. There is no agreement regarding which personality model is best, but some are more commonly used, such as the Big Five model of personality [18]. In personality traits, determining the actual personality, recognizing the personality effect on other elements such as job performance, and building a predictive model are the main areas of study [17]. Many researchers have recently focused on developing a new system to acquire a deep understanding of human behavior according to personality traits [15]. The sophisticated, automatic extraction of personality traits has different applications, including analyzing deception, finding the right match in online dating, and finding the right person to fill a job vacancy [16].

The Big Five personality framework, a personality trait ranked model, divides personality into five different traits.

- Extraversion: The first trait is extraversion, which is frequently associated with being outgoing, talkative, energetic, enthusiastic, active, and assertive. Extraversion involves positive emotions and a sociable tendency. In terms of job performance, the extraversion dimension is a good indicator for managers and salespeople [20].
- Agreeableness: The second trait is agreeableness, which is most often associated with being kind, sympathetic, forgiving, generous, and appreciative. Agreeableness involves trusting and cooperative tendencies.
- Conscientiousness: Conscientiousness is the third trait, which is associated with being responsible, playful, reliable, efficient, and organized. Conscientiousness involves scrupulous and diligent

tendencies. In terms of job performance, the conscientiousness dimension is the best indicator of performance in every job type [20].

- **Neuroticism:** The fourth trait is neuroticism, or being unstable, worried, tense, touchy, anxious, and self-pitying. Neuroticism involves danger sensitivity and psychological distress tendencies. Neuroticism shows more anxiety than other traits [21]. This type shows low emotional stability and lower stress tolerance, and has a tendency to experience negative emotions [22].
- **Openness to experience:** The fifth trait is openness to experience, or being original, widely interested, imaginative, insightful, artistic, and curious. It involves a willingness to think about other options and alternatives, and a tendency to curiosity [23]. In terms of job performance, the openness to experience dimension is a good indicator of training proficiency [20].

The Myers-Briggs Type Indicator (MBTI) personality model as an occupation personality assessment tool is based on the eight poles theory and is considered the world's most authoritative and famous personality test approach [24]. The MBTI personality theory includes four dimensions as a combination of introversion vs. extraversion, judging vs. perceiving, thinking vs. feeling, and sensing vs. intuition [24]. Each dimension has two opposite poles; through combining the eight dimensions with two poles, 16 types of psychological aspects are created [24] as follows: introversion sensing thinking judging (ISTJ), introversion sensing feeling judging (ISFJ), introversion intuition feeling judging (INFJ), introversion intuition thinking judging (INTJ), introversion sensing thinking perceiving (ISTP), introversion sensing feeling perceiving (ISFP), introversion intuition feeling perceiving (INFP), introversion intuition thinking perceiving (INTP), extraversion sensing thinking perceiving (ESTP), extraversion sensing feeling perceiving (ESFP), extraversion intuition feeling perceiving (ENFP) extraversion intuition thinking perceiving (ENTP), extraversion sensing thinking judging (ESTJ), extraversion introversion sensing feeling judging (ESFJ), extraversion intuition feeling judging (ENFJ), and extraversion intuition thinking judging (ENTJ) [24].

Dark triad personal characteristics have three subcategories: (1) narcissism, which involves leadership, authority, exploitativeness, entitlement, and self-focus; (2) psychopathy, which involves aggression, callous affect, erratic lifestyle, asocial behavior, and impulsiveness; and (3) Machiavellianism, which involves a lack of empathy, manipulation, and controversial and multi-dimensional behavior [25].

The main objective of the present study was to review and discuss the primary methods for identifying and predicting human behavior through the mining of unstructured textual data. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) were selected as structured guidelines to ensure reliable and meaningful study results [26]. This review is symmetrically structured as follows. The methodology section discusses the inclusion and exclusion criteria, and the risk of bias. The results section provides the outputs of the literature search. The discussion section describes data-based approaches among selected records.

2. Methodology

The PRISMA guidelines were followed for this systematic literature review [26]. Three main steps of the systematic review are developing the research question, determining the search strategy, and addressing the risk of bias [26]. The research question was formulated as follows:

RQ. What are the main approaches for identifying and predicting human behavior through the mining of unstructured textual data?

To answer the above research question, we developed a search strategy to identify and review all relevant scientific articles. The search strategy included (1) defining keywords and identifying all relevant materials, (2) removing duplicates, (3) filtering the remaining articles, as performed by three authors through reading titles, abstracts, and in some cases full text, and (4) resolving conflicts through meetings with other authors [26,27].

The first step in this review was developing keywords. According to the stated research question, the keywords were divided into three groups. The first group comprised human behavior and

personality traits. The second group comprised data-based approaches, multivariate analysis, big data methods, artificial intelligence, and machine learning. The third group comprised textual data, textual features, and textual indicators. A combination of three groups was used as keywords in searching the articles, as represented in Table 1. Web of Science, IEEE Xplore, and Science Direct were used as database search tools for this review.

Table 1. The keywords used in the present review.

Row	Keywords
Test set 1	“human behavior” OR “personality traits”
Test set 2	“data-based approach” OR “multivariate analysis” OR “big data methods” OR “artificial intelligence” OR “machine learning”
Test set 3	“textual data” OR “textual feature” OR “textual indicator”
Search 1	#1 AND #2 AND #3

The keywords were used, and 823 articles with relevant content were identified and added to the main database. After developing the main database and identifying relevant articles, we applied a formal screening process to the database on the basis of the exclusion and inclusion criteria. The inclusion criteria were articles associated with the objective and research question, articles written in English, and articles published between 2000 and 2019. The exclusion criteria were articles written in other languages, book chapters or articles from secondary sources that were not free or open access, letters, newspaper articles, viewpoints, presentations, anecdotes, duplicated studies, and posters. The screening of the titles, abstracts, conclusions, and keywords in the identified records after removing duplication ($n = 591$) resulted in excluding 504 articles. Among the excluded papers, 226 were not associated with the objective and research question, 93 papers focused on opinion and sentiment analysis, 62 papers were published before 2000, 56 papers related to mobility behavior and learning behavior, 48 papers were book chapters, letters, newspaper articles, viewpoints, presentations, anecdotes, duplicated studies, and posters, and 19 papers were in other languages. The remaining articles ($n = 87$) were read in full against the eligibility principle and all articles were included. Among included articles, 17 articles were identified from IEEE Xplore, 21 articles from Web of Science, and 49 from Science Direct.

The articles selected over time and the PRISMA guidelines are shown in Figures 1 and 2, respectively.

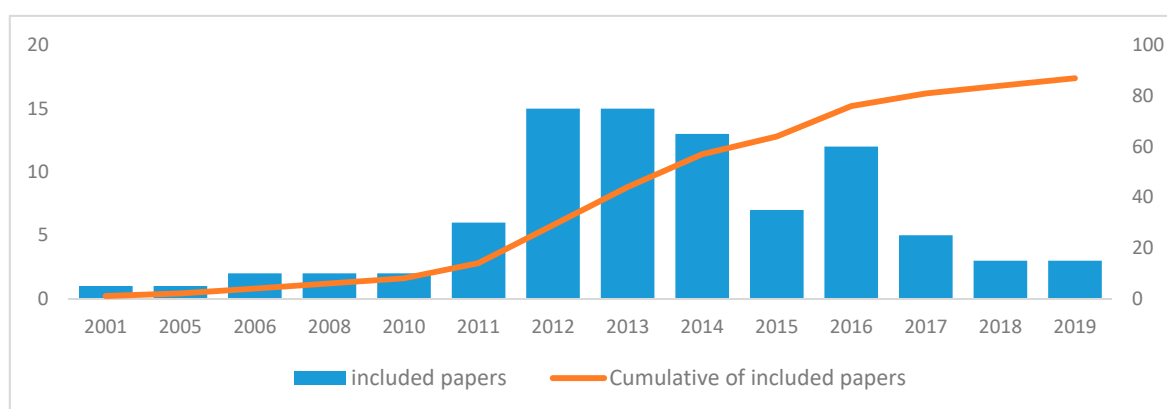


Figure 1. Included articles over time.

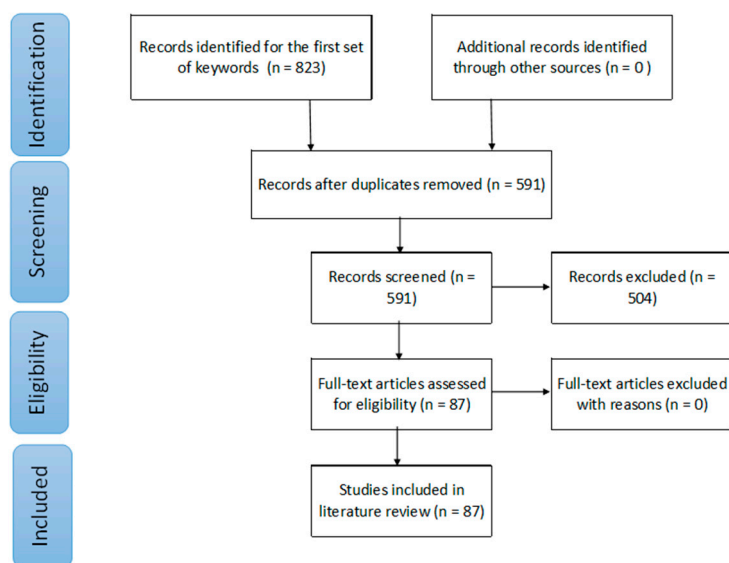


Figure 2. Chart of the selection strategy following PRISMA guidelines [26].

