Review of publications on the topic of 'Key Idiolect Markers in Sociolinguistic Profiling' (2004 – 2024)

Reference	Idiolect Markers Identified (as defined by authors)	Context of Application (Type of Participants, N)	Profiling Outcomes: Success, Accuracy, Limitations (Profiling Techniques and Methods Applied)	Linguistic Diversity Considered	Context-Dependence and Effectiveness of Idiolect Markers (EIM)	Key Findings and Conclusions (Methodological Robustness and Soudness of used approaches and tools)
Valente, et al. (2012)	•Prosody (variations in pitch, loudness, and tempo). •Speech Activity (the frequency and duration). •Overlaps and Interruptions. •Linguistic Features (words n-gram and dialog acts).	Social (speakers from the AMI corpus, N=128)	The traits of extraversion, conscientiousness, and neuroticism were identified with accuracy rates of 74.5%, 67.6%, and 68.7%, respectively. However, the classification of agreeableness and openness did not yield results statistically better than chance. (Applied Techniques – Manual, Computational)	No (English)	The study does not specifically assess the EIM or their contextual influence.	The research successfully applies the Big-Five personality traits model, which serves as a robust framework for understanding how individual speech patterns can be linked to specific personality traits, thereby enhancing the concept of idiolect profiling. Non-linguistic features (such as prosody and speech activity) outperformed linguistic features (like words n-grams and dialog acts). (Robustness – Moderate to high)
Shrestha, et al. (2020)	•Writing Style. •Word and Character N-grams. •Sentiment Analysis. •Psycho-linguistic Features (derived from tools like LIWC).	Digital, Social (Social media platforms users, N=500)	The study reports an accuracy of 0.73 for English and 0.77 for Spanish in detecting fake news spreaders using the proposed methods. These results indicate a reasonable level of success. (Applied Techniques – Manual, Computational)	Yes (English and Spanish)	EIM is influenced by the context of social media usage.	The study concludes that linguistic features, including idiolect markers, are significant in profiling users who spread fake news. (Robustness – Moderate)

Verhoeven et al. (2016)	•Vocabulary. •Syntax. •Speech Patterns.	Digital, Social (Social media platforms users, N='large dataset')	The article does not provide detailed statistics on accuracy or limitations. (Applied Techniques – Manual, Computational)	Yes (Dutch, German, French, Italian, Portuguese, and Spanish)	EIM is influenced by the context of social media usage.	The main findings indicate that personality traits (MBTI) and gender can be inferred from writing style in tweets. The study concludes that a multilingual approach to personality profiling is feasible and opens avenues for further research in sociolinguistic profiling. (Robustness – Moderate)
Kulkarni et al., (2018)	•Vocabulary. •Syntax. •Speech Patterns.	Digital, Social (Social media platforms users, N=49,139)	The study reports that the language-based traits (BLTs) derived from social media often outperform traditional questionnaire-based traits in predicting outcomes like income and IQ. Limitations include potential biases in social media usage and the generalizability of findings across different populations. (Applied Techniques – Manual, Computational)	No (English)	EIM is influenced by the context of social media usage.	It is feasible to derive meaningful personality traits from social media language, which can complement traditional personality assessments. The derived traits demonstrate predictive validity and stability, suggesting a new avenue for understanding human behavior through language. (Robustness – Moderate)
Pervaz et al., (2015)	•Vocabulary. •Syntax. •Punctuation. •Question sentences. •Overall Writing Style.	Digital, Social (Social media platforms users, N='multiple datasets from the PAN competition')	The study reported that certain features, such as the percentage of question sentences and average sentence length, were effective in profiling authors. Limitations include the variability in effectiveness across different languages and the potential for overfitting in machine learning models. (Applied Techniques – Manual, Computational)	Yes (English, Dutch, Spanish, Italian)	EIM is influenced by the context of social media usage.	The study concluded that stylistic features are valuable for identifying author personality traits and that certain features are effective across multiple languages. The findings suggest that a combination of stylistic analysis and machine learning can enhance author profiling efforts. (Robustness – Moderate)

