Score and Rank Level Fusion Algorithms for Social Behavioral Biometrics

SANJIDA NASREEN TUMPA AND MARINA L. GAVRIOLOVA, (Senior Member, IEEE)
Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada
Corresponding author: Sanjida Nasreen Tumpa (sanjidanasreen.tumpa@ucalgary.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada by the Discovery Grant (DG) on Machine Intelligence for Biometric Security under Grant 10007544, in part by the Strategic Partnership Grant on Biometric-Enabled Identity Management and Risk Assessment for Smart Cities, under Grant 10022972, and in part by the IDEaS Collaborative Project, under Grant 10027075.

ABSTRACT The goal of a biometric system is to recognize individuals based on their unique physiological or behavioral traits. Online Social Networking (OSN) platforms have become an integral part of the daily life of individuals, where they leave a recognizable trail of behavioral information. Social Behavioral Biometric (SBB), being an emerging trend, focuses on such trails to distinguish between individuals. This research investigates the impact of users’ writing profiles on OSN to conclude whether such profiles contribute to SBB. The distinctiveness of the SBB features that are extracted from the social behavioral data of Twitter is studied. A person identification system that relies on users’ writing profiles, reply, retweet, shared weblink, trendy topic networks and temporal profiles is proposed. Score and rank level weighted fusion algorithm performance is compared on a social interaction database of 241 Twitter users. The experimental results establish that the users’ writing profiles have the highest impact over other social biometric features and that score level fusion algorithms perform better than rank level fusion on SBB. The proposed system has achieved recognition rate of 99.45% at rank-1 after cross-validation using genetic algorithm based score level fusion algorithm. The system outperformed all prior researches on SBB in terms of identification accuracy.

INDEX TERMS Biometric security, cognitive systems, decision support system (DSS), linguistic model, rank level fusion, score level fusion, social behavioral biometric (SBB).

I. INTRODUCTION
The decision support system analyzes data to make various types of predictions. Physical and behavioral biometric traits can be used to support such system. An intelligent biometric design support system was proposed as part of the Physical Access Control System in [1]. Online Social Networks (OSN) data is very helpful to augment the domain of decision support systems. For example, Twitter-driven analytics is used in a crime predictive decision support system [2]; users’ web browsing behaviors are used in user identification [3], etc. The artificial intelligence approaches, such as cognitive systems, fuzzy logic, genetic algorithms, intelligent agents, case-based reasoning, evolutionary computing and artificial neural networks (ANN) can be incorporated with traditional decision support systems to augment their capabilities and improve their adaptation [4]–[7]. Adaptation reduces the uncertainty of the machine reasoning biometric systems.

Social behavioral features can be also used to develop socially intelligent and cognitive robots, which improve the interaction between humans and robots [8], [9].

Social Behavioral Biometric (SBB) research has huge potential in the areas of cybersecurity, intelligent adaptive systems, defense robots and public services [10]. Analysis of online social interaction plays a significant role to identify fake users and forgery activities. Researchers focus on social behaviors, preferences, interests, aesthetics, etc. to mine idiosyncratic features to observe important patterns in users’ online behavior [10]. If the credential of an online account is accidentally compromised, with the help of emerging social behavioral biometric research, the OSNs will be able to identify suspicious activities automatically. This can assist online network users in securing their accounts through continuous authentication. In addition, anonymous criminal activities in the OSN can be identified through this research area. This paper focuses on the impact of users’ writing profiles expressed via OSN on SBB performance. Also, it aims to analyze the performance of fusion algorithms to integrate
different SBB traits. We have set the following research questions:

1) Can users’ writing profiles of OSN be integrated with Social Behavioral Biometric?
2) Do users’ writing profiles have better-discriminating capability than other SBB features?
3) Which information fusion method improves the recognition rate of the SBB the most?

The contributions of this paper are:

1) A new SBB trait is proposed to generate the users’ writing profiles from the aggregated tweets and replies of a period based on the word importance.
2) The proposed trait is fused with five other SBB traits to improve the performance of a SBB system.
3) The impacts of score level fusion and rank level fusion algorithms on our proposed SBB system are analyzed.
4) The performance of the proposed system is cross-validated with six combinations of train-test datasets, generated from a proprietary dataset. The proposed method achieves higher recognition accuracy than other SBB systems using genetic algorithm based score level fusion algorithm.

The rest of the paper is organized as follows: Section II discusses related research on SBB. The proposed methodology is presented in Section III and the fusion techniques are discussed in Section IV. The experimental results are discussed in Section V. Finally, Section VI concludes the paper and presents directions for future research.

II. LITERATURE REVIEW

Behavioral biometric has become a subject of extensive research because of its strenuous imitative capability [11]. Many studies focused on identifying users using different behavioral traits extending from real-world actions to virtual world social interactions. Cognitive biometrics use the nerve responses for authentication, which includes eye tracking, mouse dynamics, and keystroke dynamics [12]. Body and muscle movement, smile and social interactions of persons also contain sufficient behavioral information, which can be used as personal signature of individuals [10], [13], [14]. This section provides a brief review of the behavioral biometric systems that deal with the online social behaviors of the users.