The risk of bias has been divided into external and internal biases. External bias can occur through (1) applying inclusion/exclusion criteria and (2) identifying aspects of human behavior, types of data-based approaches, and textual features. Internal bias relates to assessing the quality of the research among the selected articles. To address the first type of bias, three researchers separately reviewed the title, abstract, and conclusion to select the appropriate articles for full-text review. They compared the selected articles to produce a unified list. After analyzing the selected articles, the authors determined whether the article was appropriate for inclusion. The authors agreed on each article's inclusion before its addition to the main database. Disagreements among the three authors as to an article's inclusion or exclusion were resolved in sessions with other authors. In the next step, three authors separately summarized the data-based approaches and textual features in the selected articles, then compared the results and resolved disagreements by consulting with other authors.

For article quality assessment, the National Heart, Lung, and Blood Institute (NHLBI) Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies was used [28]. This methodology was validated by Frost and Rickwood, [29], who have reported good agreement among independent evaluators regarding classification into good/fair/poor categories of quality [29]. In the present study, following the protocol reported by [30], three researchers independently assessed the methodological quality of the reviewed articles, compared the results, and resolved disagreements by consulting with other authors. Of a total of 87 articles selected for this study, 63 articles were observational and cross-sectional studies. These articles were classified as intermediate quality (63%), poor quality (28%), or high quality (9%). A lack of exposure assessment, inadequate blinding of outcome assessors, and small samples that overrepresented young students were the main limitations of the articles classified as "poor" quality.

3. Results

All identified articles were categorized and stored in the main database according to year, source of publication, data-based approach, source of input data, and textual features. On the basis of these elements, selected articles were categorized as (1) articles that attempted to identify the main indicators of human behavior in unstructured textual data, and articles that focused on developing more accurate data-based approaches to better predict human behavior; (2) articles with manual feature selection and articles that used different computerized techniques for feature extraction and selection; and (3) articles that designed and developed new models for detecting human behavior, and articles that used

well-known data-based approaches. The list of included articles with data-based approaches is shown in Table 2. The reviewed articles were published in 39 journals (Figure 3).

Table 2. Data-based approaches among included articles.

Reference	Methods for Identifying Indicators of Human Behavior	Methods for Predicting Human Behavior
Adali & Golbeck [31]	Correlation analysis	Gaussian Process and ZeroR regression algorithms
Agarwal & Sureka [32]		Naïve Bayes, random forest, and decision tree classifiers
Alam et al. [33]		Support vector machine, Bayesian logistic regression, binary logistic regression, and multinomial naïve Bayes sparse modeling
Alsadhan & Skillicorn [34]		New approach based on the frequency and similarity of the words among each of the Big Five personality traits
Amichai-Hamburger & Vinitzky [23]	Regression analysis, two-way ANOVA	
Annisette & Lafreniere [35]	Correlation analysis and hierarchical multiple regression	
Argamon et al. [36]		Sequential minimal optimization, support vector machine algorithms
Ashton [37]	Correlation analysis	
Bachrach et al. [38]	Correlation analysis	Multivariate linear regression, support vector machine algorithms, and decision stump algorithms
Bai et al. [39]		Support vector machine, naïve Bayes, decision tree algorithms
Bai et al. [40]	Correlation analysis	Multi-task regression and incremental regression
Ben-Ari & Hammond [41]		Random forest algorithm
Bhattacharya et al. [42]	Correlation analysis	
Celli & Poesio [43]	Correlation analysis	Unsupervised personality recognition system
Celli & Polonio [44]	Correlation analysis	Unsupervised personality recognition system
Celli & Rossi [45]	Correlation analysis	Unsupervised personality recognition system
Chapsky [46]		Bayesian network
G. Chittaranjan et al. [47]	Correlation and multiple regression analysis	Decision trees and support vector machine classifiers
Gokul Chittaranjan et al. [48]	Correlation and multiple regression analysis	Support vector machine classifier
Devaraj et al. [49]	Correlation and multiple regression analysis	
Farnadi et al. [50]		Support vector machine, nearest neighbor with $k = 1$, and naïve Bayes algorithms
Farnadi et al. [8]	Correlation analysis	Decision tree algorithm and support vector machine algorithm

Table 2. Cont.

Reference		Methods for Identifying Indicators of Human Behavior	Methods for Predicting Human Behavior
Fatima et al.	[9]		Decision trees, random forest-based method, and support vector machine classifier
Gao et al.	[51]		Gaussian process, M5' rules, and Pace Regression
Golbeck et al.	[21]	Correlation analysis	M5' rules and Gaussian process algorithms
Golbeck	[52]		Receptiviti API
Gupta & Chatterjee	[16]		Rough sets and LEM algorithm
Hammond & Laundry	[53]		Software based on support vector conditional random field classifier
He et al.	[54]		Text classification algorithm named product score model
He et al.	[55]		Logistic regression and classification tree
Holtgraves	[56]	Correlation analysis	
Hu et al.	[57]		New model based on detecting the relationship between personality and topic
Iacobelli & Culotta	[58]		Conditional random fields model, sequential minimal optimization, and naïve Bayes algorithms
Jenkins-Guarnieri et al.	[59]	Correlation analysis	
Kaati et al.	[60]		Machine learning algorithms including Adaboost and classification trees
Kalghatgi et al.	[61]		Neural network algorithm
Kartelj et al.	[62]		Automated Personality Classification model based on linear regression, M5' classification tree, M5' regression tree, and support vector machine
Kermanidis	[63]		Support vector machine classifier
Kern et al.	[64]	Correlation analysis	
Kosinski et al.	[65]		Dimensionality reduction and linear regression analysis
Kosinski et al.	[66]		Logistic regression classifier
Krämer & Winter	[67]	Correlation analysis	
Krishnamurthy et al.	[68]	Correlation analysis	
Lima & de Castro	[69]		Bayesian personality predictor model based on naïve Bayes algorithm
Lima & de Castro	[70]		Personality prediction in social media data (PERSOMA) approach
Maria Balmaceda et al.	[71]	Correlation analysis, Apriori algorithm, and Knime tool	
Markovikj et al.	[14]	Pearson correlation analysis	Support vector machines, simple minimal optimization, and Boost algorithms
Moore & McElroy	[72]	Hierarchical regression and correlation analysis	

Table 2. Cont.

Reference	Methods for Identifying Indicators of Human Behavior	Methods for Predicting Human Behavior
Neuman & Cohen	[73]	New vectorial semantics approach
Neuman et al.	[74]	Correlation analysis Pedesis tool
Nie et al.	[75]	Local linear semi-supervised regression algorithm
Nokhbeh Zaeem et al.	[76]	Approach based on statistical analysis and the Identity Threat Assessment and Prediction (ITAP) algorithm
Oberlander & Nowson	[77]	Support vector machines, naïve Bayes algorithms
Ortigosa et al.	[19]	TP2010 application Naïve Bayes, classification tree algorithms
Ou et al.	[78]	ANOVA Support vector machine
Pabón et al.	[79]	IBM Watson personality insights
Panicheva et al.	[80]	Correlation analysis
Park et al.	[81]	Correlation analysis Regression model
Peng et al.	[82]	Support vector machine
Pramodh & Vijayalata	[83]	Data-based approach based on word counts
Pratama & Sarno	[84]	Naïve Bayes, K-nearest neighbors, and support vector machine algorithms
Preoțiu-Pietro et al.	[85]	Logistic regression classifiers
Qiu et al.	[86]	Correlation analysis
Quercia et al.	[87]	Correlation analysis M5' rules algorithm
Quercia et al.	[88]	Correlation analysis Simple regression model
Reips & Garaizar	[89]	Text mining with Iscience maps web service
Santos & Paraboni	[90]	Naïve Bayes classification and logistic regression
Schwartz et al.	[91]	Correlation analysis and standardized linear regression Support vector machine and regression analysis
Seibert & Kraimer	[92]	Correlation analysis
Seidman	[93]	Correlation analysis
Skues et al.	[94]	Correlation analysis
Souri et al.	[95]	Naïve Bayes, neural network, decision tree, and support vector machine classifiers
Srividya & Sowjanya	[96]	Simple correlation analysis and comparison of the frequency with a threshold
Sumner et al.	[25]	Support vector machine using sequential minimal optimization and a polynomial kernel, random forest, and naïve Bayes classifier
Tazghini & Siedlecki	[97]	Hierarchical regression and correlation analysis
Uddin	[98]	Series of algorithms and models called PERFECT Algorithm Engine (PAE)
Wald et al.	[22]	Numeric prediction models including linear regression, Reptree, and decision tables

Table 2. Cont.

