THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2022 Vol. 22 No. 1 (ISSN 1727-5997)

Research Articles

Selected PhD Thesis Abstracts

Announcements

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee
2) Feature Articles
3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/~ley/db/index.html).*

A Survey on Clinical Time Series Forecasting Methods

Simi Job
University of Southern Queensland, Australia

***Abstract*—Clinical time series forecasting is gaining research interest in recent times owing to its applicability in clinical decision support and personalised patient treatment. While conventional time series approaches such as ARIMA are regularly used for processing temporal data, deep learning methods are increasingly being adopted due to its ability in handling non linear data. Advances in the usage of temporal information has also led to the expansion of research in multivariate time series analysis. This work identifies the various methodologies adopted in time series analysis and the future development potential for researching temporal data.**

***Index Terms*—ime series Multivariate Recurrent Neural Networks Convolutional Neural Networks.ime series Multivariate Recurrent Neural Networks Convolutional Neural Networks.T**

I. Introduction

The application of artificial intelligence systems in personalised healthcare is increasingly being researched to develop enhanced clinical decision support systems. The integration of various clinical data obtained from medical records, patient sensor devices, lab and diagnostic reports provide a pathway for developing a predictive and diagnostic system for personalised patient treatment. Efficiently harnessing the volumes of data available from these resources impart a well developed AI-enabled system for clinicians to act as a supplementary system for supporting personalised medicine.

Temporal data is used in various fields of study including Finance (D. Cheng, Yang, Xiang, & Liu, 2022), Power usage (Gasparin, Lukovic, & Alippi, 2019), Health (Yee, Narain, Akmaev, & Vemulapalli, 2019), (Staffini, Svensson, Chung, & Svensson, 2021), (Albers et al., 2018) and Climate (Afrifa-Yamoah, Mueller, Taylor, & Fisher, 2020). The information obtained from this data is useful for time series analysis, particularly in forecasting or classification tasks. Time series analysis of clinical data involves risk prediction of disease, forecasting of physiological values, etc, on the basis of past clinical values. Analysing temporal data provides insights into underlying trends and patterns in data over specific time periods.

Intensive care units are sources of high volumes of patient monitoring data. Utilisation of this information is minimal since it is physically impossible for healthcare workers to monitor and manually analyse variations of multiple physiologic variables. The substantial advances in artificial intelligence pave the way to develop intelligent systems that can assimilate ICU monitoring data along with lab data and other significant clinical values to provide forward insights for early interventional measures as well as forecasting resource demands. The availability of ICU data as temporal observations enable forecasting of physiological values based on historic data. These future values can then be assimilated to determine indicators of organ dysfunction or deterioration in terms of respiratory failure, cardiac distress, liver failure or progress towards septic shock. Several studies have utilised temporal ICU data for predicting sepsis shock (Yee et al., 2019), (Thao, Tra, Son, & Wada, 2018), cardiac stress (Yoon et al., 2019), survival (Thao et al., 2018), forecasting various physiological measures such as heart rate (Staffini et al., 2021), (Oyeleye, Chen, Titarenko, & Antoniou, 2022), blood pressure (E. Huang, Wang, Chandrasekaran, & Yu, 2020), blood glucose levels (Albers et al., 2018) or forecasting multiple physiologic values (Hamidi, Borzu, Maroufizadeh, & Amini, 2021).

The review is consolidated after researching relevant journals in the past five years. Articles before the year 2017 are excluded from this review, and therefore detailed information on conventional forecasting methodologies is not provided. Around three hundred articles were studied and the primary journals of search includes IEEE, Elsevier, ACM etc. From these research articles, 100 plus articles related to the topic were selected and reviewed.

This work seeks to study the most relevant models that are recently being used for time series modelling that are appropriate for the clinical domain. In addition to determining the recent approaches to time series analysis, the work also proposes to identify the various challenges associated with clinical time series forecasting. The review is expected to provide researchers with a direction towards methodologies adopted for clinical temporal analysis. The paper is organised as follows: Section II outlines the definition of clinical time series. Section III discusses the various methodologies in time series forecasting. Section IV presents the various challenges and issues that are encountered in time series forecasting. Section V discusses the trends and potential research directions. The survey is summarized in Section VI with Conclusion.

II. Time Series Analysis of Clinical Data

Researchers have used ICU data for time series classification (Karim, Majumdar, Darabi, & Harford, 2019) or forecasting (Staffini et al., 2021), (Oyeleye et al., 2022), (E. Huang et al., 2020), (Albers et al., 2018), (Hamidi et al., 2021). For instance, temporal analysis of various physiologic parameters such as blood pressure and lab data can be used to classify whether a patient is diabetic or not. Similarly, this data is also useful for forecasting values for adopting interventional measures.

Most of these studies are based on time series analysis with a univariate approach (Staffini et al., 2021), (Oyeleye et al., 2022), (E. Huang et al., 2020), (Albers et al., 2018), with some of the recent research focusing on multivariate time series (Hamidi et al., 2021). In multivariate time series analysis, a value is predicted based on the values of multiple variables. For this purpose, it is essential to identify correlation and dependencies among these variables.

Fig. 1 shows a representation of the process involved in forecasting multiple physiological values with reference to ICU data. Multiple variables or multiple time series are used for forecasting the variables. Training data is extracted from a set of data for an initial time period. Feature engineering is performed on this data and passed through the neural networks for learning the time series patterns. This learned information is tested on a latter part of time series data followed by forecasting for a future predefined period of time.

1) Problem Definitions: Time Series is a sequential set of observations measured on time points which may or may not be uniformly spaced. Univariate time series is the simplest form of time series and considers only one time-dependent variable. Multivariate time series analysis involves investigating inter-dependencies between multiple variables in a time series or between multiple related time series. It is considered to be more powerful in prediction due to the incorporation of relationship between different variables in the time series. Let $y_t = (y_{1t}, y_{2t}, \dots y_{nt})$ be an n dimensional ICU time series. If y_{1t} and y_{2t} are the heart rate and blood pressure of a patient recorded at time point t respectively, the temporal dependence between these two variables can be analysed for the purpose of predicting the future heart rate value of the patient. The different variables in the time series represent patient information such as physiological, lab data or interventional measures captured at a specific time point. With a focus on accommodating to the interactivity between variables in a multivariate scenario, discretization of time series is often adopted (J. M. Lee & Hauskrecht, 2021). This involves converting the numbers in a time series into a group of discrete elements.

The focus of this review is on multivariate time series, wherein multiple physiological parameters such as blood pressure, heart rate etc. may be forecast based on past values. Multivariate time series analysis involves detecting periodic trends, identifying dependencies between variables or multiple time series, and anomaly detection. For two variables y_1 and y_2, forecasting for time t can be performed based on past n values. Computation of $y_1(t), y_2(t), y_n(t)$ considers the historic values of both y_1 and y_2. This can be represented as shown in Eqn. 1 and Eqn. 2. C_1 and C_2 are constants. z_1 is a coefficient and e_1, e_2 represent error factors.

$$y_1(t) = C_1 + z_1 11 * y_1(t-1) + z_1 12 * y_2(t-1) + e_1(t-1) \quad (1)$$

$$y_2(t) = C_2 + z_1 21 * y_1(t-1) + z_1 22 * y_2(t-1) + e_2(t-1) \quad (2)$$

2) Motivation: The primary motive for ICU temporal analysis is to establish a personalised treatment process for the patient. A large section of research in clinical time series forecasting has focused on univariate analysis. However, to obtain a reliable forecasting outcome it is imperative to incorporate multiple variables into the time series analysis process. As seen in Fig. 2, patient data includes ICU monitoring data, lab test results, diagnostic reports, medications administered and so on. While it is possible to forecast blood pressure values based only on historic blood pressure data, for developing a definitive personalised treatment, it is essential to include other data such as heart rate, glucose levels etc. (Seid et al., 2019) conducted a study to determine neonatal mortality risks with both univariate and multivariate approaches. While the univariate approach revealed that sepsis and hypothermia were not leading causes for mortality, multivariate analysis indicated these two factors were mortality predictors. Fig. 2 depicts the relevance of ICU and clinical data time series forecasting in personalised medicine. Significant information from ICU monitoring data and patient medical records are utilised for forecasting physiological values or probable lab results. These predicted values are availed to provide personalised treatment for a patient, such as adopting early intervention measures.

The review encompasses the univariate and multivariate approaches adopted in clinical temporal analysis and utilises the analytical results for developing a prediction system. The predicted values are used methodically by clinicians to assist in reliable decision making aided by AI enabled systems.

III. Methodologies in Time Series Forecasting

Time series forecasting has been conventionally performed with univariate statistical models such as ARMA, ARIMA or seasonal variations of these models. For multivariate time series, ARIMAX, VARMA etc. are used. However, statistical models are incapable of adapting to non-linearity. As a solution, deep learning methods are being increasingly used for multivariate time series analyses owing to their adeptness in handling non-linearity. This section discusses the various time series pre-processing methods and linear and non-linear methods used in time series analysis.

1) Time Series Decomposition: Time series data consists of trends, levels, seasonality factors and noise. Deconstructing the information presented in temporal data into several components aid in segregating the several patterns associated with each of these factors. The decomposition of time series as an additive model can be represented as Eqn. 3 where T_t, C_t, S_t and N_t represent the trend, cyclicity, seasonality and noise factors respectively.

$$y_t = T_t + C_t + S_t + N_t \quad (3)$$

The classical approach for time series decomposition is the Moving Average method in which the trend is computed by averaging the values in a specified time period. (Abdollahi, 2020) used empirical mode decomposition to build a hybrid model for price forecasting. The components were extracted as volatile and nonlinear classes. Markov-based model applied on the volatile components and SVM based forecasting on the nonlinear components were observed to attain the best performance results. The LSTM encoder based model

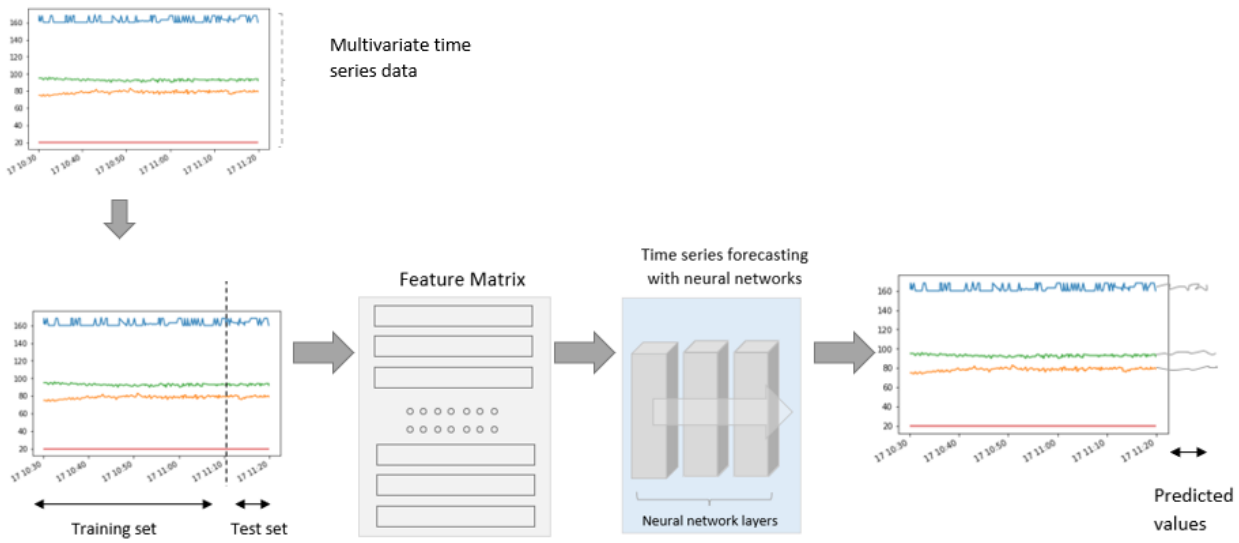

Fig. 1: Physiological value Prediction task pipeline

proposed by (Bedi & Toshniwal, 2020) incorporates variational mode decomposition technique for detecting significant features in conjunction with error variance modelling for improved performance. Decomposition is a methodology for enabling component wise analysis of temporal data and is used in various studies for highlighting the underlying temporal patterns.

2) Statistical Methods: Time series forecasting has been mostly conducted with statistical models such as ARIMA (AutoRegressive Integrated Moving Average (Aryee et al., 2018) and SARIMA (seasonal ARIMA) (Samal, Babu, Das, & Acharaya, 2019). The ARIMA model handles univariate time series and is frequently used in research because of the combined advantage of autoregressive sliding average model. The Multivariate time series is handled with VAR (Vector Autoregression) or VAR-based models such as VARMAX (Jamdade & Jamdade, 2021) and seasonality is incorporated with SARIMAX. (Q. Cheng et al., 2021) proposed a SARIMAX model with external regressors for emergency department occupancy prediction. However, the model showed reduced ability to predict with large time gaps and forecast with better accuracy for one hour predictions as compared to four hours. Recent research incorporates statistical methods with machine learning methods or deep neural networks for improved performance.

3) Machine Learning Methods: In addition to statistical methods which uses a parametric approach, researchers use Machine Learning methods such as Support Vector Regression (Maldonado, Gonzalez, & Crone, 2019), Linear Regression (Ciulla & D'Amico, 2019) and so on. Support Vector Regression with its kernel methods can support non linearity. Linear Regression is usually employed for its simplicity and high interpretability. (Chao, Zhipeng, & Yuanjie, 2019) proposed a SVM based model integrated with cooperative co-evolution algorithm for parametric optimization for the purpose of handling noisy data. Various kernel scale values in SVM was experimented by (Altan & Karasu, 2019) for time series forecasting with volatility estimation. However, these methods are less generalizable across datasets in terms of performance and recent studies have focused on deep learning methods for temporal forecasting. Specialised KNN variants for season-wise forecasting is proposed by (Martínez, Frías, Pérez-Godoy, & Rivera, 2018) which utilises each variant for learning different seasonal trends. This approach reduces erroneous forecasts by targeting cyclical patterns. Support Vector Regression model with polynomial cubic kernel function is proposed by (Beyca, Ervural, Tatoglu, Ozuyar, & Zaim, 2019) for monthly energy consumption forecasting. The model trains multiple input variables including seasonality and population estimates for improved accuracy results. On the other hand, (Kamir, Waldner, & Hochman, 2020) proposed a support vector regression model with radial function for forecasting wheat yields while considering multiple variables such as temperature and rainfall. Support Vector regression is the most commonly used machine learning method for time series analysis and is most often experimented with its kernel function for finding an optimal model for the specific dataset.

4) Graph Neural Networks: Graph Neural Networks (GNN) are used by researchers for modelling multivariate time series owing to its modularity, interpretability and cross modality which makes it suitable for the medical domain (Barbiero, Torné, & Lió, 2021). (Wu et al., 2020) adopted a graph neural network framework with propagation and a dilated inception layer for extracting feature relationships. A multimodal graph neural network combining attention mechanism is proposed by (D. Cheng et al., 2022) for financial time series prediction. However, the model performance is comparable to baselines in capturing market trends despite using the attention module. GNNs are considered to be weak in capturing changes in nodes as this is a sequential process which is deftly managed by recurrent neural networks (RNNs).

5) Recurrent Neural Networks: LSTM (Long short-term memory) networks which are based on RNNs are powerful in learning sequential information and this capability is well

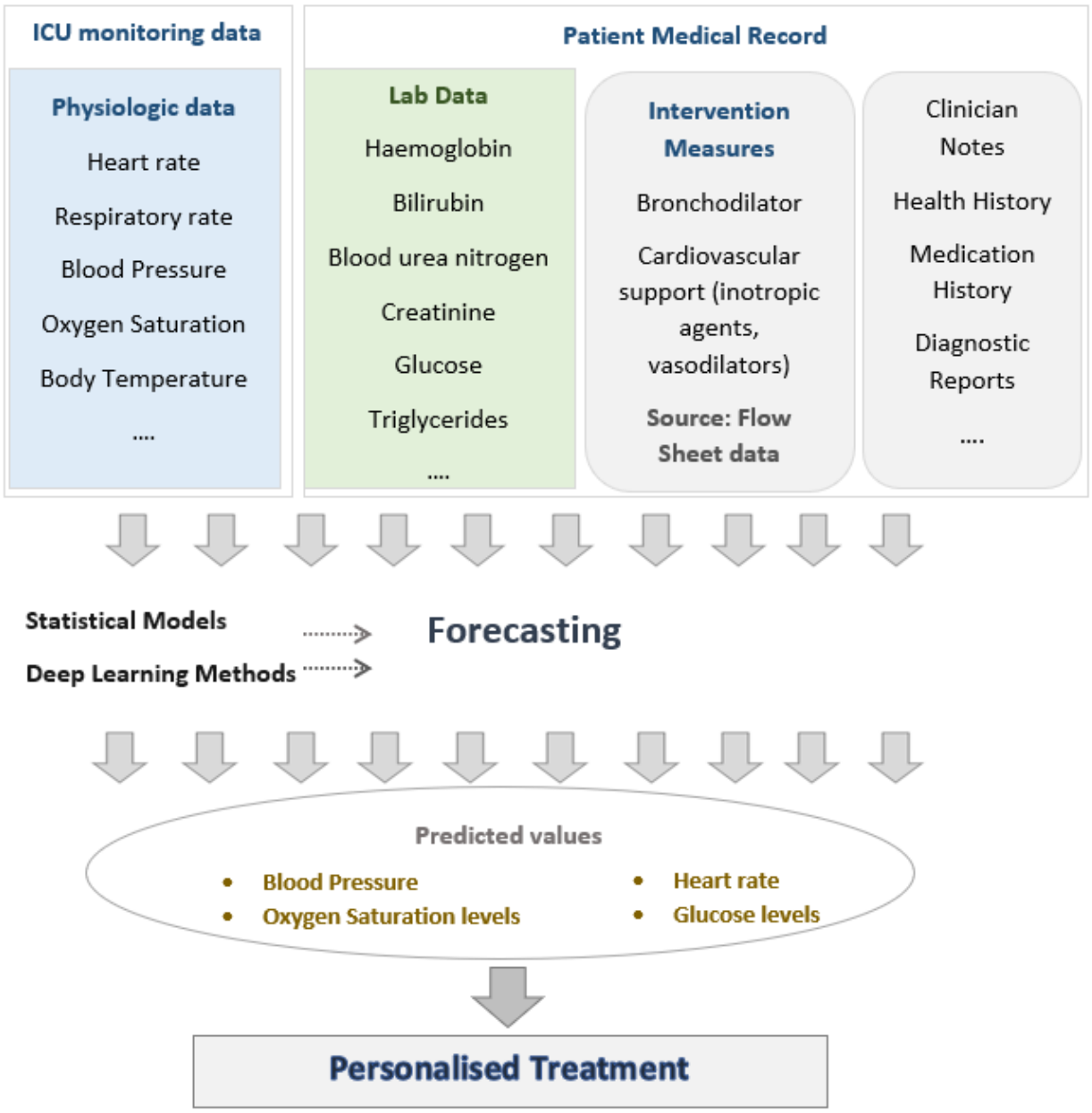

Fig. 2: Personalised medicine in ICU

suited to capture temporal sequences in large volumes of data. Multivariate clinical event forecasting modelled by (J. M. Lee & Hauskrecht, 2021) makes single step predictions with LSTM capturing distant information. Moreover, periodicity is incorporated based on probability distributions. LSTM is useful for making multi-step predictions as demonstrated by the model in (J. Zhang & Nawata, 2018) which formed a six-layered LSTM system for predicting disease outbreaks. (Sagheer & Kotb, 2019) developed a layer wise pre-trained LSTM-based stacked autoencoder as an alternative to weight initialization for multivariate time series forecasting. The approach showed good results when compared to traditional methods, but showed only marginal improvement as compared to baseline LSTM models. A de-noising approach is used by (F. Liu, Cai, Wang, & Lu, 2019) in implementing a LSTM based model with AdaBoost used for weighted distribution in the validation phase for improving prediction results. (S. Huang, Wang, Wu, & Tang, 2019) integrated an autoregressive model with convolutional components and attention mechanism for making revenue predictions on multiple time series. The model performance results were comparable to other LSTM based models, but gave good results in comparison to the traditional models.

GRUs (Gated Recurrent Neural Networks) which is also based on recurrent neural networks are used by researchers for time series modelling. A multi-output forecasting model by (Fox, Ang, Jaiswal, Pop-Busui, & Wiens, 2018) uses GRU coupled with Autoregressive model for forecasting blood glucose levels. The study incorporated sequential dependencies for performance uniformity across data subsets. A two step GRU with imputation and a hidden state decay mechanism

is proposed by (Shi et al., 2021) for mortality prediction of ICU patients. The imputation strategies adopted are more generalised and not a customized process for clinical data. LSTMs and variants of LSTM such as stacked LSTM are the most common recurrent neural network based models adopted by time series studies. The long and short term dependencies captured by RNN-based models make it an ideal model for initial experimentation with temporal data.