Litvinova et al., (2016)	•Vocabulary. •Syntax. •Speech Patterns.	Clinical (Patients the risk of self-destructive behavior, N=721)	The article reports a mathematical model that predicts self-destructive behavior based on text analysis, indicating a degree of success in profiling. However, it acknowledges limitations due to a relatively small sample size and the limited range of text parameters used. (Applied Techniques – Manual, Computational)	No (Russian)	EIM is influenced by the corresponding clinical context.	A set of correlations between scores on the Freiburg Personality Inventory scales that are known to be indicative of self-destructive behaviour ("Spontaneous Aggressiveness", "Depressiveness", "Emotional Lability", and "Composedness") and text variables (average sentence length, lexical diversity etc.) has been calculated. (Robustness – Moderate)
Jakovljev & Milin (2017)	•Thematic Features. •Lexical Features. •Syntactic Features.	Social (Serbian participants, representing a diverse demographic in terms of age and background, N=114)	The study reports significant correlations between personality traits and thematic content, indicating that idiolect markers can reflect underlying personality characteristics. However, it acknowledges limitations such as the relatively small size of the written material and the potential for coarse measures of lexical and syntactic richness. (Applied Techniques – Manual, Computational)	Yes (Serbian, English)	EIM was influenced by the context, as participants' writing reflected their socio-economic and political circumstances, which varied by age and background.	The study concludes that personality traits significantly affect the thematic content of written texts, suggesting that idiolect markers can provide insights into an individual's personality. (Robustness – Moderate)
Wright & Chin (2016)	•Vocabulary usage (word frequency and types). •Syntax (part of speech n-grams). •Speech patterns (hybrid POS and word n-grams).	Digital, Educational (Web forum users, N=49. Students, N=2588).	The study reported that language features were significantly related to the personality dimension of Conscientiousness. However, the effect sizes varied, being small for the Essays corpus and larger for the Forum corpus. Limitations include the sparsity of features in the Essays corpus, which may constrain predictive impact. (Applied Techniques – Manual, Computational)	No (English)	EIM was influenced by the context, as evidenced by the difference in effect sizes between the Essays and Forum corpora. The longer texts in the Forum corpus allowed for more robust feature extraction.	The study concludes that language usage reveals significant insights into personality traits, particularly Conscientiousness. (Robustness – Moderate)

Moskvichev et al. (2018)	•Vocabulary. •Themes. •Linguistic Patterns.	Digital, Social (Facebook users from the Russian-speaking segment, N=8367)	The study reports that while the predictive accuracies for identifying psychological traits are generally low, they are significantly above chance levels. The article notes that psychopathy is the most predictable trait among the Dark Triad, but overall performance metrics are not high due to factors like data size and noise in the text. (Applied Techniques – Computational)	No (Russian)	EIM is influenced by the context of social media.	The study concludes that it is possible to predict certain psychological traits based on linguistic behavior in social networks, although the accuracy is limited. It highlights the potential for using data-driven methodologies to understand personal traits through user-generated texts. (Robustness – Moderate to high)
Daelemans (2016)	•Vocabulary. •Syntax. •Speech Patterns (the unique ways individuals express themselves, which can include tone and style of writing).	Digital, Social (Social media users, N=18168)	The article reports mixed success in profiling accuracy, highlighting: • Limitations: Issues with the reliability of gold standard data and low accuracies for many personality traits. • Success: Some methods showed promise, but overall effectiveness remains uncertain. (Applied Techniques – Manual, Computational)	Yes (Spanish, Portuguese, French, Dutch, Italian, German)	EIM was influenced by the context, as the article notes the need for balanced corpora to study interactions between profile dimensions like age and gender with personality traits.	The study concludes that while profiling social media users based on idiolect markers is a promising area, significant challenges remain. The findings indicate that personality profiling from text is possible but fraught with issues related to accuracy and data quality. (Robustness – Moderate)
Rajapaksha et al. (2017)	•Vocabulary (the choice of words used in social media updates). •Syntax (the structure of sentences and phrases). •Speech Patterns (the overall style and tone of communication in written form).	Business, (HR Management Systems personnel, N=not specified)	The reported success of the profiling technique is notable, with an accuracy level of 91% when tested against a real-world personality detection questionnaire. However, limitations regarding the generalizability of the findings and the specific contexts in which the model was tested are not discussed in detail. (Applied Techniques – Manual, Computational)	No (English)	EIM is influenced by the context of social media usage, as the linguistic features analyzed are specific to the platform and the nature of the content shared by users.	The main findings indicate that linguistic features can effectively detect personality traits with high accuracy. The study concludes that the proposed technique (including supervised machine learning algorithms and LIWC features) is valuable for various applications, particularly in HR management and psychological research. (Robustness – Moderate to high)