Reference	Methods for Identifying Indicators of Human Behavior	Methods for Predicting Human Behavior
Wang	[99]	Correlation analysis
Wei et al.	[100]	Proposed Heterogeneous Information Ensemble framework
Winter et al.	[101]	Correlation analysis and hierarchical regression analyses
Liu et al.	[24]	Naïve Bayes classification
Yarkoni	[102]	Correlation analysis
Yoon et al.	[103]	Term frequency analysis
Zhou et al.	[104]	Medical information extraction (medie) system using decision tree-based text classification
Kumar & Gavrilova	[105]	XGBoost and support vector machine algorithms
Yang & Huang	[106]	Personality Recognizer tool based on linear regression, M5' model tree, M5' regression tree, and support vector machine algorithms
Zheng & Wu	[107]	Semi-supervised learning methods called pseudo multi-view co-training

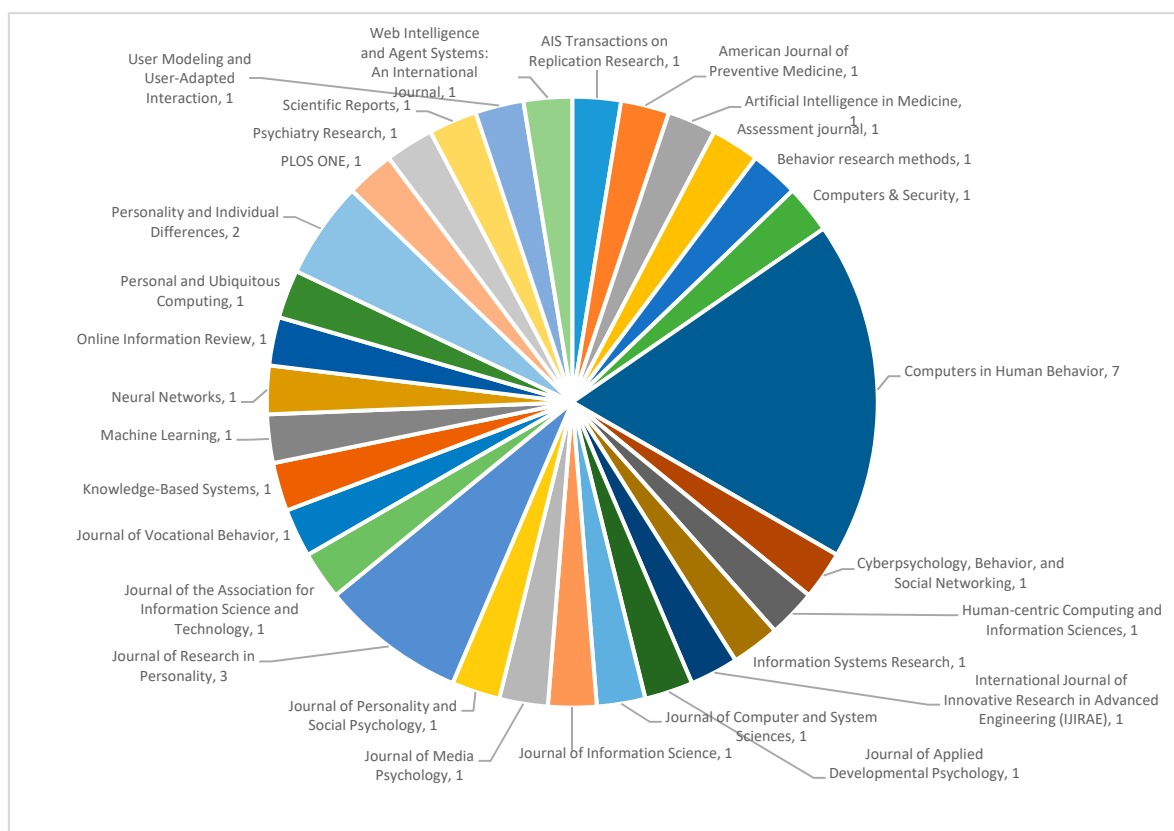


Figure 3. Journal sources of the included articles.

To obtain a better picture, we analyzed the titles and abstracts of the included articles. Figure 4 shows a co-occurrence map based on the analysis of abstracts in the form of a bubble chart.

For developing this figure, the TextRank algorithm (implemented in Gensim 3.8.3 [<https://pypi.org/project/gensim/>]) was used to extract the main keywords of abstracts of included papers. The keywords were fed into VOSviewer software to visualize the result. In Figure 4, the nodes correspond to specific textual terms, and their sizes represent the frequency of occurrence. The co-occurrence of the textual terms in different publications is represented by a link between two nodes. Frequently co-occurring textual terms create clusters and appear closer to each other with the same color. Figure 4 reflects the main cluster (purple color) with terms such as extraversion, agreeableness, openness, conscientiousness, and neuroticism. The next cluster (green color) contains several items such as user personality, social network, Facebook user, and questionnaire.

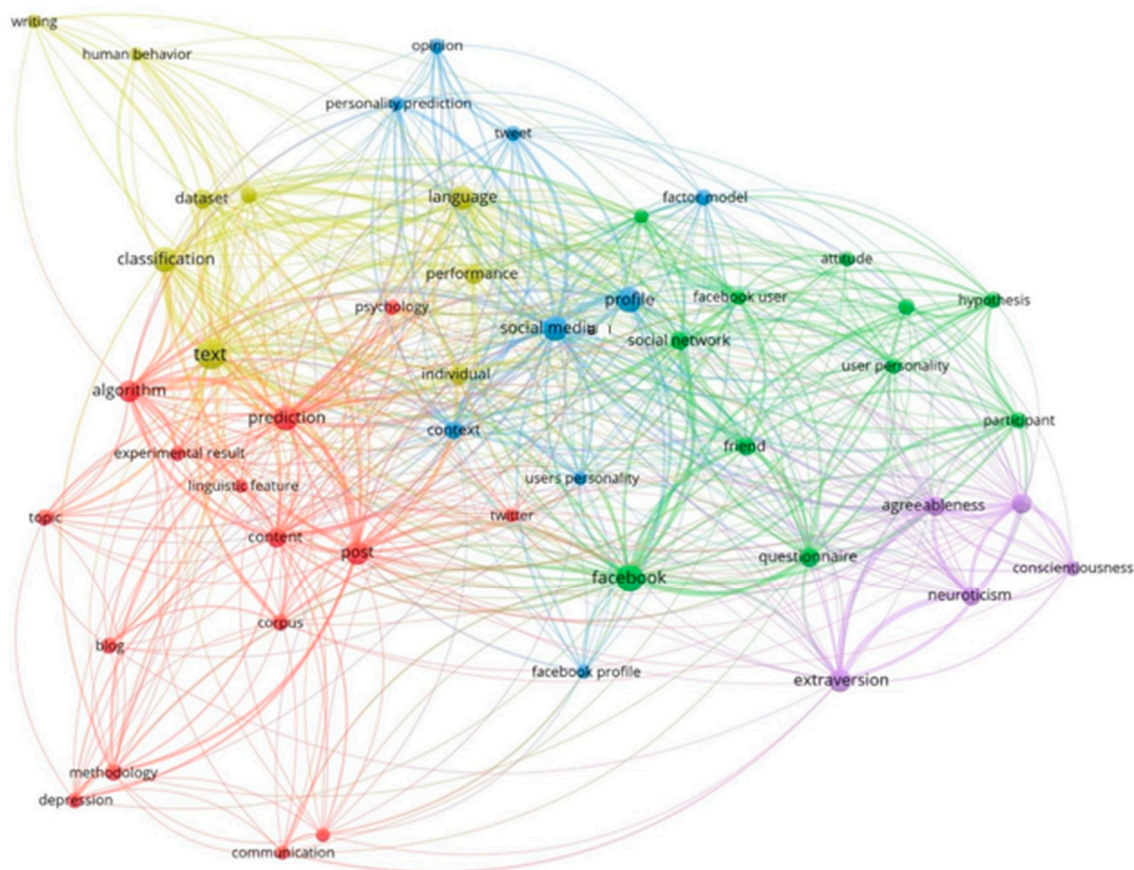


Figure 4. The map of co-occurrence of terms in titles and abstracts of the included articles.

4. Discussion

Various approaches have been reported for identifying and predicting human behavior through the mining of unstructured textual data. A primary concept is that textual features are significantly correlated with individuals' behaviors and qualities [8]. These methods can be categorized according to elements, including data sources, feature sets, and techniques.

In social networks, individuals reveal substantial information regarding different topics and items in the form of status updates, self-descriptions, and interests [21]. The raw data in social media include religion, educational history, user name, birthday, gender, relationship status, hometown, personal written information, and lists of favorite things. The main source for social network data is Facebook, which provides useful information for identifying human behavior. On Facebook, individuals reveal their identification and opinions by expressing a variety of aspects of themselves [46]. On the basis of data from Facebook, life outcomes, socio-economic status, disorder behaviors, mobility behaviors, and cultural preferences can be predicted [46]. Other data sources for human behavior studies are Twitter, LinkedIn, Myspace, Foursquare, and mobile phone data.

Feature extraction and selection are important steps, which involve removing unnecessary words and information from documents, and building the derived values to facilitate successive interpretations and learning [108]. Features in selected articles can be divided into two groups with (1) pre-defined, manually selected features, and (2) use of different methods for feature extraction and selection. The main types of features among the included articles are as follows:

- Facebook's pre-defined features include personal information, work information, contact information, education, time spent on Facebook, frequency of use, number of statuses, number of friends, number of groups, number of likes, number of photos, and number of tags.
- Twitter's pre-defined features include the number of followings, followers, retweets, hashtags, and links.
- In other social media, pre-defined features include personal information and time spent on Instagram, Sina Weibo, or LinkedIn.
- In Linguistic Inquiry and Word Count (LIWC) features, words are tagged in different sections including linguistic (e.g., adjective, pronoun, or noun), relativity (e.g., past, time, or future), personal process (e.g., family, home, or job) and psychological process (e.g., emotions).
- In Natural Language Toolkit (NLTK) features, words are tagged into 89 categories to mainly extract grammatical features.
- Word frequency-based features include frequency of depression phrases, frequency of the first and/or second and/or third-person pronouns, 1000 most frequent words, and frequency of positive or negative words.
- Character frequency-based features include frequency of question marks, punctuation marks, exclamation marks, negative and positive emoticons, or capitalized letters.
- Term frequency-inverse document frequency (TFIDF) features based on the number of times a word appears in the document and the number of documents in the corpus that contain the word [109].
- Latent Dirichlet Allocation (LDA) features based on assigning topics to documents and generates topic distributions over words given a collection of texts [110].
- Linguistic features are based on pre-processing (removing stop words, stemming, and word segmentation tools) and semantic analysis.

