SemUserProfiling: A Hybrid Knowledge Centric Approach for Semantically Driven User Profiling

Rituraj Ojha and Gerard Deepak

Abstract Information is shared by a large number of people around the globe in the form of chats, posts, blogs, tweets, and news. For information to reach the right audience and for companies to target the right people, user profiling and term profiling are much needed. In this paper, SemUserProfiling which is an entity enrichment and structural topic modeling (STM)-based approach is proposed. The proposed approach uses the Twitter dataset, and user profile *Up* as an input. The Twitter dataset is preprocessed, and scenario is retrieved for each term in the preprocessed dataset. Furthermore, the entity enrichment takes place using linked open data (LOD) cloud and classification of the entity set happens using eXtreme gradient boosting (XGBoost) algorithm using categorical domain ontologies. Similarly, user profile *Up* is preprocessed and subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and normalized Google distance (NGD) under frog leap algorithm. The proposed SemUserProfiling yields a high accuracy of 99.02% and a negligible false discovery rate of 0.011.

Keywords Entity enrichment · Scenario retrieval · Topic modeling · User profiling

1 Introduction

Social media is a digital platform that helps to create and share information, interests, ideas, and to develop new social connections with other Internet profiles. Users generally access social media sites through desktop apps, mobile apps, or Wweb-based apps. Few famous social media platforms include Twitter, Facebook, Quora,

R. Ojha
Department of Metallurgical and Materials Engineering, National Institute of Technology, Tiruchirappalli, India

G. Deepak (✉)
Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_29

Instagram, Snapchat, Reddit, and LinkedIn. Professionals use these platforms to share knowledge, express ideas, and resolve doubts.

On social media, users usually form groups to share and receive knowledge of their interests. The number of groups increase as the number of users with different interests increase on the platform. The user posts his ideas in the group, which has its id and timestamp. The platform provides features such as likes and comments on every post to encourage and provide feedback. The user's post reach depends on the relevance of the post with the people reading, and thus, categorizing the user is essential for social media to recommend the right post to the right audience.

Motivation: There may be times during gathering of users when discussions or debates start on different range of topics. Therefore, profiling becomes vital to attract similar audiences and encourage knowledge. There is no framework which achieves profiling in social media with high accuracy. Also, in the era of semantic Web, there is a need for a better profiling algorithm which is semantic driven, and thus, building this user profiling approach was a major motivating factor.

Contribution: The SemUserProfiling based on entity enrichment and STM for profiling users is proposed. The Twitter dataset drives the proposed approach. The dataset is preprocessed, and scenario is retrieved for each term. Moreover, the entity enrichment takes place using the LOD cloud and classification of the entity set happens using XGBoost algorithm using categorical domain ontologies. Similarly, user profile *Up* is preprocessed and subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and NGD under frog leap algorithm. The values of metrics like precision, accuracy, recall, and *F*-measure of the proposed framework are increased.

Organization: The flow of the remaining paper is as follows. A condensed summary of the related works is provided in Sect. 2. Section 3 depicts the architecture of the proposed system. Section 4 describes the architecture implementation. Section 5 presents the evaluation of results and performance. The conclusion of the paper is presented in Sect. 6.

2 Related Works

Several people have done research in the area of profiling users. López-Monroy et al. [1] have proposed a methodology for profiling documents. The proposed approach captures discriminative and sub-profile information of terms, and then, these representations are accumulated to represent the content of the document. Van Dam and Van De Velden [2] have proposed a framework to learn and understand the user's data shared on social media platform Facebook. The proposed approach helps individuals to find the profiles of the target users based on their interest toward a particular field.

Chen et al. [3] have proposed a technique for estimating the location of users present on Twitter social media platform. The proposed approach can find location up to city level using the user's network, data shared by them, and tie strength. Greco and Polli [4] have proposed a methodology to analyze the textual data available on social media platforms and identify the sentiments and opinion of users on the particular topic. This paper proposes a better tool focused mainly on brand management.