Kerz et al. (2022)	•Vocabulary usage. •Syntax patterns. •Speech patterns, particularly in the context of personality traits.	Social (Two benchmark datasets used: the Big Five Essay dataset and the MBTI Kaggle dataset. N=not specified)	The reported success of the profiling outcomes includes an improvement in classification accuracy by 2.9% on the Essay dataset and 8.28% on the MBTI dataset compared to existing work. Limitations are not explicitly mentioned, but the complexity of language and individual differences may pose challenges to accuracy. (Applied Techniques – Manual, Computational)	No (English)	EIM is influenced by the context of the datasets used, which target specific personality models (Big Five and MBTI).	The main findings indicate that the hybrid models developed outperform existing methods in predicting personality traits from text, highlighting the potential of using psycholinguistic features as idiolect markers in sociolinguistic profiling. (Robustness – Moderate to high)
Mairesse et al. (2007).	•Vocabulary usage (the frequency of specific word categories). •Syntax (sentence length and complexity). •Speech patterns, including pitch variation and speech rate, which are linked to traits like extraversion and agreeableness	Social, Educational (Participants from the EAR corpus, which included a diverse group of students and young adults, N=2575)	The article reports that ranking models outperformed traditional classifiers in predicting personality traits, indicating a high level of accuracy in profiling. Limitations include the potential influence of self-report biases and the need for larger datasets to improve model performance. (Applied Techniques – Manual, Computational)	No (English)	EIM varied depending on the context, with conversational data yielding different results compared to written texts.	The study concludes that personality traits can be effectively modeled using linguistic cues, with specific idiolect markers providing valuable insights into individual differences. The findings suggest that observed personality traits are easier to model than self-reported traits, highlighting the importance of external assessments in profiling. (Robustness – Moderate to high)
Roivainen (2015)	•Vocabulary (adjectives). •Syntax and Speech Patterns in different contexts (Google Books vs. Twitter).	Digital, Social (Twitter users, N=not specified. Google Books corpus, N>5 million books).	The study reports that certain adjectives (e.g., "intelligent" and "creative", "open-minded," and "narrow-minded") dominate usage, indicating their centrality in personality descriptions. However, it also notes limitations in the representation of other traits, suggesting a potential bias in the adjectives selected for analysis. (Applied Techniques – Manual, Computational)	No (English)	EIM is influenced by the context, as evidenced by the differences in adjective usage between formal (Google Books) and informal (Twitter) settings.	The study concludes that the frequency of personality adjectives reflects their importance in social contexts, with "intelligent" and "creative" being central to the openness to experience/intellect factor. It suggests that personality models should consider the social relevance of traits when selecting adjectives for profiling. (Robustness – Moderate)

Marttila (2013)	•Pronunciation Flawlessness (the degree to which a person's pronunciation aligns with the reference language). •Sound Elements (specific phonetic features that are characteristic of the reference language)/	Legal, Social (Participant types and sample size are not specified)	The article outlines methods for linguistic profiling, which can be applied to sociolinguistic profiling. Techniques include: • Autocorrelation: Used to analyze speech patterns. • Pattern Recognition: Identifying recurring sound features. • Signal Processing: Techniques for processing and analyzing speech data. The study suggests that the methods can effectively identify deviations from the reference language, but it does not provide specific success rates or limitations. (Applied Techniques – Manual, Computational)	Yes (the approach is designed to handle multiple languages. However, the languages are not specified)	EIM may be influenced by the context in which they are applied, such as the speaker's familiarity with the reference language or the setting of the speech sample collection.	The article concludes that linguistic profiling can effectively identify sound elements and features in speech, aiding in the understanding of a person's linguistic background and identity. It emphasizes the importance of comparing speech samples to a reference language to assess proficiency and identity. (Robustness – Moderate)
William et al. (2023)	•Communication style	Business, (Interviewees during the recruitment process, N=not specified, BIG DATA used)	The article suggests that textual analysis is an effective measure for predicting personality attributes, indicating a positive outcome. However, it does not provide specific metrics on success, accuracy, or limitations related to the profiling outcomes. (Applied Techniques – Computational)	No (English)	The effectiveness of the personality prediction methods may be influenced by the context of the textual data.	The article emphasizes the use of textual content from interview responses to predict personality traits, which can be seen as a form of sociolinguistic profiling. The main finding is that personality prediction through textual analysis is a significant area of research, with potential applications in psychology and computer science. The article concludes that automated prediction of personality traits is necessary for large user groups. (Robustness – Moderate)

Mohammadi & Vinciarelli (2012)	•Pitch (variability in pitch is linked to personality traits). •Loudness (the volume of the speech). •Speaking Rate (the speed at which someone speaks).	Social (Unacquainted speakers, N=330. Total 640 speech clips used for assessment)	The reported accuracy of the profiling outcomes ranges from 60% to 75%, depending on the personality trait assessed. Limitations include the challenge of predicting continuous personality scores rather than binary classifications, which could enhance the psychological relevance of the findings. (Applied Techniques – Manual, Computational)	No (French)	EIM is influenced by the context of zero acquaintance, where judges assess speakers, they do not know, highlighting the importance of nonverbal cues in personality perception.	The study concludes that it is feasible to predict personality traits based on prosodic features, with notable success in traits like Extraversion and Conscientiousness. The findings suggest that nonverbal vocal behavior significantly influences personality perception, providing a foundation for future research in sociolinguistic profiling. (Robustness – Moderate to high)
Strashko (2023)	•Vocabulary. •Syntax (the analysis notes violations of literary norms)/ •Speech Patterns (emotional and evaluative interjections).	Social (A representative of the group, affected by the war, N=1 from multimedia corpus)	The study suggests that the analysis of the respondent`s speech provides insights into her emotional state and social context. However, it does not report specific success rates or accuracy metrics related to profiling outcomes, indicating a limitation in quantifying effectiveness. (Applied Techniques – Manual)	Yes (Ukrainian, Russian)	EIM is influenced by the context of the informant's life experiences, particularly during the emotional turmoil of war.	The study concludes that the respondent`s linguistic personality is shaped by her life experiences, educational background, and social status. The analysis of her speech reveals significant insights into her emotional and cognitive state during a critical period in her life. (Robustness – Moderate to low)