6) Convolutional Neural Networks: Recurrent Neural Networks are adept at capturing sequence based information, but weak in representing periodicity. Convolutional Neural Networks (CNNs) compensate for this shortcoming and multiple CNNs with residual mapping is employed by (Wan, Mei, Wang, Liu, & Yang, 2019) for making optimal multivariate time series predictions. Convolutional Neural Networks adopt a batch learning process when compared to RNNs' sequential approach. Additionally, CNNs are more suitable for handling missing data in time series. (X. S. Zhang, Tang, Dodge, Zhou, & Wang, 2019) implemented a CNN-based model called MetaPred which accepted a multivariate time series matrix as input into an embedding layer. This layer is passed on to the convolutional layers which output a vector representation which is subsequently passed to three MLP layers for outputting risk probabilities for each patient. The model is observed to perform well for predicting mild cognitive impairment, but not adequate for alzheimers and parkinson's disease prediction. (Ismail, Du, Martinez, & He, 2019) used multi-channel CNNs for forecasting severity of Parkinson's disease with a multi-step time series analysis. Prediction results were generated by the model learning a mapping function over each time period. Convolutional layers are efficient in extraction of both the spatial and temporal aspects present in time series data. (Xiao et al., 2021) proposed a convolutional based LSTM model with dual attention layers for effective extraction of the spatio-temporal features. Additionally, the model is also functional at handling the exogenous input features to the temporal data. The Conv-LSTM model proposed by (S. W. Lee & Kim, 2020) efficiently learns from high dimensional time series data and is capable of attaining high predictive power without extensive feature extraction. Additionally, the study used trend sampling for capturing real time data trends. The convolutional layers have the capacity to capture the spatial component which is indirectly associated with temporal data. This capacitates even baseline models to attain high predictive results and is frequently used by researchers in conjunction with RNN-based models such as LSTM.

7) Attention Mechanism: Attention Mechanisms are useful for particularly focussing on certain information which are considered significant in a given problem domain. Attention Mechanisms which are adept at handling sequences have been shown to perform well in time series forecasting. (Eom et al., 2020) used attention mechanism combined with CNN and BiGRU for blood pressure estimation. The performance of attention based model is observed to be higher than the baseline model for single inputs, but similar results are obtained for multiple inputs. (Hu & Zheng, 2020) used a multistage attention network for capturing mutation information from time sequences. As emphasised in the study by (Xiao et al., 2021), adding multiple layers of attention networks contribute to the explicit capturing of interdependent exogenous variables that are relevant in sequential prediction. (Assaf & Schumann, 2019) proposed a CNN model for creating attention-based feature maps with a gradient approach for making multivariate time series predictions. Temporal pattern attention mechanism is employed on CNN by (Shih, Sun, & Lee, 2019) which showed substantial improvement in results as compared to CNN without attention. (F. Liu, Lu, & Cai, 2020) proposed a stacked LSTM model with multi-level attention mechanism, wherein the stacked layers were used as the encoder for extracting the temporal dependencies between the various features. The encoded information from the encoder is transferred to the multi-layered attention module for enhanced predictive performance. A Stacked LSTM model with attention mechanism proposed by (Girkar et al., 2018) is found to attain good prediction results combined with high clinical interpretability in predicting blood pressure in hypotensive patients. However, the study did not consider covariates such as medication, ventilation parameters etc. as input into their model. Attention layers are most often observed to be capable of improving predictive results of temporal data, however, these results are not consistently distributed for all aspects of time series tasks.

8) Generative Adversarial Networks: Generative adversarial networks (GANs) are primarily used in image and video domains. For capturing temporal dependencies in time series data, (Yoon et al., 2019) proposed a time-series version of GAN called TGAN. TGAN uses supervised loss and an embedding network for dimensionality reduction. A hybrid GAN-LSTM is proposed by (Yazdanian & Sharifian, 2021) for forecasting cloud workload assimilating the complexity and volatility of the load traces. The multi-step prediction is performed with a dual layered convolutional network acting as the discriminator. GANs are most commonly used for the purpose of anomaly detection in time series. However, research studies such as (Koochali, Dengel, & Ahmed, 2021) employed a GAN-based model with adverserial training for transforming a deterministic model into a probabilistic model for multivariate time series forecasting.

9) Forecasting Strategies: MIMO (Multiple Input Multiple Output) and DIRMO (Direct Multiple Output) are two forecasting strategies used in time series modelling. The MIMO strategy forecasts values in a single step (Gasparin et al., 2019), however the DIRMO strategy retains dependencies which enables it to perform better in forecast modelling. A comparison of LSTM, BiLSTM and CNN was conducted by (Masum, Chiverton, Liu, & Vuksanovic, 2019) by adopting MIMO (Multiple Input Multiple Output) and DIRMO (Direct Multiple Output) forecasting strategies. The BiLSTM-DIRMO model had better performance than the others in forecasting blood pressure values. However, (Gasparin et al., 2019) employed RNN-based models with Recursive and MIMO forecasting strategies for predicting electricity load, owing to the high computational requirement of the DIRMO strategy. MIMO-based model exhibited good performance results.

The dataset under consideration is significant in deciding the superiority of deep learning methods in time series forecasting.

A comparison of standard machine learning approaches with LSTM for glucose level prediction by (J. Xie & Wang, 2018) indicated that linear regression and support vector regression methods performed better than LSTM. This points to the significance of building LSTM models with approaches such as attention mechanism (X. Zhang et al., 2019) or variants of it (Y. Li, Zhu, Kong, Han, & Zhao, 2019).

10) Hybrid Methods: Statistical methods have been used in conjunction with deep learning to leverage the benefits of both methodologies. (Mathonsi & van Zyl, 2021) presented a multivariate exponential smoothing method with LSTM with results comparable to that of baseline LSTM. However, the coverage of the model was not consistent across groups of observations. (Domingos, de Oliveira, & de Mattos Neto, 2019) proposed a hybrid system with linear models used for series forecasting and a non linear approach for error forecasting. The model uses ARIMA, Multi-layer Perceptron and Support Vector Regression and combines the forecasts for better outcome. A hybrid method with ARIMA and BPNN (back propagation neural network) was proposed by (Hadwan, Al-Maqaleh, Al-Badani, Khan, & Al-Hagery, 2022) for forecasting number of cancer patients. Though the hybrid model resulted in significant error reduction, the BPNN model is more adept at capturing the overall pattern of the series. A ARIMA-ANN hybrid method proposed by (Büyükşahin & Ertekin, 2019) employed time series decomposition followed by merging of the linear and non-linear temporal components for improved time series forecasting. (Caliwag & Lim, 2019) employed hybrid VARMA and LSTM for forecasting multiple cycles with prediction at one cycle ahead that aided in lowering error rates. Building hybrid methods is an optimal approach for combining the linear patterns captured by statistical methods with the non-linear patterns inherent to the time series data for achieving distinguishable predictive results.

11) Ensemble Methods: Many studies use ensemble methods for combining predictions of independent models through a voting mechanism. (D.-R. Liu, Lee, Huang, & Chiu, 2020) proposed a LSTM-Attention layer for predicting air pollution by forecasting PM2.5 concentrations. The predictions provided by the Attention layer is applied with an ensemble method with extreme Gradient Boosting for making secondary predictions. An ensemble method combining decision trees, random forests and gradient boosted trees are proposed by (Galicia, Talavera-Llames, Troncoso, Koprinska, & Martínez-Álvarez, 2019) for multi-step forecasting power consumption. (H. Chen, Guan, & Li, 2021) proposed a multi-factor model integrating LSTM-Attention with XGBoost regression for air quality forecasting. The optimal subtree nodes are used for obtaining the final prediction results. Ensemble methods enable the expression of the most optimal results from various methods. Boosting and bagging methods can lower the prediction errors that standalone methods are prone to when making predictions with temporal data.

Table I summarises some of the research performed in time series forecasting. While forecasting pollutant concentration, (T. Li et al., 2020) demonstrated that multivariate models performed better than univariate methods. Additionally, a hybrid CNN-LSTM has higher predictive power than compared to standalone LSTM model. However, it is observed that the hybrid model is prone to error when a single day's data is input to the model. The error rate is higher when the input length is increased to 14 days. This shows the importance of employing hybrid model rather than a simple architecture such as the model proposed by (Asante, Walker, Seidu, Kpogo, & Zou, 2022), wherein an ARIMA model is observed to be predicting better as compared to a basic LSTM model. It is expected that the performance would be much higher if more layers are added with LSTM as demonstrated by (T. Wang et al., 2020) or a hybrid version such as CNN-LSTM employed by (Parashar et al., 2020). Additionally, with deep learning methods a high performing base layer can be formed, which can be used as a generalised layer and combined with any other layers. (Gu et al., 2020) proposed a generalised model using CNN with dimension reduction, which attains consistent predictive results in conjunction with additional layers such as RNN, GRU or LSTM. This demonstrates the potentiality of deep learning methods when implemented in the most optimal sequence and combination. (Tazarv & Levorato, 2021) predicted blood pressure based on two datasets with CNN-LSTM-MLP layers. The model showed excellent results in predicting blood pressure values from the vital signs dataset. However, the prediction results were similar to other comparable works when predicting with the MIMIC-II dataset. In conjunction with the methodology used, the prediction horizon and sampling period selected plays an important role in improving prediction results as demonstrated by (Song et al., 2021), (T. Wang et al., 2020), (T. Li et al., 2020). A sampling period of 60 seconds is more efficient in forecasting blood pressure levels in a model with LSTM and CNN as demonstrated by (Song et al., 2021). Similarly smaller prediction horizon is observed to give better forecasting estimates for CNN-LSTM model (T. Li et al., 2020) as well as stacked LSTM model (Meng et al., 2020), (T. Wang et al., 2020) indicating the significance of adding more efficient components to the model for predicting long term values.

As outlined in this section, statistical methods, machine learning methods, deep neural networks, hybrid or ensemble methods etc. are systematically used in time series analysis by various studies. While statistical methodology is the most classical approach, more studies are adopting deep neural networks for time series forecasting owing to its high performance. Moreover, recent studies are increasingly using hybrid methods combining statistical and deep neural networks or adding several layers of different variations of neural networks in order to improve predictive power of forecasting models.

IV. Challenges and Issues

Research with ICU data largely relate to predictive modelling of sepsis (Spaeder et al., 2019), cardiac arrests (Matam, Duncan, & Lowe, n.d.), mortality prediction (Ge et al., 2018) and physiologic value forecasting (J. Xie & Wang, 2018). Few of the issues observed in clinical time series modelling are unevenly spaced temporal observations, inconsistent recording of values and missing values. (De Brouwer, Simm, Arany, & Moreau, 2019) adopted a Bayesian update network for

TABLE I: Summary of research in Time Series forecasting with deep learning.

Model	Summary	References
CNN-RNN-AR model	Captures long-term and short-term historical data by handling both linearity and non-linearity	(S. Li et al., 2021)
LSTM-Attention	Weighted Attention mechanisms are adept at handling long term dependencies preserved by sequences captured with LSTM	(Hu & Zheng, 2020), (Yuan et al., 2020), (T. Zhang et al., 2021), (X. Zhang et al., 2019)
Stacked LSTM	Multiple layers of stacked LSTM perform better than standalone LSTM. Performance varies based on the defined prediction horizon	(Meng et al., 2020), (Sagheer & Kotb, 2019), (T. Wang et al., 2020)
Stacked LSTM-ATTN	Temporal attention combined with multiple LSTM capture dependencies effectively and enable multi-step predictions	(Gangopadhyay et al., 2018), (Girkar et al., 2018), (F. Liu et al., 2020)
LSTM-ATTN-XGBoost	Attention mechanisms handle long term dependencies and preliminary predictions are passed to an ensemble learning method	(H. Chen et al., 2021), (D.-R. Liu et al., 2020)
CNN-dimension reduction	Attains interpretable results and allows for consistent interaction between high dimensional features	(Ali et al., 2019), (Gu et al., 2020)
Conv-LSTM	Effective in handling high dimensional time series data. Manages process effectively when output gate is closed	(Essien & Giannetti, 2020), (Indrawan et al., 2021), (S. W. Lee & Kim, 2020), (Shastri et al., 2020)
Conv-LSTM-ATTN	Manages spatial feature extraction effectively for sequenctial data with efficient capturing of long term dependencies	(Singh et al., 2020), (Xiao et al., 2021), (Zheng et al., 2020)
ARIMA-CNN-LSTM	Non-linearity is captured by CNN and LSTM. While Arima captures the linearity, seasonality is incorporated with SARIMA	(Dwivedi et al., 2021), (Ji et al., 2019)
CNN-LSTM	Preserves deterministic feature representations and captures relevant sequencial patterns	(T. Li et al., 2020), (Lin et al., 2017), (Parashar et al., 2020), (H. Xie et al., 2020)
CNN-LSTM-MLP	The combination of forward and backward propagation combined with feature extraction improves the performance of the model	(Phyo & Byun, 2021), (Tazarv & Levorato, 2021)
LSTM-CNN-decomposition	Decomposition of time series data into various frequency spectra aid in compact sequence analysis in tractable sections.	(Asadi & Regan, 2020), (Rezaei et al., 2021), (Song et al., 2021)

processing inconsistent values. This is specifically a challenging factor while handling clinical time series. The data is recorded in uneven time periods, with few of the recordings registered only once per day. Though the lack of data arising due to intermittent recordings cannot be treated as missing data, it remains challenging to address the irregular nature of information. Additionally, the method is less ideal for large volumes of data.

A challenge faced when forecasting with Multivariate Time series is the handling of missing values across multiple features or multiple series. A common approach for handling missing information is imputation, which is more reliable for univariate series. However, for ensuring reliable prediction results it is inevitable to consider global temporal information of the multivariate data. (Tang et al., 2020) attempts a partial solution to the issue by proposing a LSTM based model which generates gradient feedback with discriminator to form a memory module for capturing the local and global temporal dynamics of the data. The global temporal information is fully captured by incorporating adversarial training which contributed to further error reduction of the model.

Time series data owing to its continuous nature is limited in prediction capacity when smaller datasets are involved. Data augmentation is an approach adopted by researchers to combat this limitation, which often involves generating synthetic time series. (Bandara, Hewamalage, Liu, Kang, & Bergmeir, 2021) adopted three techniques for data augmentation viz. GRATIS, moving block bootstrap and dynamic time warping barycentric averaging. The considerable dissimilarity of the newly synthesised series in comparison with the real data is indicative of the contentious performance results.

Additionally, owing to the huge volume of temporal data it is necessary to adopt dimensionality reduction and noise reduction techniques. A dimension reduction strategy employed for forecasting by (Jahangir et al., 2020) is demonstrative of the ability of models to achieve consistent predictive results for all samples. The study also utilised stacked denoising autoencoders to reduce noise typically associated with time series data. The approach is observed to attain high performance results when compared to CNN, LSTM and support vector based methods.

V. Trends and Potential directions

Deep learning approaches such as LSTM, Hybrid methods with GRU and decomposition methods (C. Wang, Liu, Wei, Chen, & Zhang, 2021), MLP, decomposition, ARIMA and SVR (W. Chen, Xu, Chen, & Jiang, 2021), Random Forest with LSTM and decomposition methods (Karijadi & Chou, 2022) are few approaches adopted in recent research whose transferability across domains needs to be studied further. Additionally, introduction of multi-stage attention layers in temporal studies are observed to considerably improve prediction results. (Yin et al., 2021) added internal attention, spatial attention and temporal attention into their proposed model which demonstrated high predictive power than baseline models including an attention based CNN model. Few of the

potential directions of the research in time series analysis are addressed here.

Sequence Transformer networks is progressively being utilised for clinical time series research for its ability to capture the various types of invariances specific to the clinical domain. The transformer architecture enable context specific sequence learning and are capable of learning multiple periodic patterns in temporal data. (Zerveas, Jayaraman, Patel, Bhamidipaty, & Eickhoff, 2021) proposed a transformer based model for multivariate time series forecasting with unsupervised learning of the series. A pre-trained transformer model displayed substantial performance capabilities in a unsupervised learning environment. However, the sequence transformer proposed by (Oh, Wang, & Wiens, 2018) is limited to applying transformations uniformly for all features in the series. Consequently, feature-specific learning is restricted leading to exclusion of invariances in the process.

Cross learning from multiple time series is another approach adopted by researchers for efficient information extraction from several variables from various interdependent time series. Feature based and hybrid cross learning methods proposed by (Semenoglou, Spiliotis, Makridakis, & Assimakopoulos, 2021) extracted relevant information from several time series containing varying temporal data. Generalising of cross learning techniques for multiple domains is a deficient process compared to series-wise training. However, in some specialised cases the approach is consistently observed to be attaining high performance results.

In multivariate time series analysis detecting and quantifying correlation among the most significant features is an important step in capturing inter-dependencies between multiple time series or features. While multidimensional recurrence quantification analysis is performed for quantifying the recurrence properties of single time series, multidimensional cross-recurrence quantification analysis proposed by (Wallot, 2019) allows the approach to be utilised for bivariate series. Additionally, the output from cross-recurrence analysis is used for constructing a diagonal cross recurrence profile for extracting time lagged features from multiple time series.

Spiking neurons is an approach adopted in time series research wherein neural networks interact with each other through a series of spikes. The information passed through these spikes, the spike frequency and the distance between the spikes determine the predictive capacity of the models. (Mateńczuk et al., 2021) proposed a standalone spiking neural network (SNN) and a LSTM based spiking neural network (SLSTM) for financial time series forecasting. The optimal number of layers for SNN was estimated to be four, with additional layers not contributing to the performance. The iterative approach adopted by the model resulted in longer training time, but less appreciable performance increase. Based on the study, further research is required to develop an optimised version of spiking neural networks in time series analysis. (Wei, Wang, Niu, & Li, 2021) proposed a convolutional spiking neural network model for extracting temporal features from output generated from the primary gated recurrent neural network layer. The GRU is initially employed on a deconstructed series of temporal data. Grey wolf optimization is combined with the spiking mechanism for determining optimal weights for each of the decomposed time series.

Recent research has adopted various novel approaches for univariate and multivariate time series forecasting. These approaches have manifested good predictive results albeit with limitations. Further studies are required to accentuate the predictive power of the models with additional techniques. While some methods showed dataset-specific or domain-specific performance capabilities, other approaches such as spiking neurons require enhancement with optimisation methodologies to attain replicable results across domains.

VI. Conclusion

The review provides an outline of the methodologies adopted for time series analysis. The conventional methods are based on statistical methods which are linear in nature. However, more research studies are adopting deep learning methods for time series forecasting and classification. The advantage of deep learning methods for time series analysis is that non linear relationships are addressed seamlessly. Additionally, deep learning methods are more adept at making multi-step forecasts, feature engineering, addressing sequential nature of time series data and learning temporal dependencies. Moreover, early studies involved univariate time series analysis. Recent studies have recognised the importance of adopting a multivariate approach in time series analysis for improved performance results. Incorporating of multiple variables or multiple time series in temporal studies have demonstrated high predictive power as compared to single variables.

This work has comprehensively reviewed the various time series analysis methods adopted by researchers including their limitations. This review also provides an indication of challenges faced in the field as well as future trends in time series forecasting, for the development of novel methodologies in temporal analysis. The specific methodology adopted by a researcher is dependent on the research requirement and the domain under study. It is necessary to identify the advantages offered by each method so as to construct a model with a combination of algorithms enabling the model to enhance the predictive potential. Each layer of algorithm is commissioned to handle the different aspects and requirements of the dataset features, thereby qualifying a model with an all encompassing capability for manoeuvring the various facets of the forecasting problem. While many research studies focus on univariate temporal analysis, it is imperative to adopt multivariate analysis for enabling the development of high level real-time temporal forecasting solutions. The review highlights the capability of deep neural networks in handling such complex temporal information for achieving reliable prediction results.

References

Abdollahi, H. (2020). A novel hybrid model for forecasting crude oil price based on time series decomposition. *Applied Energy*, *267*, 115035.

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S., & Fisher, A. (2020). Missing data imputation of high-resolution

temporal climate time series data. *Meteorological Applications*, *27*(1), e1873.

Albers, D. J., Levine, M. E., Stuart, A., Mamykina, L., Gluckman, B., & Hripcsak, G. (2018). Mechanistic machine learning: how data assimilation leverages physiologic knowledge using bayesian inference to forecast the future, infer the present, and phenotype. *Journal of the American Medical Informatics Association : JAMIA*, *25*(10), 1392-1401.

Ali, M., Jones, M. W., Xie, X., & Williams, M. (2019). Timecluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer*, *35*(6), 1013–1026.

Altan, A., & Karasu, S. (2019). The effect of kernel values in support vector machine to forecasting performance of financial time series. *The Journal of Cognitive Systems*, *4*(1), 17–21.

Aryee, G., Kwarteng, E., Essuman, R., Nkansa Agyei, A., Kudzawu, S., Djagbletey, R., ... Forson, A. (2018). Estimating the incidence of tuberculosis cases reported at a tertiary hospital in ghana: a time series model approach. *BMC Public Health*, *18*(1), 1–8.

Asadi, R., & Regan, A. C. (2020). A spatio-temporal decomposition based deep neural network for time series forecasting. *Applied Soft Computing*, *87*, 105963.

Asante, D. O., Walker, A. N., Seidu, T. A., Kpogo, S. A., & Zou, J. (2022). Hypertension and diabetes in akatsi south district, ghana: Modeling and forecasting. *BioMed Research International*, *2022*.