The concept of identifying users based on their social interactions was introduced in [10]. A formal definition of SBB has been presented in this work together with the unimodal and multimodal frameworks of SBB. Person identification based on the social interactions of online social networking sites has been proposed in [10]. Authors mined Twitter data to obtain features that can be used as a personal signature using frequency-based and knowledge-based SBB features. The feasibility of the SBB features was further analyzed by the same authors in [10]. A set of social behavioral features from the online social interactions of 241 Twitter users has been identified in [15]. Authors found that only ten recent tweets are enough to recognize 58% of users at rank-1 and established the stability of SB features over time for both frequent and infrequent OSN users. In [16], authors proposed a system based on the temporal information of the users, extracted from OSN.

One of the first mentioning of stylometry for behavioral biometric authentication was in the context of online course student authentication [17]. Researchers also worked on the forensic applications using stylometric features of the OSN [18], [19]. Authors used Gaussian-Bernoulli Deep Belief Network on stylometric features for authorship verification of emails and tweets in [18]. The machine learning and deep learning classifiers were investigated on different combinations of stylometric features for continuous authentication of emails and tweets in [19].

A unimodal biometric system frequently suffers from poor performance due to low data quality, missing information, limited scalability, lack of distinctiveness and other factors [20]. To overcome these issues, the concept of the multimodal biometric system has been introduced [21]. The multimodal biometric system strives to improve the recognition performance incorporating more than one biometric traits in a single system. The performance of a multimodal system highly depends on the fusion algorithm that combines the matchers’ information. Score level fusion and rank level fusion are quite popular in multimodal biometric researches [22]. In [23], authors used weighted score level fusion technique to fuse the matching scores from face and iris modalities. Authors proposed a finger multimodal biometric system combining finger vein, fingerprint, finger shape and finger knuckle print features in [24]. They used score level fusion approach based on triangular norm to fuse these modalities. In [25], a rank level fusion algorithm was used to combine the fingerprint and iris biometrics. A performance comparison was also presented between Borda count and Logistic regression rank level fusion algorithms. In [26], authors used rank level fusion to combine multiple palmprint representations. A comparative study on different rank level fusion algorithms were conducted in this research. Authors integrated face, iris and ear biometric features using Markov chain model based rank level fusion method in [27]. In [28], fingerprint and gait biometrics were fused using contourlet derivative weighted rank level fusion method.

Researchers combined SBB with other traditional biometrics using different types of fusion viz; face, iris, ear and signature. Authors linked social network relationships with geo-analytical behavioral for identifying unique behaviors in [29]. In [30], authors combined social network relationships with face, ear and signature modalities. In [31], SBB traits were fused with face and ear biometrics to enhance the performance of the conventional biometric systems. In [32], social behavioral data from five social networks were combined to continuously verify the users on mobile devices.

However, no research fused the users’ writing profiles generated from the aggregated tweets and replies to SBB along with other social behavioral profiles and networks. The number of studies comparing rank level and score level
fusion algorithms on multimodal biometrics systems are limited [33], [34]. No comparative analysis between rank level and score level fusion on SBB has been conducted. To the best of our knowledge, this is the first study of the users’ writing profiles integration to SBB, where the writing profiles are generated basing on the word importance aggregating tweets and replies. The comparative analysis of score level fusion and rank level fusion on SBB is also an untouched area.

III. METHODOLOGY

Social interaction varies from person to person in OSN. Twitter is one of the most popular social media platforms, accommodating millions of people. Regardless of the geographic locations, Twitter has an approximate of 554.7 million active users around the globe, who post 58 million tweets per day and 9100 tweets per second [35]. Tweets are restricted to 280 characters [36]. However, Twitter profiles and tweets are overloaded with information about the individuals and their social communication. We have used the following types of information in our proposed system:

1) Textual content based data is usually obtained from the tweets in terms of vocabulary, lexicons, abbreviations, capitalization, punctuation, structure of sentences, misspelled words, etc.
2) Contextual data is mined from the shared weblinks, hashtags, emoticons, check-ins, etc.
3) Interpersonal communication data reflects the social interactions between users, such as mention, reply, retweet, etc.

This information needs to be organized to generate distinguishable profiles for users. For example, a reply profile for a particular user can be constructed from the frequent connections in the user’s replies or mentions. Trending topic networks can be constructed using the hashtags that users share.

In this section, a new SBB trait is proposed to identify users based on their writing profiles on Twitter. This proposed trait is integrated with five other SBB traits based on shared posts, acquaintance networks, interest in trendy topics, temporal information and shared weblinks. Different types of fusion algorithms are used to combine the SBB traits and the performance of the system is observed. Fig. 1 demonstrates the design of the integrated system. At first, the writing profiles, trendy topic networks, reply networks, retweet networks, temporal profiles and shared weblink networks are generated for all users of the dataset. Then, a fusion algorithm is applied to fuse these social behavioral biometric traits. Finally, the fused scores are used to identify the same individuals from a different set of data. In this paper, we have applied different types of fusion algorithms on the proposed system to combine the six mentioned SBB traits and analyzed the performance of the system.