Mishra et al. [5] have proposed an approach for profiling users for detection of abusive and hateful comments on social media platforms. The proposed technique uses community-based profiling characteristics for the social media users. Mishra et al. [6] have proposed SNAP-BATNET, a deep learning framework to find the suicidal tendency of the user. The proposed approach uses the data in the form of social graph embeddings. It also profiles users based on features from their previous data.

Wisniewski et al. [7] have proposed a methodology to suggest Facebook users with a set of distinct privacy options based on the user profiling. The proposed approach uses advanced factor analysis methods to present several privacy management strategies. Kosmajac and Keselj [8] have proposed a technique for bot and gender identification for Twitter using feature extraction, user behaviors fingerprints, syntactic information, and transformation methods.

Singh et al. [9] have proposed a user behavior profiling approach to detect threats present within an organization. The proposed approach uses an ensemble hybrid machine learning model. This machine learning model uses multistate long short-term memory and CNN. Chen et al. [10] have proposed a methodology for classifying the sentiments of a user in a document for a particular product. This is done using their proposed hierarchical neural network.

Guimaraes et al. [11] have proposed a technique for determining characteristics and age of a person by analyzing user profile and historical data. The proposed techniques use several approaches for this problem and the deep convolutional neural network archives the best accuracy. Menini et al. [12] have proposed a system for identifying cyberbullying on social media platforms like Instagram and Twitter by combining classification and social network analysis techniques. Papers [13–19] have proposed several approaches based on ontologies and knowledge bases.

3 Proposed System Architecture

The architecture for user profiling is depicted in Fig. 1. The socially aware user profiling takes place in several steps. Initially, the Twitter dataset (tweet dataset) is preprocessed using tokenization, lemmatization, stop word removal, and named entity recognition (NER). Tokenization is a process of splitting the texts in the dataset into pieces called tokens. Byte pair encoding (BPE) is used for the tokenization process. Lemmatization involves grouping together several inflected kinds of the same word so that they can be analyzed as a single term. During stop word removal, the common ubiquitous words are removed as they add no value for the analysis and