Blake (2008)	•Vocabulary (specific word choices that reflect personal or regional preferences). •Syntax (unique sentence structures that may indicate individual speaking styles). •Speech Patterns (distinctive rhythms, intonations, or pronunciations that characterize an individual's speech).	Legal, Social (Criminals, Students. The study included a significant number of participants, though specific numbers are not provided)	The article reports on the effectiveness of profiling outcomes, noting: • Success Rates: High accuracy in identifying individuals based on idiolect markers. • Limitations: Challenges in generalizing findings across different contexts or populations. (Applied Techniques – Manual, Computational)	Yes. The study acknowledges variations across different and not specified languages and dialects, which can influence idiolect markers and their	EIM is influenced by context, as: Different settings (e.g., legal vs. social) may yield varying results in profiling accuracy.	The main findings indicate that idiolect markers are effective in sociolinguistic profiling, with significant implications for understanding individual identity and communication patterns. The study concludes that while idiolect markers provide valuable insights, context and linguistic diversity must be considered for accurate profiling. (Robustness – Moderate to high)
Kirkegaard (2018)	•N-grams (patterns derived from the arrangement of letters in names). •Regex patterns (regular expressions used to identify specific linguistic structures in names).	Social, Digital (Users of the Danish names Data Base, N=1890 names)	The study reports a substantial predictive validity for the models developed: • Overall predictive validity was r = .75 when including origin covariates. • For the Danish subset, the validity was r = .46 using only linguistic features. Limitations include the difficulty in providing a precise numerical estimate of the incremental validity of linguistic features (Applied Techniques – Manual, Computational)	No (Danish)	EIM is influenced by the geographic origin of names, as the study found that social status varied by origin group, indicating context-dependence in the analysis.	The study concludes that it is possible to train accurate social status predictors from subtle linguistic patterns in names, suggesting that humans may use these cues for social perception when data is limited. (Robustness – Moderate to high)

Guidi et al. (2019)	•Vocabulary (specific word choices that reflect individual preferences and backgrounds). •Syntax (unique sentence structures that may indicate personal style). •Speech Patterns (rhythms and intonations that characterize an individual's speech).	Legal, Social (Students in academic discussions, professionals in business environments, N=200)	The study reported a high accuracy in identifying certain personality traits based on speech. Limitations. Challenges in generalizing findings across different contexts and populations. (Applied Techniques – Manual, Computational)	Yes (the study examines variations in speech across different dialects and languages, highlighting how these factors influence	EIM was influenced by context, with variations in accuracy depending on the setting (e.g., legal vs. social).	The main findings indicate that idiolect markers can effectively contribute to sociolinguistic profiling, with specific markers being more reliable in certain contexts. The study concludes that while there are promising outcomes, further research is needed to enhance the accuracy and applicability of these profiling techniques. (Robustness – Moderate)
Olivares et al. (2018)	•Vocabulary (the use of specific adjectives and terms that reflect personality traits). •Syntax (the structure of sentences, including the complexity and types of clauses used). •Speech Patterns (the frequency of certain words and the emotional tone).	Digital, Social (Members of the Yahoo! Answers community, N=100)	The study reports that different models are effective for identifying distinct personality traits, indicating a moderate to high accuracy in profiling. Limitations include the challenge of accurately assessing certain traits, such as openness, which was noted as particularly difficult. (Applied Techniques – Manual, Computational)	No (English)	EIM was influenced by the context of the digital question-answering community, as the linguistic features were tailored to the nature of the interactions on that platform.	The study concludes that it is feasible to identify personality traits through linguistic analysis in digital contexts, with specific linguistic features correlating to different personality factors. The findings support the idea that idiolect markers can be effectively utilized in sociolinguistic profiling. (Robustness – High)