4.1. Identifying Human Behavior

Multiple articles have attempted to identify the main indicators of human behavior in unstructured textual data. Data-based approaches of these articles have three main parts: (1) collecting self-reported survey data, (2) collecting social media data and extracting features, and (3) analyzing the relationship between two sets of data, as shown in Figure 5.

Text messages in social media can be good indicators of users' personality. Adali and Golbeck [31] have extracted network bandwidth and message content features from Facebook and Twitter data, and analyzed the correlation between these features and data from the Big Five inventory [31]. The study concluded that linguistic features are useful in identifying personality. In another study, Annette and Lafreniere [35] have tested the shallowing hypothesis, in which constantly using social networking sites can lead to a significant decrease in daily reflective thought. Participants were asked to complete five measures including texting and social media use; 44 items in the Big Five Inventory to assess the levels of five personality dimensions; 58 items in the life goals inventory to assess life goals; the 12 item reflection scale from the Rumination-Reflection Questionnaire to assess tendencies to engage in self-reflective states; and a demographic questionnaire to assess participants' background information. The study used correlation analysis and concluded that participants who constantly use social networking sites place less value on life goals [35]. The relationship between text messages and human behavior has been investigated by other researchers. Holtgraves [56] has studied the relationship between language variances in text messaging and personality traits.

Maria Balmaceda et al. [71] have investigated users' personality through evaluating text messages in social network, then verified the stability of the identified personality. Panicheva et al. [80] have investigated the link between the dark triad personality traits and Russian linguistic features in social networking texts.

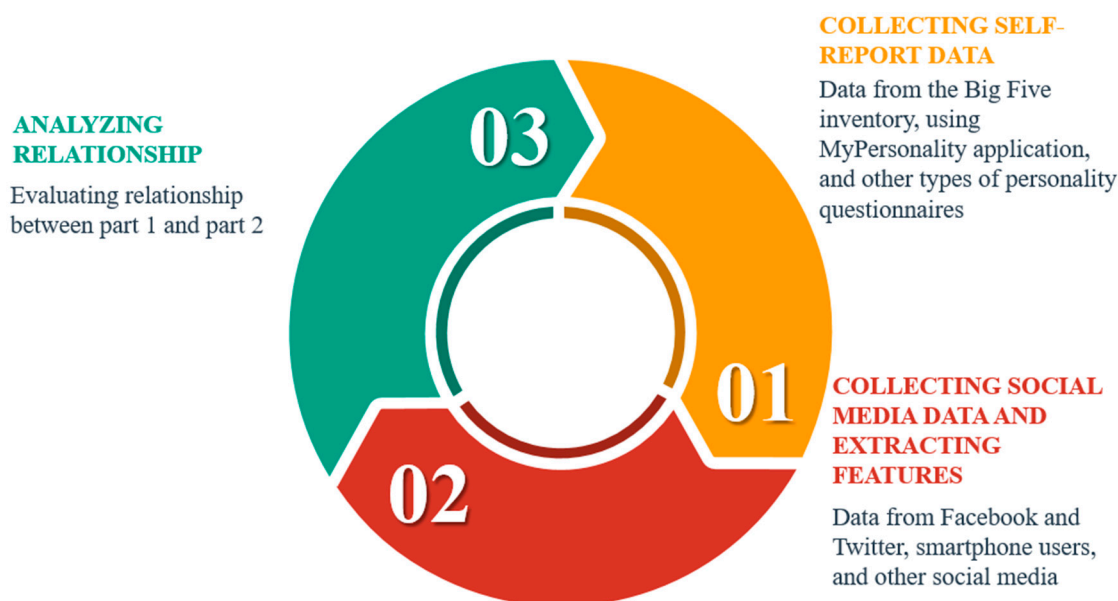


Figure 5. Main parts of data-based approaches for identifying human behavior.

Smartphone data can be indicators of personality. Chittaranjan et al. [47,48] have collected data from smartphone users in three categories, namely call, text, and application, to automatically extract different features regarding applications and communication usage on the phone. The study used correlation analysis among these features and collected data from an online ten item personality inventory questionnaire.

Multiple articles have indicated that Facebook data can be significant indicators of user personality. Amichai-Hamburger and Vinitzky [23] have used analysis of covariance (ANCOVA) to investigate the link between personality traits and Facebook features [23]. The study was conducted in three sequential research phases of (1) evaluating personality by asking participants to complete the Revised NEO Personality Inventory as a self-reported measure; (2) collecting users' information from Facebook and dividing it into four dimensions of basic information, contact information and education, personal information, and work information; and (3) applying ANCOVA. The study developed several hypotheses about the relationship between traits of the Big Five model and behavior on Facebook, and highlighted a strong connection between Facebook behavior and the personality of users [23]. In this regard, Bachrach et al. [38] have shown a significant relationship between user personality and the information on their Facebook profiles, such as the number of uploaded photos, size of their friendship network, number of events attended, density of their friendship network, and number of group memberships. The study used a dataset containing the Facebook profiles and personality profiles of 180,000 users and applied correlation analysis to evaluate the link between personality and Facebook content [38]. On this topic, Schwartz et al. [91] have conducted a survey from 75,000 volunteers, extracted 700 million words and phrases from Facebook messages of participants, and used correlation analysis to identify the main indicators of personality traits in Facebook messages.

Beyond Facebook content, the Facebook status updates [101] and Facebook usage [93,94,97,99], as measured by the Facebook intensity scale and the Facebook use scale, have been used to detect personality. On this topic, Jenkins-Guarnieri et al. [59] have collected data from 463 participants and investigated the relationship between Big-Five personality traits and Facebook usage [59]. Moore and McElroy [72] have conducted a survey from Facebook users and have used Facebook data to detect

why some users are more active than others. The authors have indicated that personality can explain differences among Facebook users [72].

Several articles have used other methods for collecting personality data rather than conducting questionnaire-based surveys. Kern et al. [64] have collected millions of posts from Facebook users and used the MyPersonality application to evaluate personality scores. The study used correlational analysis to examine the relationship between the extracted features from posts and personality scores and accordingly distinguished words and phrases representing each Big Five personality trait.

Data from other social media can be good indicators of human behavior. Quercia et al. [87,88] have used Pearson product-moment correlation to investigate the relationship between personality traits and different characteristics of Twitter users, such as the numbers of users followed and of followers. In another study, Krämer and Winter [67] have investigated the relationship between personality traits and the manner of self-presentation and self-esteem in social media. Participants were asked to complete two measures of the Revised NEO Personality Inventory to measure extraversion, and Mielke's questionnaire to measure self-presentation. The study used multivariate analysis of variance to analyze the relationship between extraversion and self-presentation.

For identifying human behavior on the basis of unstructured textual data, it is important to remember that (1) there are no significant differences between predicting all personality traits of users at once or identifying each trait of personality separately; (2) selecting only correlated features does not necessarily improve the performance of the predictive model; and (3) discovering the smallest feature set without decreasing the performance of the human behavior predictive model is a main goal for the future research [8].

4.2. Predicting Human Behavior

Most articles have focused on developing more accurate models to predict human behavior through the mining of unstructured textual data. The included articles used different data-based approaches to accurately predict human behavior, as shown in Figure 6.

Several studies have developed language-independent methods for predicting human behavior from unstructured textual data. Alsadhan and Skillicorn [34] have developed an approach based on word counts to predict both the Big Five and the Myers-Briggs personality traits from small amounts of text [34]. The proposed method is language independent, does not require particular lexicons, and has been successfully applied to different languages [34]. To develop this method, the 1000 most frequent words labeled with Big Five personality traits and Myers-Briggs personality types were selected (without removal of stop words or performing stemming) to build a model for each personality trait. The developed models were compared with posts and tweets on Facebook and Twitter to predict user personality [34]. On this topic, Pramodh and Vijayalata [83] have predicted the Big Five personality traits of authors through their writings and essays. First, two datasets containing positive and negative terms corresponding to each Big Five personality trait were created. For predicting personality after collecting data, pre-processing including tokenizing the input textual data, removing stopwords, stemming, scaling, and scoring was performed [83]. In the final step, the stemmed data were compared with the datasets, and the matched percentages of the data were calculated [83].

Several studies have used language-independent features in developing their data-based approaches. Celli et al. [43–45] have developed an unsupervised personality recognition system using language-independent features to predict Big Five personality traits from unstructured textual data. For developing language-independent features, the studies have used a list of pre-defined linguistic cues from published research [43,45]. A processing pipeline has been developed, including preprocessing, processing, and evaluation modules [43,45]. In the preprocessing module, the average occurrence of each feature is determined by randomly selecting samples of posts [43,45]. By matching features and using correlations, the system creates one personality hypothesis per post [43,45]. For a single user, the system evaluates all hypotheses created for all posts and generates one personality model per user

with a confidence level [43,45]. The proposed system has been tested on English and Italian Twitter posts and shown to have an accuracy of approximately 65.0.