Previously, only the contextual traits were explored in SBB research. The vocabulary set generation from the aggregating tweets and replies, and integrating it to SBB is a new concept. The vocabulary set of a user will contain only those words, which contribute to the identification of that user. Fig. 2 shows the workflow diagram to generate users’ writing profiles from the aggregated tweets and replies. The raw tweets and replies need some linguistic pre-processing steps to remove the noisy data and tokenization. Then, the vocabulary sets are generated through a feature extraction algorithm. The extracted features are used to identify persons based on their writing profiles with the help of a machine learning based classification algorithm.

A. GENERATION OF WRITING PROFILES

The writing profile is generated for each user automatically, aggregating the recent collection of tweets and replies. This profile deals with the words that bear high importance for a particular user. The words used by different users vary to some extent.

1) NOISE REMOVAL

Punctuation and special characters are necessary for human understanding of tweets, but they are not useful for classification algorithms. After pre-processing, all punctuation marks, bad characters, nonASCII and emoji characters are removed from the data.

2) TOKENIZATION AND REMOVAL OF STOP WORDS

Normalization is required to get meaningful data. Therefore, the noise-free data is converted into lowercase alphabets and tokenized into tokens. The English stop-words are eliminated after tokenization. Articles, prepositions and other frequently occurring words that do not carry any importance are considered as stop-words. We keep the misspelled words, abbreviations and slang as we are working on texts written mostly in informal languages. The matrix of tokens is now grouped by the users and this is considered as the vocabulary sets for the users. Each user has a separate vocabulary set, which is denoted as $UV$.

3) LABEL ENCODING OF THE USERS

Twitter handles are unique identifiers for the users on Twitter. In the dataset, there are 250 handles to identify approximately 200,000 interactions. The machine learning models require the numeric value to perform an operation. Therefore, these handles are encoded with unique numeric values.

4) FEATURE EXTRACTION

The process of converting textual content into meaningful numerical representation is known as vectorization [37]. Many feature engineering techniques exist in literature to vectorize texts, such as Bag-of-Words, Word2Vec, FastText, Glove and Term Frequency-Inverse Document Frequency (TF-IDF) [38], [39]. Bag-of-Words, one of the popular vectorization techniques, provides information related to the presence of words in a corpus. It generates a vocabulary set using all words present in a set of documents. Another popular vectorization technique is TF-IDF, which assigns higher
value to the words of higher importance and gives lower value to the common words in a dataset. Both techniques generate vocabulary sets based on the words. However, TF-IDF identifies unique and less common words with high priority in the vocabulary set for a user and holds more information than Bag-of-Words. As we aim to represent the unique behaviors of the users, TF-IDF approach is more beneficial.

In our model, we use this technique to vectorize the users’ vocabulary set generated from the data.

If a user’s vocabulary set $U_V$ contains $T$ number of words, then the weight for the word $W_i$ will be calculated in two steps. Here, $i = 1, 2, 3, \ldots, T$. The first step computes $WT_1$, a ratio to estimate the occurrence of $W_i$ in $U_V$ using (1) and (2). The second step determines how rare $W_i$ is, compared to all users’ vocabulary sets using (3), which is denoted as $WT_2$. The final weight ($WT_1$) for $W_i$ is calculated by multiplying both weights $WT_1$ and $WT_2$ as shown in (4). In this way, the weights will be calculated for all words present in all users’ vocabulary sets.

\[
WT_1 = \frac{No \ of \ times \ W_i \ appears \ in \ U_V}{Total \ no \ of \ words \ in \ U_V}
\]

\[
WT_1 = \begin{cases} 
1 + \log_{10} t_{W_i} & \text{if } t_{W_i} > 0 \\
0 & \text{Otherwise}
\end{cases}
\]

**FIGURE 1.** Architecture of the proposed system to identify users using their social interactions on OSN.
FIGURE 2. Workflow of the proposed SBB trait to identify users using their writing profiles of OSN.

\[ WT_2 = \log_{10} \frac{\text{No of users}}{\text{No of UV in which } W_i \text{ appears}} \]  
\[ WT_i = WT_1 \times WT_2 \]  (3)  (4)

5) SELECTION OF CLASSIFICATION ALGORITHM
One of the most important parts of a classification problem is choosing the best classifier. Recent studies showed that Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) are well suited for multiclass problems based on textual data [40]. Naïve Bayes algorithms use the traditional technique for text classification, namely the generative model. The multiclass variation of Support Vector Machine is also widely used for text classification problems. Therefore, we have built a predictive model using these two classification algorithms. This model is trained with one session’s data and tested with another session’s data of \( N \) users. The outcome of this model is an \( N \times N \) matrix of predicted probabilities based on users’ writing profiles. Each test sample has \( N \) predicted probabilities for the same number of classes. These probabilities are used to make classification decisions, as illustrated in the next section.

