A Semantic Analysis Approach to Extract Personality Traits from Tweets (X)

Original Scientific Paper

Marouane Echhaimi*
Ibn Tofail University,
Science and Engineering, ENSA
(National School of Applied Sciences)
Kenitra, Morocco
Marouane.echhaimi@uit.ac.ma

Khadija Lekdioui
Ibn Tofail University,
Science and Engineering, ENSA
(National School of Applied Sciences)
Kenitra, Morocco
khadija.lekdioui@uit.ac.ma

*Corresponding author

Youness Chaabi
Royal Institute of Amazigh Culture,
CEISIC
Rabat, Morocco
chaabi@ircam.ma

Tarik Boujiha
Ibn Tofail University,
Science and Engineering, ENSA
(National School of Applied Sciences)
Kenitra, Morocco
tarik.boujiha@uit.ac.ma

Abstract *– The utilization of social networks has experienced a substantial surge in the past decade, with individuals routinely exchanging and consuming personal data. This data, subject to analysis and utilization across diverse contexts, has spurred scholarly interest in discerning the personality traits of social network users. Personality, as an intrinsic characteristic, distinguishes individuals in terms of cognition, emotion, and behavior, thereby influencing social relationships and interactions. Among the extensively studied frameworks elucidating personality variance is the Five Factor Model, commonly referred to as the "Big Five," encompassing Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (OCEAN). Personality assessment holds practical utility across domains such as education, security, marketing, e-learning, healthcare, and personnel management. Prior investigations have demonstrated the feasibility of automatic text analysis in personality discernment. This paper introduces a multi-agent methodology grounded in semantic similarity metrics for personality trait recognition via automatic text analysis of Tweets. Our approach leverages WordNet and information content-based semantic similarity measures to analyze tweet content and classify users' personality traits. Experimental results demonstrate the effectiveness of our method, achieving a remarkable 96.28% accuracy in identifying personality traits from Tweets. This high success rate underscores the potential of our semantic analysis approach in accurately profiling social media users' personalities, offering valuable insights for various applications in behavioral analysis and personalized services.*

Keywords: *Big data, personality traits, big five personality, semantic similarity, Tweets*

Received: March 28, 2024; Received in revised form: September 28, 2024; Accepted: September 23, 2024

1. INTRODUCTION

The utilization of social networking platforms on the internet has experienced a substantial surge over the past decade, with platforms like Facebook and Twitter gaining widespread popularity for information dissemination and social interaction purposes. The online activities of users on these platforms offer valuable insights into their personalities, encompassing individual differences in cognition, behavior, experiences, emotions, opinions, and interests [1]. Understanding personality entails grasping how various aspects of an individual coalesce into a cohesive whole, representing a blend of characteristics and behaviors across diverse situations [2]. Moreover, personality plays a pivotal role in influencing decision-making processes across various domains [3]. It significantly impacts interpersonal interactions, relationships, and one's immediate surroundings, showcasing relevance in diverse contexts such as job satisfaction, professional success, and user preferences [4].

Personality delineation holds significance in numerous processes, including personnel recruitment, digital marketing, psychological interventions, educational mentoring, teaching methodologies, and health ad-

visory services. Hence, several applications stand to benefit from insights into personality traits, prompting organizational interest in profiling individuals' personalities. The existing literature presents a multitude of approaches to personality identification [5-7]. However, a common issue arises when these approaches overlook semantic similarity metrics, which are crucial for achieving concrete results in text-based semantic comparisons. Incorporating semantic similarity measures into personality identification frameworks is imperative for enhancing the accuracy and reliability of personality assessments based on textual data.

This article introduces a multi-agent system designed to analyze messages from social networking platforms and extract personality traits of Internet users utilizing semantic similarity measures. The methodology employed in this approach is grounded in the "Big Five" factor theory [8], which is currently the most widely acknowledged personality model within the scientific community. The Big Five model has gained prominence through numerous independent studies [9], culminating in its widespread acceptance and adoption as a comprehensive model for understanding personality traits [10-11].

2. STATE OF THE ART

The field of personality recognition has experienced a notable increase in research activity over recent years [12-13]. The pervasive presence of social media platforms has incentivized researchers to leverage these platforms for valuable insights that can aid in personality prediction. Numerous studies have highlighted the correlation between personality traits and online behavior [14-15]. Quercia et al. [16] were among the pioneers in investigating the association between personality traits and Twitter usage. They proposed a model capable of accurately inferring user personalities based solely on three publicly available metrics from profiles: followers, following, and listed counts. Similarly, Jusupova et al. [17] utilized demographic and social activity data to predict personalities, particularly focusing on children.

Liu et al. [18] introduced a deep learning approach to construct hierarchical systems for word and sentence representations, enabling the inference of user personalities across three languages: English, Italian, and Spanish. Van de Ven et al. [19] conducted analyses using LinkedIn, a platform primarily used for job-related decision-making, and found notable correlations with personality traits, particularly Extraversion. Furthermore, YouYou et al. [20] demonstrated the potential for computerized assessments to surpass human judgments in accuracy, particularly when sufficient data are available, surpassing judgments made by friends, spouses, and even individuals themselves.