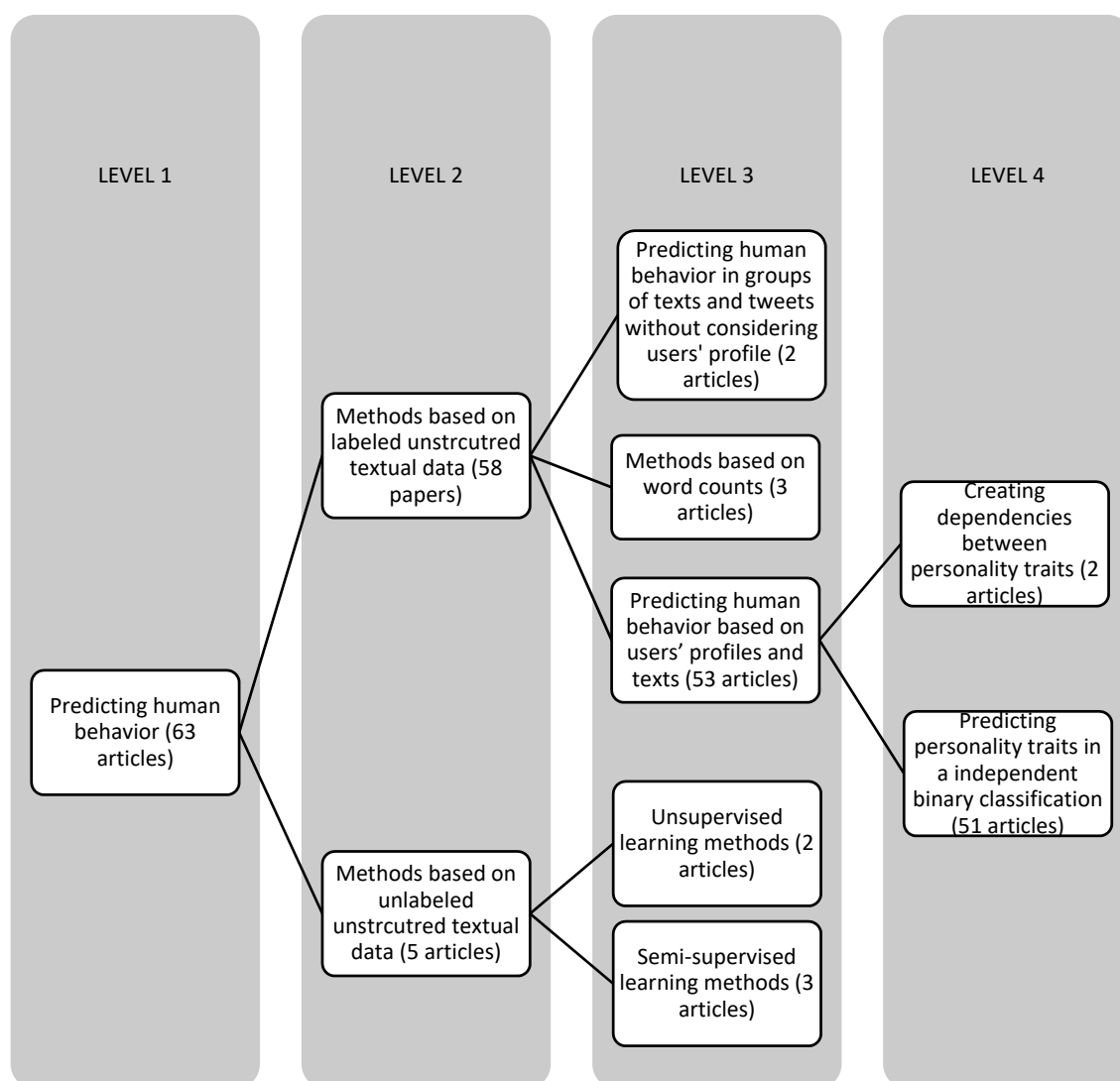


Figure 6. Data-based approaches used for predicting human behavior.

Several studies [75,107] have described improperly labeled samples in published research and have developed semi-supervised learning methods to evaluate the personality traits by using unlabeled samples. Nie et al. [75] have extracted 47 features for each user in the categories of the user personal profile, social circles, social activities, and social habits. Next, stepwise regression was used to select important features for each trait of personality. Then, the study conducted a small survey and calculated the score for each personality trait from completed Big Five personality questionnaires. Two datasets were developed: (1) a small labeled dataset $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$, where X_i is a dimensional feature vector, and Y_i is personality score for user i , and (2) the main unlabeled dataset $\{X_{n+1}, \dots, X_m\}$. Finally, a local linear kernel regression algorithm and a local linear semi-supervised regression algorithm were used to predict personality in the unlabeled dataset.

Most included articles have developed supervised machine learning-based models to predict human behavior from unstructured textual data. Although in many published articles, aspects of human behavior have been predicted according to extracting data from user profiles, texts, and tweets, several studies have attempted to predict human behavior in groups of labeled texts and tweets without taking user profiles into account. For example, Lima and de Castro [69,70] have developed a

multi-label classifier algorithm based on the naïve Bayes algorithm to predict personality in texts and tweets [69].

Most studies have predicted personality traits through independent binary classification (modeling each personality trait in isolation). However, Iacobelli and Culotta [58] have used conditional random fields and structured classification to model the dependencies between personality traits. The authors have concluded that there is a correlation between neuroticism and agreeableness traits and that considering this correlation in a classification model can help improve accuracy for classification of the agreeableness trait [58].

Studies have additionally developed supervised learning methods to predict human behavior on the basis of labeled data and consideration of user information without considering dependencies between personality traits. For example, Park et al. [81] noting the small sample size of previous articles, have developed a model based on a sample of 66,000 participants. The authors collected data from Facebook users and their Big Five personality trait questionnaires [81]. The study generated thousands of linguistic features including multiword phrases, single words, and clusters of semantically related words [81]. After using a variety of dimensionality-reduction methods, the authors used a regression model to predict personality traits [81]. The main differences between articles were (1) feature selection and extraction, and (2) classification method, as shown in Table 3.

Table 3. Features and classification techniques among articles with supervised learning methods.

Reference	Features	Classification Method (Best Performance Method)	Accuracy (Best Case)
[31]	Pairwise network bandwidth features	Gaussian Process and zeroR regression algorithms	
[32]	Linguistic features	Random forest algorithm	78.0
[33]	TF-IDF features	Multinomial naïve Bayes sparse modeling	61.8
[36]	Linguistic features	Sequential minimal optimization algorithm	58.0
[38]	Facebook’s pre-defined features	Multivariate regression	
[39]	Facebook’s pre-defined features	Support vector machine, naïve Bayes, and decision tree algorithms	81.1
[46]	Facebook’s pre-defined features	Bayesian network	
[50]	Facebook’s pre-defined features, LIWC features	Support vector machine with a linear kernel, naïve Bayes algorithms	66.0
[9]	LIWC features	Random forest classification	
[47]	Sequential backward feature selection algorithm	Support vector machine classifier	83.0
[51]	Word-based features, character-based features, and LIWC features	M5’ rules, Pace regression and Gaussian process	
[21]	Facebook’s pre-defined features	M5’ rules and Gaussian process algorithms	65.0
[16]	LIWC and NLTK features	Rough sets and LEM algorithm	84.67
[53]	Linguistic features	Software based on support vector conditional random fields classifier	
[54]	Using chi-square selection algorithm to select top keywords	Text classification algorithm named product score model	80.0

Table 3. Cont.

Reference	Features	Classification Method (Best Performance Method)	Accuracy (Best Case)
[55]	Textual features of posts on Facebook	Logistic regression and classification tree	62.9
[60]	Psychologically meaningful features, according to LIWC	Adaboost algorithm	92.2
[65]	Facebook's pre-defined features	Logistic/linear regression	78.0
[66]	Facebook's pre-defined features	Logistic/linear regression	75.0
[63]	Modern Greek textual features	Support vector machine classifier	86.0
[14]	Facebook's pre-defined features	Multiboostab and adaboostM1 algorithms	
[73]	Textual features	Vectorial semantics approach (tree-based classification model)	64.0
[19]	Facebook's pre-defined features	Naïve Bayes and classification tree algorithms	82.8
[81]	Linguistic features	Regression model	
[82]	Linguistic features	Support vector machine algorithm	73.5
[85]	Demographic features and linguistic features	Binary logistic regression classifiers with elastic net regularization	81.9
[87]	Facebook's pre-defined features	M5' rules algorithm	
[95]	Facebook's pre-defined features	Boosting-decision tree classifier	82.0
[25]	LIWC features	Support vector machine, random forest, naïve Bayes classifiers	64.7
[22]	Facebook's pre-defined features	Linear regression, REPTree, and decision table algorithms	75.0
[105]	TF-IDF features and GloVe word embedding	XGBoost and support vector machine algorithms	85.0
[106]	Linguistic features	Personality Recognizer tool based on linear regression, M5' model tree, M5' regression tree, and support vector machine algorithms	

5. Conclusions

Identification of human behavior can provide valuable information across multiple job spectra, including sales (developing recommendation systems), hiring (predetermining potential from resumes and writing samples), marketing (improving stakeholder management and enhancing individuals' communication strategies, negotiations (analyzing a rival or head of a successful organization), detecting terrorists and criminals, discovering depression and disorders, jury selection, and creating personal and professional relationships. This article provides a systematic review of the published research relevant to identifying and predicting human behavior through the mining of unstructured text data. A total of 87 published articles that met the predefined inclusion criteria were included in the review. The following research question has been explored:

RQ. What are the main approaches to identify and predict human behavior through the mining of unstructured textual data?