Assaf, R., & Schumann, A. (2019). Explainable deep neural networks for multivariate time series predictions. In *Ijcai* (pp. 6488–6490).

Bandara, K., Hewamalage, H., Liu, Y.-H., Kang, Y., & Bergmeir, C. (2021). Improving the accuracy of global forecasting models using time series data augmentation. *Pattern Recognition*, *120*, 108148.

Barbiero, P., Torné, R. V., & Lió, P. (2021). Graph representation forecasting of patient's medical conditions: Toward a digital twin. *Frontiers in genetics*, *12*.

Bedi, J., & Toshniwal, D. (2020). Energy load time-series forecast using decomposition and autoencoder integrated memory network. *Applied Soft Computing*, *93*, 106390.

Beyca, O. F., Ervural, B. C., Tatoglu, E., Ozuyar, P. G., & Zaim, S. (2019). Using machine learning tools for forecasting natural gas consumption in the province of istanbul. *Energy Economics*, *80*, 937–949.

Büyükşahin, Ü. Ç., & Ertekin, Ş. (2019). Improving forecasting accuracy of time series data using a new arima-ann hybrid method and empirical mode decomposition. *Neurocomputing*, *361*, 151–163.

Caliwag, A. C., & Lim, W. (2019). Hybrid varma and lstm method for lithium-ion battery state-of-charge and output voltage forecasting in electric motorcycle applications. *IEEE Access*, *7*, 59680–59689.

Chao, L., Zhipeng, J., & Yuanjie, Z. (2019). A novel reconstructed training-set svm with roulette cooperative coevolution for financial time series classification. *Expert Systems with Applications*, *123*, 283–298.

Chen, H., Guan, M., & Li, H. (2021). Air quality prediction based on integrated dual lstm model. *IEEE Access*, *9*, 93285–93297.

Chen, W., Xu, H., Chen, Z., & Jiang, M. (2021). A novel method for time series prediction based on error decomposition and nonlinear combination of forecasters. *Neurocomputing*, *426*, 85–103.

Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern recognition*, *121*, 108218-.

Cheng, Q., Argon, N. T., Evans, C. S., Liu, Y., Platts-Mills, T. F., & Ziya, S. (2021). Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, *48*, 177–182.

Ciulla, G., & D'Amico, A. (2019). Building energy performance forecasting: A multiple linear regression approach. *Applied Energy*, *253*, 113500.

De Brouwer, E., Simm, J., Arany, A., & Moreau, Y. (2019). Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in Neural Information Processing Systems*, *32*.

Domingos, S. d. O., de Oliveira, J. F., & de Mattos Neto, P. S. (2019). An intelligent hybridization of arima with machine learning models for time series forecasting. *Knowledge-Based Systems*, *175*, 72–86.

Dwivedi, S. A., Attry, A., Parekh, D., & Singla, K. (2021). Analysis and forecasting of time-series data using s-arima, cnn and lstm. In *2021 international conference on computing, communication, and intelligent systems (icccis)* (pp. 131–136).

Eom, H., Lee, D., Han, S., Hariyani, Y. S., Lim, Y., Sohn, I., ... Park, C. (2020). End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism. *Sensors*, *20*(8), 2338.

Essien, A., & Giannetti, C. (2020). A deep learning model for smart manufacturing using convolutional lstm neural network autoencoders. *IEEE Transactions on Industrial Informatics*, *16*(9), 6069–6078.

Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R., & Wiens, J. (2018). Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1387–1395).

Galicia, A., Talavera-Llames, R., Troncoso, A., Koprinska, I., & Martínez-Álvarez, F. (2019). Multi-step forecasting for big data time series based on ensemble learning. *Knowledge-Based Systems*, *163*, 830–841.

Gangopadhyay, T., Tan, S. Y., Huang, G., & Sarkar, S. (2018). Temporal attention and stacked lstms for multivariate time series prediction.

Gasparin, A., Lukovic, S., & Alippi, C. (2019). Deep learning for time series forecasting: The electric load case. *arXiv preprint arXiv:1907.09207*.

Ge, W., Huh, J.-W., Park, Y. R., Lee, J.-H., Kim, Y.-H., & Turchin, A. (2018). An interpretable icu mortality prediction model based on logistic regression and re-

current neural networks with lstm units. *AMIA Annual Symposium proceedings*, *2018*, 460-469.

Girkar, U. M., Uchimido, R., Lehman, L.-W. H., Szolovits, P., Celi, L., & Weng, W.-H. (2018). Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *arXiv preprint arXiv:1812.00699*.

Gu, K., Dang, R., & Prioleau, T. (2020). Neural physiological model: A simple module for blood glucose prediction. In *2020 42nd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 5476–5481).

Hadwan, M., Al-Maqaleh, B. M., Al-Badani, F. N., Khan, R. U., & Al-Hagery, M. A. (2022). A hybrid neural network and box-jenkins models for time series forecasting. *CMC-Computers Materials & Continua*, *70*(3), 4829–4845.

Hamidi, O., Borzu, S. R., Maroufizadeh, S., & Amini, P. (2021). Application of multivariate generalized linear mixed model to identify effect of dialysate temperature on physiologic indicators among hemodialysis patients. *Journal of Biostatistics and Epidemiology*, *7*(3), 263–271.

Hu, J., & Zheng, W. (2020). Multistage attention network for multivariate time series prediction. *Neurocomputing*, *383*, 122–137.

Huang, E., Wang, R., Chandrasekaran, U., & Yu, R. (2020). Aortic pressure forecasting with deep learning. In *2020 computing in cardiology* (pp. 1–4).

Huang, S., Wang, D., Wu, X., & Tang, A. (2019). Dsanet: Dual self-attention network for multivariate time series forecasting. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 2129–2132).

Indrawan, R., Saadah, S., & Yunanto, P. E. (2021). Blood glucose prediction using convolutional long short-term memory algorithms. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *7*(2).

Ismail, N. H., Du, M., Martinez, D., & He, Z. (2019). Multivariate multi-step deep learning time series approach in forecasting parkinson's disease future severity progression. In *Proceedings of the 10th acm international conference on bioinformatics, computational biology and health informatics* (pp. 383–389).

Jahangir, H., Tayarani, H., Baghali, S., Ahmadian, A., Elkamel, A., Golkar, M. A., & Castilla, M. (2020). A novel electricity price forecasting approach based on dimension reduction strategy and rough artificial neural networks. *IEEE transactions on industrial informatics*, *16*(4), 2369-2381.

Jamdade, P. G., & Jamdade, S. G. (2021). Modeling and prediction of covid-19 spread in the philippines by october 13, 2020, by using the varmax time series method with preventive measures. *Results in Physics*, *20*, 103694.

Ji, L., Zou, Y., He, K., & Zhu, B. (2019). Carbon futures price forecasting based with arima-cnn-lstm model. *Procedia Computer Science*, *162*, 33–38.

Kamir, E., Waldner, F., & Hochman, Z. (2020). Estimating wheat yields in australia using climate records, satellite image time series and machine learning methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, *160*, 124–135.

Karijadi, I., & Chou, S.-Y. (2022). A hybrid rf-lstm based on ceemdan for improving the accuracy of building energy consumption prediction. *Energy and Buildings*, *259*, 111908.

Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate lstm-fcns for time series classification. *Neural Networks*, *116*, 237–245.

Koochali, A., Dengel, A., & Ahmed, S. (2021). If you like it, gan it—probabilistic multivariate times series forecast with gan. In *Engineering proceedings* (Vol. 5, p. 40).

Lee, J. M., & Hauskrecht, M. (2021). Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artificial Intelligence in Medicine*, *112*, 102021.

Lee, S. W., & Kim, H. Y. (2020). Stock market forecasting with super-high dimensional time-series data using convlstm, trend sampling, and specialized data augmentation. *Expert Systems with Applications*, *161*, 113704.

Li, S., Huang, H., & Lu, W. (2021). A neural networks based method for multivariate time-series forecasting. *IEEE Access*, *9*, 63915–63924.

Li, T., Hua, M., & Wu, X. (2020). A hybrid cnn-lstm model for forecasting particulate matter (pm2. 5). *IEEE Access*, *8*, 26933–26940.

Li, Y., Zhu, Z., Kong, D., Han, H., & Zhao, Y. (2019). Ea-lstm: Evolutionary attention-based lstm for time series prediction. *Knowledge-Based Systems*, *181*, 104785.

Lin, T., Guo, T., & Aberer, K. (2017). Hybrid neural networks for learning the trend in time series. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 2273–2279).

Liu, D.-R., Lee, S.-J., Huang, Y., & Chiu, C.-J. (2020). Air pollution forecasting based on attention-based lstm neural network and ensemble learning. *Expert Systems*, *37*(3), e12511.

Liu, F., Cai, M., Wang, L., & Lu, Y. (2019). An ensemble model based on adaptive noise reducer and over-fitting prevention lstm for multivariate time series forecasting. *IEEE Access*, *7*, 26102–26115.

Liu, F., Lu, Y., & Cai, M. (2020). A hybrid method with adaptive sub-series clustering and attention-based stacked residual lstms for multivariate time series forecasting. *IEEE Access*, *8*, 62423–62438.

Maldonado, S., Gonzalez, A., & Crone, S. (2019). Automatic time series analysis for electric load forecasting via support vector regression. *Applied Soft Computing*, *83*, 105616.

Martínez, F., Frías, M. P., Pérez-Godoy, M. D., & Rivera, A. J. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with knn. *Expert systems with applications*, *103*, 38–48.

Masum, S., Chiverton, J. P., Liu, Y., & Vuksanovic, B. (2019). Investigation of machine learning techniques

in forecasting of blood pressure time series data. In *International conference on innovative techniques and applications of artificial intelligence* (pp. 269–282).

Matam, B. R., Duncan, H., & Lowe, D. (n.d.). Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit: Prediction of cardiac arrests. *Journal of clinical monitoring and computing*, *33*(4), 713-724.

Mateńczuk, K., Kozina, A., Markowska, A., Czerniachowska, K., Kaczmarczyk, K., Golec, P., ... others (2021). Financial time series forecasting: Comparison of traditional and spiking neural networks. *Procedia Computer Science*, *192*, 5023–5029.

Mathonsi, T., & van Zyl, T. L. (2021). A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. *Forecasting*, *4*(1), 1–25.

Meng, X., Liu, M., & Wu, Q. (2020). Prediction of rice yield via stacked lstm. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)*, *11*(1), 86–95.

Oh, J., Wang, J., & Wiens, J. (2018). Learning to exploit invariances in clinical time-series data using sequence transformer networks. In *Machine learning for healthcare conference* (pp. 332–347).

Oyeleye, M., Chen, T., Titarenko, S., & Antoniou, G. (2022). A predictive analysis of heart rates using machine learning techniques. *International Journal of Environmental Research and Public Health*, *19*(4), 2417.

Parashar, A., Mohan, Y., & Rathee, N. (2020). *Forecasting covid-19 cases in india using deep cnn lstm model* (Tech. Rep.). EasyChair.

Phyo, P. P., & Byun, Y.-C. (2021). Hybrid ensemble deep learning-based approach for time series energy prediction. *Symmetry*, *13*(10), 1942.

Rezaei, H., Faaljou, H., & Mansourfar, G. (2021). Stock price prediction using deep learning and frequency decomposition. *Expert Systems with Applications*, *169*, 114332.

Sagheer, A., & Kotb, M. (2019). Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems. *Scientific reports*, *9*(1), 1–16.

Samal, K. K. R., Babu, K. S., Das, S. K., & Acharaya, A. (2019). Time series based air pollution forecasting using sarima and prophet model. In *proceedings of the 2019 international conference on information technology and computer communications* (pp. 80–85).

Seid, S. S., Ibro, S. A., Ahmed, A. A., Akuma, A. O., Reta, E. Y., Haso, T. K., & Fata, G. A. (2019). Causes and factors associated with neonatal mortality in neonatal intensive care unit (nicu) of jimma university medical center, jimma, south west ethiopia. *Pediatric health, medicine and therapeutics*, *10*, 39.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., & Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, *37*(3), 1072–1084.

Shastri, S., Singh, K., Kumar, S., Kour, P., & Mansotra, V. (2020). Time series forecasting of covid-19 using deep learning models: India-usa comparative case study. *Chaos, Solitons & Fractals*, *140*, 110227.

Shi, Z., Wang, S., Yue, L., Pang, L., Zuo, X., Zuo, W., & Li, X. (2021). Deep dynamic imputation of clinical time series for mortality prediction. *Information Sciences*, *579*, 607–622.

Shih, S.-Y., Sun, F.-K., & Lee, H.-y. (2019). Temporal pattern attention for multivariate time series forecasting. *Machine Learning*, *108*(8), 1421–1441.

Singh, S. P., Sharma, M. K., Lay-Ekuakille, A., Gangwar, D., & Gupta, S. (2020). Deep convlstm with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal*, *21*(6), 8575–8582.

Song, X., Zhu, L., Feng, X., Wu, H., & Li, Y. (2021). Combined forecast model of lstm-cnn hypertension based on eemd. In *2021 4th international conference on signal processing and machine learning* (pp. 117–122).

Spaeder, M. C., Moorman, J. R., Tran, C. A., Keim-Malpass, J., Zschaebitz, J. V., Lake, D. E., & Clark, M. T. (2019). Predictive analytics in the pediatric intensive care unit for early identification of sepsis: capturing the context of age. *Pediatric research*, *86*(5), 655-661.

Staffini, A., Svensson, T., Chung, U.-i., & Svensson, A. K. (2021). Heart rate modeling and prediction using autoregressive models and deep learning. *Sensors*, *22*(1), 34.

Tang, X., Yao, H., Sun, Y., Aggarwal, C., Mitra, P., & Wang, S. (2020). Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 5956–5963).

Tazarv, A., & Levorato, M. (2021). A deep learning approach to predict blood pressure from ppg signals. In *2021 43rd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 5658–5662).

Thao, P. T. N., Tra, T. T., Son, N. T., & Wada, K. (2018). Reduction in the il-6 level at 24 h after admission to the intensive care unit is a survival predictor for vietnamese patients with sepsis and septic shock: a prospective study. *BMC Emergency Medicine*, *18*(1), 1–7.

Wallot, S. (2019). Multidimensional cross-recurrence quantification analysis (mdcrqa)–a method for quantifying correlation between multivariate time-series. *Multivariate behavioral research*, *54*(2), 173–191.

Wan, R., Mei, S., Wang, J., Liu, M., & Yang, F. (2019). Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics*, *8*(8).

Wang, C., Liu, Z., Wei, H., Chen, L., & Zhang, H. (2021). Hybrid deep learning model for short-term wind speed forecasting based on time series decomposition and gated recurrent unit. *Complex System Modeling and Simulation*, *1*(4), 308–321.

Wang, T., Li, W., & Lewis, D. (2020). Blood glucose forecasting using lstm variants under the context of open source artificial pancreas system.

Wei, D., Wang, J., Niu, X., & Li, Z. (2021). Wind speed forecasting system based on gated recurrent units and convolutional spiking neural networks. *Applied Energy*, *292*, 116842.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. (2020). Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 753–763).

Xiao, Y., Yin, H., Zhang, Y., Qi, H., Zhang, Y., & Liu, Z. (2021). A dual-stage attention-based conv-lstm network for spatio-temporal correlation and multivariate time series prediction. *International Journal of Intelligent Systems*, *36*(5), 2036–2057.

Xie, H., Zhang, L., & Lim, C. P. (2020). Evolving cnn-lstm models for time series prediction using enhanced grey wolf optimizer. *IEEE Access*, *8*, 161519–161541.

Xie, J., & Wang, Q. (2018). Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge. In *Khd@ ijcai.*

Yazdanian, P., & Sharifian, S. (2021). E2lg: a multiscale ensemble of lstm/gan deep learning architecture for multistep-ahead cloud workload prediction. *The Journal of Supercomputing*, *77*(10), 11052–11082.

Yee, C. R., Narain, N. R., Akmaev, V. R., & Vemulapalli, V. (2019). A data-driven approach to predicting septic shock in the intensive care unit. *Biomedical informatics insights*, *11*, 1178222619885147.

Yin, X., Wu, G., Wei, J., Shen, Y., Qi, H., & Yin, B. (2021). Multi-stage attention spatial-temporal graph networks for traffic prediction. *Neurocomputing*, *428*, 42–53.

Yoon, J. H., Mu, L., Chen, L., Dubrawski, A., Hravnak, M., Pinsky, M. R., & Clermont, G. (2019). Predicting tachycardia as a surrogate for instability in the intensive care unit. *Journal of Clinical Monitoring and Computing*, *33*(6), 973–985.

Yuan, Y., Lin, L., Huo, L.-Z., Kong, Y.-L., Zhou, Z.-G., Wu, B., & Jia, Y. (2020). Using an attention-based lstm encoder–decoder network for near real-time disturbance detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 1819–1832.

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021). A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining* (pp. 2114–2124).

Zhang, J., & Nawata, K. (2018). Multi-step prediction for influenza outbreak by an adjusted long short-term memory. *Epidemiology & Infection*, *146*(7), 809–816.

Zhang, T., Zheng, X.-Q., & Liu, M.-X. (2021). Multiscale attention-based lstm for ship motion prediction. *Ocean Engineering*, *230*, 109066.

Zhang, X., Liang, X., Zhiyuli, A., Zhang, S., Xu, R., & Wu, B. (2019). At-lstm: An attention-based lstm model for financial time series prediction. In *Iop conference series: Materials science and engineering* (Vol. 569, p. 052037).

Zhang, X. S., Tang, F., Dodge, H. H., Zhou, J., & Wang, F. (2019). Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2487–2495).

Zheng, H., Lin, F., Feng, X., & Chen, Y. (2020). A hybrid deep learning model with attention-based conv-lstm networks for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *22*(11), 6910–6920.

A Two-stage Approach for Detecting Spammers in Online Social Networks

Bandar Alghamdi[1,2] ,Yue Xu[1] and Jason Watson[1]

[1] Faculty of Science and Engineering, Queensland University of Technology Australia Brisbane City 4000, Australia.
bandar.alghamdi@hdr.qut.edu.au
{yue.xu, ja.watson}@qut.edu.au
[2] Institute of Public Administration Riyadh City 11141, Saudi Arabia
alghamdib@ipa.edu.sa

***Abstract*— The phenomenon of evolving behavior by spammers in social networks has received consistent attention from different researchers to combat this challenge. Twitter is an example of a micro-blogging where spammers take place and change their spamming strategies and behavioral patterns to evade detection. Several approaches have been put forward to fight this problem, nevertheless they lost their effectiveness. The main limitation of existing methods they use unified features to characterize spammers' behavior without considering the fact that spammers behave differently, and this results in distinct patterns and features. In this research project, we approach the challenge of spammer's behavior by utilizing the level of focused interest patterns of users to uncover the differences between spammers and legitimate user. We propose quantity methods using three topical features: topic entropy, standard deviation of topic distributions, and Local Outlier Standard Score (LOSS) to measure the change in user's interest and then determine whether the user has a focused-interest or a diverse-interest. We develop a framework by combining unsupervised and supervised learning to differentiate between spammers and legitimate users. The results of this experiment show that our proposed approach can effectively differentiate between spammers and legitimate users regarding the level of focused interest. It also demonstrates the similarity level between the explicit user's interest and implicit tweets content. Compared with other detection methods, our method has better performance. To the best of our knowledge, our study is the first to provide a generic and efficient framework to represent user-focused interest level that can handle the problem of the evolving behavior of spammers.**

***Index Terms*— Spam, Behavior, Spammers detection, User interest, Online social networks, Machine learning**

I. INTRODUCTION

Recent developments in the field of online social networks have led to the integration of OSNs into nearly all aspects of everyday activity; however, spammers take advantage of these services for malicious purposes. With the increase in the influence of OSNs among users, a large platform has been established that spammers use to spread spam messages [2]. In Twitter, Spam tweets refer to unsolicited tweets containing malicious links that direct victims to external sites containing malware downloads, phishing scams, drug sales, etc.[1] . Spammers utilize different methods in spreading spam content, either using compromised accounts with already established reputations and exploiting the inherent trust of these accounts to spread malicious messages [2, 3] or creating fake accounts that appear to be legitimate to mimic legitimate user behavior by posting spam content and normal content [4].

Existing approaches address the detection of spam and malicious content on social networks through the use of language patterns and content-based metadata [5, 6]. Some works employ the user's profile in detecting compromised and fake accounts [3, 7]. Some recent additions to the literature have offered valuable findings about spammers' behavior, using networked communities or developing a hybrid approach for spam detection using multiple views [8-11]. Further studies have discussed communities and cooperative spammers [10, 12] for spam detection. Although most of the aforementioned detection methods detect spammers, a major limitation is that they characterize spammers' behavior with unified features, without considering the fact that spammers behave differently, and this results in distinct patterns and features for different spammers with different purposes.