B. REPLY AND RETWEET NETWORK CREATION
The reply network and retweet network are implemented similarly to [31]. The list of replied, mentioned and retweeted acquaintances are parsed from the dataset. The networks contain a set of nodes and weighted edges. The nodes of the reply network are created from the users and the list of acquaintances whom the users frequently reply to and mention in the comments. The edges are formed between the users based on the reply and mention relationship. The weights are calculated according to the occurrences of the relationship. Frequently replied and mentioned nodes gain higher weights. Sometimes, users repost their own tweets or tweets from someone else and share it with the followers. In the retweet network, reposting own tweets will not be considered. The network constructing process is similar to the reply network. The set of nodes is generated from the users and the acquaintances, whose tweet the users retweet often. The edges are constructed based on the retweeting relationship, and weights are determined based on the frequencies.

C. SHARED WEBLINKS NETWORK CREATION
A lot of users find it convenient to combine the weblinks with the tweets instead of writing the whole idea to share with the followers. The lists of shared weblinks represent the users’ preferred domains that contain the sharing pattern of the users. The shared URLs are parsed from the written contents of the dataset, preceded by “http://” or “https://”. The edges are constructed based on the sharing relationship. People share common websites, such as IEEE, YouTube, Facebook and Google often, which do not help to figure out the unique sharing pattern of the users. Therefore, the unusual websites get higher weights on the edges. There are four weblink networks assembled for four sessions of data. The shared weblinks network is implemented using the URL network algorithm [31].

D. TRENDY TOPIC NETWORK CREATION
Twitter users tend to use small words preceded by a hashtag (#) to categorize and connect their posts with the trends. These hashtag words can be considered as the main topic of the tweets. A trendy topic network or hashtag network can be generated from this hashtag relationship. The set of nodes contains the users and their shared hashtags, and edges between them are decided based on the hashtag relationship. Among the shared hashtag words, the popular ones do not help much to explore the unique behaviors of the users. Self-made hashtag words contribute more in this regard. Therefore, fewer weights are assigned on the edges that are connected to the popular and mostly used hashtag words. This network is implemented using the hashtag network algorithm of [31].
E. TEMPORAL PROFILE
The concept of generating temporal profile was first introduced in [15]. The temporal profile of users reveal their posting patterns in social network. The temporal profile of each user can be created by extracting the features from the timestamps of users’ profiles [16], such as average probability of tweeting per day, average probability of tweeting per hour, average probability of tweeting per week, seven days interval period, seven days tweeting period, average probabilities of original tweet, retweet, and reply/mention per day, etc.

IV. FUSION TECHNIQUES
The performance of a multimodal biometric system depends on an effective fusion technique that combines different modalities. In this research, we have selected the score level fusion and rank level fusion algorithms to combine different traits of SBB because these algorithms have performed well on multimodal biometric systems [27], [31].

Rank level fusion is performed after comparing the test sample with the templates based on the individual matcher’s score. For each testing sample and an individual matcher, every possible template in the database has an associated rank. Each template has a list of ranks provided by the $M$ matchers for a single testing sample, where $M$ is the number of matchers in the system. The rank level fusion algorithm is applied to the list of ranks to get the consolidated ranks for each sample. The lowest rank resembles the best match for the test sample with the corresponding template.

There are some popular rank level fusion algorithms, namely Weighted Borda Count and Modified Highest Rank Method [26], [27], [41]. Weighted Borda Count is a variation of Borda Count rank level fusion algorithm. This method calculates the sum of the ranks obtained from individual matchers multiplied by the associated weights using (5). Here, $R_{ic}$ is the consolidated rank for individual identity $i$, $M$ is the number of matchers and $W_m$ is the associated weight for a matcher $m$. The weights to the different matchers can be assigned by evaluating their recognition performance with a number of experimentation or using logistic regression method. Weighted Borda Count is useful when the matchers have significant differences in their recognition performances. We have selected this method because the individual recognition accuracy of the six SBB trait matchers are not uniform. The Highest rank method considers the highest rank returned by the matchers. For a testing sample, every possible identity in the database receives a list of $M$ ranks from the $M$ matchers. This algorithm computes the fused rank for every user template from the minimum of these $M$ ranks. The user identities are then sorted to get the combined rank list. One drawback of this method is sometimes there can be rank ties among user identities due to having the same combination of ranks. The chance of such ties is relatively less if the system deals with a large number of users and a small number of matchers. A small number of users and a large number of matchers increase the possibility of ties. As the proposed SBB system identifies a large number of users and the number of SBB trait matchers is six, we have selected this algorithm to experiment with our proposed system. We have used the modified version of highest rank method shown in (6) to avoid the occurrence of a tie. The ranks are determined based on the obtained decimal values.

$$R_{ic} = \sum_{m=1}^{M} R_{im} W_m$$  \hspace{1cm} (5)

$$R_{ic} = \min(R_{i1}, R_{i2}, \ldots, R_{iM}) + \frac{\sum_{m=1}^{M} R_{im}}{100}$$  \hspace{1cm} (6)

Score level fusion is carried out on the matching scores which are obtained from the comparison of two samples. The matching score can be of two types, the similarity score demonstrates how close the matched samples are, while the dissimilarity score shows the distance between the samples. For the score level fusion, an individual SBB trait matcher provides a vector of $N$ probabilities for every sample in the testing set, where $N$ is the total number of possible identities present in the database. The highest matching score is considered to be the best match with the sample. Hence, decision scores from all six SBB trait matchers are calculated for all samples. Each sample has a list of decision scores obtained from six matchers for every person in the database, which are combined using a score level fusion algorithm to get the final decision scores. The final decision scores are used to identify the best match with the sample.