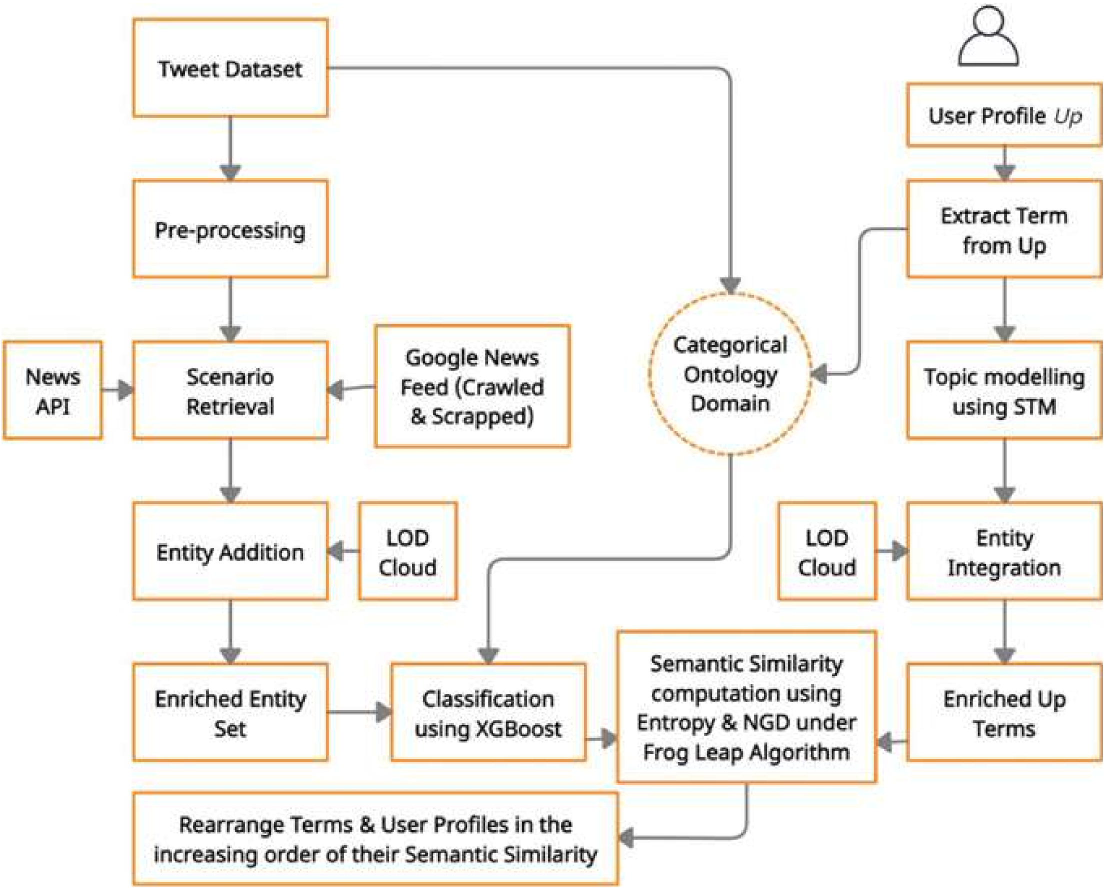
Tweet Dataset
Pre-processing
News API
Scenario Retrieval
Google News Feed (Crawled & Scrapped)
Entity Addition
LOD Cloud
Enriched Entity Set
Classification using XGBoost
Categorical Ontology Domain
User Profile Up
Extract Term from Up
Topic modelling using STM
LOD Cloud
Entity Integration
Enriched Up Terms
Semantic Similarity computation using Entropy & NGD under Frog Leap Algorithm
Rearrange Terms & User Profiles in the increasing order of their Semantic Similarity

Fig. 1 Proposed system architecture

only increase the dimension of the feature set. NER is the process of finding and categorizing the data or entity into predefined categories.

The next step involves retrieving scenarios for the preprocessed tweets. The individual terms from the preprocessed tweets are taken and for these individual terms, the scenario is obtained by crawling and scraping Google news feed and other news APIs, namely mediastack, news API, and Webhose. Furthermore, real world similar entities are also added from the LOD cloud. After these processes, we obtain the enriched entity set. The next step involves classifying the enriched entity set using categorical domain ontologies using the XGBoost algorithm. XGBoost is an ensemble machine learning model that is based on the decision tree. Features of XGBoost include handling of missing values automatically, supports parallel processing, tree pruning, and regularized boosting to prevent overfitting. Equation (1) presents how gradient tree boosting works.

$$F_m(X) = F_{m-1}(X) + \propto_m h_m(X, r_{m-1}) \tag{1}$$

where $\propto_i$ and r_i represent the regularization parameters and residuals calculated with the ith tree. The function for predicting residuals, r_i using X for the ith tree, is represented by h_i. To calculate the $\propto_i$, we use r_i and calculate the following: arg arg =

$\sum_{i=1}^{m}\left[L(Y_i, F_{i-1}(X_i)+ \propto h_i(X_i, r_{i-1}))\right]$ where $L(Y, F(X))$ is a differentiable loss function.

The user profile *Up* is preprocessed, and the terms are extracted from it. We generate the categorical domain ontologies between the tweet dataset and the extracted terms from user profile, and these categorical domain ontologies are used for the classification process using XGBoost algorithm. Furthermore, STM is applied using document corpus and extracted user profile data. The document corpus is crawled using beautiful soup API. This helps our model to get more topics that are similar to the user profile terms. Topic modeling is a technique to discover different topics present in the document corpus and get information about the hidden patterns exhibited by the document corpus. The STM is used for topic modeling that accommodates corpus structure through covariates at document-level. This model consolidates and expands three models: the Dirichlet-multinomial regression topic model, the correlated topic model, and the sparse additive generative topic model [20].