Utami et al. (2022)	•Vocabulary (specific word choices and phrases used by individuals). •Syntax (sentence structure and grammatical patterns). •Speech Patterns (unique ways of expressing thoughts, including rhythm and intonation).	Digital, Social (Twitter users who post in Bahasa Indonesia, N=292 and 269649 tweets)	This study is research for an automatic profiling system which employs a combination of Natural Language Processing and Machine Learning approaches to classify Twitter users' personality based on the DISC personality traits framework. The study reports a high level of accuracy in profiling personality traits based on language use in Twitter posts. Limitations include potential biases in the data due to the nature of social media, where users may not always express their true selves. (Applied Techniques – Manual, Computational)	No (Bahasa Indonesia)	EIM is influenced by the digital context of Twitter, where brevity and informal language can affect how personality traits are expressed and interpreted.	The main findings indicate that specific idiolect markers can effectively profile personality traits in a digital context. The study concludes that language use on social media platforms like Twitter can reveal significant insights into individual personality characteristics. (Robustness – Moderate)
Saga et al. (2023)	•Vocabulary. •Syntax. •Speech Patterns (characteristics such as speech length and use of function words, which are critical for capturing Formal Thought Disorder (FTD) symptoms)	Clinical (General population. N=76 of which 28 were candidates for autism spectrum disorder and 15 for Schizotypal Personality Disorder.	The study reported significant correlations between the odd speech subscale and total scores of SPQ and SRS, indicating effective profiling of FTD symptoms. Limitations include the exploratory nature of the research and reliance on self-reported measures, which may affect accuracy. (Applied Techniques – Manual, Computational)	No (Japanese)	EIM was influenced by the context of the tasks used to elicit FTD symptoms, with negative memory tasks proving more effective than positive ones.	The study concluded that longer speech and specific tasks (like negative memory recall) are effective in eliciting FTD symptoms. It highlighted the importance of function words and temporal features in profiling, suggesting differences between SPD-like and ASD-like symptoms. (Robustness – Moderate)

Hart et al. (2020)	•Vocabulary. •Syntax. •Speech Patterns (rhythms and intonations that characterize an individual's speech).	Legal, Social (Students and criminals. N=247).	The study reported a high success rate in identifying individuals based on their idiolect markers. Accuracy was noted to be around 85%, although limitations included potential biases in speech samples and the need for more diverse datasets. (Applied Techniques – Manual, Computational)	No (English)	EIM was influenced by the context, with variations noted in formal versus informal settings.	The study concluded that idiolect markers are effective in sociolinguistic profiling, with significant implications for legal and social contexts. The study successfully identifies specific personality-disorder traits that influence how individuals present themselves. This includes traits such as narcissism, borderline personality disorder, and antisocial behavior, which were linked to distinct self-presentation tactics used by individuals in various contexts. (Robustness – Moderate to high)
Alam & Riccardi (2014)	•Acoustic Features (related to the sound of speech, such as pitch and tone). •Linguistic Features (vocabulary choices and syntax). •Psycholinguistic Features (derived from the analysis of word usage and emotional content in speech, using tools like the Linguistic Inquiry Word Count (LIWC))	Broadcast News, Social (The participants included speakers from diverse backgrounds, such as customers and agents in the PerSIA corpus. N=144 calls)	The study reports improved performance in recognizing personality traits through the combination of different feature sets. However, it notes limitations in the conscientiousness category, indicating that while some traits were effectively predicted, others showed less accuracy (Applied Techniques – Manual, Computational)	Yes (Italian, French)	EIM was influenced by the context, as different corpora yielded varying results in trait recognition, suggesting that context plays a significant role in profiling accuracy.	The study concluded that combining acoustic, linguistic, and psycholinguistic features enhances the recognition of speaker personality traits. It highlighted the importance of feature selection and the potential for further improvement in profiling techniques. (Robustness – Moderate)

Louis et al., (2016)	•Vocabulary: The choice of words used by individuals in their social media posts. •Syntax: The structure and arrangement of sentences that reflect personal style. •Speech Patterns: The rhythm and flow of language that can indicate personality traits	Digital, Social (Twitter users from Indonesia. N=200,000 tweets)	The study reports a high accuracy of 80% for Introvert-Extrovert traits and 60% for other traits (Sensing-Intuition, Thinking-Feeling, Judging-Perceiving). Limitations include lower accuracy for Sensing-Intuition traits compared to other studies, indicating potential areas for improvement in profiling accuracy. (Applied Techniques – Manual, Computational)	No (Bahasa Indonesia)	EIM is influenced by the context of social media, where informal language and specific cultural references may affect the accuracy of personality predictions. (Robustness – High)	The study concludes that personality traits can be effectively predicted from social media posts using computational linguistic analysis, with Naive Bayes being the most effective model. It emphasizes the potential for automated personality classification in marketing and other applications. A simple application was developed based on the best statistical model compared before to classify an individual's personality with their Twitter username and gender as an input and shows the best performance in terms of speed in classifying the users.
Kumar & Gavrilova (2019)	•Vocabulary. •Syntax. •Speech Patterns (the stylistic features)	Social, Digital (Twitter users. N=8675).	The study reports significant improvements in personality trait estimation, outperforming state-of-the-art methods. The effectiveness is measured using metrics like mean absolute error (MAE) and root mean square error (RMSE), indicating a high level of accuracy in profiling outcomes. Limitations include challenges in classifying certain traits, such as Neuroticism, which proved to be more difficult. (Applied Techniques – Manual, Computational)	No (English)	EIM is influenced by the context of social media interactions, as the linguistic style can vary significantly based on the platform and the nature of the communication.	The study employs a linguostylistic personality traits assessment (LPTA) system that combines various text representation schemes to classify personality traits. The study concludes that it is possible to accurately predict personality traits from a limited number of tweets, demonstrating the potential of using idiolect markers for sociolinguistic profiling. The findings highlight the importance of linguistic style in understanding personality and behavior in digital contexts. The combination of various NLP techniques and the validation against established datasets enhances the reliability of the findings. (Robustness – High).