3. BIG FIVE MODEL

The Five Factor Model of personality is a cornerstone in psychological research [21-23]. These factors are not theoretically derived but have been empirically identified through natural language analyses and psychological assessments, aiming to capture personality traits independently while providing a comprehensive description of personality. The five primary traits, known as OCEAN [24], are as follows:

- Openness: Individuals scoring high on Openness exhibit a penchant for learning new things and embracing novel experiences. This trait encompasses qualities like insightfulness, imagination, and diverse interests.
- Conscientiousness: Those with high conscientiousness levels are characterized by reliability and punctuality. Traits associated with Conscientiousness include organization, methodicalness, and rigor.
- Extroversion: Extroverts derive energy from social interactions, contrasting with introverts who draw energy from within. Extroversion involves traits such as dynamism, talkativeness, and assertiveness.
- Agreeableness: Individuals high in Agreeableness display friendliness, cooperativeness, and compassion. Conversely, lower scores in Agreeableness may indicate a more distant demeanor. Traits associated with Agreeableness include kindness, affection, and sympathy.
- Neuroticism: Also known as emotional stability, Neuroticism refers to an individual's emotional steadiness and the presence of negative emotions. High Neuroticism scores are often associated with emotional instability and heightened negative emotions. Traits linked to Neuroticism include moodiness and tenseness.

4. PROPOSED APPROACH

The proposed methodology involves automatic analysis of tweet content to determine the personalities of individual Internet users. The primary challenge lies in identifying the personality traits of Twitter users through automated semantic analysis of tweet content. Each tweet undergoes a sequence of treatments [25].

This section outlines five major personality profiles that have been identified and characterized based on various criteria. Four treatments are executed to establish a profile. Initially, tweets are retrieved, followed by simplification through the removal of unnecessary information as the second treatment. The third treatment involves linguistic analysis for word normalization, while the fourth treatment entails semantic analysis of tweets to ascertain a profile based on the Big Five personality traits for each user within the system. The architectural depiction of this process is illustrated in Fig. 1.

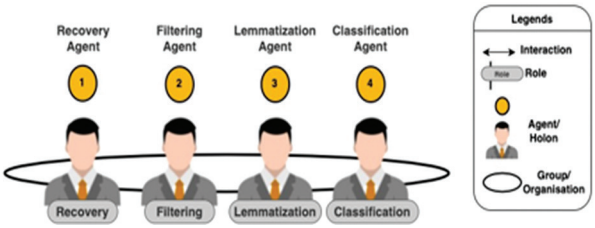

Fig. 1. General architecture of the system

4.1. RECOVERY AGENT

Initially, a retrieval agent employing the Tweepy algorithm [26] is utilized to extract tweets from Twitter and prepare them for subsequent processing steps. In our methodology, the first treatment applied to the tweet corpus involves correcting spelling and grammar errors. Such errors can significantly impact text analysis, both for human comprehension and software algorithms. A single misspelled word or sentence can drastically alter the analysis outcomes. Spelling and grammar correction is achieved using a dictionary corpus integrated with an algorithm that considers language variations, including verbal conjugations, nouns, and adjectives. This process involves comparing words in the tweets with the dictionary corpus, while also considering the context of sentences.

However, it's important to note that while the spelling and grammar checker can be beneficial, it should not replace a thorough manual review for accuracy and precision.

4.2. FILTERING AGENT

Once the retrieval task is completed, the subsequent step involves filter processing to remove words that contribute little to the information conveyed in text messages. These words, termed "empty words," are automatically filtered out for each language [27].

The most commonly occurring words in a corpus typically belong to the category of empty grammatical words, also known as stop words. These include articles, prepositions, linking words, determiners, adverbs, indefinite adjectives, conjunctions, pronouns, and auxiliary verbs, among others [27]. While these words constitute a significant portion of the text, they do not significantly contribute to the overall meaning of the text as they are ubiquitous across all texts. As per Zipf's law [28], removing these empty words during corpus pre-processing streamlines the modeling and analysis process by saving time and reducing computational complexity.

4.3. LEMMATIZATION

After the filtering step, the message undergoes linguistic analysis for word normalization, a process that involves transforming words into their canonical forms through stemming [29]. This normalization process leads to a notable reduction in the lexicon sample size [30]. Lemmatization rules are applied to various words in the corpus to unify morphological variants into a common form, such as converting verbs to their infinitive form and eliminating plural forms. Morphological variants of a word are grouped under the same lemma, allowing them to be treated as a single element (term or concept) during analysis. By reducing the total number of distinct terms, lemmatization contributes to simplifying the complexity of the analyzed text, providing significant advantages to the system.

In many languages, words can exist in multiple forms. For instance, in French, the verb "marcher" may appear as "marche," "marchait," "marchent," or "marchaient." The base form "marcher," typically found in dictionaries, is referred to as the lemma of the word. The combination of the base form with its grammatical properties is often termed the lexeme of the word.

4.4. CLASSIFICATION AGENT

The classification agent evaluates the semantic similarity of a newly acquired tweet and identifies its corresponding personality category (openness, conscientiousness, extroversion, agreeableness, and neuroticism) based on the ratio of training tweets associated with each category.

4.4.1. Semantic similarity measure

In various research domains like psychology, linguistics, cognitive science, and artificial intelligence, assessing semantic similarity among words stands as a critical concern [31]. Semantic similarity, also known as semantic proximity, refers to a measure applied to a set of messages or terms, where the concept of distance between them is predicated on the similarity of their semantic meanings or contents [32]. Conversely, syntactic similarity pertains to a different type of similarity that can be gauged based on the syntactic structures of terms.