Topic-based features proposed by Liu et al [13] discriminate human-like spammers from legitimate users based on user's content interest, which is represented by the user's topic distribution. Liu used the same set of features to classify users as spam or legitimate. However, using only one set of features is insufficient to differentiate spam users from legitimate users because both spam and legitimate users can have focused or diverse information interests in terms of information content. Users with a wide scope of interests are called diverse users, and users with limited scope of interest are called focused users. The study in [13] indicates that spam users can have either very focused interests or diverse interests. However, their research was not specifically designed to analyse this overlap or to characterize spammers with focused interests to separate them from legitimate users whose interests are focused too. Likewise, spam and legitimate users with multiple topics of interest cannot be classified by their approach.

Before describing our study in detail, we will provide the motivation behind our work and the assumptions used in our approach. As mentioned earlier, evolving behaviors by

spammers on online social networks continue to be a big challenge. Most existing approaches characterize users on the basis of features that are used commonly for all spammers, whereas spammers change their spamming strategies and behave differently, which requires us to consider this difference. Our study was conducted under two assumptions.

Assumption1: Spammers can behave differently, and this results in distinct patterns and features that need to be considered. The assumption of there being different behavior models for spammers has drawn attention recently. Some approaches have been proposed for addressing the difference [14, 15] where the users' features are noticed across different groups of users. Their study indicates that some users interact with others with less mentions, whereas the users in another group use more hashtags. In our study, we assume that this pattern used by spammers must be reflected in some features that may be good for a certain type of spammer yet that is not applicable to another type. We extract different features to represent users in two different groups, focused user group and diverse user group. In each group, more effective features allow a more accurate classifier by applying classification techniques

Assumption 2: The integration of both content and profile features is effective to properly understanding users' behavior and interest through combining implicit and explicit information. We assume that there is a need to combine the relevant features of the user's profile and content messages from the user's interest perspective in order to obtain a comprehensive understanding of spammer behavior. Therefore, in this paper, we propose a feature which takes account of users' self-descriptions which explicitly reflect the user's interest and the relation to their tweets. This can be used as a unique feature for identifying spammers that mimic legitimate user's behavior.

In this paper, we propose to take user information interests as a key factor for spammer detection since the engagement of users in any activity is driven by their interests. In online social networks, users tend to post messages that are interesting to them. However, since spammers intend to propagate spam messages or malicious URLs, their interests change frequently so that they can exploit any event that is trending or that has active users. Therefore, users' information interest alone is insufficient for identifying spammers from legitimate users given the fact that spammers could be focused users or diverse users in terms of information interest. This has motivated us to deeply understand users' behavior in terms of topics of interest in order to separate users into different groups so that we can use different features to build a classifier for each group in order to classify spammers more accurately. In order to separately analyze users with different behaviors, we propose to split users into two different groups by using clustering techniques in terms of the scope of their content interest. To this end, we propose a novel two-stage approach to detect spammers in online social networks.

The purpose of the first stage is to separate users into two different groups: Focused-Interest who have focused information interests, and Diverse-Users who have diverse information interests. Three topic-based features are proposed to assess the focuses level of user's interest: topic entropy, standard deviation of topic distributions and Local Outlier Standard Score (LOSS) [13]. Based on the topical features, we uncover a clear distinction between focused-interest group and diverse-interest group using clustering algorithm. In the second stage, different sets of features are proposed for characterizing the users in the focused cluster and users in the diverse cluster separately. Based on the features for each cluster, a separate classifier can be generated. With this approach, spammers with focused interests and diverse interests can be more accurately classified.

The main contribution of this study can be summarized as follows:

- We propose a novel two-stage approach for detecting spammers. In the first stage, users are grouped into two clusters based on the content diversity of their posts, i.e., focused cluster and diverse cluster. In the second stage, a classification technique is used to classify spammers from legitimate users for each of the clusters.
- Based on the level of content diversity, we represent users in one cluster with the features that are different from the features used for the users in the other cluster. Therefore, the classification accuracy can be greatly improved
- We propose a new feature to represent users for differentiating spammers from legitimate users. The new feature measures the consistency between a user's self-stated interest and the content of the user's posts.

The remainder of this paper is organized as follows: Section 2 introduces rated work. Section 3 explains the proposed method: the novel application of focused and diverse user's interest based on topical features. Section 4 discusses the experiment, evaluation and results. Section 5 contains a discussion. Section 6 finishes this paper by presenting a conclusion and future work.

II. Related Work

Many studies have been conducted to investigate spammers' behavior in online social networks, and researchers have shown an increased interest in this regard. A survey of potential solutions and challenges on spam detection in online social network has been proposed by [16]. Previous works have focused on characterizing spammers' behavior using different features and approaches [6, 8, 12, 15, 17].

A considerable amount of literature has been published on spam detection using content-based features [5, 14]. The statistical analysis of language, such as linguistics evolution, self-similarity and vocabulary, are the primary features used for spam detection. Although they perform well in detecting spam tweets, their limitation relies in the fact that content features alone cannot be used to properly analyse spammer behavior. Spammers mainly utilize the tactics of mixing normal tweets

and posting heterogeneous tweets. Therefore, the inclusion of other features than content features would result in higher accuracy in detection and would provide extensive range in understanding spammer behavior. Some works consider only user's profile [18] to detect spam users, without considering content features due to the idea that this is a fast and effective way. They capture users' behavior and identify certain patterns from the profile to detect spammers and compromised accounts. Alternatively, [19] combined profile features with some content features to identify suspended accounts and spam campaigns. [7] has reported that spam and non-spam profiles overlap, which can make it a challenge to identify spam users across a network. However, certain characteristics are noticeable among spam profiles, including young accounts, tweets with a higher succession rate, tweets with greater status and tweets that contain spam words. However, these studies were limited to characterizing spammer behaviors in regard to a few aspects, and they showed a lack of classification accuracy as their approaches are not sophisticated.

It has conclusively been shown that combining content and profile features provides a comprehensive understanding of spammer behavior [8, 11, 17]. They determine that there is a strong and consistent correlation between the profile and content for all suspicious accounts. Such a combination shows the fundamental characteristics of spammers from different views and provides a different level of detection rates. [8] proposed dynamic metrics to measure the change in user activities and to identify abnormal behavior with a combination of some user profile features. [6] presented a detailed analysis of 14 million tweets with a focus on hashtags and tweet content. They observed that spam detection at tweet-level can be made more accurate by combining user- level. In our present study, we extend this combination by considering content features and user demographic data with a focus on the user's interests.

The social graph is one of the most widely used approaches for spam detection [8, 10]. In social networks, users are connected with each other to form network communities that share similar characteristics, such as interests, location or past common history. Analysing the underlying structure of the network community provides insight in detecting the outlier or spammer that drifts from the community or that behaves abnormally. Despite the efficacy of this method, analysing community networks requires effort and time, and spammers work cooperatively to form communities that are a challenge to identify through the network graph approach [12]. In addition, spammers change their behavior and strategies to evade detection [20], which makes this technique not very effective.

To meet the challenge of the evolving behavior of spammers, subsequent approaches have been proposed using topics features to detect spammers. [21] introduced word-, topic- and user-based features, using the Labeled Latent Dirichlet Allocation (L-LDA) model to model discriminated topics and words to detect spam comments in YouTube comments. Another study by [13] performed an experiment using the standard Latent Ditichlet Allocation (LDA) approach to measure the degree of change in user's interest to detect human-like spammers. After generating a number of topic probabilities for each user, they calculate two topical features: Local Outlier, which captures the user's interest, and Global Outlier, which reveals user's interest in comparison with the interests of other users. The results of this study indicate that spam users either concentrate on certain topics or have interests in some topics. Similarly, legitimate users mainly focus on limited topics. The main limitation of this study, however, is that they did not provide a separation between legitimate and spam users who have focused on different topics or who have focused on certain topics; also, the topical features proposed by this study showed a low detection rate when we applied them without the integration of profile features.

Alternatively, [12] proposed a distinctive approach using retweeting behavior to discover anomalous topics among trending topics on Twitter. Their aim was to detect cooperative spammers who hijacked topics by analysing the change in the topology of characteristics of their retweeting networks. Another sophisticated approached offered by [10] to detect malicious messages is by inspecting the way in which the messages spread on online social networks. Nilizadeh et al [10] identified different communities that share similar topics of interest and inspected the dissemination path to predict the pattern of posting within and outside of the community in order to detect malicious messages. They argued that each community has normal messages between members within that community, reflected by intra-community communication and inter-community exchanges between structural communities, and malicious messages do not match these normal message patterns. Nevertheless, this approach is scalable and successfully detects spam messages; it involves multiple phases that make it complicated. The study would have been less complicated and more efficient if it had considered user's interest rather than community interest.

The limitations of the above approaches are that they do not address the critical issue, as they still characterize spammers with unified features, whereas spammers behave differently and should be represented by different features too. To address this problem, we propose a metric to describe changing patterns in user's interest and develop a detection method for spammers. In our present study, we extend the combination of content and profile features by considering user's interest as a novel way that is different from previous works.

III. Proposed Method

We propose a two-stage approach: unsupervised learning stage, and supervised learning stage. The components of our approach are shown in Fig. 1. Stage one has the following components :1)- Modelling users' interest, 2) topic-based features extraction, 3) clustering of focused and diverse users. Stage two consists of: 4) feature extraction for representing users in each cluster, 5) classification of spam and legitimate

users. The idea behind this approach is to be able to model focused and diverse users who usually behave differently. The crucial part of this approach is to extract appropriate features for clustering and for classification, especially the different features for the users in the two different clusters. We propose to model users' information interests using topics generated from users' posts by using topic modelling techniques. In the following sections, we explain each component in detail.

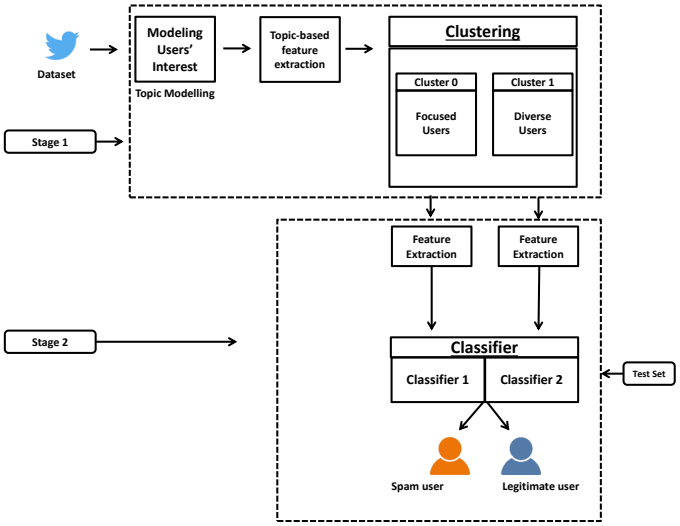

Fig. 1: Proposed Framework: Spam user detection through user's interests, employing unsupervised and supervised machine learning to classify spam users.

A. *Unsupervised Learning Stage*

1) Modelling user's interests based on LDA topic models: Latent Dirichlet Allocation (LDA) was first introduced by Blei [22] as an example of a topic model. Each document d_i is represented as a bag of words $W = \{w_1, w_{2.....}w_M\}$, and M is the number of words. Each word is attributable to one of the document's latent topics $Z = \{z_1, z_{2.....}z_k\}$, and k is the number of topics. φ_j is a multinomial distribution over words for topic z_j, $\varphi_j =< p(w_1|z_j), \dots\dots, p(w_M|z_j) >$, $\sum_{i=1}^{M} p(w_i|z_j) = 1$. φ_j is called the topic representation for topic z_j. θ_i is another multinomial distribution over topics for document d_i. $\theta_i =< p(z_1|d_i), p(z_2|d_i), \cdots, p(z_k|d_i) >$, and $p(z_j|d_i)$) indicates the proportion of topic z_j in document d_i. θ_i is called the topic distribution for document d_i.

We considered each user's tweets as one document. The document collection contains all users' tweets. The user's information interest is reflected in the tweet content, and we need to model the user's interest using LDA. So, we apply the LDA Topic Model to generate k topics for each user and get the topic probabilities for each single user. From these topic distribution values, we extract three topic-based features, which are discussed in next section, to measure the user's interest in order to distinguish between users who have focused interests and users who have diverse interests.

2) Topic-based features to depicting users' interest focus level: The separation of users based on interest concentration is motivated by the observation that users with focused interests should have different features from those who have diverse interests. In this paper, we propose to cluster users into two groups based on their content interest described by topic distribution generated from their tweets. Clustering in this research project is different from previous studies [8, 23], as previous studies used clustering techniques to group spammers into a cluster with similar spamming behavior, whereas in this research we utilize clustering to identify focused users and diverse users, and then we extract features that are more representative for each cluster for spammer detection. The following section details three topic-based features for clustering users as focused and diverse: Topic Entropy, Standard Deviation of Topic Distribution, and Local Outlier Standard Score.

We mentioned previously that user's interest is a reliable feature that is difficult for spammers to evade and that can therefore be used for detection. After generating topics from users' documents by using LDA, we used topic entropy to measure the diversity of topics for each user, using the following equation:

$$\mathrm{H(u)} = -\sum_{i=1}^{k} p(z_i|u) \log_2 p(z_i|u) \quad (1)$$

$p(z_i|u)$ is the topic distribution for user u. A user with low topic entropy is more likely concentrated meaning that the user is interested on limited topics, while a user with higher entropy is more likely to have wide interest spreading somewhat evenly over many topics. $H(u)$ can help us later in the clustering stage to get users with different levels of focus. The example in Table 1 shows two users with different values of entropy and topic distributions, where $k = 5$. The two users are from the Honeypot dataset [24] which is a popularly used labelled spam dataset for spam detection research. User 1 has higher entropy values comparing with User 2, and the topic distribution for this user shows that this user looks interested on many topics. User 2, however, **has** very lower entropy value, and the topic distribution shows that this user has very strictly focused on Topic 0, which can clearly indicate by the uneven topic probability distribution.

Standard deviation of topic distribution is also a good indicator for differentiating focused and diverse users in

TABLE I: Two users have different topic distributions, entropy and standard deviation.

	Topic Entropy	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Std
User 1	**1.604808**	0.203534	0.199607	0.187173	0.176047	0.23363	0.021694
User 2	0.015792	0.998186	0.000454	0.000454	0.000454	0.00045	**0.446199**

addition to the topic entropy. Using the standard deviation, we have a 'standard' way of knowing how spread out the topics are from the mean of a given user. This demonstrates the degree of change in topics for a particular user. From Table 1 we see that User 2 has a higher degree of deviation (Std) than User 1.

Local outlier standard score was first proposed by [13] to discriminate human-like spammers from legitimate users using topic distribution. Liu used this feature for the purpose of classifying spammers and legitimate users, but we use it for clustering users into focused and diverse groups. This feature measures the degree of interest of user in respect to a certain topic, using the following equation:

$$\mu(u_i) = \frac{\sum_{j=1}^{k} p(z_j|u_i)}{k}$$

$$LOSS(u_{il}) = \frac{x_{il} - \mu(x_i)}{\sqrt{\sum_{j=1}^{k}\left(x_{ij} - \mu(x_i)\right)^2}} \quad (2)$$

Where, $\mu(u_i)$ is the average interesting degree for all topics for a certain user. If we extract k topics for each user, we will end up with a vector of k features for each user, $LOSS(u_{i1}), \ldots\ldots, LOSS(u_{ik})$. In our experiment discussed in Section 4, we generate 5 topics for each user, we will have a vector of 5 LOSS features for each user.

In this paper, we propose to use topic entropy, standard deviation of topics distributions, and the vector of LOSS values as the features to cluster users. Next section will discuss the clustering process to cluster users. Since clustering is a typical type of unsupervised learning technique, this clustering process is considered as the unsupervised learning stage in the proposed spam detection approach.

3) User clusters with focused interest and diverse interest: The purpose of using topic-based features in this research is to model user's interest and then use this to identify two groups of users in terms of their interest concentration: focused user (who mainly is interested in a few topics) and diverse user (who have a wide range of interests).

A common opinion is that spammers do not have clear information interest and thus their tweets involve a wide range of topics meaning that they show diverse information interest. Some existing classification-based detection methods [6, 17, 25] use the number of hashtags as a feature to classify spammers from legitimate users because it is considered that spammers use more hashtags than legitimate users. However, in reality, some spammers could be focused such as content polluters for promoting some specific commercial product. To get an idea of the importance of level of interest concentration for the purpose of spam detection, we provide in Table 2 an example of two spam users having different levels of interest concentration. For example, User 1 shows one topic of interest, which is "American Football" across all tweets, and a link associated with each tweet. User 2 on the other hand, has diver interest with @mention and hashtags in most of the tweets. Both users are spammers, but they show different features and behavior, and if we consider, for example, the number of @mention or the number of hashtags as features to differentiate them, we will have an error of misclassification, because these features may not be applicable to classify User 1 as a spammer.

It would be ineffective to look for unified features of spammers to detect smart spammers, and it would be more useful to analyze them from the interest-level perspective to extract the most effective features to detect spammers properly.

TABLE II: Example of two spam users with different messages and interest.

User 1	**"**American Football. NFC North Winner, Divisional Markets. Detroit Lions is decimal odds of 18.5 to win. [_RUL]" "American Football. AFC North Winner, Divisional Markets. Cincinnati "Bengals is decimal odds of 4.1 to win. [_RUL]" "American Football. AFC South Winner, Divisional Markets. Houston Texans is decimal odds of 5.5 to win. [_RUL]" "American Football. Super Bowl Winner, NFL Season 2010-11. New "Orleans Saints is decimal odds of 13.0 to win. [_RUL]" "American Football. AFC Conference Winner, Conference Markets. Baltimore Ravens is decimal odds of 8.4 to win. [_RUL]"
User 2	"RT @beisick306: Custom @Lemarvelous23 #NTD #S-10 #calgarystampeders [_RUL]" "@yasminhasann6 I'm so Wavesy [_RUL]" "When u combine wine and dinner the new word is winner" "RT @DRUGRANGE:DRU - Don't Be Afraid Teaser [_RUL] via @YouTube" "RT @DRUGRANGE: #NowPlaying [_RUL]" "I don't want all these other apps to have snapchat, too much stuff"

Therefore, it is desirable for this research to establish a way to determine focused users and diver users first before classifying spammers from legitimate users. We want to examine the effects of using multiple topics generated from each user's tweets and then quantify the change in these topics with different degree assessments to demonstrate that such assessments may reveal an important difference between focused-interest and diverse-interest users.

Based on our observation we find that any of the above three topic-based features alone is not sufficient in identifying focused and diverse users when we have somewhat low variance between topic probabilities. Therefore, we combine them together to construct a unified feature vector for each user. Formally, for each user u_i, let k be the number of topics, we can calculate a total of $N = k + 2$ topical features which include k $LOSS$ features $LOSS(u_{i1}), \ldots\ldots, LOSS(u_{ik})$, topic entropy $H(u_i)$ and standard deviation $Std(u_i)$. Each user is represented by a N-dimensional feature vector $V_i =< v_{i,1} \ldots\ldots v_{i,N} >$.

From the tweet dataset Honeypot [24], we generate a topic model with k = 5, then generate the topical features based on the topic model for each user in the dataset. By applying a clustering method, we generate two clusters based on the topical features. The results in Figure 2 indicate that both clusters contain spammers and non-spammers. Table 3 shows the average values of each feature over the users in each cluster.

TABLE III: Average values of each feature over the users in each cluster (K=5).

Attribute	Cluster0	Cluster1
Topic 0 LOSS	0.1126	0.1774
Topic1 LOSS	0.7051	0.2255
Topic 2 LOSS	0.0617	0.1644
Topic 3 LOSS	0.0247	0.2092
Topic 4 LOSS	-0.2888	-0.3005
Std of topics dis	0.2987	0.2321
Topic Entropy	0.8468	1.0584

Based on the average feature values, we can decide that the users in Cluster 0 are more focused than the users in Cluster 1. This is because Cluster 0's topic LOSS distribution is much uneven than that of Cluster 1, and Cluster 0's topic entropy is less than that of cluster 1, where Standard deviation of cluster 0 is higher than that of cluster 1. All these comparisons indicate that Cluster 0 contains focused users while Cluster 1 contains diverse users. Table 3 shows the overall average values for each feature of the clustering output. The concentricity and diversity of the two clusters are further discussed in the next subsection.

The size of a cluster is the number of users in the cluster. According to our observation of the clusters, we have 2540 users in Cluster 0, with a total of 2263 legitimate users and 287 spam users, which shows that the distribution of legitimate and spam users is unbalanced with 94% of the users in Cluster 0 being legitimate and only 6% being spammers. Even the number of spammers is low, but it also shows that spammers can be of focused.

In contrast, Cluster 1 is a balanced cluster, which composes of 2891 spam users and 3611 legitimate users as showed in Fig. 2, which indicates that a diverse user could be a legitimate user or a spammer with similar probability. The result in Cluster 1 shows that legitimate users tend to have diverse interests. For the spam users in this cluster, they are compromised accounts or fake accounts that randomly try to mimic legitimate account behavior to avoid detection by Twitter. As this study set out with the aim of assessing the importance of focused level, we want to distinguish between spammer and legitimate users for both groups. We hope that we can represent each group with divergent features and then utilize this to classify spammers and legitimate users.