Furthermore, the user profile data are enriched with the linked data from LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and normalized Google distance under the frog leap algorithm, which is the preferred optimization metaheuristics. Entropy is represented by Eq. (2) and it represents the uncertainty or surprise that a variable can output. NGD between search terms x and y is represented by Eq. (3) and it is a semantic similarity measure. The calculation is derived by measuring Google returning the number of hits for a set of keywords. The NGD produces results such that keywords with similar meaning tend to be close as compared to keywords with dissimilar meaning, which are far apart. The frog leap algorithm is an optimization algorithm based on the actions observed in a family of frogs when they search for the place that has the highest quantity of available food. Their population consists of the group of frogs that are further divided into subsets called memeplexes. Each of the memeplexes perform a local search, and the ideas are later passed between them during shuffling. The shuffling process and the local search are continued until a defined convergence criterion is fulfilled [21].

$$\text{Entropy}(P) = -\sum_{i=1}^{N} P_i \log_2 P_i \tag{2}$$

$$\text{NGD}(x, y) = \frac{\max\{\log\log f(x), \log\log f(y)\} - \log f(x, y)}{\log\log N - \min\{\log\log f(x), \log\log f(y)\}} \tag{3}$$

Finally, after all these steps, the terms and user profiles are rearranged. The rearrangement takes place in the increasing order of their semantic similarity.

4 Implementation

The algorithm for the proposed approach is depicted in Algorithm 1, which takes Twitter dataset and user profile as input. The Twitter dataset is preprocessed using tokenization, lemmatization, stop word removal, and NER. The preprocessing is done using several Python libraries. The scenario is retrieved for each term using Google news feed API and several other news APIs, namely mediastack, news API, and Webhose. Moreover, the entity enrichment takes place using LOD cloud and classification of entity sets happen using XGBoost algorithm using categorical domain ontologies. Similarly, user profile *Up* is preprocessed and is subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Eventually, the semantic similarity is calculated between the enriched user profile terms and the classified enriched entity set using entropy and NGD under frog leap algorithm, to obtain the rearranged user profiles and terms in increasing order of the calculated semantic similarity. The proposed SemUserProfiling was successfully implemented and evaluated using Windows 10 operating system, equipped with 8th generation Intel Core i5 and 16 GB RAM, in Google collaboratory environment with Nvidia graphics card.

Algorithm 1: Proposed SemUserProfiling Algorithm

```
Input:     Twitter Dataset & User Profile Up
Output:    Rearranged User profiles & Terms
begin
Step 1:    Twitter dataset D is subjected to preprocessing. D is tokenized,
           lemmatized, and stop word removal and NER is performed on it.
Step 2:    while (D.next()!=NULL)
           D set <- scenario is retrieved using Google News Feed API and News
           API
           end while
Step 3:    while (D.next()!=NULL)
           D set <- relevant entities using LOD Cloud
           end while
Step 4:    User profile Up is pre-processed and the terms are extracted from it
Step 5:    Categorical Domain Ontologies generated using Up terms and Twitter
           dataset
Step 6:    6.1: STM (document corpus, user profile data Up)
           6.2: Up set.append(terms relevant to Up)
Step 7:    while (Up.next()!=NULL)
           Up set <- relevant entities using LOD cloud
           end while
Step 8:    Classification of D set using XGBoost algorithm (using Categorical
           Domain Ontologies)
Step 9:    SemanticSimilarity(D set, Up set)
Step 10:   Rearranged User profiles & Terms in increasing order of the calculated
           semantic similarity.
end
```

5 Results and Performance Evaluation

The performance of the proposed SemUserProfiling is measured by considering precision, recall, and accuracy. Other measures including false discovery rate, F-measure, and normalized discounted cumulative gain are also used. The performance is evaluated for 6129 queries, and the ground truth has been collected.