Mathematical methodologies are employed to gauge the degree of semantic association between linguistic units, concepts, or instances, through numerical representations. This quantification is achieved by comparing the information that underpins their meanings or describes their essence. Topological similarity can be defined to estimate semantic similarity, utilizing ontologies to determine the distance between terms or concepts [33]. For instance, a basic metric for comparing concepts organized in a partially ordered set and depicted as nodes in a directed acyclic graph (e.g., a taxonomy) could be the shortest path connecting the two concept nodes.

Furthermore, semantic proximity among language units such as words and sentences can also be assessed using statistical techniques like vector space models to correlate words and textual contexts derived from an appropriate text corpus.

4.4.2. Taxonomy

The concept of semantic similarity is more narrowly focused compared to kinship or semantic relationship because the latter encompasses concepts like antonymy and meronymy, whereas similarity does not. However, there is considerable interchangeability in the literature regarding these terms, including semantic distance [34-35]. Fundamentally, semantic similarity, semantic distance, and semantic proximity address the question: "How similar are terms A and B?" The response to this query typically yields a numerical value between -1 and 1, or 0 and 1, where 1 signifies exceedingly high similarity.

4.4.3. Measuring Topological Similarity

There are primarily two approaches for calculating topological similarity between ontological concepts:

- Edge-based approach: This approach utilizes edges and their types as the primary data source [36]. It focuses on the relationships represented by the edges connecting different concepts within the ontology.

- Information content approach: In contrast, the information content approach relies on nodes and their properties as the main data sources [37-38]. It places emphasis on the inherent characteristics and attributes associated with each node or concept in the ontology.

These approaches offer distinct methodologies for evaluating topological similarity within ontological structures, with each approach leveraging different aspects of the ontology's structure and content.

4.4.4. Semantic similarity

Semantic similarity or semantic relationship refers to the measurement of closeness between terms or documents based on their meaning. There are two distinct methods for calculating semantic similarity. One method involves defining topological similarity using ontology to establish a distance metric between words. The other method relies on statistical techniques, such as the vector space model, to correlate words and their textual contexts extracted from a suitable text corpus. In this study, we concentrate on the first approach, utilizing the WordNet ontology for semantic similarity computation [39]. This approach computes similarity by considering the shared and distinct characteristics of objects as the basis for similarity assessment.

4.4.5. WordNet

WordNet is a lexical ontology designed for the English language, serving as a semantic network developed by Princeton University [40]. It structures lexical knowledge in a taxonomic hierarchy, comprising three separate databases: one for nouns, one for verbs, and one for adverbs and adjectives. Within WordNet, terms and concepts are organized into Synsets, which are lists of synonymous terms or concepts. The core component of WordNet is the Synset, which gathers synonyms associated with a specific concept. These Synsets are interconnected through various relationships, such as hypernymy (type of), meronymy (part of), and antonymy (opposite word) [41].

Semantic similarity within WordNet can be computed using two main methods: path length and information content. The path length method calculates the number of nodes or relationships between nodes within the taxonomy. This method offers advantages as it is not reliant on the static distribution of the corpus or word distributions. In our study, we focus on two concepts, "relation" and "name," within the WordNet hierarchy. We utilize WordNet 2.1, which encompasses nine distinct name hierarchies. It's worth noting that in some instances, the path between two concepts may not exist in this version of WordNet (refer to Figure 2). To address this, we introduce a root node labeled "Entity" (refer to Fig. 2), encompassing all nine provided hierarchies within WordNet.

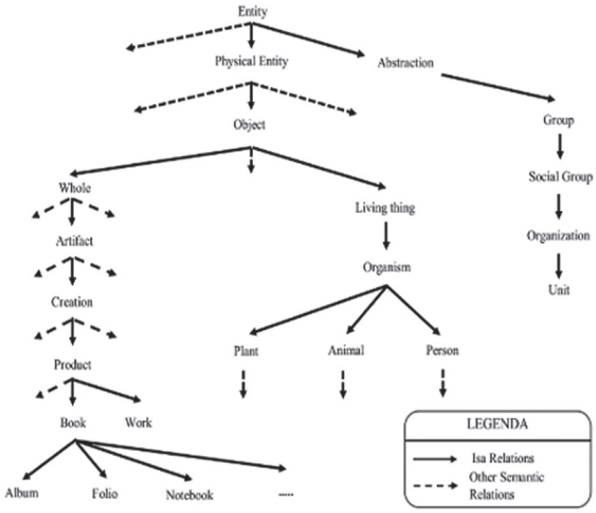

Fig. 2. Extract from the nominal hierarchy in WordNet [40].

4.4.6. Semantic similarity measurement process

The classification agent facilitates a comprehensive processing sequence, as depicted in Figure 3, for semantic similarity computation. This process comprises three distinct phases:

- **Phase 1**: Temporary construction module.
- **Phase 2**: Semantic computation module.
- **Phase 3**: Semantic similarity measurement procedures.

In Phase 1, the temporary construction module sets the groundwork for subsequent semantic computations. Phase 2 involves the actual computation of semantics, while Phase 3 encompasses procedures for measuring semantic similarity between entities.