Among all fusion algorithms, score level fusion algorithms, namely Median rule, Product rule and Sum rule are the frequently chosen because of its simplicity of implementation [22]. Median rule, Product rule and Sum rule algorithms combine the list of matching scores received from the SBB trait matchers using (7), (8) and (9), respectively. Here, $S_{ic}$ is the combined score and $S_{im}$ is the score received from individual $m$ matcher for each identity, for $i^{th}$ testing sample. These algorithms perform more effectively for the systems where the matchers’ accuracies are uniform. In our proposed system, the recognition performance of all SBB traits are not equal. Therefore, we have experimented the proposed system with the Weighted sum rule algorithm using (10) also, where $W_i$ is the corresponding weight for $m$ matcher [42]. In this algorithm, different weights are assigned to the SBB trait matchers to calculate the fused score. The weights are chosen using the Genetic Algorithm (GA) [43] to maximize the recognition rate. The initial population has been chosen randomly for the GA. To generate successive populations, single point crossover and two points uniform mutation operators have been used.

$$S_{ic} = \text{Median}(S_{im})$$ \hspace{1cm} (7)

$$S_{ic} = \prod_{m=1}^{N} S_{im}$$ \hspace{1cm} (8)

$$S_{ic} = \sum_{m=1}^{N} S_{im}$$ \hspace{1cm} (9)
TABLE 1. Performance of multinomial Naïve Bayes and support vector machine algorithms using the users’ writing profiles.

| Method                  | Train-test set | Accuracy | Precision | Recall | F-measure | Average of all train-test sets | Accuracy | Precision | Recall | F-measure |
|-------------------------|---------------|----------|-----------|--------|-----------|--------------------------------|----------|-----------|--------|-----------|
| Multinomial Naïve Bayes | $Set_{1,2}$  | 94.61%   | 92.05%    | 94.61% | 92.88%    | 91.70%                         | 88.37%   | 91.70%    |        | 89.38%    |
|                         | $Set_{1,3}$  | 90.63%   | 85.03%    | 89.63% | 86.47%    |                                 |          |           |        |           |
|                         | $Set_{1,4}$  | 85.48%   | 80.73%    | 85.48% | 82.07%    |                                 |          |           |        |           |
|                         | $Set_{2,3}$  | 92.95%   | 90.66%    | 92.95% | 91.27%    |                                 |          |           |        |           |
|                         | $Set_{2,4}$  | 92.95%   | 89.42%    | 92.95% | 90.59%    |                                 |          |           |        |           |
|                         | $Set_{3,4}$  | 94.61%   | 92.32%    | 94.61% | 93.02%    |                                 |          |           |        |           |
| Support Vector Machine  | $Set_{1,2}$  | 92.53%   | 89.76%    | 92.53% | 90.66%    | 91.42%                         | 88.28%   | 91.42%    |        | 89.32%    |
|                         | $Set_{1,3}$  | 92.12%   | 89.00%    | 92.12% | 90.04%    |                                 |          |           |        |           |
|                         | $Set_{1,4}$  | 86.31%   | 81.22%    | 86.31% | 82.88%    |                                 |          |           |        |           |
|                         | $Set_{2,3}$  | 94.19%   | 92.55%    | 94.19% | 93.08%    |                                 |          |           |        |           |
|                         | $Set_{2,4}$  | 91.29%   | 87.97%    | 91.29% | 89.07%    |                                 |          |           |        |           |
|                         | $Set_{3,4}$  | 92.12%   | 89.21%    | 92.12% | 90.18%    |                                 |          |           |        |           |

$$S_{ic} = \sum_{m=1}^{N} S_{im} W_m$$ (10)

V. EXPERIMENTAL RESULTS

The proposed system is implemented using Python 3.6. All experiments have been carried out on the Windows 10 operating system, 1.8 GHz Quad-Core Intel Core i5 processor with 8GB RAM. The dataset consists of 250 users, whose profiles are publicly available [15], [31]. The dataset contains four sessions of social interactions with the time interval of four to six weeks between each of the sessions for the same set of users. The interactions consist of tweets, mentions, replies, retweets, shared hashtags, shared URLs, timestamps, etc. The sessions are denoted as $S_i$, where $i$ is the session number in the range from 1 to 4. Each session $S_i$ contains data of 250 users, which are denoted as $j$ and $j = 1, 2, 3, \ldots, 250$. For each user $j$, there are approximately 200 social interactions. Accordingly, the data of user $j$ from $i^{th}$ session is denoted as $S_{ij}$.

The performance of the system is established by cross-validation [44]. We have generated six different combinations of training and testing data from the dataset: $Set_{i,j}$, where $i$ is training session and $j$ is testing session: $Set_{1,2}$, $Set_{1,3}$, $Set_{1,4}$, $Set_{2,3}$, $Set_{2,4}$, $Set_{3,4}$.