$$\text{Precision} = \frac{\text{Retrieved} \cap \text{Relevant}}{\text{Retrieved}} \tag{4}$$

$$\text{Recall} = \frac{\text{Retrieved} \cap \text{Relevant}}{\text{Relevant}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{Proportion Corrects of each query passed ground truth test}}{\text{Total number of queries}} \tag{6}$$

$$F\text{-Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{7}$$

$$\text{False Discovery Rate} = 1 - \text{Positive Predictive Value} \tag{8}$$

$$\text{nDCG} = \frac{\text{DCG}_{\propto}}{\text{IDCG}_{\propto}} \tag{9}$$

$$\text{DCG} = \sum_{i=1}^{\propto} \frac{\text{rel}_i}{\log_2(i+1)} \tag{10}$$

Equations (4–6) represent precision, recall, and accuracy, respectively. Furthermore, Eqs. (7–10) represent F-measure, false discovery rate (FDR), normalized discounted cumulative gain (nDCG), and discounted cumulative gain, respectively. The reason for considering accuracy, precision, recall, and F-measure metrics for evaluation is because they measure the relevance of the results, and FDR computes the number of false discoveries made by the proposed system. The nDCG represents the diversity of the relevant results.

Table 1 presents the performance comparison of the proposed SemUserProfiling model with Wisely et al. [22], PSMU [23], and other similar baseline models. It is evident from the table that the proposed approach achieves the highest average accuracy with the precision of 98.87%, recall of 99.87%, accuracy of 99.02%, F-measure of 99.36%, and nDCG value of 0.97. Also, the FDR is least with the value of 0.011. Wisely et al. achieves a low precision of 84.89%, recall of 92.32%, accuracy of 87.32%, F-measure of 87.52%, high FDR of 0.151, and a low nDCG value of 0.81. Furthermore, PSMU achieves a low precision of 95.32%, recall of 98.31%, accuracy of 96.71%, F-measure of 96.79%, high FDR of 0.047, and low nDCG value of 0.84.

Figure 2 represents the precision % versus number of recommendations for

Table 1 Performance comparison of the proposed SemUserProfiling with other approaches

Search technique	Average precision %	Average recall %	Accuracy %	*F*-measure %	FDR	nDCG
Wisely et al. [22]	84.89	90.32	87.32	87.52	0.151	0.81
PSMU [23]	95.32	98.31	96.71	96.79	0.047	0.84
Fuzzy C-means clustering + LDA	90.12	93.33	92.24	91.69	0.099	0.90
content-based filtering using max entropy classifier	87.32	91.14	88.71	89.19	0.127	0.86
Proposed SemUserProfiling	98.87	99.87	99.02	99.36	0.011	0.97