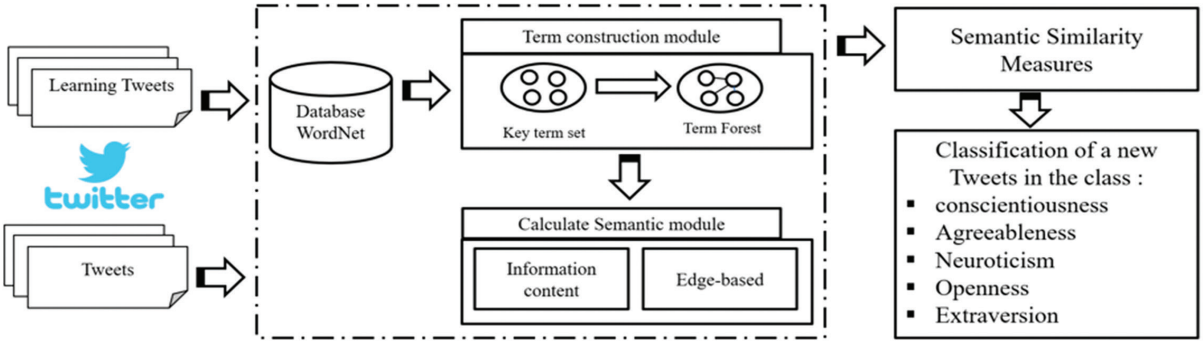

Fig. 3. Semantic Similarity Computation Diagram

- **Phase 1: Term Construction Module**

The primary goal of this module is to identify all words within the tweets that are present in WordNet and to establish the relationships between these words [42]. WordNet is leveraged to enhance the representation of text by incorporating a broader range of semantic information. Specifically, this module utilizes the hypernyms provided by WordNet as valuable features for text analysis. Therefore, the module aims to extract all tweet words found in WordNet and ascertain the relationships between these words based on the hypernym relationships provided by WordNet.

- **Phase 2: Semantic calculation module**

Philip Resnik [42] and Sun Microsystems laboratories propose an alternative to pathfinding in semantic hierarchies by introducing the concept of information content. The information content is a measure of specificity assigned to each concept within a hierarchy based on evidence extracted from a corpus. A concept with high information content is considered highly specific, whereas concepts with low information content are associated with more general ideas. The information content of a concept is calculated by tallying its frequency in a large corpus, as well as the frequency of all subordinate concepts in the hierarchy. The probability of a concept is determined through maximum likelihood estimation, and its information content is derived from the negative logarithm of this probability.

Resnik's similarity measure establishes a semantic relationship between two concepts based on the extent of shared information between them. This shared information is determined by the information content of the least specific concept in the hierarchy that encompasses both concepts.

The similarity between words based on information content:

- Relies on the structure of the thesaurus.
- Improves the path-based approach by normalizing based on hierarchy depth.
- Represents the distance associated with each edge in the hierarchy.
- Integrates probabilistic information derived from a corpus.

The probability that a random word belongs to a concept is calculated as follows (Equation 1) [43]:

$$p(c)=(\sum w \in w_{(c)}\ count\ (w))/\ N \quad (1)$$

Here:

Words (c) represent the set of words subsumed by the concept c.

N is the total number of words in the corpus and the thesaurus.

P(root) = 1 since all words are subsumed by the root concept.

Furthermore, it's worth noting that the probability decreases as the concept descends lower in the hierarchy, reflecting the decreasing specificity and generality as we move down the hierarchy levels.

We need two more definitions:

1) Information Content of a Concept ($IC(c)$) [43]:

$$IC(c)=-log\ P(c) \quad (2)$$

This equation represents the information content of a concept c based on the probability $P(c)$ that a random word from the corpus belongs to the concept c. It quantifies the specificity of the concept within the hierarchy.

2) Lowest Common Subsumer (LCS($c1$, $c2$)) [43]:

The LCS($c1$, $c2$) refers to the lowest node in the hierarchy that serves as a hypernym of both concepts $c1$ and $c2$. It denotes the most specific common ancestor shared by the two concepts.

3) Resnik Similarity Measurement (simResnik($c1$, $c2$)) [43]:

$$simResnik\ (c1,c2) = -log\ P\ (LCS(c1,c2)) \quad (3)$$

This equation calculates the Resnik similarity between concepts $c1$ and $c2$. It estimates the shared amount of information between the concepts by utilizing the information content of their lowest common subsumer (LCS).

- **Phase 3: Semantic similarity measures**

Semantic vectors for $T1$ and $T2$ can be constructed using T statistics and corpus information. The process of deriving semantic vectors for $T1$ (Equation 4) can be described as follows:

Given a word w, let us define [43]:

$$Sim(W_1, W_2) = max_{c1,c2}[smin(c1, c2)]$$
$$Sim(T1, T2) = \sum_{i=1}^{n} \left(\frac{sim(Wi,Wi+1)}{n}\right) \quad (4)$$

We obtain measurement values of semantic similarity for Resnik between Tweet 1 and Tweet 2 (5) [43]:

$$Sim\ Resnik\ (T1,T2) = value\ 2 \quad (5)$$

Tweets are comprised of words, hence it is rational to represent a Tweet using the words it contains. Unlike conventional methods that utilize pre-compiled word lists with numerous words, our approach dynamically constructs semantic vectors solely based on the Tweets being compared. Recent advancements in semantic analysis focus on automatically extracting semantic word vectors for sentences [40]. Given two Tweets $T1$ and $T2$, a word set is formed with (Equation 6) [43]:

$$T=T1 \cup T2 = \{W_1, W_2, ..., W_n\} \quad (6)$$

The word set T encompasses all distinct words from $T1$ and $T2$. Inflectional morphology may lead to a word appearing in various forms within a message, each form having a specific meaning in a given context. Therefore, we consider the word form as it appears in the Tweet for our analysis.