We have conducted six types of experiments to answer the research questions stated in Section I. The first two experiments are related to the proposed SBB trait users’ writing profiles and the other four experiments are conducted to evaluate the performance of the proposed multitrait SBB system.

A. EVALUATING THE PERFORMANCE OF PROPOSED WRITING PROFILES

The goal of this experiment is to establish the stand-alone recognition ability of the proposed users’ writing profiles. In this experiment, the performance of the proposed SBB trait, users’ writing profiles is cross-validated with six combinations of train-test sets. We have generated rank-1 accuracy to rank-10 accuracy for all train-test sets using both SVM and MultinomialNB classifiers. For SVM, we have used the polynomial kernel with degree 1 and the gamma value is optimized automatically by the algorithm. The average rank-1 identification rates obtained from MultinomialNB and SVM are 91.70% and 91.42%, respectively, which are very close. Table 1 shows the performance of the proposed trait in terms of accuracy, precision, recall and f-measure after applying both algorithms. The writing style of the users varies with time. Consequently, it can be observed the consecutive train-test sessions have achieved higher recognition rates. The highest rank-1 accuracy is 94.61%, achieved from the train-test set $Set_{1,2}$ and $Set_{3,4}$ using MultinomialNB classifier.

B. PERFORMANCE COMPARISON OF PROPOSED WRITING PROFILES WITH OTHER SBB TRAITS

This experiment aims to prove that the proposed SBB trait has the better discriminating ability than other SBB traits. We have re-implemented other SBB traits mentioned in literature, such as reply network, retweet network, hashtag network, URL network and temporal features, and compared with the identification ability of users’ writing profiles.

The reply network, retweet network, URL network, hashtag network and temporal profile for the 241 users are generated similarly to [15], [31]. Fig. 4 demonstrates the CMC curves of users’ writing profiles, hashtag network,
network have achieved 59.54%, 54.36%, and 49.52% recognition accuracy at rank-1, respectively. The lowest recognition rate of 21.16% at rank-1 is achieved from the temporal profiles of the users. This shows that many users have similar tweet posting patterns, which is not distinguishable enough to impact the identification system.

Table 2 shows the performance comparison of users’ writing profile with the re-implemented version of existing SBB traits in literature in terms of average rank-1 accuracy, precision, recall and f-measure after cross-validating with the same train-test sets, and Table 3 shows the performance comparison of users’ writing profiles with the reported in literature accuracy of the same SBB traits. Fig. 5 demonstrates the average number of misclassified samples at rank-1 to rank-10 for all SBB traits mentioned above. The proposed trait has better-discriminating capability and lower average number of misclassification at rank-1 than other SBB traits. Furthermore, it can be observed that the users’ writing profile tends to converge to the solution with no misclassification faster than other existing SBB traits. In light of the above discussion, this experiment answers the research question of integrating users’ writing profiles to SBB. The users’ writing profiles are distinct enough to contribute to the SBB systems.
The goal of this experiment is to demonstrate the performance of score level fusion algorithms on our proposed SBB system. In this experiment, the score level fusion algorithms, namely Median rule, Product rule, Sum rule and Weighted Sum rule are applied to the proposed system to fuse all six SBB traits, and the performance is cross-validated with the aforementioned six train-test sets. These algorithms have never been used in multitrait SBB systems. We have generated rank-1 accuracy to rank-10 accuracy for all train-test sets. The average ranked accuracies of six train-test datasets are considered as the final rank-1 to rank-10 accuracy. Table 4 presents the performance of the proposed model in terms of average accuracy, precision, recall and f-measure. The CMC curves of all train-test sets for the mentioned algorithms are demonstrated in Fig. 6. The behavior of the CMC curves is consistent as the accuracies change over ranks similarly for all train-test sets. The system achieved the highest rank-1 recognition rate of 99.45% when utilizing the GA based weighted sum rule score level fusion algorithm. Fig. 7 demonstrates that recognition rates for all train-test sets became stable within 200 generations. Hence, the weights are assigned by the genetic algorithm after executing the algorithm till 200 generations. The second highest rank-1 recognition rate of 97.72% is achieved using the Weighted sum rule algorithm with weights .52, .40, .03, .02, .02, .01 for users’ writing profiles, retweet, reply, trendy topic, shared weblinks networks and temporal profile, respectively. The rank-1 recognition accuracy of 94.88% is achieved from the Sum rule algorithm. The rank-1 accuracies of the proposed system after fusing with Product rule and Median rule algorithm are close to each other. The rank-1 accuracies are 89.90% and 89.70% for Product rule and Median rule, respectively. Though the precision of the Median rule algorithm is higher than the Product rule algorithm, the Product rule algorithm converges faster than the Median rule algorithm. The Product rule algorithm achieves the same rank-9 accuracy as the Sum rule algorithm. Weighted sum rule, Sum rule and Product rule algorithms performed well with our proposed model for all train-test sets.