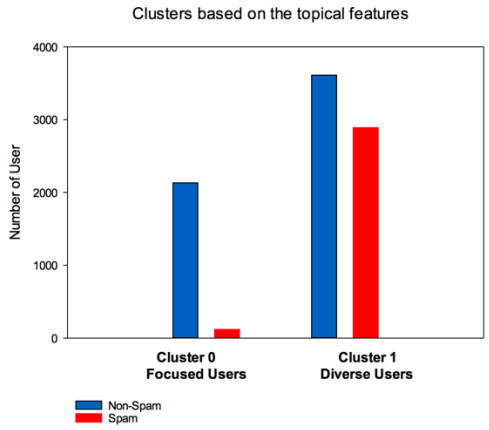

Fig. 2: Two clusters of users based on topical features.

Topic entropy can help to determine how focused a user is on different topics. In order to make this feature easier to understand, we presented early two distinct users in Table 1. We can see that User 1 has high entropy values, and the document topic distributions for this user show that this user does not have focused topics, whereas User 2 has very low entropy value, and document topic distributions show that this user has a very strict focus on Topic 0. For the two clusters generated from the Honeypot dataset, as showed in Table 3, the average value of entropy for Cluster 0 is 0.8468 and 1.0584 for Cluster 1, which indicates that the users in Cluster 0 are more focused than the users in Cluster 1.

The standard deviation of topic distribution shows how much the topic of a given user differs from the mean value for the other topics for the user. This standard deviation with topic entropy can measure the change pattern of user's topics. Fig. 3

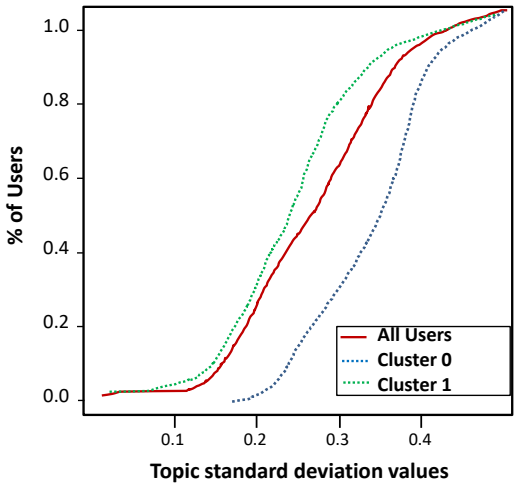

Fig. 1: Cumulative distribution function of standard deviation of topic distribution for each cluster

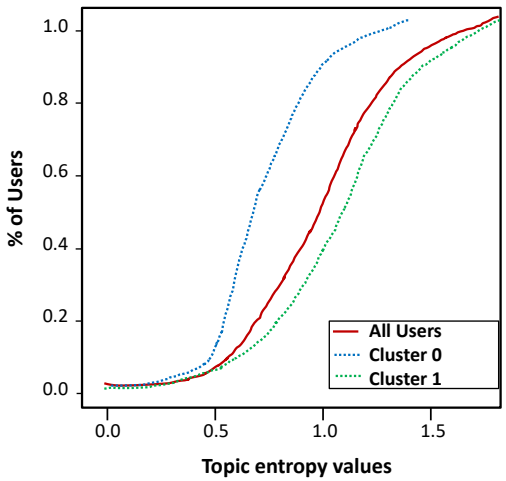

Fig. 2: Cumulative distribution function of topic entropy for each cluster

and Fig. 4 show the cumulative distribution function (CDF) of standard deviation and topic entropy for both Cluster 0 and Cluster 1. From the values in both figures, we can see that for the same percentage of users, the topic standard deviation of Cluster 1 is always smaller than that of Cluster 0, and the topic entropy of Cluster 1 is always larger than that of Cluster 0. Therefore, Cluster 1 contains diverse users and Cluster 0 contains focused users. The focused users have higher standard deviation values than those who are diverse as Fig 3 shows.

For LOSS feature, we use this feature to measure the degree

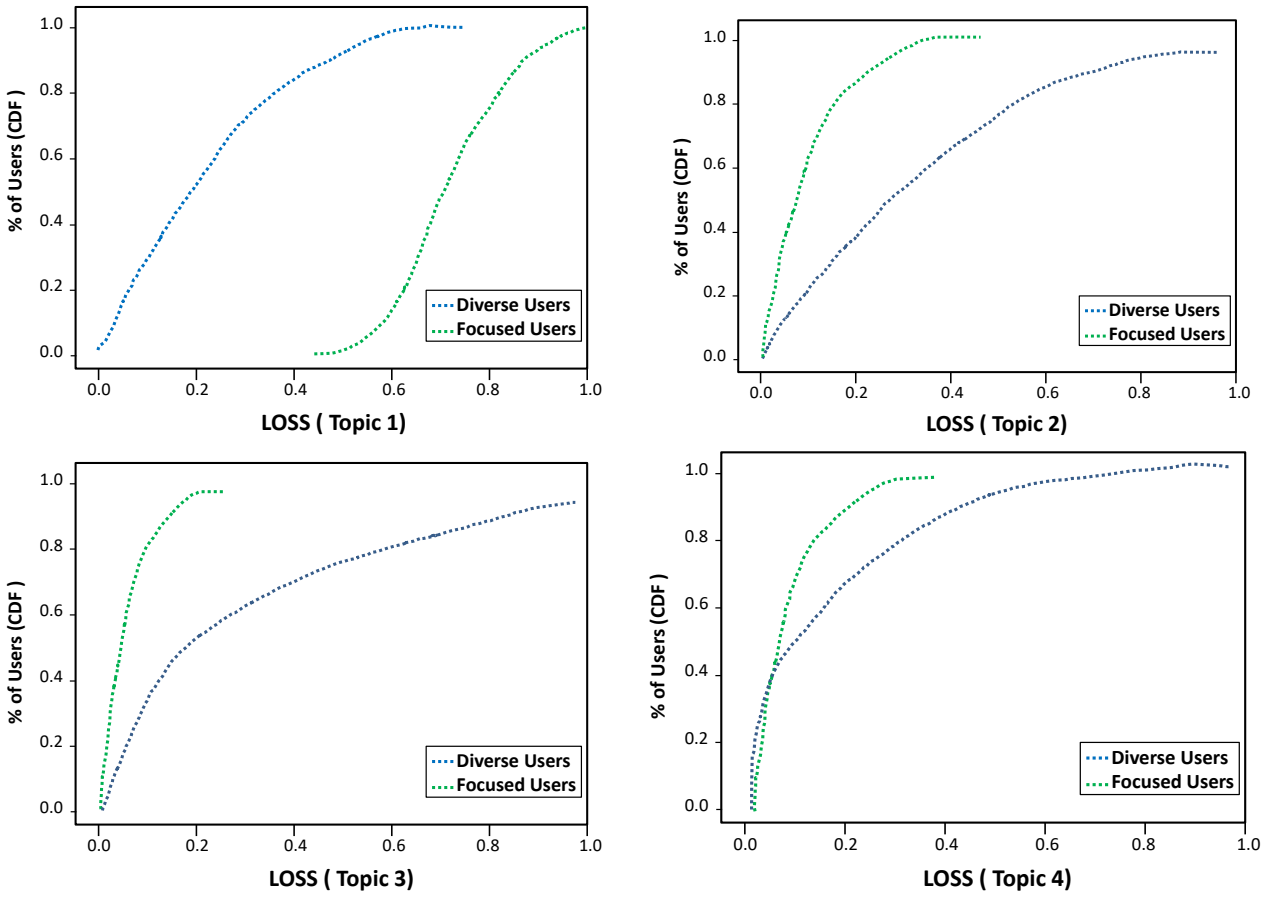

Fig. 3: Cumulative distribution function (CDF) for LOSS features grouped by cluster type.

of user's interest with respect to the 5 topics that we generated for each user. Fig 5 shows cumulative distribution function (CDF) for LOSS features for each cluster. It is clear that LOSS is also able to discriminate the two type of users, focused and diverse interest.

A. *Supervised learning stage*

While clustering provides a division of the observed topics and divides users based on focused and diverse interests, classification will be used to identify spammers in each of the clusters. As mentioned above, users in both groups, i.e., focused-interest and diverse-interest, can be spammers or legitimate users. For differentiating spammers from legitimate users, in this paper we propose new features based on users' self-description which describes what they are interested and their post content. We also use other existing features to represent each group.

1) *Features for classification:* As described in previous sections, the first stage of our proposed approach is to divide users into two clusters, focused-interest and diverse-interest, based on users' topic distribution. In the classification stage, we proposed a novel feature to differentiate spammers from legitimate users based on users' self-description and their posts content. This feature represents the consistence between a user's self-description and his/her posts content. For the classification, we also use some of the existing features proposed by different researches, which are listed in Table 4.

In social networks such as Twitter, users are allowed to describe their interest in the description statement contained in their profile. The users' descriptions provide explicit information about their interests, whereas the users' tweets reveal their interest implicitly. As the aim of this study is to detect spammers, we try to understand spammers' behavior through explicit and implicit behavior from user's interest point of view. The comparison between a user's description and his/her tweets can help in understanding their behavior. Therefore, the user can be assessed by analyzing the user's description in relation to the tweets that they have posted. The description in a user's profile in Twitter is called a Bio, which gives the users up to 160 text characters to tell the others about themselves, plus 30 bonus characters for location as well as an opportunity to give a backlink to their own site. This field contains a few simple sentences describing the user's interest to give people a first impression about a user, and many attackers use it to attract more followers. All existing studies so far, however, failed to take the content of Bio into consideration for detecting spammers, which we cover in this research.

The basic assumption about the user's description is that a user's self-description is generally consistent with the content of the user's posts. It is worth to mention that this assumption doesn't mean that a user's self-description should match every single tweet he/she posts. For example, if a user's description states that she/he is a "specialist in children with special needs", this user should have a large number of tweets that relate to the topic of "children with special needs", but every single tweet does not have to necessarily match with the description.

We propose a new feature that is the similarity between a user's self-description (i.e., user Bio) and the user's posts or messages such as tweets. We propose this feature to find the relationship between users' explicit statement in their profile and implicit behavior in their tweet content. This feature reveals how consistent the users' self-stated interest is with the interest showed in their tweets. Spammers often create a fake account or use a compromised account and they try to mimic legitimate behavior to avoid detection. However, because spammers aim to spread unsolicited or harmful messages to as many users as they can, very often the content of the messages does not match their self-description. The primary hypothesis of this feature was described earlier in Assumption 2 that the integration of both post and profile features are effective to properly understand users' behavior. We believe that this behavior of inconsistent interest must be reflected in the evolution pattern of the tweets content and could help detecting spammers.

Cosine similarity can be used to measure the similarity between a user's self-description and the tweet content represented as vectors. Given two vectors and the cosine similarity is calculated as follows, where A and B are vectors representing the user's self-description and the tweet content:

$$Sim(A,B) = cos(\theta) = \frac{A * B}{\|A\|\|B\|}$$

We generated a vector for the self-description of each user, and a vector that represents each of the user's tweets. We used term frequency–inverse document frequency TF-IDF values to produce these vectors, where each tweet and each user's self-description is treated as a document. We then calculated the similarities between a user's description and each of the user's tweets and got the average similarity to represent the interest consistency of this user. Usually most of a legitimate user's tweets are relevant to his/her interest described in his/her profile. However, this type of behavior is not always pretested in spammer behavior because most spammers do not have clear information interest, or they use compromised accounts. For this feature, legitimate users showed that averagely the content of their posts has relatively higher similarity to their self-description in comparison with that of spammers. The following existing features proposed in [6, 8, 13, 17] are also chosen in the classification stage.

2) Number of unique words: It has been proved that legitimate normal accounts are more innovative in their use of language, while spammers may repeat themselves more often, since they usually have a specific agenda or target to achieve [14]. The unique words feature can reflect the innovative pattern of using language for a particular user, but it is important to note that this is not applicable to all spammers. In our dataset,

we have found that a number of legitimate users in Cluster 0 (focused users) somehow exhibit similar use of the same word and do not post a large number of unique words. This is why we cluster users into two different groups based on the topical features to characterize each cluster from topics point of view. The number of unique words is more effective for diverse users than for focused one, because users are interested in different topics and the use of unique words can differentiate spammers from non-spam users. For the focused user, however, this feature is not applicable, as all users in this group have almost the same behavior of using words and posting similar tweets most of the time. Although users in the diverse-interest group have different interests and normally use new words, spammers have limitations in the use of unique words, which is a significant feature when distinguishing spammers from legitimate users in this group.

2) Average count of "@username" per tweet: The insertion of @username is essentially used to deliver the tweet to the username's account, even if the user has no relationship with the intended target. This is very common behavior by spammers and has been examined in previous studies [24, 26]. Users with this type of behavior use @username in order to attract new followers or to harm the user. In all cases, the use of @username is found to be a good feature for diverse- users, to discriminate spam from non-spam users. The reason for this is that users with focused-interest generally exhibit similar behavior of posting similar content and posting @username very often. [9] in their research, categorized harvested spammers into different groups such as duplicate group and promoters group, both of which make use of @ very often in their tweets. However, we found that most focused users (both spam and non-spam) have somehow similar behavior in this regard with relatively small notable differences between them, and this feature is more useful for diverse-users than focused-users.

3) Number of links: This feature is very similar to the use of @username for both groups. Among focused users, spammers and non-spammers post a large number of links to target users. If we consider this feature as a unified feature for spam detection, we would have misclassification of legitimate users who have focused-interest. We have noticed that there are a number of bots among focused users that post the same content with links that take users to disreputable web pages such as phishing sites or drag sellers. In contrast, diverse-interest users vary in term of using links, and spammers show a higher usage of links in their tweet than normal users do, and this feature is more applicable in diverse- users. We may also attribute the high usage of links among diverse-interest users to the idea that spammers exploit reputable accounts and seek to harm existing followers that already have a trusted relationship with the owner of the account.

4) Number of Following and followers: Number of following as feature can be used in both clusters, focused-interest and diverse-interest. The number of Following is abused by spammers to gain access to many targeted users. This behavior is a common characteristic of spammers and has been extensively used for spammer detection by [9, 10, 15, 27] and [11]. It is worth mentioning that the number of followers as a feature is not effective for in diverse users, whereas this feature has a high contribution to the focused users. As one of the contributions from this paper is to demonstrate that both groups, focused and diverse, are characterized by different features, the number of followers is not a significant feature in diverse-interest users. Spam users in the diverse group are hidden as legitimate accounts that have a good reputation and that do not seek to have more followers in order to appear as legitimate accounts. Interestingly, this feature is significant in focused users, as spammers tend to appear as legitimate users, and they use third parties to get more followers as Lee [9] reported that the number of following and followers fluctuated significantly over the time of the spam users. They lose or gain followers quickly, and this can be reflected in our findings that most of this type of behavior is among focused users, not diverse users.

In addition to this feature, three other features closely linked with the following and follower features are also significant for both groups, which are standard deviation of following, ratio of following and followers, and change rate of following [9]. These are temporal features that show how often a user follows others. In this paper, we confirm that these features are applicable to both focused and diverse users, with the absence of followers as a single feature for diverse users.

In summary, the approach of having two different clusters based on the level of focus interest suggested that characterizing spammers with distinct features for users in different clusters is the key point of detecting spammers with higher detection rate. The effectiveness of using different features for different clusters is demonstrated with some example comparisons showed in Fig 6. The top two figures show the comparison of the feature @username for the focused and diverse clusters using the cumulative distribution function (CDF) of the features representing focused users in left figure and diverse users in the right. We can see from Fig. 6_(a) that, for focused users, the feature value of spammers is lower than that of normal users for some of the users, but higher for some other users, indicating that the feature @username is not consistent over all users and thus is not effective for focused users. However, for diverse users, the feature value of spammers is consistently smaller than that of normal users, indicating that this feature can be used to differentiate spammers from normal users for diverse users.

The unique word feature show differences to the two clusters in the amount of unique words used by both groups. Although this feature shows that normal users are more innovative in their use of language [28] comparing with spam users in both clusters, it is more effective with diverse users. With the nature of diversity in the user's interest for diverse groups, there will be more unique words in their content, however spammers in this group still show less ability to use new words. For the focused cluster the number of spammers is much smaller than that of normal users, which might be why the value of normal users is almost the same as the average since normal users dominate. The figure shows that the unique word feature can be used for both clusters, but it is more effective for diverse group than focused group. The CDF curves of these sample features have proven the assumption that the focus level of user's interest

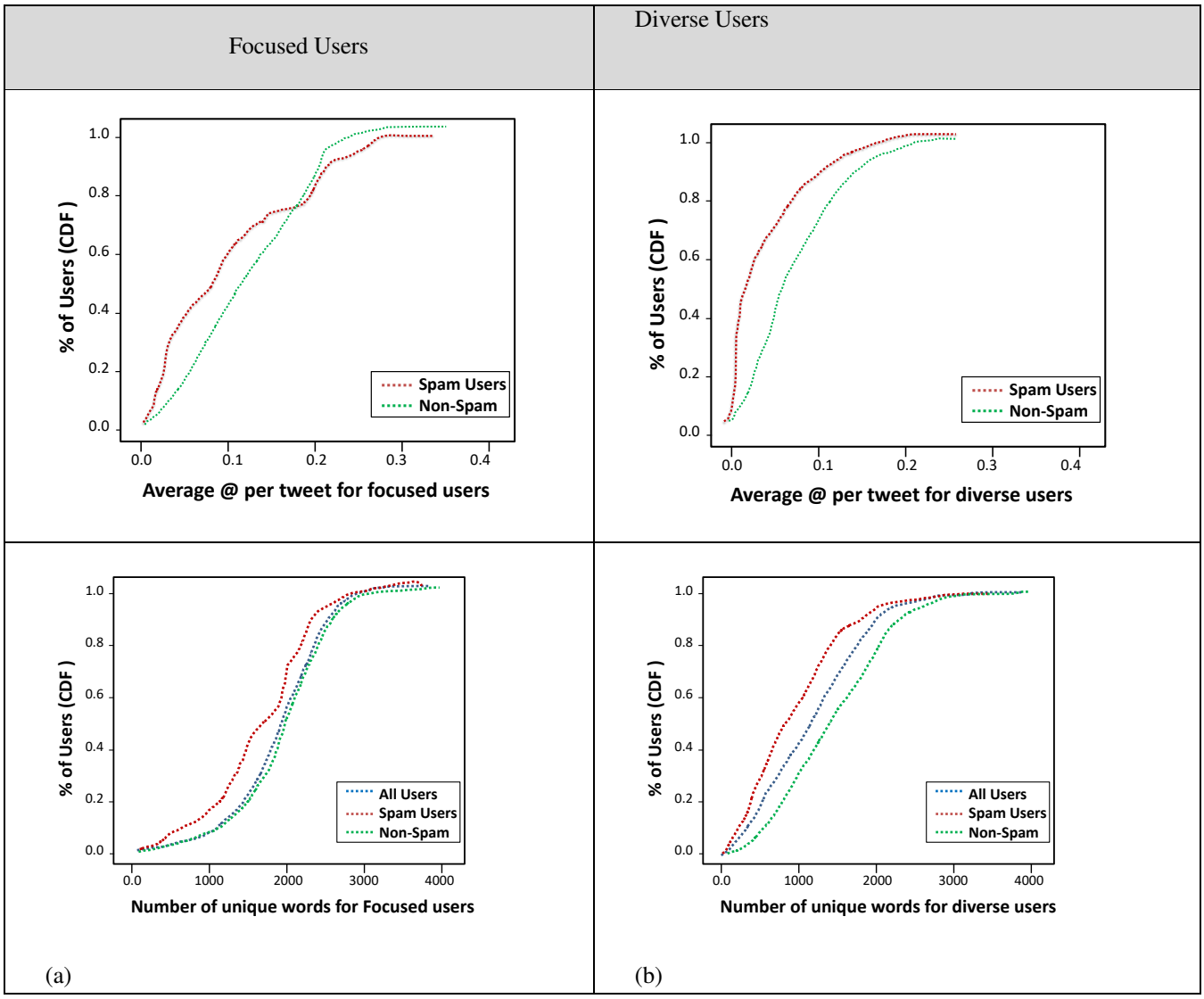

Fig. 4: The comparison between focused users and diverse users in terms of the content features

plays a central role in characterizing spammer behavior and content features are not effective in the focused group.

5) Feature selection: For our approach, an important task is to select the most effective features for each cluster. Here, we select features that best classify users using the correlation-based feature (CFS) algorithm [29]. Then a supervised machine learning module is adapted to train a classifier that is used to make a decision on each user in the testing dataset on whether the user is a spammer or a legitimate user. Table 4 shows the selected features for classification from existing features and our proposed features. We organized features into four categories: *content features*, such as unique words, number of links and number of @ signs; user *demographic features*, such as number of followers and number following; topical features such as LOSS features that has been used for clustering stage; and our proposed feature, the consistency of user interest. We assume that some features are not suitable for focused users while they are effective for diverse users. By using the CFS algorithm, the most effective features for each cluster are selected. This algorithm evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The subsets of features that are highly correlated with the class while having low inter-correlation with class are preferred [29]. Irrelevant features should be ignored because they will have low correlation with the class. The CFS's feature subset evaluation function is as follows:

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (3)$$

Where Ms is the correlation between the class and the features in feature subset s containing k features. $\overline{r_{cf}}$ is the mean feature-to-class correlation over the features in s (i.e., $f \in s$, c is a class), the higher the better. $\overline{r_{ff}}$ is the average inter-correlation between features in s, the lower the better. Overall, the higher the Ms, the better the feature set s is selected. We used this evolution algorithm to select the best features that have high correlation with the class and low inter-correlation between the features in each cluster. _ Since the aim of clustering is to divide users into different groups according to their content interest, the clustering stage was designed based on topical features derived from tweets content. As a result, we expect the features for each cluster, especially the content features, to be potentially different, and we also expect there might be overlaps in user demographic features.