5. EXPERIMENTS AND RESULTS

The objective of this research is to automate the process of identifying the personalities of Internet users by conducting a semantic analysis of their Tweets. To achieve this, we conducted a comparative study between human evaluation and the results produced by our model.

We performed experiments involving intuitive analysis of Tweets from Internet users based on notes from a test corpus. Our focus was on both qualitative and quantitative analyses conducted with the input of three experts. We compiled a corpus of Tweets from a sample of 10 Internet users, each contributing 100 Tweets to our dataset. The intuitive analysis of these Tweets included assigning a personality to each user and then identifying the language acts that contribute to determining the personality traits.

For the identification of personality traits, we utilized the MyPersonality database as the learning base for our system. This database served as the foundation for training our model to accurately classify and infer personality traits based on the semantic analysis of Tweets.

5.1. MYPERSONNALITY DATABASE

To test our approach, we utilize a dataset derived from the MyPersonality project. This dataset was curated for research purposes by David Stillwell and Michael Kosinski through a Facebook application designed to administer a personality test and gather diverse personal information and activities from the profiles of consenting Facebook users. The MyPersonality application operated from 2007 to 2012, accumulating a substantial volume of data.

Our study is built upon a subset of the original MyPersonality dataset, which has been made publicly available [12]. This subset comprises 9913 English status updates extracted from 250 users, with their identities anonymized. The dataset is further annotated with scores for personality traits and includes basic statistics describing the users' social networks.

5.2. TEST CORPUS

In order to have a suitable test base, more than 1.5 million tweets were retrieved using Twitter's Tweepy algorithm [25]; then 10 users were selected with at least 100 tweets per person, the base is in CSV format and each line has 6 fields:

0 - the polarity of the Tweet

1 - Tweet id

2 - the date of the Tweet

3 - the request. If there is no query, then this value is NO_QUERY.

4 - the user who tweeted

5 - the text of the Tweet

The text of the Tweets was compared with MyPersonality learning base taking into account the semantic similarity measure to extract personality traits more accurately.

5.3. EXECUTION RESULTS

The first set of results from the analysis on Tweets was monitored by 3 experts. Each expert analysis was done in two steps. The first step consisted of assigning a profile to each internet user according to their personality traits. The second step consisted of analyzing the Tweets exchanged by internet users. The experts were asked to classify the Tweets of internet users into five personality traits (Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness) by analyzing their content, i.e. identifying the speech acts that characterize Tweets (see Table 1).

When the same Tweets are submitted between internet users to the automatic analysis system that is proposed, the results shown in Tables 1 and 2 are obtained, for the same internet users.

The analysis of these results in light of the characteristics of the internet users' profiles allows associating a personality profile to each user. Seen the results of the semantic analysis to calculate the percentage of each personality type for the users JBnVFCLover78, five broad personality traits emerge: extraversion, agreeableness, openness, conscientiousness, and neuroticism. However, by analyzing the resulting percentages of each personality, we find that the percentage of neuroticism personality is important and characterizes the personality of JBnVFCLover78.

Table 1. Results of the intuitive analysis

Twitter users	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
User 1	29.15 %	16%	16.21 %	16.2 %	22.44 %
User 2	42.01 %	8.9 %	8.6 %	20.82 %	19.67 %
User 3	27%	11.4%	14.13 %	31.47 %	16%
User 4	10.54 %	18%	21.56 %	21.06%	28.84 %
User 5	27.17 %	7.05 %	37.22 %	20.8 %	7.76 %
User 6	65.75 %	15.27 %	3.98 %	13.15 %	1.85 %
User 7	4%	45.4 %	10.38 %	32.56 %	7.66 %
User 8	14.46 %	5.54 %	51.58 %	10.69 %	15.73 %
User 9	10.85 %	11.09 %	68.21 %	4.53 %	5.32 %
User 10	5.96%	6.15%	13.65 %	69.06 %	5.18 %

The table below represents the results of running the system on the test corpus:

Table 2. Results of the execution of the system on the test corpus

Twitter users	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
User 1	34.33 %	14.91 %	13 0/0	17.93 %	19,83 %
User 2	46.41 %	4.57 %	10,11 %	22.22 %	16.69 %
User 3	31.62 %	0.68 %	12,42 %	33.51 %	21.77 %
User 4	06.12 %	24.56 %	17.26	16.22 %	35.84 %
User 5	33.08 %	09.00 %	46.62 %	20.00 %	6.57 %
User 6	60.15 %	12.14 %	05.98 %	17.73 %	04.00 %
User 7	07.56 %	51.12 %	07.00 %	28.00 %	06.32 %
User 8	11.00%	08.00 %	56.27 %	14.00 %	10.73 %
User 9	08.00 %	13.16 %	71.06 %	02.78 %	05.00 %
User 10	07.13 %	04.22 %	08.65 %	80.00 %	02.56 %

Table 3. Comparison between analysis system and intuitive analysis

Twitter users	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
User 1	94,82%	98,91%	96,79%	98,27%	97,39%
User 2	95,6%	95,67%	98,49%	98,6%	97,02%
User 3	95,38%	89,28%	98,29%	97,96%	94,23%
User 4	95,58%	93,44%	95,7%	95,16%	93%
User 5	94,09%	98,05%	90,6%	99,2%	98,81%
User 6	94,4%	96,87%	98%	95,42%	97,85%
User 7	96,44%	94,28%	96,62%	95,44%	98,66%
User 8	96,54%	97,54%	95,31%	96,69%	95%
User 9	97,15%	97,93%	97,15%	98,25%	99,68%
User 10	98,83%	98,07%	95%	89,06%	97,38%
Average	95,88%	96,00%	96,19%	96,40%	96,90%
Final result			**96,28 %**		