Based on collected data, all reviewed articles were divided into two categories: (1) articles that attempted to establish a clear connection between textual data features and aspects of human behavior, and (2) articles that focused on developing more accurate data-based approaches to predict human behavior. In the first category, data-based approaches had three main parts: (1) collecting self-reported survey data, (2) collecting data from social media and extracting different textual features, and (3) using correlation analysis to evaluate the connection between two sets of data. In the second category, different data-based approaches were used to predict human behavior in unstructured textual data. These data-based approaches can be divided into two categories: (1) methods based on labeled unstructured textual data and (2) methods based on unlabeled unstructured textual data. In methods based on labeled data, human behavior can be predicted according to data extracted from user profiles or from groups of texts and tweets. In addition, methods based on unlabeled data can be divided into semi-supervised learning methods and unsupervised learning methods. Extracted features in selected articles include Facebook's pre-defined features, Twitter's pre-defined features, pre-defined features of other social media, LIWC features, NLTK features, word frequency-based features, character frequency-based features, TFIDF features, LDA features, and part-of-speech tagging features. The main sources of input data include Facebook, Twitter, LinkedIn, Myspace, Foursquare, and mobile phone data.

This systematic literature review has some limitations. One of the main limitations is the timeframe for article discovery and the timetable for published articles. Article discovery was finished at the end of August 2019, and only articles published between 2000 and August 2019 were included. The second limitation is the inability to discover and include individual relevant papers arising from inclusion and exclusion criteria, a limited number of keywords, and a limited number of search databases for article discovery. Therefore, based on the developed research strategy, some highly cited articles that applied deep learning-based personality detection were not included in this literature review. For example, Majumder et al. [111] used a deep convolutional neural network to extract different essays' features. Sun et al. [112] developed a model as a fusion of bidirectional long short term memory networks with a convolutional neural network to predict users' personality using structures of texts. Su et al. [113] tried to indicate nature in a dyadic conversation through using a recurrent neural network (for modeling short term temporal evolution of a dialog) and coupled hidden Markov model for predicting the personalities of two speakers [114].

It should be noted that deep learning techniques have performed effectively in predicting human behavior among textual data. Because of generating vast amounts of unstructured textual data, these techniques with new and complex architectures will take momentum in the near future. Also, numerous other fertile research areas can be applied to study through mining textual data. Such areas include human behavior during disaster and the impact of behavioral signature [115], human behavior concerning resource management [116], patient behavior [117], and big data retrieval in social action [118].

Author Contributions: Methodology, writing—original draft and revisions, M.R.D.; conceptualization, funding acquisition, writing—review and revisions, editing, supervision, and project administration, W.K.; review and revisions, editing, methodology, E.G.; review and revisions, editing, project administration, K.F.; writing, review and revisions, G.W.; review and revisions, editing, R.T.; review and editing, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by a research grant from the Office of Naval Research N000141812559 awarded to the University of Central Florida, Orlando, Florida, USA.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fournier, M.A.; Moskowitz, D.S.; Zuroff, D.C. Integrating dispositions, signatures, and the interpersonal domain. *J. Pers. Soc. Psychol.* **2008**, *94*, 531–545. [[CrossRef](#)] [[PubMed](#)]

2. Maddock, J.; Starbird, K.; Al-Hassani, H.J.; Sandoval, D.E.; Orand, M.; Mason, R.M. Characterizing online rumoring behavior using multi-dimensional signatures. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Vancouver, BC, Canada, 14–18 March 2015; pp. 228–241.
3. Makeig, S.; Gramann, K.; Jung, T.-P.; Sejnowski, T.J.; Poizner, H. Linking brain, mind and behavior. *Int. J. Psychophysiol.* **2009**, *73*, 95–100. [[CrossRef](#)] [[PubMed](#)]
4. Shen, Z.; Su, J. Web service discovery based on behavior signatures. In Proceedings of the 2005 IEEE International Conference on Services Computing (SCC'05) Vol-1, Orlando, FL, USA, 11–15 July 2005; Volume 1, pp. 279–286.
5. Shoda, Y. Behavioral expressions of a personality system. *Coherence Personal. Soc. Cogn. Bases Consistency Var. Organ.* **1999**, *29*, 155–181.
6. Shoda, Y.; Wilson, N.L.; Whitsett, D.D.; Lee-Dussud, J.; Zayas, V. The person as a cognitive-affective processing system: From quantitative idiography to cumulative science. *Handb. Personal. Process. Individ. Differ.* **2014**, *4*, 491–513.
7. Sticha, P.J.; Weaver, E.A.; Tatman, J.A.; Mahoney, S.M.; Buede, D.M. Reading the Behavior Signature: Predicting Leader Personality from Individual and Group Actions. In Proceedings of the AAAI Spring Symposium: Technosocial Predictive Analytics, Stanford, CA, USA, 23–25 March 2009; pp. 130–136.
8. Farnadi, G.; Sitaraman, G.; Sushmita, S.; Celli, F.; Kosinski, M.; Stillwell, D.; Davalos, S.; Moens, M.-F.; De Cock, M. Computational personality recognition in social media. *User Model. User-Adapt. Interact.* **2016**, *26*, 109–142. [[CrossRef](#)]
9. Fatima, I.; Mukhtar, H.; Ahmad, H.F.; Rajpoot, K. Analysis of user-generated content from online social communities to characterise and predict depression degree. *J. Inf. Sci.* **2018**, *44*, 683–695. [[CrossRef](#)]
10. Azucar, D.; Marengo, D.; Settanni, M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personal. Individ. Differ.* **2018**, *124*, 150–159. [[CrossRef](#)]
11. Fiok, K.; Karwowski, W.; Gutierrez, E.; Reza-Davahli, M. Comparing the Quality and Speed of Sentence Classification with Modern Language Models. *Appl. Sci.* **2020**, *10*, 3386. [[CrossRef](#)]
12. Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 2544–2558. [[CrossRef](#)]
13. Boumi, S.; Vela, A.; Chini, J. Quantifying the relationship between student enrollment patterns and student performance. *arXiv* **2020**, arXiv:2003.10874.
14. Markovikj, D.; Gievaska, S.; Kosinski, M.; Stillwell, D. Mining Facebook Data for Predictive Personality Modeling. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA, 8–11 June 2013; pp. 23–26.
15. Gou, L.; Zhou, M.X.; Yang, H. KnowMe and ShareMe: Understanding automatically discovered personality traits from social media and user sharing preferences. In Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems—CHI '14, Toronto, ON, Canada, 26 April–1 May 2014; ACM Press: Toronto, ON, Canada, 2014; pp. 955–964.
16. Gupta, U.; Chatterjee, N. Personality Traits Identification Using Rough Sets Based Machine Learning. In Proceedings of the 2013 International Symposium on Computational and Business Intelligence, New Delhi, India, 24–26 August 2013; pp. 182–185.
17. Vinciarelli, A.; Mohammadi, G. A Survey of Personality Computing. *IEEE Trans. Affect. Comput.* **2014**, *5*, 273–291. [[CrossRef](#)]
18. Staiano, J.; Lepri, B.; Aharony, N.; Pianesi, F.; Sebe, N.; Pentland, A. Friends don't lie: Inferring personality traits from social network structure. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 321–330. [[CrossRef](#)]
19. Ortigosa, A.; Carro, R.M.; Quiroga, J.I. Predicting user personality by mining social interactions in Facebook. *J. Comput. Syst. Sci.* **2014**, *80*, 57–71. [[CrossRef](#)]
20. Barrick, M.R.; Mount, M.K. The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Pers. Psychol.* **1991**, *44*, 1–26. [[CrossRef](#)]
21. Golbeck, J.; Robles, C.; Turner, K. Predicting personality with social media. In Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems—CHI EA '11, Vancouver, BC, Canada, 7–12 May 2011; ACM Press: Vancouver, BC, Canada, 2011; pp. 253–262.