Table 5 shows the chosen features for representing the users in each of the two clusters. From Table 5 we can see that both content features and demographic features are chosen to represent the diverse users in cluster 1, which means that both types of features are effective for diverse users. However, for the focused users in cluster 0, only user demographic features are chosen and none of the content features seems effective to be used to represent focused users. Both clusters contain spammers and legitimate users as well. This confirms our

TABLE IV: Features selected for each cluster using the CFS filtering algorithm.

Reference	Category	Features	Cluster 0 (focused users)	Cluster 1 (Diverse users)
[6] [8, 24, 17]	**Content features**	Num of hashtag	—	—
		Num unique word	_	✓
		Num links	—	✓
		Num unique links	—	—
		Num of at@	—	—
		Num of unique at@	—	—
		Aver links/tweet	—	—
		Aver unique link/tweet	—	—
		Aver at@ per/tweet	—	✓
		Aver unique at@/tweet	—	—
[17] [24]	**Demographic Features**	Num of followers	✓	—
		Num of followings	✓	✓
		len about me	—	—
		len username	—	—
		Std following	✓	✓
		Ratio following & followers	✓	✓
		Change rate of following	✓	✓
Our proposed features	**Consistency of User's interest**	Max of Similarity	—	—
		Ave of Similarity	—	✓
		STD of Similarity	—	—
		Min of Similarity	—	—
Our proposed features	**Topical Features**	Topic Entropy	**Used for clustering stage**	
		Std of Topic Distributions		
[13]		LOSS 0		
		LOSS 1		
		LOSS 2		
		LOSS 3		
		LOSS 4		

Assumption 1 that spammers behave differently, and this results in distinct patterns and features. In general, our results indicate that the variances of features between different types of spammers are existing and need to be considered. In Fig. 6, we showed samples of features that are effective for cluster 1 but it is not suitable for cluster 0 conversely. [9] and [14] point out that the strength of classification lies mainly in the choice of features, and we try to model this phenomenon utilizing the level of focused interest in our current research project in order to detect spammers, with higher rates of detection.

TABLE V: The chosen features for the two clusters.

	Cluster 0 (Focused interest)	Cluster 1 (Diverse interest)
Demographic Features	Number of Followings. Number of Followers. Std of Following. Ratio of Following and Followers. Change rate of following.	Number of Followings. Std of following. Ratio of Following and Followers. Change rate of following.
Content features	None	*Number of unique words.* *Average at@ per tweet.* *Number of links.* *Average value of cosine similarity.*

As we have established the importance of correlating demographic and content features for spammer detection, separating users based on the level of interest diversity supports this association. Nevertheless, the selection of the most discriminative features is necessary in order to detect spam users. Demographic features play a key role in characterizing spammers' behavior for focused users, while on the other hand, content features in line with demographic features are more suitable for diverse users to uncover spammer behavior.

6) Spam detection by classification: Using the selected features for the focused user cluster and the diverse user cluster, a separate classifier can be constructed by applying a classification algorithm for each of the clusters, as shown in Fig. 1. The two classifiers together form an overall classifier which can be used to classify new users into spammers or legitimate users. Based on the similarity between a new user's feature vector and the centroid of the focused cluster and the centroid of the diverse cluster, the user can be considered as a focused user if he/she is more similar to the focused cluster, a diverse user otherwise. Then the corresponding classifier will be used to determine whether the user is a spammer or a legitimate user.

IV. Experiment and Evaluation

In this section, we first describe the implementation of our detection approach. We then introduce the dataset and the ground truth for evaluation. For the evaluation, we conduct several empirical studies to reveal the difference between spammers and legitimate users in terms of topical evolution patterns and some existing features. The results are found to conform to our assumptions. Finally, we evaluate the performance of our spammer detection method using the standard metrics.

A. *Overview*

Fig. 1 illustrates the framework of our proposed method. After we extract topical features of users' tweets content, we cluster users into two different groups, focused and diverse. Then we apply feature selection to select the most effective features for each cluster, as described in Section 3.A.5. We use these features to train our supervised learning classifier. Note that we use Weka machine learning framework [30] to conduct the experiment and evaluation. We use default values for parameters of the chosen clustering and classification methods, and 10-fold cross-validation where the original sample dataset is divided into 10 sub-sample sets, and 10 training and testing steps are performed. For the training, nine sub-sample sets are used, and the remaining sub-sample set is used for testing. The final evaluation result is the average of the 10 testing results.

B. *Data Set*

We chose the Honeypot dataset [24], which uses 60 honeypot accounts in Twitter to attract spammers and crawl any account that follows them. The data was collected from December 30, 2009 to August 2, 2010. We used the profile IDs in this dataset to crawl users' descriptions for each profile. It is worth mentioning that the dataset was reduced due to the limited number of users with descriptions in their profile or limited tweets that were not enough to understand the user's interest. We ended up with 9750 spam users, and 7167 legitimate users.

Before directly conducting the experiment on the employed dataset, we performed pre-processing steps. This involved deleting accounts that had few tweets, because a sufficient number of tweets are necessary to extract information on the user's interest and topics. Each of the remaining users has at least 20 tweets. We removed punctuations, stopwords and non-ASCII words and applied stemming. The ultimate dataset contained 5875 spam users with a total of one million tweets, and 3178 non-spam users with 572,040 tweets.

C. *Clustering*

In the first step of the experiment, we considered each user's tweets as one document and generate 5 topics from user's tweets document collection, then calculated topical features based on the topic model described in Section 3.A.2. We built the feature vectors with 7 features for clustering the users in our dataset. We used K-mean algorithm [31] and clustered users into two different groups. K-means is an iterative technique using a centroid-based method that takes the number of instances around which the clusters are built. Instances are

assigned to clusters based on similarities or distances. To evaluate the quality of the clustering result and verify that our clustering process can effectively divide focused and diverse users into different clusters, we collected statistics on the fractions of topic distributions. Differences between focused users and diverse users are shown using the topical features.

In general, our results indicate that the standard deviation of user's' topic distributions is higher for focused users than for diverse users because focused users have uneven topic distributions, while their topic entropy is low because of the same reason. Fig. 7 shows the comparison using the Honeypot dataset. From Fig. 7, we can see that the average topic probabilities (i.e., topic distribution) for focused users indicated in blue are very different, e.g., the probability of topic 1 is close to 1 and the other 4 topics have very small probabilities. In contrast, the probabilities of the 5 topics for diverse users indicated in red are very similar, all around 0.2. The mean value of topic entropy for focused users is found to be approximately 0.8468, whereas for diverse users it is 1.0584.

The decision to categorize a user as focused or diverse needs an effort and cannot be quantified easily. However, our proposed topical features provide quantity values which indicate that the topic interests of diverse users are more uncertain than those who are focused users because diverse users' topic probabilities are evenly distributed as Fig 7 shows. This behavioral difference between the two types of users is clearly represented by the quantity values in the topical features.

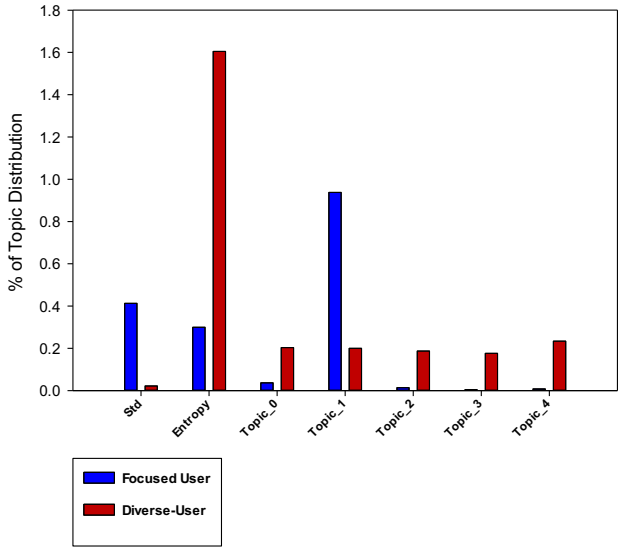

Fig. 7: Comparison between focused users and diverse users in terms of topic distribution, topic entropy and standard deviation of topics distributions.

In the clustering stage, three topical features are used: topic entropy, standard deviation of topic distribution and topic distribution. We observe that although users may have two topics of interest, their distributions can be entirely different and not close to each other. This type of user still has a concentration on one topic, with a notably higher value than for other topics. For example, user 2 in Table 6 has concentration around 71% on one particular topic, yet the user also shows interest in other topics, but with less concentration. Diverse users primarily have different topics of interest with different levels of focus, as user 4 shows in Table 6. We further calculated the topical change rate for each user obtained by topic distributions as follows:

$$topical\ change\ rate = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}|t_{i+1} - t_i|}$$

where n is the total number of topics, and ti is the topic distribution values. Most of focused users center on the vicinity of the average change value (i.e. 0.14), whereas diverse users are primarily distributed in a higher values (0.22) as boxplot shows in Fig 8.

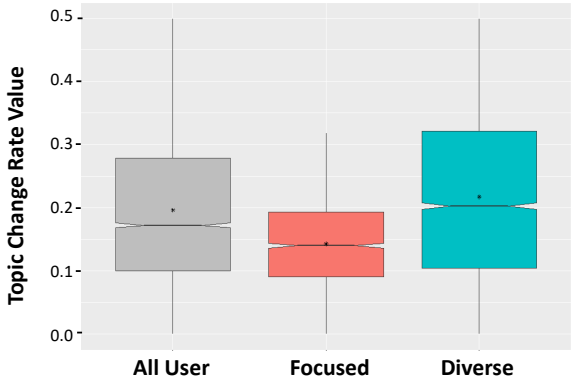

Fig. 8: Topic change rate grouped by clusters where diverse users experience higher change rate than focused users.

This verification experiment successfully reveals the difference between the two kinds of users in terms of focused and diverse interests. The result is roughly consistent with our assumptions and makes an excellent foundation for our subsequent experiment. By analyzing our clustering results, we conclude that our clustering-based features and process can distinguish focused and diverse users effectively.

V. Spam Detection Evaluation

In this section, we evaluate the performance of the proposed two-stage spammer detection approach and the contribution of the proposed two features. For the evaluation metrics, accuracy, precision, recall and F1-score are used to measure the performance. The metrics are defined below.

TABLE VI: Focused and diverse users with different values of topics distributions.

User	Std	Topic Entropy	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
Focused User 1	0.4446	0.0355	0.0023	0.0007	0.0007	0.9953	0.00077
Focused User 2	0.2985	0.8444	0.1994	0.7167	0.0412	0.0066	0.03590
Diverse User 3	0.1537	**1.3066**	0.1048	0.2951	0.0035	0.2017	0.39466
Diverse User 4	0.0216	**1.6048**	0.2035	0.1996	0.1871	0.1760	0.2336

Accuracy: it is one of the evaluation metrics for classification models, which is the total number of correct predictions divided by the number of users in the testing dataset:

$$Accuracy(\%) = \frac{TP + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} * 100$$

Where TP (True Positives) is the amount of correctly classified spam users, and FN (False Negatives) is the amount of spam users misclassified as legitimate users. FP (False Positives) is the amount of legitimate users incorrectly classified as spam users, and TN (True Negatives) is the number legitimate users correctly classified.

Precision: it is another metric used for evaluating classification model. It is the number of correctly classified spam users divided by the total number of users who are classified as spammers:

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

Recall: we measured the recall of the spam users, which is the number of correctly classified spam users, divided by the number of spam users in the testing dataset:

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

F-measure: is calculated based on precision and recall as follows:

$$F - measure = 2\ \frac{Precision * \mathrm{Recall}}{Precision + \mathrm{Recall}}$$

Four classification algorithms were used in the experiment, which are Support Vector Machine (SVM), J48, Decision tree and Random Forest. The evaluation results are given in Table 7, from which we can see that Random Forest achieves the best performance. This algorithm has also shown strong results in different spam detection researches [1] [32]. For comparison, we also show the detection performance using the random forest algorithm for each cluster as shown in Table 8.

TABLE VII: Comparisons of different classification algorithms.

Method	Precision	Recall	F1-Score	Accuracy
SVM	0.882	0.877	0.862	87.75%
J48	0.954	0.953	0.954	95.37%
Decision Tree	0.949	0.95	0.949	94.92%
Random Forest	**0.962**	**0.963**	**0.963**	**96.25%**

The results in Table 7 and Table 8 are obtained by using our two-stage approach. To evaluate the performance of the two-stage approach, we conducted another experiment which does not include the clustering stage. In this experiment, we classify users without clustering them in order to determine the detection rate without our proposed method of clustering users based on the topical features. We trained the data using the Random Forest algorithm as one group (without clustering stage) using the existing features and our proposed features, and we got an accuracy of 94.65% as shown in Table 9, which is worse than the accuracy 96.25% produced by using clustering. The results indicate that our proposed method performs well. It shows that spammers' behavior cannot be characterized with unified features, and the technique of grouping users based on

TABLE VIII: Detection result for each cluster using random forest algorithm.

Focused users			**Diverse user**		
Correctly Classified Instances	2469	**96.8235 %**	Correctly Classified Instances	6222	**95.6789 %**
Incorrectly Classified Instances	81	3.1765 %	Incorrectly Classified Instances	281	4.3211 %
Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
0.967	**0.968**	**0.968**	**0.957**	**0.957**	**0.957**

focused and diverse interest has confirmed that characterizing users in different groups with different features is crucial to detecting spammers with higher accuracy. This finding of the current study is consistent with that of [14] and [9], who found that the strength of classification relies mainly on the selection of the most appropriate features for spammer detection.

TABLE IX: Detection results with clustering stage and without clustering stage.

	Precision	Recall	F-Measure	Accuracy
With clustering Stage	**0.962**	**0.963**	**0.963**	**96.25%**
Without Clustering Stage	0.947	0.947	0.947	94.65%

A remarkable observation is that, on diverse group, spammers show similar behavior to that of legitimate users. This behavior of having different topics of interest is very common for a large number of legitimate users and spammer try to mimic this behavior by post different messages and topics. Our proposed method of measuring similarity between the user's description and tweets has uncovered this evasive behavior with less effort than existing approaches do [2, 18]. To test the effectiveness of this proposed feature, we train the data without this feature for diverse users and the detection result decreased by 1% as Table 10 shows.

TABLE X: The effectiveness of the Interest consistency feature for classification results.

	Precision	Recall	F-Measure	Accuracy
With our proposed feature	**0.957**	**0.957**	**0.957**	**95.67%**
Without the proposed feature	0.947	0.947	0.947	94.66%

Existing features shown in Table 4 were used to test the classification system. These features have been widely adapted in many previous methods. To further evaluate the effectiveness of the proposed feature, the methods proposed in [13, 24] are chosen as the baselines. We selected these baselines because they used some of the existing features and also used the same dataset as us. The comparison results are provided in Table 11.

VI. Discussion

The strong relationship between spammers' behavior and features has been described in the literature. However, most of these previous works use unified features to represent spammers, without considering that spammers behave differently and that this results in distinct patterns and features that need consideration. The present study aimed to integrate existing features and new features into a framework from the perspective of user level of interests in order to increase the level of detection and provide more reliable features that cannot be easily evaded by spammers.

We believe that our work provides good suggestions for micro-blogging systems to consider focused-interest and diverse-interest users. However, effective features need to be defined to determine focused-interest and diverse-interest users with reliable measurements. We used the LDA topic model to model user information interest using users' tweet content and then we applied cosine similarity between user's description and tweets to cluster users into two separate groups, i.e., focused and diverse users. However, more information about the user's interest can be used in addition to the user's description. Using heuristics, for example, the underlying page posted by the user or considering changes of profile description, would help to establish a greater understanding of users' interests.

The Twitter dataset has restrictions and imposes certain constraints on data collection. The size of the dataset has been slashed. We could not have accessed all Twitter accounts when we crawled the user descriptions. Also, users with insufficient numbers of tweets were excluded as it is difficult to understand user interests from a limited number of tweets. These restrictions may affect the quality of our approach, but our proposed approach performed well in detecting spam users. The experiment conducted in this research project considered the fact that spammers cannot be represented by constant features, and the level of focused interest enabled additional inferences about the effective features. The recommended method and the experiment conducted in this study to detect spammers presented the following strengths:

- This work considered user interests as playing a key role, since the engagement of users in any activity is driven by their interests; further, this feature is difficult for a spammer to manipulate, given that the behavior of spammers lacks a focused interest.
- The framework can handle smart spammers, given that spammers tend to set up fake accounts that appear to be legitimate or to compromise legitimate accounts to hide behind such account. However, the proposed method can identify this tricky behavior implicitly and explicitly through tweet content and profile description.

VII. Conclusion and Future Work

Due to the ability of spammers to use different strategies to evade detection, we conducted an extensive study of user

TABLE XI: The effectiveness of the Interest consistency feature for classification results.

Models	Precision	Recall	F1-Score	Accuracy
Reference [24]	Not provided	Not provided	0.888	88.98%
Reference [13]	0.895	0.951	0.922	Not provided
Our model	**0.962**	**0.963**	**0.963**	**96.25%**

interest evolution patterns. We propose a method to quantify changes in user interest and to depict user topic evolution patterns to understand the degree of focus interest. Based on the level of focus interest among users, we put forward a framework that combines the clustering algorithm with supervised machine learning to detect spammers in online social networks. Our experiment, based on a real-world dataset, reveals the differences between spammers and legitimate users in terms of focus level of interest and shows that user interest evolution patterns are indeed sufficient to represent and detect spammers with different features.

There are many potential directions for future work on this research project. It would be interesting to explore user interest in a dynamic way through different activities to characterize user interest evolution patterns comprehensively. In addition, our detection approach is offline, so it would also be interesting to extend it as an online real-time detection system that is deployed on online social network. Moreover, our proposed method includes a supervised learning stage that is needed to obtain labelled data to train the classification model. However, as it is hard to get enough labelled data because of certain factors, it would be good development to add the ability to perform detection without training data. The main idea of our proposed method is that it does not view social network spammers through constant behavior or represent them with constant features. If we model the user's behavior from a user interest perspective, then the differences between legitimate users and spammers become more evident. This is the most important feature of our work in the design of spammer detection systems in online social networks.

REFERENCES

[1] Benevenuto, F., et al. Detecting spammers on twitter. in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. 2010.

[2] Egele, M.S., Gianluca Kruegel, Christopher Vigna, Giovanni, COMPA: Detecting Compromised Accounts on Social Networks. *NDSS*. 2013, San Diego, CA United States: NDSS.

[3] Egele, M., et al., Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing*, 2017. 14(4): p. 447-460.

[4] Boshmaf, Y., et al., Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs. *Computers & Security*, 2016. 61: p. 142-168.

[5] Martinez-Romo, J. and L. Araujo, Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 2013. 40(8): p. 2992-3000.

[6] Sedhai, S. and A. Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2015. Santiago, Chile: ACM.

[7] Hua, W.Z., Yanqing. Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter. in *Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on*. 2013. Beijing, China: IEEE.

[8] Fu, Q., et al., Combating the evolving spammers in online social networks. *Computers & Security*, 2018. 72: p. 60-73.

[9] Lee, K., B.D. Eoff, and J. Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. in *International Conference on Weblogs and Social Media ICWSM*. 2011. AAAI.

[10] Nilizadeh, S., et al. POISED: Spotting Twitter Spam Off the Beaten Paths. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017. Dallas, Texas, USA: ACM.

[11] Shen, H., et al., Discovering Social Spammers from Multiple Views. *Neurocomputing*, 2016. 255: p. 49-57.

[12] Dang, Q., et al., Detecting cooperative and organized spammer groups in micro-blogging community. *Data Mining and Knowledge Discovery*, 2017. 31: p. 573–605.

[13] Liu, L., et al., Detecting" Smart" Spammers on Social Network: A Topic Model Approach. *arXiv preprint arXiv*:1604.08504, 2016.

[14] Alfifi, M. and J. Caverlee. Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter. in *International Conference on Social Informatics*. 2017. Cham: Springer

[15] Almaatouq, A., et al., If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 2016. 15(5): p. 475-491.

[16] Kaur, R., S. Singh, and H. Kumar, Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. *Journal of Network and Computer Applications*, 2018. 112: p. 53-88.

[17] Sedhai, S. and A. Sun, Semi-Supervised Spam Detection in Twitter Stream. *IEEE Transactions on Computational Social Systems*, 2018. 5(1): p. 169-175.

[18] Ruan, X., et al., Profiling Online Social Behaviors for Compromised Account Detection. *Information Forensics and Security, IEEE Transactions on*, 2016. 11(1): p. 176-187.