Table 3 shows the margin of error between the intuitive analysis of the expert and the system analysis. This error margin gives a confidence degree for results validation. A result of 100% means that the system is perfectly aligned with the human expert. In the example below, from the intuitive analysis the internaut « Bigeny » emerges as an Agreeableness personality with 29,15% (see table 2) and 34,33 % according to the results of the system (see table 1). We have considered the result of the intuitive analysis expert as a reference; we can see that the error rate is 3.72 %.

Table 4. A comparison of the proposed approach with alternative methods.

Methode	Accuracy				
	Agreeableness	Conscientiousness	Extraversión	Neuroticism	Openness
Wang et al [44]	76.8 %	75%	85%	70%	79%
Zheng andWu [45]	65%	62%	71%	68%	70%
Xue et al [46]	78.57 %	77.20 %	82.35 %	83.08 %	79.55 %
BOW [47]	88.75 %	87.86 %	88.15 %	89.52 %	87.94 %
Skip-Vec [47]	88.81 %	88.19 %	88.39 %	89.71 %	88.27 %
Our method	**95.88 %**	**96.00 %**	**96.19 %**	**96.40 %**	**96.90 %**

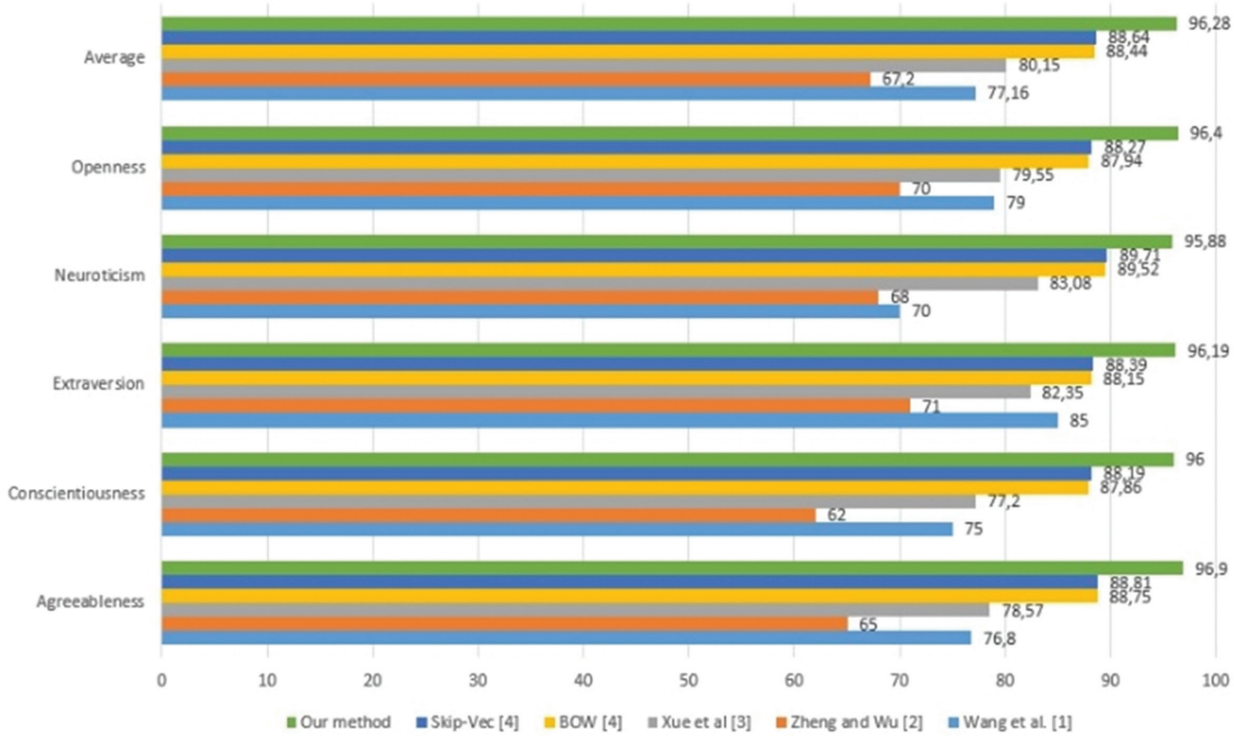

Fig. 4. Graphical representation comparing the proposed approach with alternative methods

The comparison results demonstrate that our method significantly outperforms other evaluated approaches, achieving an average precision of 96.28%. This exceptional performance is particularly evident in traits such as Openness and Neuroticism, showcasing its superior ability to capture nuanced textual cues associated with these personality dimensions.

In contrast, alternative methods exhibit varying levels of performance. Wang et al.'s use of graph convolutional networks for text encoding achieves an average precision of 77.16%, indicating moderate effectiveness [44]. Zheng and Wu's approach, employing semi-supervised learning on Facebook status data, shows a lower precision of 67.2%, suggesting limitations in leveraging social media for precise personality trait recognition [45].

Xue et al.'s method, which employs semantically-enhanced sequential modeling, improves upon these results with an average precision of 80.15% [46]. This method excels particularly in traits like Agreeableness and Extraversion, highlighting its ability to capture contextual relationships within texts.