22. Wald, R.; Khoshgoftaar, T.; Sumner, C. Machine prediction of personality from Facebook profiles. In Proceedings of the 2012 IEEE 13th International Conference on Information Reuse Integration (IRI), Las Vegas, NE, USA, 8–10 August 2012; pp. 109–115.
23. Amichai-Hamburger, Y.; Vinitzky, G. Social network use and personality. *Comput. Hum. Behav.* **2010**, *26*, 1289–1295. [[CrossRef](#)]
24. Liu, Y.; Liu, T.; Wang, Y.J. Research on micro-blog character analysis based on Naïve Bayes. In Proceedings of the Seventh International Conference on Digital Image Processing (ICDIP 2015); International Society for Optics and Photonics, Los Angeles, CA, USA, 9–10 April 2015; Volume 9631, pp. 96312F1–96312F5.
25. Sumner, C.; Byers, A.; Boochever, R.; Park, G.J. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In Proceedings of the 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012; Volume 2, pp. 386–393.
26. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Med.* **2009**, *6*, e1000100. [[CrossRef](#)]
27. Davahli, M.R.; Karwowski, W.; Taiar, R. A System Dynamics Simulation Applied to Healthcare: A Systematic Review. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5741. [[CrossRef](#)]
28. National Heart, Lung, and Blood Institute (NHLBI) Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies. Available at: National Heart, Lung, and Blood Institute, Bethesda, MD, USA. Available online: <https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools> (accessed on 24 May 2019).
29. Frost, R.L.; Rickwood, D.J. A systematic review of the mental health outcomes associated with Facebook use. *Comput. Hum. Behav.* **2017**, *76*, 576–600. [[CrossRef](#)]
30. Carbia, C.; López-Caneda, E.; Corral, M.; Cadaveira, F. A systematic review of neuropsychological studies involving young binge drinkers. *Neurosci. Biobehav. Rev.* **2018**, *90*, 332–349. [[CrossRef](#)]
31. Adali, S.; Golbeck, J. Predicting Personality with Social Behavior. In Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 26–29 August 2012; pp. 302–309.
32. Agarwal, S.; Sureka, A. Role of Author Personality Traits for Identifying Intent Based Racist Posts. In Proceedings of the 2016 European Intelligence and Security Informatics Conference (EISIC), Uppsala, Sweden, 17–19 August 2016; p. 197.
33. Alam, F.; Stepanov, E.A.; Riccardi, G. Personality Traits Recognition on Social Network—Facebook. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA, USA, 8–11 July 2013.
34. Alsadhan, N.; Skillicorn, D. Estimating Personality from Social Media Posts. In Proceedings of the IEEE 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 350–356.
35. Annisette, L.E.; Lafreniere, K.D. Social media, texting, and personality: A test of the shallowing hypothesis. *Personal. Individ. Differ.* **2017**, *115*, 154–158. [[CrossRef](#)]
36. Argamon, S.; Dhawle, S.; Koppel, M.; Pennebaker, J.W. Lexical predictors of personality type. In Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America, St. Louis, MO, USA, 8–12 June 2005; pp. 1–16.
37. Ashton, M. Personality and job performance: The importance of narrow traits. *J. Organ. Behav.* **1998**, *19*, 289–303. [[CrossRef](#)]
38. Bachrach, Y.; Kosinski, M.; Graepel, T.; Kohli, P.; Stillwell, D. Personality and patterns of Facebook usage. In Proceedings of the 4th Annual ACM Web Science Conference; ACM: New York, NY, USA, 2012; pp. 24–32.
39. Bai, S.; Zhu, T.; Cheng, L. Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. *arXiv* **2012**, arXiv:1204.4809.
40. Bai, S.; Yuan, S.; Hao, B.; Zhu, T. Predicting personality traits of microblog users. *Web Intell. Agent Syst. Int. J.* **2014**, *12*, 249–265. [[CrossRef](#)]
41. Ben-Ari, A.; Hammond, K. Text mining the EMR for modeling and predicting suicidal behavior among US veterans of the 1991 Persian Gulf War. In Proceedings of the IEEE 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2015; pp. 3168–3175.

42. Bhattacharya, S.; Yang, C.; Srinivasan, P.; Boynton, B. Perceptions of presidential candidates' personalities in twitter. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 249–267. [[CrossRef](#)]
43. Celli, F.; Poesio, M. PR2: A Language Independent Unsupervised Tool for Personality Recognition from Text. *arXiv* **2014**, arXiv:1402.2796.
44. Celli, F.; Polonio, L. Relationships between Personality and Interactions in Facebook. In *Social Networking: Recent Trends, Emerging Issues and Future Outlook*; Nova Science Publishers: Hauppauge, NY, USA, 2013; pp. 41–54.
45. Celli, F.; Rossi, L. The Role of Emotional Stability in Twitter Conversations. In Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, France, 12 April 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 10–17.
46. Chapsky, D. Leveraging Online Social Networks and External Data Sources to Predict Personality. In Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, Taiwan, 25–27 July 2011; pp. 428–433.
47. Chittaranjan, G.; Blom, J.; Gatica-Perez, D. Who's Who with Big-Five: Analyzing and Classifying Personality Traits with Smartphones. In Proceedings of the 2011 15th Annual International Symposium on Wearable Computers, San Francisco, CA, USA, 12–15 June 2011; pp. 29–36.
48. Chittaranjan, G.; Blom, J.; Gatica-Perez, D. Mining large-scale smartphone data for personality studies. *Pers. Ubiquitous Comput.* **2013**, *17*, 433–450. [[CrossRef](#)]
49. Devaraj, S.; Easley, R.F.; Crant, J.M. How Does Personality Matter? Relating the Five-Factor Model to Technology Acceptance and Use. *Inf. Syst. Res.* **2008**, *19*, 93–105. [[CrossRef](#)]
50. Farnadi, G.; Zoghbi, S.; Moens, M.-F.; De Cock, M. Recognising personality traits using Facebook status updates. In Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAI conference on weblogs and social media (ICWSM13), AAI, Boston, MA, USA, 11 June 2013.
51. Gao, R.; Hao, B.; Bai, S.; Li, L.; Li, A.; Zhu, T. Improving user profile with personality traits predicted from social media content. In Proceedings of the 7th ACM Conference on Recommender Systems—RecSys '13, Hong Kong, China, October 2013; ACM Press: Hong Kong, China, 2013; pp. 355–358.
52. Golbeck, J. Predicting Personality from Social Media Text. *AIS Trans. Replication Res.* **2016**, *2*, 1–10. [[CrossRef](#)]
53. Hammond, K.W.; Laundry, R.J. Application of a Hybrid Text Mining Approach to the Study of Suicidal Behavior in a Large Population. In Proceedings of the 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 6–9 January 2014; pp. 2555–2561.
54. He, Q.; Veldkamp, B.P.; Vries, T. Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Res.* **2012**, *198*, 441–447. [[CrossRef](#)]
55. He, Q.; Glas, C.A.W.; Kosinski, M.; Stillwell, D.J.; Veldkamp, B.P. Predicting self-monitoring skills using textual posts on Facebook. *Comput. Hum. Behav.* **2014**, *33*, 69–78. [[CrossRef](#)]
56. Holtgraves, T. Text messaging, personality, and the social context. *J. Res. Personal.* **2011**, *45*, 92–99. [[CrossRef](#)]
57. Hu, Z.; Liu, Y.; Zhang, C.; Xu, Y. The analysis of topic's personality traits using a new topic model. In Proceedings of the 2017 IEEE 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2–4 June 2017; pp. 1079–1083.
58. Iacobelli, F.; Culotta, A. Too Neurotic, Not Too Friendly: Structured Personality Classification on Textual Data. In Proceedings of the Seventh International AAI Conference on Weblogs and Social Media, Cambridge, MA, USA, 8–11 July 2013.
59. Jenkins-Guarnieri, M.A.; Wright, S.L.; Hudiburgh, L.M. The relationships among attachment style, personality traits, interpersonal competency, and Facebook use. *J. Appl. Dev. Psychol.* **2012**, *33*, 294–301. [[CrossRef](#)]
60. Kaati, L.; Shrestha, A.; Sardella, T. Identifying warning behaviors of violent lone offenders in written communication. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 12–15 December 2016; pp. 1053–1060.
61. Kalghatgi, M.P.; Ramannavar, M.; Sidnal, N.S. A neural network approach to personality prediction based on the big-five model. *Int. J. Innov. Res. Adv. Eng. IJIRAE* **2015**, *2*, 56–63.
62. Kartelj, A.; Filipović, V.; Milutinović, V. Novel approaches to automated personality classification: Ideas and their potentials. In Proceedings of the 2012 35th International Convention MIPRO, Opatija, Croatia, 21–25 May 2012; pp. 1017–1022.

63. Kermanidis, K.L. Mining authors' personality traits from modern greek spontaneous text. In Proceedings of the Workshop on Corpora for Research on Emotion Sentiment & Social Signals, in conjunction with LREC, Istanbul, Turkey, 26 May 2012; pp. 90–93.
64. Kern, M.L.; Eichstaedt, J.C.; Schwartz, H.A.; Dziurzynski, L.; Ungar, L.H.; Stillwell, D.J.; Kosinski, M.; Ramones, S.M.; Seligman, M.E.P. The Online Social Self: An Open Vocabulary Approach to Personality. *Assessment* **2014**, *21*, 158–169. [[CrossRef](#)] [[PubMed](#)]
65. Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [[CrossRef](#)] [[PubMed](#)]
66. Kosinski, M.; Bachrach, Y.; Kohli, P.; Stillwell, D.; Graepel, T. Manifestations of user personality in website choice and behaviour on online social networks. *Mach. Learn.* **2014**, *95*, 357–380. [[CrossRef](#)]
67. Krämer, N.C.; Winter, S. Impression management 2.0: The relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. *J. Media Psychol.* **2008**, *20*, 106–116. [[CrossRef](#)]
68. Krishnamurthy, M.; Mahmood, K.; Marcinek, P. A hybrid statistical and semantic model for identification of mental health and behavioral disorders using social network analysis. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 1019–1026.
69. Lima, A.C.E.S.; de Castro, L.N. Multi-label Semi-supervised Classification Applied to Personality Prediction in Tweets. In Proceedings of the 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, Ipojuca, Brazil, 8–11 September 2013; pp. 195–203.
70. Lima, A.C.E.S.; de Castro, L.N. A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Netw.* **2014**, *58*, 122–130. [[CrossRef](#)]
71. Maria Balmaceda, J.; Schiaffino, S.; Godoy, D. How do personality traits affect communication among users in online social networks? *Online Inf. Rev.* **2014**, *38*, 136–153. [[CrossRef](#)]
72. Moore, K.; McElroy, J.C. The influence of personality on Facebook usage, wall postings, and regret. *Comput. Hum. Behav.* **2012**, *28*, 267–274. [[CrossRef](#)]
73. Neuman, Y.; Cohen, Y. A Vectorial Semantics Approach to Personality Assessment. *Sci. Rep.* **2014**, *4*, 4761. [[CrossRef](#)]
74. Neuman, Y.; Cohen, Y.; Assaf, D.; Kedma, G. Proactive screening for depression through metaphorical and automatic text analysis. *Artif. Intell. Med.* **2012**, *56*, 19–25. [[CrossRef](#)] [[PubMed](#)]
75. Nie, D.; Guan, Z.; Hao, B.; Bai, S.; Zhu, T. Predicting Personality on Social Media with Semi-supervised Learning. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 11–14 August 2014; Volume 2, pp. 158–165.
76. Nokhbeh Zaeem, R.; Manoharan, M.; Yang, Y.; Barber, K.S. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Comput. Secur.* **2017**, *65*, 50–63. [[CrossRef](#)]
77. Oberlander, J.; Nowson, S. Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, 17–18 July 2006; Association for Computational Linguistics: Sydney, Australia, 2006; pp. 627–634.
78. Ou, G.; Li, J.; Guo, J.; Cai, Z.; Lu, M. *The Bloggers' Personality Traits Categorizing Algorithm Based on Text Features Analysis*; Atlantis Press: Paris, France, 2016; pp. 1–6.
79. Pabón, O.H.P.; González, F.A.; Aponte, J.; Camargo, J.E.; Restrepo-Calle, F. Finding Relationships between Socio-Technical Aspects and Personality Traits by Mining Developer E-mails. In Proceedings of the 2016 IEEE/ACM Cooperative and Human Aspects of Software Engineering (CHASE), Austin, TX, USA, 16 May 2016; pp. 8–14.
80. Panicheva, P.; Ledovaya, Y.; Bogolyubova, O. Lexical, morphological and semantic correlates of the dark triad personality traits in russian facebook texts. In Proceedings of the 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL), St. Petersburg, Russia, 10–12 November 2016; pp. 1–8.
81. Park, G.; Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; Kosinski, M.; Stillwell, D.J.; Ungar, L.H.; Seligman, M.E.P. Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* **2015**, *108*, 934–952. [[CrossRef](#)] [[PubMed](#)]
82. Peng, K.; Liou, L.; Chang, C.; Lee, D. Predicting personality traits of Chinese users based on Facebook wall posts. In Proceedings of the 2015 24th Wireless and Optical Communication Conference (WOCC), Taipei, Taiwan, 23–24 October 2015; pp. 9–14.