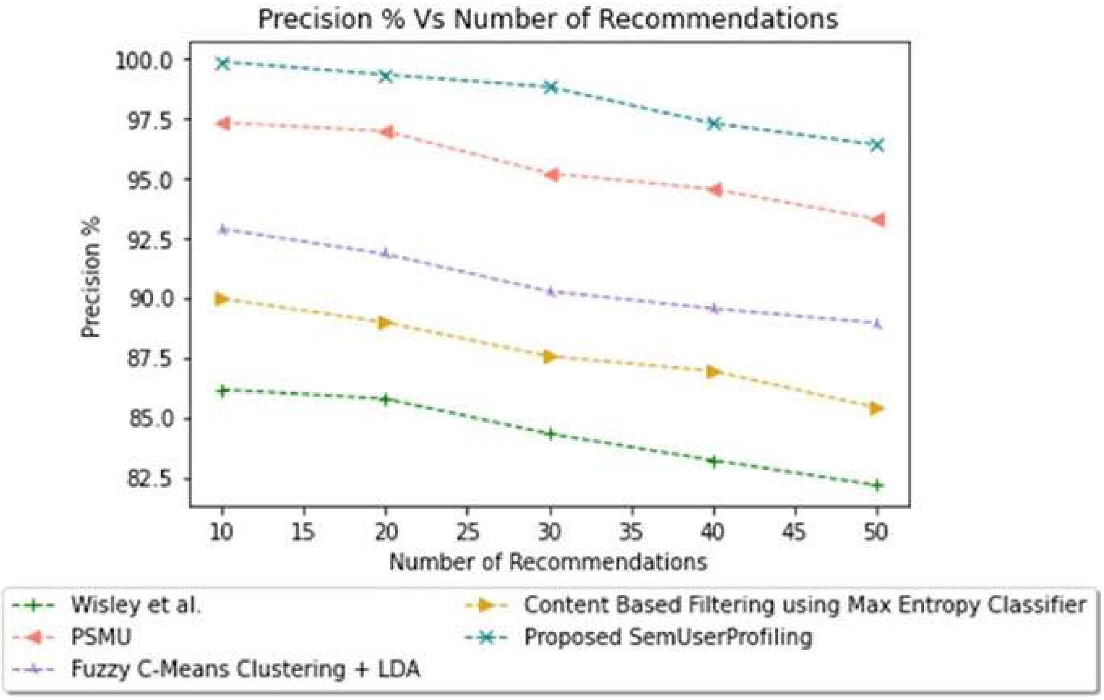

Fig. 2 Precision % versus number of recommendations

each model. The proposed SemUserProfiling approach is better than other baseline approaches due to several reasons. Firstly, it is a classification-based approach. The precision, recall, accuracy, and *F*-measure are high mainly because we are enriching the entities. Entity enrichment happens using news API and Google news feed API along with LOD cloud. Furthermore, the entity classification takes place using XGBoost algorithm and the semantic similarity is computed using entropy and NGD under the frog leap algorithm. Also, the user profile is subjected to topic modeling and entity integration. These are the main reasons for better performance of the proposed approach. Higher nDCG value means higher diversity in results. The reason for the high nDCG value of the proposed approach is due to topic modeling using STM, entity integration using LOD cloud, and scenario retrieval using Google news feed API and other news APIs.

Content-based filtering using max entropy classifier approach is not as good as the proposed approach because it does not have any entity integration approach, and the relevant entities are not added. Also, there is no scenario retrieval. Fuzzy C-means clustering + LDA used the topic modeling approach but the fuzzy C-means clustering alone is not sufficient to reach such high accuracy as the proposed approach. In the PSMU [23] approach, the author is using the NER technique to cluster keywords which is later used to cluster the users. There is no process to further enrich the entity set, and also there is no procedure for scenario retrieval for latest information from several news APIs. Similarly, the Wisely et al. [22] approach uses the user's browsing history to crawl the content with no entity enrichment or scenario retrieval techniques. Hence, the baseline models are not as good as the proposed approach.

6 Conclusions

In the world where social media plays an important role in information sharing, user profiling is much needed. SemUserProfiling has been proposed where a user profiling model which is a semantically driven approach and is based on entity enrichment and STM. The Twitter dataset is preprocessed, and scenario is retrieved for each term using Google news feed API and several other news APIs. Moreover, the entity enrichment takes place using the LOD cloud and classification of the entity set happens using XGBoost algorithm using categorical domain ontologies. Similarly, user profile *Up* is preprocessed and subjected to topic modeling using STM, and entity integration is achieved using LOD cloud. Eventually, the semantic similarity is calculated between the enriched user profile terms and the classified enriched entity set using entropy and NGD under frog leap algorithm, to obtain the rearranged user profiles and terms in increasing order of the semantic similarity. The proposed approach achieves the precision of 98.87%, recall of 99.87%, accuracy of 99.02%, *F*-measure of 99.36%, false discovery rate of 0.011, and nDCG of 0.97. The overall accuracy of the proposed approach is much better than the existing approaches.