The Bag of Words (BOW) and Skip-Vec methods achieve average precisions of 88.44% and 88.64%, respectively, demonstrating solid performance but still trailing behind our approach [47]. Skip-Vec slightly outperforms BOW, likely due to its superior incorporation of contextual relationships [47].

Our method clearly surpasses others, demonstrating superior efficacy in personality trait recognition from texts. These findings underscore the robustness and accuracy of our approach, even outperforming newer and more sophisticated techniques in the field.

6. CONCLUSION AND PERSPECTIVES:

Personality traits significantly influence decision-making processes, interpersonal interactions, and individual success. Understanding people's personalities is essential for various applications, such as job candidate selection, targeted marketing, and security measures. Our study focused on detecting personality traits by analyzing Tweets using semantic similarity measures and a learning base grounded in the Big Five model. The experimental results demonstrated a high accuracy rate of 96.28% in identifying personality traits, underscoring the potential of our approach in accurately profiling social media users' personalities.

We have incorporated recent literature to contextualize our findings, highlighting the alignment and relevance of our methodology with current research trends in automated personality detection. The integration of semantic similarity measures, particularly using WordNet and information content-based similarity measures, played a crucial role in enhancing the accuracy and reliability of personality assessments based on textual data.

Future work will focus on improving the system's execution time and expanding the test base to include more comprehensive user information. Additionally, we plan to explore the use of deep learning methods and generative AI to further optimize the accuracy of personality trait detection. By incorporating advanced new techniques, we aim to enhance the robustness and applicability of automated personality profiling in various domains.

7. REFERENCES

[1] E. Kazdin, Encyclopedia of Psychology: 8 Volume Set, American Psychological Association and Oxford University Press, 2000.

[2] B. Y. Pratama, R. Sarno, "Personality classification based on Twitter text using Naïve Bayes, KNN and SVM", Proceedings of the International Conference on Data and Software Engineering, Yogyakarta, Indonesia, 25-26 November 2015, pp. 170–174.

[3] I. Cantador, I. Fernández-Tobías, A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains", Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization, Vol. 997, 2013.

[4] J. Golbeck, C. Robles, M. Edmondson, K. Turner, "Predicting personality from Twitter", Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust and the IEEE Third International Conference On Social Computing, Boston, MA, USA, 9-11 October 2011, pp. 149-156.

[5] E. Kafeza, A. Kanavos, C. Makris, P. Vikatos, "T-PICE: Twitter personality based influential communities extraction system", Proceedings of the IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June - 2 July 2014, pp. 212-219.

[6] Y. Zheng, "Identifying Dominators and Followers in Group Decision Making based on The Personality Traits", Proceedings of the IUI Workshops, 2018.

[7] N. Hawi, M. Samaha, "Identifying commonalities and differences in personality characteristics of Internet and social media addiction profiles: traits, self-esteem, and self-construal", Behaviour & Information Technology, Vol. 38 No. 2, 2019, pp. 110-119.

[8] S. V. Paunonen, M. C. Ashton, "Big five factors and facets and the prediction of behavior", Journal of Personality and Social Psychology, Vol. 81, No. 3, 2001, p. 524.

[9] M. B. Donnellan, F. L. Oswald, B. M. Baird, R. E. Lucas, "The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality", Psychological Assessment, Vol. 18, No. 2, 2006, p. 192.

[10] J. M. Digman, "Higher-order factors of the Big Five", Journal of Personality and Social Psychology, Vol. 73 No. 6, 1997, p. 1246.

[11] B. De Raad, "The Big Five Personality Factors: The psycholexical approach to personality", Hogrefe & Huber Publishers, 2000.

[12] F. Celli, B. Lepri, J. Biel, D. Gatica-Perez, G. Riccardi, "The workshop on computational personality recognition 2014", Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 2014, pp. 1245–1246.

[13] M. Tkalcic, B. D. Carolis, M. D. Gemmis, A. Odi´c, A. Košir, "Preface: EMPIRE 2014-2nd Workshop Emotions and Personality in Personalized Services", Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization Aalborg, Denmark, 7-11 July 2014.

[14] D. Hughes, M. Rowe, M. Batey, A. Lee, "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage", Computers in Human Behavior, Vol. 28, 2011, pp. 561–569.

[15] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, D. Stillwell, "Personality and patterns of Facebook usage", Proceedings of the 4th Annual ACM Web Science Conference, Evanston, IL, USA, 22-24 June 2012.

[16] D. Quercia, M. Kosinski, D. Stillwell, J. Crowcroft, "Our Twitter profiles, our selves: Predicting personality with Twitter", Proceedings of the IEEE Third International Conference on Privacy, Security, Risk and Trust and the IEEE Third International Conference on Social Computing, Boston, MA, USA, 9-11 October 2011, pp. 180-185.

[17] A. Jusupova, F. Batista, R. Ribeiro, "Characterizing the Personality of Twitter Users based on their Timeline Information", Proceedings of the Atas da 16 Conferência da Associacao Portuguesa de Sistemas de Informação, Porto, Portugal, Vol. 16, 22-24 December 2016, pp. 292-299.

[18] F. Liu, J. Perez, S. Nowson, "A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts", arXiv:1610.04345, 2016.