### TABLE 3. Performance comparison of other SBB traits in literature with the proposed users’ writing profile.

| SBB Trait                | Accuracy  |
|--------------------------|-----------|
| Retweet Network [15]     | 89.00%    |
| URL Network [15]         | 45.00%    |
| Reply Network [15]       | 62.00%    |
| Hashtag Network [15]     | 52.00%    |
| Temporal Network [15]    | 13.00%    |
| Retweet Network [31]     | 90.00%    |
| URL Network [31]         | 53.00%    |
| Reply Network [31]       | 68.00%    |
| Hashtag Network [31]     | 61.00%    |
| Writing Profile (Proposed Trait) | 91.70% |
TABLE 4. Performance of the score level fusion algorithms with the proposed model.

| Algorithm                  | Accuracy | Precision | Recall | F-measure |
|----------------------------|----------|-----------|--------|-----------|
| Weighted sum rule + GA     | 99.45%   | 99.17%    | 99.45% | 99.26%    |
| Weighted sum rule          | 97.72%   | 96.58%    | 97.72% | 96.96%    |
| Sum rule                   | 94.88%   | 92.48%    | 94.88% | 93.26%    |
| Product rule               | 89.90%   | 85.52%    | 89.90% | 86.92%    |
| Median rule                | 89.70%   | 85.70%    | 89.70% | 86.94%    |

D. EXPERIMENTATION ON RANK LEVEL FUSION ALGORITHMS FOR SBB

This experiment intends to show the performance of rank level fusion algorithms on our proposed SBB system. In this experiment, the rank level fusion algorithms, namely Modified Highest Rank (MHR) and Weighted Borda Count (WBC) are applied to the proposed system to fuse six SBB traits, and the performance is cross-validated with the same six sets of train-test data. The average ranked accuracies of six train-test datasets are considered as the final rank-1 to rank-10 accuracy. Table 5 presents the performance of the proposed model in terms of average accuracy, precision, recall and f-measure for different rank level fusion algorithms. The CMC curves for all train-test sets after applying the mentioned rank level fusion algorithms are demonstrated in Fig. 8. For all train-test sets, the curves show a steady behavior. We have experimented on the Weighted Borda Count method with all possible weight combinations and reported five significant variations, which achieved the accuracy above 80%. The weight combination 0.54, 0.40, 0.03, 0.02 and 0.01 was used for writing, retweet, reply, shared weblinks and trendy topic network, respectively in WBC1, which achieved the highest rank-1 accuracy of 84.92% and rank-10 accuracy of 95.23%. This proves higher accurate traits should be getting higher weights to get a better performance in rank level fusion. The other rank level fusion algorithm, Modified Highest Rank (MHR), did not perform well on the proposed system. It can be concluded from this experiment, the rank level fusion performs better if majority trait accuracies are strong.

TABLE 5. Performance of the rank level fusion algorithms on the proposed model.

| Algorithm | Accuracy | Precision | Recall | F-measure |
|-----------|----------|-----------|--------|-----------|
| WBC1      | 84.92%   | 79.02%    | 84.92% | 80.79%    |
| WBC2      | 84.79%   | 78.75%    | 84.79% | 80.87%    |
| WBC3      | 84.72%   | 78.68%    | 84.72% | 80.90%    |
| WBC4      | 82.37%   | 75.59%    | 82.37% | 77.60%    |
| WBC5      | 80.36%   | 73.03%    | 80.36% | 75.16%    |
| MHR       | 73.51%   | 65.28%    | 73.51% | 67.49%    |

FIGURE 7. Number of generations vs recognition rates of the proposed system using genetic algorithm based weighted sum rule algorithm.

E. PERFORMANCE COMPARISON OF PROPOSED MULTITRAIT SBB SYSTEM WITH PRIOR RESEARCHES

From the previous experiments, it can be observed that the GA based weighted sum rule score level fusion algorithm achieved the highest accuracy among all other fusion algorithms. Therefore, we have used this algorithm to combine the matching scores obtained from writing profiles, reply networks, retweet networks, trendy topic networks, shared weblinks networks and temporal profiles. In this experiment, we have compared the performance of our multitrait SBB system with prior researches conducted on this dataset and Table 6 demonstrates this comparison. While Sultana et al. [15] reported 94% accuracy, after re-implementation we obtained 87.48% and 88.17% recognition rates excluding and including SBB trait temporal profile, respectively. Another recent research conducted on the same benchmark dataset, Rocha et al. [45] compared the performance of various classifiers varying a set of bag-of-words features. This bag-of-words model along with the Power Mean Support Vector Machine (PMSVM) classifier was re-implemented and cross-validated with the aforementioned six train-test sets. The re-implemented system obtained 68.74% accuracy on the aggregated tweets. The proposed system achieved the highest recognition rate of 99.45% at rank-1, surpassing all other previously developed systems.