References

1. A.P. López-Monroy, M. Montes-y-Gómez, H.J. Escalante, L. Villasenor-Pineda, E. Stamatatos, Discriminative subprofile-specific representations for author profiling in social media. Knowl. Based Syst. **89**, 134–147 (2015)
2. J.W. Van Dam, M. Van De Velden, Online profiling and clustering of Facebook users. Decis. Support Syst. **70**, 60–72 (2015)
3. J. Chen, Y. Liu, M. Zou, Home location profiling for users in social media. Inf. Manag. **53**(1), 135–143 (2016)

4. F. Greco, A. Polli, Emotional text mining: customer profiling in brand management. Int. J. Inf. Manag. **51**, 101934 (2020)
5. P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, in *Author Profiling for Abuse Detection.* Proceedings of the 27th International Conference on Computational Linguistics (2018 August), pp. 1088–1098
6. R. Mishra, P.P. Sinha, R. Sawhney, D. Mahata, P. Mathur, R.R. Shah, in *SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (2019 June), pp. 147–156
7. P.J. Wisniewski, B.P. Knijnenburg, H.R. Lipford, Making privacy personal: profiling social network users to inform privacy education and nudging. Int. J. Hum Comput Stud. **98**, 95–108 (2017)
8. D. Kosmajac, V. Keselj, in *Twitter User Profiling: Bot and Gender Identification.* International Conference of the Cross-Language Evaluation Forum for European Languages (Springer, Cham, 2020 September), pp. 141–153
9. M. Singh, B. M. Mehtre, S. Sangeetha, in *User Behavior Profiling Using Ensemble Approach for Insider Threat Detection.* 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA) (IEEE, 2019 January), pp. 1–8
10. H. Chen, M. Sun, C. Tu, Y. Lin, Z. Liu, in *Neural Sentiment Classification with User and Product Attention.* Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016 November), pp. 1650–1659
11. R.G. Guimaraes, R.L. Rosa, D. De Gaetano, D.Z. Rodriguez, G. Bressan, Age groups classification in social network using deep learning. IEEE Access **5**, 10805–10816 (2017)
12. S. Menini, G. Moretti, M. Corazza, E. Cabrio, S. Tonelli, S. Villata, in *A System to Monitor Cyberbullying Based on Message Classification and Social Network Analysis.* Proceedings of the Third Workshop on Abusive Language Online (2019 August), pp. 105–110
13. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. Procedia Comput. Sci. **167**, 737–746 (2020)
14. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. Multimedia Tools Appl. 1–25 (2021)
15. K. Vishal, G. Deepak, A. Santhanavijayan, in *An Approach for Retrieval of Text Documents by Hybridizing Structural Topic Modeling and Pointwise Mutual Information.* Innovations in Electrical and Electronic Engineering (Springer, Singapore, 2021), pp. 969–977
16. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. J. Discrete Math. Sci. Crypt. **23**(1), 157–165 (2020)
17. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, *OntoQuest: An Ontological Strategy for Automatic Question Generation for e-Assessment Using Static and Dynamic Knowledge.* 2019 Fifteenth International Conference on Information Processing (ICINPRO) (IEEE, 2019 December), pp. 1–6
18. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. Comput. Commun. **160**, 284–298 (2020)
19. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization.* European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
20. M.E. Roberts, B.M. Stewart, D. Tingley, E.M. Airoldi, in *The Structural Topic Model and Applied Social Science.* Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, vol. 4 (2013 December), pp. 1–20
21. M. Eusuff, K. Lansey, F. Pasha, Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. Eng. Optim. **38**(2), 129–154 (2006)

22. W.L. Dennis, A. Erwin, M. Galinium, in *Data Mining Approach for User Profile Generation on Advertisement Serving*. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE) (IEEE, 2016 October), pp. 1–6
23. G.U. Vasanthakumar, D.R. Shashikumar, L. Suresh, in *Profiling Social Media Users, a Content-Based Data Mining Technique for Twitter Users*. 2019 1st International Conference on Advances in Information Technology (ICAIT) (2019 July), pp. 33–38