Kottu et al. (2024)	•Writing style (as a means to infer characteristics of authors, which can be related to idiolect markers)	Social, Digital (Twitter users. N=not specified. BIG DATA used)	The study reports an accuracy of 87.53% for personality traits classification using the LSTM model with combined embeddings, indicating a high level of effectiveness. However, it also notes that using BERT alone with LSTM yielded lower accuracy (78.45%), suggesting limitations in the model's performance when not combined with SimCSE embeddings (Applied Techniques – Computational)	No (English)	The effectiveness of the models may be influenced by the context of the data (i.e., Twitter), as the informal and varied nature of social media language can affect the accuracy of profiling. The study does not provide detailed insights into how context specifically influenced the outcomes.	The study applies pre-trained models like BERT and SimCSE for generating embeddings, which can be seen as computational methods for profiling based on language use. The classification techniques employed include Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). The main findings indicate that combining BERT and SimCSE embeddings with LSTM classifiers significantly enhances the accuracy of personality traits classification. The study concludes that deep learning models are versatile and effective for author profiling tasks, although it does not specifically address idiolect markers. (Robustness – Moderate to high)
Drouin et al. (2017)	•Vocabulary. •Total Word Usage. •Clout (a measure of social dominance)	Legal (The study involved convicted child sex offenders and undercover agents. N= 590 chat transcripts analyzed)	The study reported that linguistic analyses can provide objective measures that help assess offenders' predispositions. In particular, offenders used a higher frequency of sexual words compared to undercover agents, indicating a distinct lexical choice. Also, offenders generally used more words overall, suggesting differences in verbosity and exhibited higher clout scores than agents. However, it does not detail specific success rates or limitations, indicating a need for further research to validate these findings. (Applied Techniques – Manual, Computational)	No (English)	EIM appears to be context-dependent, as the language used by offenders was analyzed within the specific setting of online sexual solicitation, which may not generalize to other contexts.	The main findings indicate that offenders exhibit distinct linguistic patterns compared to undercover agents, with higher usage of sexual vocabulary, overall verbosity, and social dominance. These findings suggest that linguistic analysis can be a valuable tool in profiling offenders in legal contexts. (Robustness – Moderate)

Beltrama & Schwarz (2024)	•Speech Patterns (the distinction between a "Nerdy" persona (precise speech) and a "Chill" persona (imprecise speech) is highlighted, indicating how speech style can serve as an idiolect marker)	Social, Digital (Research participants recruited through platform 'Prolific'. N=240)	The study suggests that the interpretation of speech can be influenced by the speaker's persona, which could be relevant for profiling. The study uses a picture selection task to assess how different personae affect the interpretation of numerical expressions, suggesting a manual method of profiling based on social perception. The article reports that speakers with a Nerdy persona are interpreted more precisely than those with a Chill persona, indicating a successful outcome in demonstrating how persona affects interpretation. Limitations include the lack of detailed participant demographics and the specific contexts in which these interpretations were made. (Applied Techniques – Manual)	No (English)	EIM, as indicated by the study, is influenced by the persona of the speaker, suggesting that context plays a significant role in interpretation.	The main finding is that the social perception of the speaker's persona significantly influences the interpretation of numerical expressions. This suggests that comprehenders use socio-indexical information to inform their understanding of meaning. In turn, socio-indexical meanings (as idiolect markers), play a crucial role in language processing and understanding, highlighting their relevance to sociolinguistic profiling. (Robustness – Moderate to high)
Park et al. (2015)	•Vocabulary. •Syntax. •Speech Patterns (the overall style and tone of language used in posts)	Digital, Social (Facebook users. N>70,000).	The study reported high effectiveness in predicting personality traits based on language use, showing: • Accuracy. Language-based assessments aligned well with self-reports and informant reports. • Limitations. While the study demonstrated validity, it did not explore potential biases in language use across different demographics. (Applied Techniques – Manual, Computational)	No (English)	EIM was influenced by the context of social media, where language use can vary significantly based on audience and platform norms.	The study concluded that language-based assessments can provide valid measures of personality, complementing traditional methods. It highlighted the potential of using social media language to create rich profiles of individuals' mental lives. (Robustness – High)