[19] N. Van de Ven, A. Bogaert, A. Serlie, M. J. Brandt, J.J. Denissen, "Personality perception based on LinkedIn profiles", Journal of Managerial Psychology, Vol. 32, 2017, pp. 419-429.

[20] W. YouYou, M. Kosinski, D. Stillwell, "Computer-based personality judgments are more accurate

than those made by humans", Proceedings of the National Academy of Sciences, USA, Vol. 112, No. 4, 2014, pp. 1036-1040.

[21] S. D. Gosling, P. J. Rentfrow, W. B. Swann Jr, "A very brief measure of the Big-Five personality domains", Journal of Research in personality, Vol. 37 No. 6, 2003, pp. 504-528.

[22] M. R. Barrick, M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis", Personnel Psychology, Vol. 44 No. 1, 1991, pp. 1-26.

[23] L. M. Hough, "The 'Big Five' personality variables-construct confusion: Description versus prediction", Human performance, Vol. 5, No. 1-2, 1992, pp. 139-155.

[24] O. Plaisant, J. Guertault, R. Courtois, C. Réveillère, G. A. Mendelsohn, O. P. John "Big Five History: OCEAN of personality factors. Introduction of the French Big Five Inventory or BFI-Fr", Annales Médico-psychologiques, revue psychiatrique, Elsevier Masson, Vol. 168, No. 7, 2010, pp. 481-486.

[25] Y. Chaabi, L. Khadija, B. Mounia, "Semantic Analysis of Conversations and Fuzzy Logic for the Identification of Behavioral Profiles on Facebook Social Network", International Journal of Emerging Technologies in Learning, Vol. 14, No. 7, 2019.

[26] J. Roesslein, "tweepy Documentation", http://tweepy.readthedocs.io/en/v3, 5. (accessed: 2009)

[27] Y. Chaabi, L. Khadija, F. Jebbor, R. Messoussi, "Determination of Distant Learner's Sociological Profile Based on Fuzzy Logic and Naïve Bayes Techniques", International Journal of Emerging Technologies in Learning, Vol. 12, No. 10, 2017, pp. 56-75.

[28] H. Guiter, M. V. Arapov, "Studies on Zipf's law", Brockmeyer, 1982.

[29] J. B. Lovins, "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics, Vol. 11, No. 1-2, 1968, pp. 22-31.

[30] Y. Chaabi, R. Messoussi, V. Hilaire, Y. Ruichek, K. Lekdioui, R. Touahni, "Design of an Intelligent System to Support Tutors in Learning Communities Using Multi-Agent Systems and Fuzzy Logic", International Review on Computers and Software, Vol. 10, No. 8, 2015, pp. 845-855.

[31] Z. Wu, M. Palmer, "Verbs semantics and lexical selection", Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994, pp. 133-138.

[32] G. Zhu, C. A. Iglesias "Computing Semantic Similarity of Concepts in Knowledge Graphs", IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 1, 2017, pp. 72-85.

[33] W. Hua, Z. Wang, H. Wang, K. Zheng, X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge", IEEE transactions on Knowledge and Data Engineering, Vol. 29, No. 3, 2017, pp. 499-512.

[34] J. J. Jiang, D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", arXiv:cmp-lg/9709008, 1997.

[35] T. Pedersen, S. V. Pakhomov, S. Patwardhan, C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain", Journal of Biomedical Informatics, Vol. 40, No. 3, 2007, pp. 288-299.

[36] D. Girardi, S. Wartner, G. Halmerbauer, M. Ehrenmüller, H. Kosorus, S. Dreiseitl, "Using concept hierarchies to improve calculation of patient similarity", Journal of Biomedical Informatics, Vol. 63, 2016, pp. 66-73.

[37] H. Y. Wang, W. Y. Ma, "Integrating Semantic Knowledge into Lexical Embeddings Based on Information Content Measurement", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2, 2017, pp. 509-515.

[38] W. U. Hao, H. Huang, "Efficient Algorithm for Sentence Information Content Computing in Semantic Hierarchical Network", IEICE Transactions on Information and Systems, 2017.

[39] A. Meštrović, A. Calì, "An ontology-based approach to information retrieval", Semantic Keyword-based Search on Structured Data Sources, 2016, pp. 150-156.

[40] C. Fellbaum, G. A. Miller, "WordNet: An electronic lexical database, language, speech, and communication", MIT Press, 1998.

[41] S. Sudha, "WordNet-An On-line Lexical Database", 2016.

[42] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995, pp. 448-453.

[43] L. Meng, R. Huang, J. Gu, "A review of semantic similarity measures in Wordnet", International Journal of Hybrid Information Technology, Vol. 6, No. 1, 2013, pp. 1-12.

[44] Z. Wang, C. H. Wu, Q. B. Li, B. Yan, K. F. Zheng, "Encoding text information with graph convolutional networks for personality recognition", Applied Sciences, Vol. 10, No. 12, 2020, p. 4081.

[45] H. Zheng, C. Wu, "Predicting personality using Facebook status based on semi-supervised learning", Proceedings of the 11th International Conference on Machine Learning and Computing, Zhuhai, China, 22-24 February 2019, pp. 59-64.

[46] X. Xue, J. Feng, X. Sun, "Semantic-enhanced sequential modeling for personality trait recognition from texts", Applied Intelligence, Vol. 51, 2021, pp. 7705-7717.

[47] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, "Modeling, recognizing, and explaining apparent personality from videos", IEEE Transactions on Affective Computing, Vol. 13, No. 2, 2022, pp. 894-911.