[19] Thomas, K.G., Chris Song, Dawn Paxson, Vern. Suspended accounts in retrospect: an analysis of twitter spam. in the *2011 ACM SIGCOMM conference on Internet measurement conference*. 2011. New York, USA: ACM.

[20] Zhu, Y., et al. Discovering Spammers in Social Networks. in *Twenty-Sixth AAAI Conference on Artificial Intelligence AAAI*. 2012.

[21] Song, L., R.Y. Lau, and C. Yin. Discriminative Topic Mining for Social Spam Detection. in *Pacific Asia Conference on Information Systems PACIS*. 2014. AIS Electronic Library.

[22] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. *Journal of machine Learning research*, 2003. 3(Jan): p. 993-1022.

[23] Gao, H., et al. Towards Online Spam Filtering in Social Networks. in *NDSS*. 2012. NDSS.

[24] Lee, K.C., James Webb, Steve. Uncovering social spammers: social honeypots+ machine learning. in the *33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010. New York, USA: ACM.

[25] Yang, C., R. Harkreader, and G. Gu, Empirical evaluation and new design for fighting evolving Twitter spammers. *Information Forensics and Security, IEEE Transactions on*, 2013. 8(8): p. 1280-1293.

[26] Chu, Z., I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. in *International Conference on Applied Cryptography and Network Security*. 2012. Springer.

[27] Shen, Y., et al. Automatic fake followers detection in Chinese micro-blogging system. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2014. Cham Switzerland: Springer.

[28] Alfifi, M. and J. Caverlee. *Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter*. 2017. Cham: Springer International Publishing.

[29] Hall, M.A., Correlation-based feature selection for machine learning, in Computer Science. 1999, Waikato: Hamilton, NewZealand. p. 171.

[30] Witten, I.H., et al., *Data Mining: Practical machine learning tools and techniques*. Fourth Edition ed. 2016, United States: Morgan Kaufmann.

[31] Arthur, D. and S. Vassilvitskii. k-means++: The advantages of careful seeding. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007. Society for Industrial and Applied Mathematics.

[32] Castillo, C., et al. Know your neighbors: Web spam detection using the web topology. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.

Recent Data Augmentation Techniques in Natural Language Processing: A Brief Survey

Lingling Xu, Haoran Xie, *Senior Member, IEEE*, Fu Lee Wang, *Senior Member, IEEE*, and Weiming Wang

***Abstract*—Data augmentation has recently gained increasing interest in natural language processing (NLP) because of its excellent performance in low-resource settings, contrastive learning, and few-shot learning. Data augmentation is initially a strategy to increase the amount of data by employing semantically invariant transformations, such as back translation and synonym replacement, on the raw data. With the development of data augmentation, a variety of augmentation strategies are designed to produce samples with opposite labels to the original data or even samples with unseen categories. In this paper, we provide a comprehensive and thorough study of text data augmentation techniques. We first discuss various data augmentation methods and then classify them into three types: semantic-invariant augmentation, random augmentation, and generative augmentation. Subsequently, we highlight the main application scenarios and downstream tasks involving data augmentation. We also describe the challenges in developing text data augmentations and the work that can be further investigated in the future. To conclude, this paper aims to summarize data augmentation techniques in NLP and show how they work to further improve the performance of NLP tasks.**

***Index Terms*—Data Augmentation, Contrastive Learning, Low-resource Setting, Few-shot Learning, NLP, Survey.**

I. Introduction

DATA augmentation works mainly by making small changes to the data directly or by generating new data using some deep learning models. Data augmentation is extremely important in low-resource scenarios in which the number of training data is sparse, as it helps increase the number of training data while reducing the operational costs of annotating. In addition, data augmentation can create diverse data and enrich the semantic feature space of data, further enhancing the robustness of model. Data augmentation first appeared in the field of computer vision (CV), where studies [1], [2], [3] discovered that cropping, rotation, and scaling of image data greatly improved model performance. However, it is challenging to employ these continuous noises for text data augmentation due to the discrete nature of the text.

Despite this limitation, data augmentation for NLP has seen an increase in interest and demand. Inspired by the data augmentation methods of cropping and rotation in CV, [4] proposes two text data augmentation strategies, sentence cropping and sentence rotating, based on the dependency tree structure of NLP. Sentence cropping preserves some important words in a sentence and then removes the rest of the irrelevant words to generate a new sentence. Sentence rotating rotates the portable tree segment around the root of the dependency tree to form a synthetic sentence. Besides, data augmentation techniques such as random deletion and token cutoff can also be seen as a variant of cropping in CV. Inspired by the Mixup [5] image data augmentation strategy in CV, SeqMix [6] and MixText [7] are then proposed for text data augmentation. SeqMix [6] attempts to incorporate word embeddings and sentence embeddings from the convolutional neural network (CNN) [8] to form novel samples, whereas MixText [7] combines sentence embeddings from the BERT [9] to obtain new synthetic samples.

To address various NLP tasks, a large number of text data augmentation methods have been devised, resulting in many surveys on data augmentation in NLP. [10] explores text data augmentation for deep learning, which includes not only data augmentation in NLP but also in recommender systems. [11] focuses on data augmentation techniques used in text classification. [12] provides a systematic and empirical investigation of data augmentation in NLP with a small amount of labeled data. Both [13] and [14] discuss NLP data augmentation methods. [13] does not contain data augmentation used in contrastive learning, while [14] does not discuss in detail to which NLP tasks data augmentation can be applied. Therefore, we take the data augmentation approaches used in contrastive learning into account and present the NLP tasks involving data augmentation in detail.

In this paper, we aim to provide a systematic investigation of text data augmentation in NLP according to the form of data augmentation. We discover that some data augmentations are well-designed using prior knowledge to enable the semantic meaning of augmented data to remain unchanged, while certain data augmentations focus on generating label-conditioned sentences. In addition, we also give the specific application scenarios of data augmentation and downstream tasks that involve data augmentation. The remaining paper is organized as follows. Section II discusses commonly used text data augmentation techniques and classifies them as semantic invariant augmentation, random augmentation, and generative augmentation. Section III describes the application scenarios of data augmentation, including low-resource language, contrastive learning, and few-shot learning. Section IV analyzes the downstream tasks that use data augmentation. Section V presents challenges and future work in data augmentation for

The research has been supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19), and Lam Woo Research Fund (LWP20019) and Direct Grant (DR23B2), Lingnan University, Hong Kong.

Lingling Xu, Weiming Wang, and Fu Lee Wang are with the School of Science and Technology, Hong Kong Metropolitan University, Hong Kong SAR.

Haoran Xie is with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong.

Fu Lee Wang is the corresponding author (email: pwang@hkmu.edu.hk).

NLP. Section VI concludes the paper.

II. Text Data Augmentation Methods and Techniques

Numerous data augmentation strategies have been proposed to promote the performance of NLP tasks, as they can both increase the quantity of data and enrich its diversity. In this survey, we focus on studying how augmented sentences are generated from original sentences. After summarizing these approaches, we observe that data augmentation is mainly performed by some well-designed transformations, stochastic change, and generative models. We divide the existing text data augmentation methods into three categories: semantic-invariant augmentation, random augmentation, and generative augmentation.

Semantic invariant augmentation is usually carefully designed and implemented by exploiting prior knowledge or deep learning models. Random augmentation, on the other hand, emphasizes the randomness of the generation of the augmented samples, so that the semantics of the augmented samples do not always remain the same as the original sentences. Generative augmentation is usually done by using generative models like VAE and BART to generate sentences that are consistent with the label on top of the original sentence and the given label. In the following work, we will discuss these text data augmentation methods in detail. Particularly, we provide an overview of current data augmentation techniques in Fig. 1.

A. Semantic-invariant Data Augmentation

Semantic-invariant augmentation is an augmentation strategy that preserves the syntax and semantics of the sentence via making well-designed local modifications to the original sentence. Paraphrases and well-designed substitutions are two common types of semantic-invariant augmentation.

Paraphrasing is widely applied as a text data augmentation strategy in NLP tasks [15], [16], [17], as it can provide augmented text with more varied lexical choices and syntactic structures while maintaining the semantic meaning of the raw sentence. Back-translation [18], [19] is definitely the most popular paraphrasing method, which involves translating the sentence into a certain intermediate language and then translating it back into the original language. Other research aims to train an end-to-end model to produce meaningful translations [20] and augment sentences at the decoding stage by adding syntactic features [21], latent variables [22], or submodular targets [17].

Well-designed substitution is also a common data augmentation method, where certain words in a sentence are replaced with other words without changing the semantics of the sentence. An intuitive idea is to use the synonyms as replacement words for substitutions [23]. The synonyms can be words from a pre-defined corpus such as WordNet [24], words with high similarity to the replacement word [25], entities of the same type [26], [27], or words with the same morphology [4]. Additionally, work from [4] argues that we can also keep the semantics of a sentence intact by removing words that are not important. Moreover, Xie et al. [15] devise a replacement approach based on TF-IDF where the uninformative words in the sentence are replaced with other uninformative words. Hsu et al. [28] substitute the unimportant words with the predicted words generated by the auto-encoding model or the seq2seq model without altering the aspect-level polarity. Notably, these semantic-invariant augmentation techniques we discussed are unsupervised data augmentation and do not use the label information of sentences. However, Wang et al. [29] attempt to substitute representative words with their corresponding antonyms to obtain new sentences, which may be semantically irrelevant or even opposite to the original sentence.

TABLE I: The examples of word-level random augmentations.

Method	Text
Original	There is a little boy running in the playground.
Deletion	There is a boy in the playground.
Swapping	There is little a boy running in playground the.
Insertion	There is **great** a little **dog** boy running in the playground.
Substitution	There is a **beautiful cat** running in the playground.
Repetition	There there is a little boy boy running in the playground.

B. Random Data Augmentation

Semantic-invariant augmentation is crucial for tasks that require augmented samples to have the same semantic label as the original sentences. Random data augmentation, on the other hand, has also received extensive research attention due to its ease of implementation. Furthermore, random data augmentation can be roughly divided into word-level, token-level, and embedding-level augmentation, depending on the body of the noise added to the sentence.

Word-level augmentation means that noise is added to the words of sentences, either by random deletion, swapping, insertion, and substitution [30], or random repetition for some selected words [31]. These stochastic operations are easy to implement and do not always ensure that the semantic labels of the text remain unchanged. We give some examples to show this word-level random augmentation in Table I. Token-level augmentation includes token shuffling (shuffles the order of tokens randomly), token cutoff (erases some tokens randomly), feature cutoff (erases feature dimensions randomly), and span cutoff (erases token spans randomly) [32]. AEDA [33] is another easier random augmentation that generates augmented samples by inserting punctuation marks randomly in the original sentence.

The embedding-level random augmentation can be mainly performed by Mixup [6], [7], [34] and adversarial training. Inspired by Mixup [5], a data augmentation method that linearly interpolates two input images to obtain a target sample. Guo et al. [6] apply this method to the domain of text and proposed SeqMix, which creates augmented sentences by interpolating word embeddings and sentence embeddings linearly with CNN [8] and LSTM [35] as sentence encoders. Similarly, Chen et al. [7] use BERT as an encoder to generate sentence embeddings for sentence Mixup, and Sun et al. [34] employ a pretrained transformer as an encoder to obtain sentence embeddings for linear interpolation, further demonstrating the effectiveness

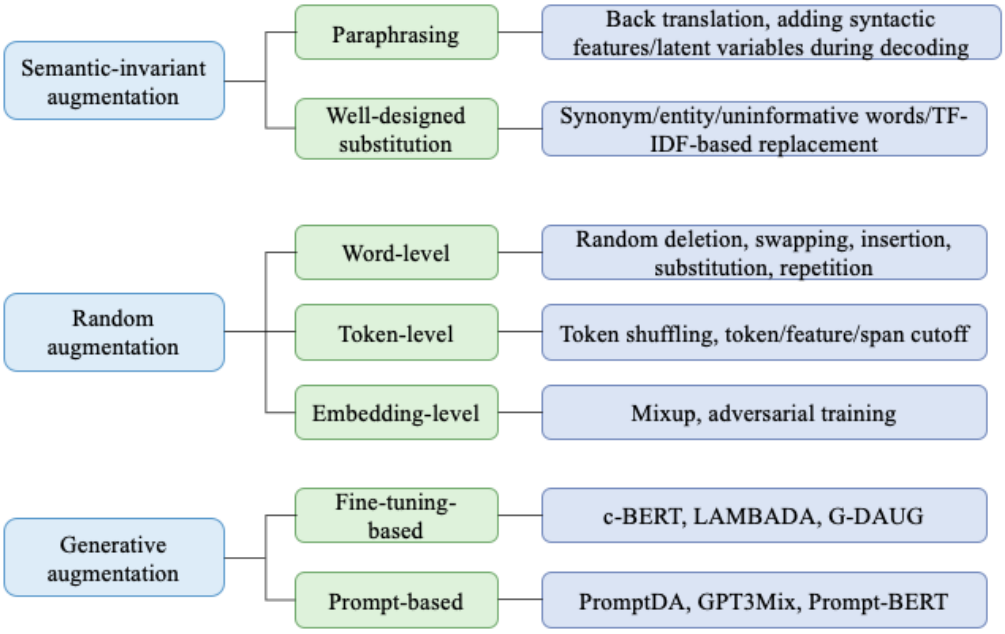

Fig. 1: An overview of recent data augmentation methods in NLP.

and generality of the Mixup augmentation in the text domain. Significantly, Mixup data augmentation requires label information to be known and is thus a supervised data augmentation technique. Assuming that the sentence embeddings of the two input sentences are e_i and e_j, and the labels are y_i and y_j, the augmented sentences and labels can be expressed as Equation 1:

$$\begin{aligned} e &= \lambda e_i + (1-\lambda)e_j, \\ y &= \lambda y_i + (1-\lambda)y_j, \end{aligned} \tag{1}$$

in which λ is sampled from the Beta distribution. Since the generated embeddings are a linear interpolation combination of two sentence embeddings, Mixup data augmentation can create semantically rich sentences. Additionally, the generated sentence labels vary because they are also an interpolation of two labels.

Adversarial training methods are commonly used to improve the robustness of models in text data [36], [37], [38]. It can also be used as a data augmentation technique to create adversarial examples using gradient-based noise. Specifically, for the input sentence embedding e_i with the label y_i, then the augmented sentence embeddings can be written as Equation 2:

$$e_i^* = e_i + \epsilon \frac{g}{\|g\|^2}, g = \nabla_{e_i} \mathcal{L}(f(e_i, y_i)), \tag{2}$$

where ϵ is random noise. Significantly, Mixup and adversarial training all require the participation of the label, thus they are supervised data augmentation approaches.

Moreover, dropout is another random augmentation method that is widely applied to contrastive learning [39], [40], which utilizes the dropout in the embedding layer and attention layer of BERT [9] to produce augmented samples. Concretely, a sentence is passed to the BERT encoder twice to obtain two different sentence representations.

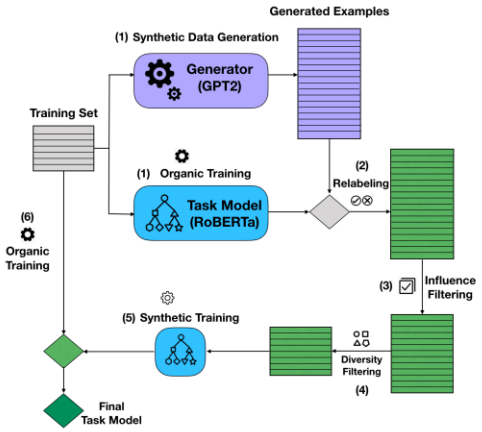

Fig. 2: The overview of data augmentation method G-DAUG^c [55].

C. Generative Data Augmentation

Deep generative models such as VAE [41], GAN [42], GPT-2 [43], GPT-3 [44], BART [45], and T5 [46] are employed in generative data augmentation methods to generate new sentences conditioned on the label. The semantic label of the augmented data obtained through generative data augmentation is thus determined by the given label and does not always maintain the same semantic label as the original data. Early generative data augmentation is typically performed on condition VAE [47], [48], [49], [50]; GANs [51], [52]; and a bidirectional RNN language model [53]. Furthermore, the benefits of developing pretrained language models (PLMs) [9], [54], two promising paradigms for data augmentation in NLP are proposed.

The first approach involves finetuning the PLMs using task-

specific data and then using the finetuned language model to generate new sentences. For example, [56], [57] use masked language modeling (MLM) mechanisms from BERT and BART, respectively, to produce new synthetic data by masking random words in the original sentences. Yang et al. [55] and Anaby-Tavor et al. [58] employ PLMs GPT-2 as the generator to capture the semantic information expressed implicitly in their training dataset to generate new synthetic sentences. With the help of pretrained language models, the novel data augmentation framework G-DAUG^c [55] (shown in Fig. 2) produces synthetic samples and chooses the most informative and varied samples for data augmentation. FliDA [59] generate augmented data using word substitution based on the pretrained T5, with a classifier to choose label-flipped data. C^{3}DA [60] adopts the T5 model as a text generator and produces new sentences based on given aspect words or sentiment labels (e.g., positive and negative) to enrich the dataset for aspect-based sentiment analysis. Despite these advancements, these PLMs could be overfitted with a small amount of task-specific data and fail to achieve excellent results.

The second type of approach utilizes the prompts, combined with the off-the-shelf PLM, to generate sentences directly without any task-specific fine-tuning. Wang et al. [61], for example, proposed PromDA, a data augmentation built on top of the T5-large model [46]. Specifically, PromDA keeps the parameters of the PLM frozen and trains only the soft prompt prepend at the beginning of the sentence, significantly reducing training resources. GPT3Mix [62] synthesizes hyper-realistic sentences from a variety of real samples by utilizing the large-scale language models of GPT-3 and the discrete prompt. Chen et al. [63] propose a label-guided data augmentation method that exploits the enriched label semantic information for data augmentation in a fashion similar to prompt-tuning. Liu et al. [64] devise a label-conditioned word substitution technique and a question-answering-based prompting approach for data augmentation. The label-conditioned technique aims to create a label-consistent example by capturing potential word-label dependencies, while the question-answering-based prompting approach focuses on generating new training data from unannotated text. Specific details of these two methods can be shown in Fig. 3. Moreover, Prompt-BERT [65] adds different discrete prompt templates to the same sentence and uses PLMs like BERT, RoBERTa to obtain different sentence representations to generate augmented examples.

III. Application Scenarios

In this section, we will discuss some application scenarios for data augmentation. Significantly, data augmentation mainly serves to raise the number of training data in low-resource scenarios, generate positive samples in contrastive learning, and synthesize unseen class samples in few-shot learning.

A. Low-resource Setting

Recent advances in large-scale neural language models [35], [9] have led to excellent performance in various NLP tasks, including machine translation [66], [67], [68] and NER [69],

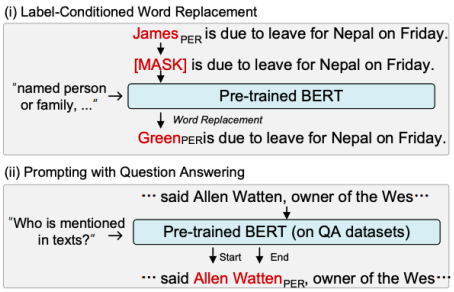

Fig. 3: Label-conditioned and prompting with question answering augmentation methods [64].

[70], [71], their accuracy largely depends on the accessibility of extensive sets of human-annotated training data. However, annotating data is time-consuming and expensive, thus data augmentation is extremely important in low-resource settings.

Inspired by the work in CV, Fadaee et al. [66] present a new data augmentation method that targets low-frequency words and produces novel sentence pairs with uncommon words in a novel synthetic context to enrich the training corpus for machine translation. Xia et al. [67] propose a generic data augmentation framework that generates parallel corpora via back-translating English to low-resource language or high-resource language as pivoting to a related high-resource language, improving the performance of low-resource translation. To improve the performance of low-resource machine translation, Li et al. [68] extend the training data via creating diverse pseudo-parallel data from the source and target sides.

Zhou et al. [71] employ masked entity language modeling (MELM) for data augmentation to obtain augmented data and alleviate the data scarcity in low-resource NER tasks. Liu et al. [70] use back translation to generate multilingual labeled data. These augmented data allow the NER model to learn the different linguistic features for cross-lingual NER tasks. For low-resource tagging tasks like NER and part-of-speech (POS) tagging, Ding et al. [69] develop a novel augmentation method using a language model trained on linearized labeled sentences to produce high-quality synthetic data. For low-resource natural language understanding (NLU) tasks, Wang et al. [61] develop synthetic data using generative augmentation techniques, reducing the effort for humans to annotate data while maintaining the quality of the synthetic data produced.

B. Contrastive Learning

Contrastive learning is a metric learning algorithm that learns useful representations by encouraging positive pairs to be closer and negative pairs to be further away. Positive pairs are typically different views of the anchor and can be generated by various data augmentations. Negative pairs are usually the remaining in-batch samples. As a result, data augmentation is essential to contrastive learning, and effective data augmentation will significantly enhance contrastive learning's performance. Notably, data augmentation can not only help generate positive pairs that share the same semantic label but also negative samples that are semantically dissimilar.