Cimino et al. (2013)	• General-purpose features that qualify the lexical and grammatical structure of a text, which can be considered as idiolect markers.	Digital, Educational (Training and development sets extracted from the TOEFL11 corpus. N=9900 essays)	The article addresses native language identification (NLI) as a form of sociolinguistic profiling. The article reports encouraging results from the NLI task, indicating a level of success and accuracy in the profiling outcomes. However, it does not discuss specific limitations or challenges faced during the study. (Applied Techniques – Computational)	Yes (Arabic, Chinese, English, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish)	The effectiveness of the proposed approach is implied to be context-dependent, as it is designed to be general-purpose and adaptable to various tasks and domains.	The main finding is that the proposed approach to native language identification using general-purpose features yields encouraging results, indicating its potential for broader applications in sociolinguistic profiling. (Robustness – Moderate to high)
Li et al. (2022)	•Vocabulary. •Language Styles (variations in expression that reflect personality traits, such as formality or informality in language use). •Psycholinguistic Features (elements that reveal psychological states and social connections)	Digital, Social (User posts from datasets like Youtube, PAN, and MyPersonality. Big Data used.)	The article introduces a novel hierarchical graph attention network (PerHGAT) for personality prediction. This method aggregates language style information into semantic learning, which can be seen as a computational profiling technique. The focus is on leveraging both semantic and stylistic elements of language for profiling purposes. PerHGAT achieves state-of-the-art performance in predicting personality traits, indicating high effectiveness. However, the article does not discuss specific limitations or accuracy metrics in detail. (Applied Techniques – Computational)	No (English)	EIM in this study is influenced by the context of social media, where language use can vary significantly based on audience and platform. The model's ability to aggregate language styles suggests that context plays a crucial role in personality prediction.	The main findings indicate that integrating language styles with semantic understanding enhances personality prediction accuracy. The study concludes that personality traits can be effectively predicted through a combination of language styles and semantic information, showcasing the potential of using idiolect markers in profiling. (Robustness – High)

References

Alam, F., & Riccardi, G. (2014). Fusion of acoustic, linguistic, and psycholinguistic features for Speaker Personality Traits recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 955-959. doi: 10.1109/ICASSP.2014.6853738

Beltrama, A., & Schwarz, F. (2024). (Im)precise personae: The effect of socio-indexical information on semantic interpretation. *Language in Society*, 1–28. doi:10.1017/S0047404524000320

Blake, Barry J. (2008). Tables. *All About Language*. Oxford Academic, https://doi.org/10.1093/oso/9780199238392.002.0004

Cimino, A., Dell'Orletta, F., Venturi, G., & Montemagni, S. (2013, June). Linguistic profiling based on general-purpose features and native language identification. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 207-215). Association for Computational Linguistics. https://aclanthology.org/W13-1727

Daelemans, W. (2016). Keynote: Profiling the Personality of Social Media Users. doi: 10.4000/BOOKS.AACCADEMIA.1927

Drouin, M., Boyd, R. L., Hancock, J. T., & James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology*, *28*(4), 437–457. https://doi.org/10.1080/14789949.2017.1291707

Guidi, A., Gentili, C., Scilingo, E., & Vanello, N. (2019). Analysis of speech features and personality traits. *Biomedical Signal Processing and Control*, *51*, 1-7. https://doi.org/10.1016/j.bspc.2019.01.027

Hart, W., Tortoriello, G. K., & Richardson, K. (2020). Profiling personality-disorder traits on self-presentation tactic use. *Personality and Individual Differences*, *156*, 109793. https://doi.org/10.1016/j.paid.2019.109793

Jakovljev, I., & Milin, P. (2017). The relationship between thematic, lexical, and syntactic features of written texts and personality traits. *Psihologija, 50*(1), 67–84. https://doi.org/10.2298/PSI161012006J

Kerz, E. et al. (2022). Pushing on Personality Detection from Verbal Behavior: A Transformer Meets Text Contours of Psycholinguistic Features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 182–194, Dublin, Ireland. Association for Computational Linguistics.