F. CONVERGENCE OF SCORE AND RANK LEVEL FUSION ALGORITHMS FOR SBB

The goal of this experiment is to analyze the convergence of score level fusion and rank level fusion algorithms. We have experimented on the same six sets of train-test datasets for cross-validation as previous experiments. Usually, with the increment of ranks, the number of misclassified samples decreases. We have taken the average number of misclassified samples generated by all train-test datasets for each ranked accuracy. For example, WBC1 misclassified 23, 32, 34, 45, 38 and 46 samples at rank-1 accuracy for the respective train-test sets. Therefore, WBC1 on average misclassified 36 samples at rank-1. In this way, we have calculated the average number of misclassified samples for score level fusion and rank level fusion algorithms used in this research from rank-1 to rank-10. Fig. 9 and Fig. 10 demonstrate the average number of misclassified samples for score level fusion and rank level fusion algorithms up to rank-10,
respectively. The Weighted sum rule score level fusion algorithm has the least amount of misclassification at rank-1 compared to other algorithms. Initially, the Median rule score level fusion and Product rule score level fusion algorithms have misclassified similar number of samples at rank-1. However, the Product rule score level fusion algorithm reduced the misclassification drastically at rank-2 and performed close to the Weighted sum rule algorithm. The Product rule, Weighted sum rule and Sum rule score level fusion algorithms converged close to 100% accuracy within rank-5, while none of the rank level fusion algorithms achieved such high accuracy even within rank-10. All the rank level fusion algorithms...
This is the first study that investigates the impacts of score level fusion and rank level fusion algorithms on Social Behavioral Biometrics. The experimentation concludes that users’ writing profiles have better discriminating capability and lower misclassification rate than other SBB traits for person identification. Also, score level fusion algorithms are more suitable for SBB systems than rank level fusion algorithms. The convergence to the 100% recognition rate of score level fusion is faster than rank level fusion for SBB systems.

Score level fusion offers the best trade-off between the available information and easy implementation for multitrait SBB systems. On the other hand, rank level fusion works better on the multimodal systems that have higher accurate modalities. The integration of users’ writing profiles to SBB will result in better recognition performance. This research contributes to the emerging trend of investigating social behavior as a biometric signature. The proposed method can be used in continuous authentication, identity theft detection, author profiling, anomaly detection and other forensic applications on both Twitter and other OSNs. Future works include the introduction of new social behavioral features and fusing them with the existing biometric traits to improve the recognition performance of the biometric systems. The system performance can be investigated on a dataset of users with similar work profiles to explore their idiosyncratic behavior over time.

VI. CONCLUSION AND FUTURE WORKS

This research analyzes the individual performance of different SBB traits as well as integrates a new trait with the existing SBB traits to generate a higher accurate SBB system.

REFERENCES

[1] S. Chague, B. Droit, O. Boulanov, S. N. Yanushkevich, V. P. Shmerko, and A. Stoica, “Biometric-based decision support assistance in physical access control systems,” in Proc. Bio-Inspired, Learn. Intell. Syst. Secur., Aug. 2008, pp. 11–16.
[2] M. S. Gerber, “Predicting crime using Twitter and kernel density estimation,” Decis. Support Syst., vol. 61, pp. 115–125, May 2014.
[3] Y. Yang, “Web user behavioral profiling for user identification,” Decis. Support Syst., vol. 49, no. 3, pp. 261–271, Jun. 2010.
[4] G. Phillips-Wren, “Intelligent decision support systems,” in Multicriteria Decision Aid and Artificial Intelligence: Links, Theory and Applications. New York, NY, USA: Wiley, 2013, pp. 25–43.
[5] M. Hou, S. Banbury, and C. Burns, Intelligent Adaptive Systems: An Interaction-Centered Design Perspective. Boca Raton, FL, USA: CRC Press, 2014.
[6] Y. Wang and Y. Tian, “A formal knowledge retrieval system for cognitive computers and cognitive robotics,” Int. J. Softw. Sci. Comput. Intell., vol. 5, no. 2, pp. 37–57, Apr. 2013.
[7] M. M. Monwar, M. Gavrilova, and Y. Wang, “A novel fuzzy multimodal information fusion technology for human biometric traits identification,” in Proc. IEEE 10th Int. Conf. Cognit. Inform. Cognit. Comput. (ICCI-CC), Aug. 2011, pp. 112–119.
[8] K. Dautenhahn, “Socially intelligent robots: Dimensions of human–robot interaction,” Phil. Trans. Roy. Soc. B, Biol. Sci., vol. 362, no. 1480, pp. 679–704, Apr. 2007.
[9] D. Bell, T. Koulouri, S. Lauria, R. D. Macredie, and J. Sutton, “Microblogging as a mechanism for human–robot interaction,” Knowl.-Based Syst., vol. 69, pp. 64–77, Oct. 2014.
[10] M. Sultanova, P. P. Paul, and M. Gavrilova, “Social behavioral biometrics: An emerging trend,” Int. J. Pattern Recognit. Artif. Intell., vol. 29, no. 8, Dec. 2015, Art. no. 1556013.
[11] A. K. Jain, K. Nandakumar, and A. Ross, “50 years of biometric research: Accomplishments, challenges, and opportunities,” Pattern Recognit. Lett., vol. 79, pp. 80–105, Aug. 2016.
[12] A. Awad and Y. Liu, “Cognitive biometrics for user authentication,” in Biometric-Based Physical and Cybersecurity Systems. Cham, Switzerland: Springer, 2019, pp. 387–399.