83. Pramodh, K.C.; Vijayalata, Y. Automatic personality recognition of authors using big five factor model. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; IEEE: Coimbatore, India, 2016; pp. 32–37.
84. Pratama, B.Y.; Sarno, R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In Proceedings of the 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, Indonesia, 25–26 November 2015; pp. 170–174.
85. Preoțiu-Pietro, D.; Eichstaedt, J.; Park, G.; Sap, M.; Smith, L.; Tobolsky, V.; Schwartz, H.A.; Ungar, L. The role of personality, age, and gender in tweeting about mental illness. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, CO, USA, 5 June 2015; pp. 21–30.
86. Qiu, L.; Lin, H.; Ramsay, J.; Yang, F. You are what you tweet: Personality expression and perception on Twitter. *J. Res. Personal.* **2012**, *46*, 710–718. [[CrossRef](#)]
87. Quercia, D.; Kosinski, M.; Stillwell, D.; Crowcroft, J. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 180–185.
88. Quercia, D.; Lambiotte, R.; Stillwell, D.; Kosinski, M.; Crowcroft, J. The Personality of Popular Facebook Users. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, Seattle, WA, USA, 11–15 February 2012; ACM: New York, NY, USA, 2012; pp. 955–964.
89. Reips, U.-D.; Garaizar, P. Mining twitter: A source for psychological wisdom of the crowds. *Behav. Res. Methods* **2011**, *43*, 635. [[CrossRef](#)]
90. dos Santos, W.R.; Paraboni, I. Personality facets recognition from text. *arXiv* **2018**, arXiv:1810.02980.
91. Schwartz, H.A.; Eichstaedt, J.C.; Kern, M.L.; Dziurzynski, L.; Ramones, S.M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M.E.P.; et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* **2013**, *8*, e73791. [[CrossRef](#)]
92. Seibert, S.E.; Kraimer, M.L. The Five-Factor Model of Personality and Career Success. *J. Vocat. Behav.* **2001**, *58*, 1–21. [[CrossRef](#)]
93. Seidman, G. Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personal. Individ. Differ.* **2013**, *54*, 402–407. [[CrossRef](#)]
94. Skues, J.L.; Williams, B.; Wise, L. The effects of personality traits, self-esteem, loneliness, and narcissism on Facebook use among university students. *Comput. Hum. Behav.* **2012**, *28*, 2414–2419. [[CrossRef](#)]
95. Souri, A.; Hosseinpour, S.; Rahmani, A.M. Personality classification based on profiles of social networks' users and the five-factor model of personality. *Hum. Centric Comput. Inf. Sci.* **2018**, *8*, 1–15. [[CrossRef](#)]
96. Srividya, K.; Sowjanya, A.M. Behavioral analysis of internet messaging and malicious activity detection. In Proceedings of the 2016 International Conference on Advances in Human Machine Interaction (HMI), Doddaballapur, India, 3–5 March 2016; pp. 1–5.
97. Tazghini, S.; Siedlecki, K.L. A mixed method approach to examining Facebook use and its relationship to self-esteem. *Comput. Hum. Behav.* **2013**, *29*, 827–832. [[CrossRef](#)]
98. Uddin, M.F. Noise Removal and Structured Data Detection to Improve Search for Personality Features. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, San Francisco, CA, USA, 18–21 August 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 1349–1355.
99. Wang, S.S. “I Share, Therefore I Am”: Personality Traits, Life Satisfaction, and Facebook Check-Ins. *Cyberpsychology Behav. Soc. Netw.* **2013**, *16*, 870–877. [[CrossRef](#)] [[PubMed](#)]
100. Wei, H.; Zhang, F.; Yuan, N.J.; Cao, C.; Fu, H.; Xie, X.; Rui, Y.; Ma, W.-Y. Beyond the Words: Predicting User Personality from Heterogeneous Information. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; ACM: New York, NY, USA, 2017; pp. 305–314.
101. Winter, S.; Neubaum, G.; Eimler, S.C.; Gordon, V.; Theil, J.; Herrmann, J.; Meinert, J.; Krämer, N.C. Another brick in the Facebook wall—How personality traits relate to the content of status updates. *Comput. Hum. Behav.* **2014**, *34*, 194–202. [[CrossRef](#)]
102. Yarkoni, T. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *J. Res. Personal.* **2010**, *44*, 363–373. [[CrossRef](#)]

103. Yoon, S.; Elhadad, N.; Bakken, S. A Practical Approach for Content Mining of Tweets. *Am. J. Prev. Med.* **2013**, *45*, 122–129. [[CrossRef](#)]
104. Zhou, X.; Han, H.; Chankai, I.; Prestrud, A.; Brooks, A. Approaches to Text Mining for Clinical Medical Records. In Proceedings of the 2006 ACM Symposium on Applied Computing, Dijon, France, 23–27 April 2006; ACM: New York, NY, USA, 2006; pp. 235–239.
105. Kumar, K.P.; Gavrilova, M.L. Personality Traits Classification on Twitter. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
106. Yang, H.-C.; Huang, Z.-R. Mining personality traits from social messages for game recommender systems. *Knowl. Based Syst.* **2019**, *165*, 157–168. [[CrossRef](#)]
107. Zheng, H.; Wu, C. Predicting Personality Using Facebook Status Based on Semi-supervised Learning. In Proceedings of the Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22–24 February 2019; Association for Computing Machinery: Zhuhai, China, 2019; pp. 59–64.
108. Irfan, R.; King, C.K.; Grages, D.; Ewen, S.; Khan, S.U.; Madani, S.A.; Kolodziej, J.; Wang, L.; Chen, D.; Rayes, A.; et al. A survey on text mining in social networks. *Knowl. Eng. Rev.* **2015**, *30*, 157–170. [[CrossRef](#)]
109. Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, Princeton, NJ, USA, 3 December 2003; Volume 242, pp. 133–142.
110. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
111. Majumder, N.; Poria, S.; Gelbukh, A.; Cambria, E. Deep learning-based document modeling for personality detection from text. *IEEE Intell. Syst.* **2017**, *32*, 74–79. [[CrossRef](#)]
112. Sun, X.; Liu, B.; Cao, J.; Luo, J.; Shen, X. Who am I? Personality detection based on deep learning for texts. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
113. Su, M.-H.; Wu, C.-H.; Zheng, Y.-T. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEEACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 733–744. [[CrossRef](#)]
114. Mehta, Y.; Majumder, N.; Gelbukh, A.; Cambria, E. Recent trends in deep learning based personality detection. *Artif. Intell. Rev.* **2020**, *53*, 2313–2339. [[CrossRef](#)]
115. Tsiropoulou, E.; Koukas, K.; Papavassiliou, S. A socio-physical and mobility-aware coalition formation mechanism in public safety networks. *EAI Endorsed Trans. Future Internet* **2018**, *4*, 154176. [[CrossRef](#)]
116. Vamvakas, P.; Tsiropoulou, E.E.; Papavassiliou, S. On controlling spectrum fragility via resource pricing in 5g wireless networks. *IEEE Netw. Lett.* **2019**, *1*, 111–115. [[CrossRef](#)]
117. Molani, S.; Madadi, M.; Wilkes, W. A partially observable Markov chain framework to estimate overdiagnosis risk in breast cancer screening: Incorporating uncertainty in patients adherence behaviors. *Omega* **2019**, *89*, 40–53. [[CrossRef](#)]
118. Wu, L.; Morstatter, F.; Hu, X.; Liu, H. Mining misinformation in social media. *Big Data Complex. Soc. Netw.* **2016**, 123–152.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).