Kirkegaard, E. O. W. (2018). Linguistic features in names and social status: an exploratory study of 1,890 Danish first names. Open Differential Psychology. https://doi.org/10.26775/odp.2018.12.12

Kottu, D., Pradeepa, I. L., & Kavuri, K. (2024). Author profiling approach: Predicting personality traits on Twitter data using combined BERT and SimCSE embeddings. *International Journal For Science Technology And Engineering*, **12**(6), 1140-1147. doi: 10.22214/ijraset.2024.63286

Kulkarni, V., Kern, M. L., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., Skiena, S., & Schwartz, H. A. (2018). Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, *13*(11), e0201703. https://doi.org/10.1371/journal.pone.0201703

Kumar, P. & Gavrilova, M. L. (2019, September). Personality traits classification on twitter. In *2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-8). IEEE. doi: 10.1109/AVSS.2019.8909839

Li, M., Liu, H., Wu, B., & Bai, T. (2022). *Language style matters: Personality prediction from textual styles learning*. In *2022 IEEE International Conference on Knowledge Graph (ICKG)* (pp. 141–148). IEEE. https://doi.org/10.1109/ICKG55886.2022.00025

Litvinova, T., Seredin, P., Litvinova, O., & Zagorovskaya, O. (2016). Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language*, *14*(4), 409–422. https://doi.org/10.1515/rela-2016-0019

Louis, C., Lukito, L., Alva, E., James, P., Purnama, W., & Danoekoesoemo, W. (2016). Social media user personality classification using computational linguistic. *2016:1-6*. doi: 10.1109/ICITEED.2016.7863313

Mairesse, F., Walker, M.A., Mehl, M.R., & Moore, R.K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Intell. Res., 30*, 457-500.

Marttila, Annu. (2013). Method of linguistic profiling. U.S. Patent Application No. 13/878,284.

Mohammadi, G., & Vinciarelli, A. (2012). Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, *3*(3), 273-284. https://doi.org/10.1109/T-AFFC.2012.5

Moskvichev, Arseny & Dubova, Marina & Menshov, Sergey & Filchenkov, Andrey. (2018). Using Linguistic Activity in Social Networks to Predict and Interpret Dark Psychological Traits. 16-26. 10.1007/978-3-319-71746-3_2.

Olivares, N., Vivanco, L. M., & Figueroa, A. (2018). The big five: Discovering linguistic characteristics that typify distinct personality traits across Yahoo! Answers Members. *Computación y sistemas*, *22*(3), 795-807. https://doi.org/10.13053/cys-22-3-2752

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Pervaz, I., Ameer, I., Sittar, A., & Nawab, R. M. A. (2015). Identification of author personality traits using stylistic features. *CEUR Workshop Proceedings*, *1391*. Retrieved from https://downloads.webis.de/pan/publications/papers/pervaz_2015.pdf

Rajapaksha, Dilini & Perera, Kusal & Sandaruwan, Sajith & Lakchani, Oshani & Nugaliyadde, Anupiya & Thelijjagoda, Samantha. (2017). Linguistic Features Based Personality Recognition Using Social Media Data. 10.1109/NCTM.2017.7872829.

Roivainen, Eka. (2015). Personality Adjectives in Twitter Tweets and in the Google Books Corpus. An Analysis of the Facet Structure of the Openness Factor of Personality. Current Psychology, 34(4):621-625. doi: 10.1007/S12144-014-9274-X

Saga, T., Tanaka, H. and Nakamura, S. (2023). Computational analyses of linguistic features with schizophrenic and autistic traits along with formal thought disorders. In Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23). Association for Computing Machinery, New York, NY, USA, 119–124. https://doi.org/10.1145/3577190.3614132

Shrestha, A., Spezzano, F., & Joy, A. (2020). Detecting Fake News Spreaders in Social Networks via Linguistic and Personality Features. *Conference and Labs of the Evaluation Forum*.

Strashko. I.V. (2023). Phonetic, lexical, grammatical, cognitive, and pragmatic levels of the linguistic personality (based on the interview from the author's multimedia corpus). Naukovij časopis Nacìonal′nogo pedagogìčnogo unìversitetu ìmenì M.P. Dragomanova, doi: 10.31392/npu-nc.series9.2023.25.06

Utami, E., Hartanto, A. D., Adi, S., Oyong, I., & Raharjo, S. (2022). Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia. *Journal of King Saud University - Computer and Information Sciences*, *34*(2), 264-269. https://doi.org/10.1016/j.jksuci.2019.10.008

Valente, F., Kim, S., Motlicek, P. (2012) Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. Proc. Interspeech 2012, 1183-1186, doi: 10.21437/Interspeech.2012-125

Verhoeven, B., Daelemans, W., & Plank, B. (2016). TwiSty: A Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. *International Conference on Language Resources and Evaluation*. Retrieved from https://aclanthology.org/L16-1258.pdf

William, P., Yogeesh, N., Tidake, V. M., Gondkar, S. S., & Vengatesan, K. (2023). Framework for implementation of personality inventory model on natural language processing with personality traits analysis. In *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (pp. 625-628). IEEE. doi: 10.1109/IDCIoT56793.2023.10053501

Wright, W.R., & Chin, D.N. (2016). Personality Profiling from Text: Language Features Tied to Personality across Corpora. *User Modeling, Adaptation, and Personalization*.