The most popular text data augmentation in contrastive learning for NLP is the dropout augmentation [72], [73], which is also the most recent state-of-the-art data augmentation strategy. Dropout augmentation utilizes the dropout in the embedding and attention layers of BERT to encode the same sentence twice to obtain two different sentence representations as positive pairs. Additionally, back translation [74], synonym replacement [29],token shuffle, and feature cutoff [32] are also used to produce positive pairs for sentence representation learning. By combining various data augmentations, Qu et al. [75] create diverse augmented examples, which are then combined with the contrastive learning objective to enhance NLU tasks. Wang et al. [60] employ generative data augmentation and contrastive learning to improve sentiment analysis.

In addition to generating positive samples, data augmentation can be used to create negative samples for contrastive learning. For example, CLINE [29] replaces words in sentences with antonyms to create negative samples for feature extraction with a triplet contrastive loss objective. MixCSE [76] produces hard negative samples by mixing the features of positive samples and negative samples randomly to further improve the performance of contrastive learning. To differentiate and uncouple semantic similarity from textual similarity, SNCSE [60] uses the Spacy[1] to perform sentence parsing to obtain the syntactic tree, lexical labels, and label stems of the sentence and then utilizes this information to transform the sentence into a syntactically correct and semantically-opposite sentence as soft negative samples. FlipDA [59] adopts generative data augmentation to generate a label-flipped augmented sample automatically, which can be considered negative samples of contrastive learning.

C. Few-shot Learning

Few-shot learning is a technique for extracting information from a small number of examples. Data augmentation techniques can assist few-shot learning by introducing different kinds of examples. Chao et al. [77] devise a novel data augmentation to address the problems of imbalanced data distribution and small samples of rare classes in few-shot learning. Arthaud et al. [78] use contextual augmentation to create new samples to train a pretrained machine translation model that can accurately translate previously unseen words on the basis of a few examples. Chen et al. [63] propose PromptDA generative augmentation to obtain multiple label words for few-shot text classification tasks. According to Wei et al. [79], data augmentation improves curriculum learning in triplet networks for few-shot text classification tasks. FlipDA [59] aims to produce label-flipped data as they found label-flipped data to be more effective than label-preserved data in enhancing the performance of few-shot learning.

IV. Downstream Tasks

In this section, we discuss some common NLP tasks involving data augmentation, i.e., sentence representation learning, text classification, question answering, and sequence tagging tasks.

[1]https://github.com/explosion/spaCy

A. Sentence Representation

Learning sentence representations has long been a fundamental and important research direction in NLP. Sentence representation aims to learn key semantic and syntactic information about sentences. Most existing work on sentence representation learning involving data augmentation is based on contrastive learning, a metric learning method that performs well in learning representations.

For example, [80] considers any two integrations of word deletion, span deletion, span swap, and synonym replacement to form a stronger augmentation for sentence representation learning. SimCSE [72] achieves excellent performance in the seven semantic textual similarity (STS) tasks using dropout augmentation. ESimCSE [31] argues that all the positive sentence embeddings constructed by SimCSE have the same length, which may mislead the model into viewing this as a distinctive feature to differentiate positives from negative instances. To address this issue, they propose a novel data augmentation method, word repetition, along with dropout augmentation, to improve the performance of sentence learning. ConSERT [32] selects randomly two data augmentation approaches for contrastive representation learning: token shuffling, token cutoff, feature cutoff, and dropout [81], with token shuffling and feature cutoff yielding the best results for positive pairs.

B. Text Classification

Text classification is the most simple and fundamental NLP task. It aims to train a text classifier that can automatically analyze text and then assign a predefined label based on the content of the text. Text classification covers a wide range of tasks such as sentiment analysis, topic detection, text matching, etc. Simple EDA augmentation [30], and AEDA augmentation [33] can both be used to produce augmented samples to improve the performance of text classification. [28] substitute the unimportant words with the predicted words generated by Auto-Encoding model or Seq2Seq model without altering the aspect-level polarity for data augmentation to improve aspect-based sentiment analysis.

For few-shot text classification, [82] investigates data augmentation methods that work in the feature space and combine supervised and unsupervised representation learning methods to improve classification performance. MEDA [83] is proposed based on meta-learning, this data augmentation framework is made up of one ball generator and one meta-learner, with the ball generator being used to increase the amount of shots per class via producing more examples, allowing the meta-learner to be trained with both original and augmented examples. Experimental results show that MEDA greatly improves the performance of meta-learning in the classification of a small number of texts.

For contrastive text classification, [60] proposes cross-channel data augmentation to raise the number of training samples and also to provide more diverse samples with multi-aspects. It employs contrastive learning to learn and capture the sentiment representations of various aspects to improve the performance of aspect-based sentiment analysis.

C. Question Answering

Question answering is the task of providing appropriate answers to given questions. It retrieves the answers to questions from a given text, which is very useful for searching for answers in documents. [84] demonstrates that the SQuAD benchmarks for reading comprehension significantly improve when contextual paraphrases are produced through back translation. [85] explores back translation based on query and context paraphrases for domain-agnostic question answering. [86] centers on data augmentation using distant supervision techniques to construct datasets that more closely resemble the types of passages readers see when reasoning to address open domain question answering. [87] propose XLDA, a cross-lingual data augmentation technique that enhances the performance of model on the SQuAD question answering task by substituting a section of the input text with the translation in different language. [88] uses labeled training data, in conjunction with logical and linguistic knowledge for augmentation, significantly improving a range of question-answering tasks. In order to improve zero-shot cross-lingual question answering, [89] makes use of question generation models to generate samples in other languages. While [90] employs back translation to convert question-answer pairings into multiple different languages to enhance the performance of cross-lingual open-retrieval question answering.

D. Sequence Tagging

Sequence tagging is a problem where the model sees a sequence of words or tokens and is expected to output a tag for each word in the sequence. To put it another way, the model is anticipated to tag the entire sequence with a suitable tag drawn from a pre-existing tag dictionary. Applications of sequence tagging in NLU include named entity recognition (NER) and part of speech (POS) tagging. NER is an information extraction technique designed to identify named entities in a given sequence of text tokens (words). POS tagging is a text data processing technique that tags words in a sentence with proper POS based on their semantic and contextual content.

Sahin et al. [4] use sentence cropping and sentence rotating to generate synthetic data for POS tagging. [69] leverage the generative augmentation with LSTM as a sentence generator on the given label for NER and POS tagging. [27] employs label-wise token replacement and synonym replacement for NER. With the help of MELM, [71] creates novel entities in high-quality augmented data, enhancing NER performance by supplying rich entity regularity knowledge. [70] translates the training data into other languages to produce augmented data in multiple languages for cross-lingual NER. [64] designs two generative data augmentation startegies for low-resource NER using the prompting approach along with the BERT model.

V. Challenges

Data augmentation has made great progress in the past few years. Despite these successes, there are still challenges that can be explored further. In this section, we discuss these challenges and suggest directions for future research.

A. Theoretical Explanation of Text Data Augmentation

The effectiveness of data augmentation in NLP has been demonstrated in a large number of experiments [30], [71], [88], [70], [34], but few studies have theoretically investigated how data augmentation works. Several recent studies [91], [92] have investigated and analyzed how data augmentation helps capture features. However, these studies have focused on images because image data can be represented by sparse coding models [93] or spike covariance models [94]. However, because the text is discrete, comparable theoretical studies in NLP are still lacking.

B. Trade-off between Computing Resources and Augmentation Effects

With the advancement of data augmentation, a variety of data augmentation methods have been proposed, especially generative augmentation strategies based on large-scale pretrained language models. These generative augmentation approaches usually show better performance than random augmentation methods in improving the model, as they are designed for specific tasks. For instance, recent studies from [60], [61] show that the generative augmentation method using large-scale PLM as the generator is obviously superior to augmentation methods like EDA and back translation in aspect-based sentiment analysis [60] and low resource NLU [61] tasks. Despite success in improving model performance, these generative augmentation approaches based on large-scale PLMs typically consume more computational resources and time. Therefore, the development of data augmentation strategies that are effective and consume little computational resources could be considered in the future.

C. Generative Augmentation without Label

The majority of current generative augmentation methods perform well in producing high-quality augmented samples. However, these generative augmentation methods usually require a label or prompt to help the generator generate appropriate sentences. This limits the application of these generative augmentation methods in the unsupervised text domain. The text data augmentation strategy proposed in Prompt-BERT [65] prepends different prompt templates at the beginning of the same sentence and then feeds them to the sentence encoder to obtain sentence representations as augmented samples. This augmentation method is a generative augmentation method that uses prompts and does not use labels. Future work could therefore consider how to develop unsupervised generative augmentation methods from this perspective.

VI. Conclusion

In this paper, we present a comprehensive and brief survey of recent data augmentation approaches in NLP. We discuss the benefits of data augmentation and common representative methods for textual data augmentation techniques, and classify these methods into three categories: semantic-invariant augmentation, random augmentation, and generative augmentation. In addition, we conclude the main application

scenarios and downstream application tasks for data augmentation. Finally, we outline the challenges in the field of textual data augmentation and show that there is still a lot of room to be further exploited. Overall, we hope that this paper will provide a novel perspective on current text data augmentation techniques and inspire more effective data augmentation approaches to be devised.

REFERENCES

[1] H. Qi, M. Brown, and D. G. Lowe, "Low-shot learning with imprinted weights," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5822–5830.

[2] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[3] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[4] G. G. Şahin and M. Steedman, "Data augmentation via dependency tree morphing for low-resource languages," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5004–5009.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[6] H. Guo, Y. Mao, and R. Zhang, "Augmenting data with mixup for sentence classification: An empirical study," *arXiv preprint arXiv:1905.08941*, 2019.

[7] J. Chen, Z. Yang, and D. Yang, "MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 2147–2157.

[8] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, June 2019, pp. 4171–4186.

[10] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, no. 1, pp. 1–34, 2021.

[11] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A survey on data augmentation for text classification," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–39, 2022.

[12] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in nlp," *arXiv preprint arXiv:2106.07499*, 2021.

[13] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Aug. 2021, pp. 968–988.

[14] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, 2022.

[15] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[16] J. Chen, Y. Wu, and D. Yang, "Semi-supervised models via data augmentationfor classifying interactive affective responses," *arXiv preprint arXiv:2004.10972*, 2020.

[17] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar, "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, June 2019, pp. 3609–3619.

[18] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2016, pp. 86–96.

[19] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 489–500.

[20] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, "Neural paraphrase generation with stacked residual LSTM networks," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Dec. 2016, pp. 2923–2934.

[21] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, July 2015, pp. 1681–1691.

[22] A. Gupta, A. Agarwal, P. Singh, and P. Rai, "A deep generative framework for paraphrase generation," in *Proceedings of the aaai conference on artificial intelligence*, vol. 32, no. 1, 2018.

[23] O. Kolomiyets, S. Bethard, and M.-F. Moens, "Model-portability experiments for textual temporal analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, vol. 2. ACL; East Stroudsburg, PA, 2011, pp. 271–276.

[24] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[25] W. Y. Wang and D. Yang, "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2557–2563.

[26] J. Raiman and J. Miller, "Globally normalized reader," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sept. 2017, pp. 1059–1069.

[27] X. Dai and H. Adel, "An analysis of simple data augmentation for named entity recognition," in *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Dec. 2020, pp. 3861–3867.

[28] T.-W. Hsu, C.-C. Chen, H.-H. Huang, and H.-H. Chen, "Semantics-preserved data augmentation for aspect-based sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 4417–4422.

[29] D. Wang, N. Ding, P. Li, and H. Zheng, "CLINE: Contrastive learning with semantic negative examples for natural language understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 2332–2342.

[30] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[31] X. Wu, C. Gao, L. Zang, J. Han, Z. Wang, and S. Hu, "ESimCSE: Enhanced sample building method for contrastive learning of unsupervised sentence embedding," in *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Oct. 2022, pp. 3898–3907.

[32] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5065–5075.

[33] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Nov. 2021, pp. 2748–2754.

[34] L. Sun, C. Xia, W. Yin, T. Liang, P. Yu, and L. He, "Mixup-transformer: Dynamic data augmentation for NLP tasks," in *Proceedings of the 28th*

International Conference on Computational Linguistics. International Committee on Computational Linguistics, Dec. 2020, pp. 3436–3440.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "Freelb: Enhanced adversarial training for natural language understanding," in *International Conference on Learning Representations*, 2020.

[37] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 2177–2190.

[38] Y. Cheng, L. Jiang, W. Macherey, and J. Eisenstein, "AdvAug: Robust adversarial augmentation for neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 5961–5970.

[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[43] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[45] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[46] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer." *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[47] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Aug. 2016, pp. 10–21.

[48] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *International conference on machine learning*. PMLR, 2017, pp. 1587–1596.

[49] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang, "Generating sentences by editing prototypes," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 437–450, 2018.

[50] N. Malandrakis, M. Shen, A. Goyal, S. Gao, A. Sethi, and A. Metallinou, "Controlled text generation for data augmentation in intelligent artificial agents," in *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Nov. 2019, pp. 90–98.

[51] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, June 2018, pp. 1875–1885.

[52] J. Xu, X. Ren, J. Lin, and X. Sun, "Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3940–3949.

[53] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, June 2018, pp. 452–457.

[54] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A robustly optimized BERT pre-training approach with post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227.

[55] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. Le Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, "Generative data augmentation for commonsense reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020, pp. 1008–1025.

[56] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu, "Conditional bert contextual augmentation," in *International conference on computational science*. Springer, 2019, pp. 84–95.

[57] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," in *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. Association for Computational Linguistics, Dec. 2020, pp. 18–26.

[58] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7383–7390.

[59] J. Zhou, Y. Zheng, J. Tang, L. Jian, and Z. Yang, "FlipDA: Effective and robust data augmentation for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 8646–8665.

[60] B. Wang, L. Ding, Q. Zhong, X. Li, and D. Tao, "A contrastive cross-channel data augmentation framework for aspect-based sentiment analysis," in *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Oct. 2022, pp. 6691–6704.

[61] Y. Wang, C. Xu, Q. Sun, H. Hu, C. Tao, X. Geng, and D. Jiang, "PromDA: Prompt-based data augmentation for low-resource NLU tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 4242–4255.

[62] K. M. Yoo, D. Park, J. Kang, S.-W. Lee, and W. Park, "GPT3Mix: Leveraging large-scale language models for text augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Nov. 2021, pp. 2225–2239.

[63] C. Chen and K. Shu, "Promptda: Label-guided data augmentation for prompt-based few shot learners," *arXiv preprint arXiv:2205.09229*, 2022.

[64] J. Liu, Y. Chen, and J. Xu, "Low-resource ner by data augmentation with prompting," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4252–4258.

[65] T. Jiang, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, H. Huang, L. Zhang, and Q. Zhang, "Promptbert: Improving bert sentence embeddings with prompts," *arXiv preprint arXiv:2201.04337*, 2022.

[66] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," *arXiv preprint arXiv:1705.00440*, 2017.

[67] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized data augmentation for low-resource translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2019, pp. 5786–5796.

[68] Y. Li, X. Li, Y. Yang, and R. Dong, "A diverse data augmentation strategy for low-resource neural machine translation," *Information*, vol. 11, no. 5, p. 255, 2020.

[69] B. Ding, L. Liu, L. Bing, C. Kruengkrai, T. H. Nguyen, S. Joty, L. Si, and C. Miao, "DAGA: Data augmentation with a generation approach for low-resource tagging tasks," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2020, pp. 6045–6057.

[70] L. Liu, B. Ding, L. Bing, S. Joty, L. Si, and C. Miao, "MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5834–5846.

[71] R. Zhou, X. Li, R. He, L. Bing, E. Cambria, L. Si, and C. Miao, "MELM: Data augmentation with masked entity language modeling for low-resource NER," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022, pp. 2251–2262.

[72] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 6894–6910.

[73] Y.-S. Chuang, R. Dangovski, H. Luo, Y. Zhang, S. Chang, M. Soljacic, S.-W. Li, S. Yih, Y. Kim, and J. Glass, "DiffCSE: Difference-based contrastive learning for sentence embeddings," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, July 2022, pp. 4207–4218.

[74] Y. Zhang, R. He, Z. Liu, L. Bing, and H. Li, "Bootstrapped unsupervised sentence representation learning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2021, pp. 5168–5180.

[75] Y. Qu, D. Shen, Y. Shen, S. Sajeev, W. Chen, and J. Han, "Co{da}: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding," in *International Conference on Learning Representations*, 2021.

[76] Y. Zhang, R. Zhang, S. Mensah, X. Liu, and Y. Mao, "Unsupervised sentence representation via contrastive learning with mixing negatives," 2022.

[77] X. Chao and L. Zhang, "Few-shot imbalanced classification based on data augmentation," *Multimedia Systems*, pp. 1–9, 2021.

[78] F. Arthaud, R. Bawden, and A. Birch, "Few-shot learning through contextual data augmentation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Apr. 2021, pp. 1049–1062.

[79] J. Wei, C. Huang, S. Vosoughi, Y. Cheng, and S. Xu, "Few-shot text classification with triplet networks, data augmentation, and curriculum learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, June 2021, pp. 5493–5500.

[80] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma, "Clear: Contrastive learning for sentence representation," *arXiv preprint arXiv:2012.15466*, 2020.

[81] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[82] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell, "A closer look at feature space data augmentation for few-shot intent classification," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Nov. 2019, pp. 1–10.

[83] P. Sun, Y. Ouyang, W. Zhang, and X. Dai, "Meda: Meta-learning with data augmentation for few-shot text classification." in *IJCAI*, 2021, pp. 3929–3935.

[84] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *arXiv preprint arXiv:1804.09541*, 2018.

[85] S. Longpre, Y. Lu, Z. Tu, and C. DuBois, "An exploration of data augmentation and sampling techniques for domain-agnostic question answering," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Nov. 2019, pp. 220–227.

[86] W. Yang, Y. Xie, L. Tan, K. Xiong, M. Li, and J. Lin, "Data augmentation for bert fine-tuning in open-domain question answering," *arXiv preprint arXiv:1904.06652*, 2019.

[87] J. Singh, B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "Xlda: Cross-lingual data augmentation for natural language inference and question answering," *arXiv preprint arXiv:1905.11471*, 2019.

[88] A. Asai and H. Hajishirzi, "Logic-guided data augmentation and regularization for consistent question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 5642–5650.

[89] A. Riabi, T. Scialom, R. Keraron, B. Sagot, D. Seddah, and J. Staiano, "Synthetic data augmentation for zero-shot cross-lingual question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2021, pp. 7016–7030.

[90] C.-C. Hung, T. Green, R. Litschko, T. Tsereteli, S. Takeshita, M. Bombieri, G. Glavaš, and S. P. Ponzetto, "ZusammenQA: Data augmentation with specialized models for cross-lingual open-retrieval question answering system," in *Proceedings of the Workshop on Multilingual Information Access (MIA)*. Association for Computational Linguistics, July 2022, pp. 77–90.

[91] Z. Wen and Y. Li, "Toward understanding the feature learning process of self-supervised contrastive learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 112–11 122.

[92] W. Ji, Z. Deng, R. Nakada, J. Zou, and L. Zhang, "The power of contrast for feature learning: A theoretical analysis," *arXiv preprint arXiv:2110.02473*, 2021.

[93] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[94] Z. Bai and J. Yao, "On sample eigenvalues in a generalized spiked population model," *Journal of Multivariate Analysis*, vol. 106, pp. 167–177, 2012.

Using Gamification to Promote Student Engagement in STEM Project-Based Learning

Palash Chhabra* and Patrick Delaney

School of Computer Science, QUT, Australia

E-mail: palash.chhabra@gmail.com

***Abstract*— This exploratory quantitative research examines how meaningful gamification could be used for student motivation and engagement in the context of STEM project-based learning (PBL) courses. A survey was developed using self-determination theory, including themes of relatedness (sense of belonging), competence and autonomy to understand student attitudes toward gamification elements being embedded in the PBL course design. We received 43 responses to the survey, which were analysed using descriptive statistics to measure agreement levels across diverse student groups. We also explored the attitudes of students identifying as strongly self-directed and not self-directed learners. Our results showed that students tend to favour practical implementations such as online forums and bonus marks over intangible gamification elements, such as personal connections and imaginary rewards. The findings are presented in the form of design recommendations that can serve as a guideline for course designers and developers of the ICT platforms on how to use gamification to promote student engagement in STEM PBL courses.**

***Index Terms*— Higher education, STEM, Student engagement, student motivation, gamification, meaningful gamification, self-determination theory**

I. INTRODUCTION

Student engagement in post-secondary classrooms is a topic of interest for universities and colleges, yet faculties and administrators still struggle to implement it effectively at a course level (Mandernach, 2015). Research by Jabbar and Felicia (2015) and Handelsman et al. (2005) suggests that effective student engagement is linked positively to desirable learning outcomes such as critical thinking, student motivation and student learning, but if neglected, can lead to disengagement, cheating and learned helplessness (O'Donovan et al., 2013). Project-based learning (PBL) courses are widely used in STEM disciplines to improve students' self-directed learning and prepare them for professional life (Mills & Treagast, 2014; Sabhaba et al., 2016). In PBL courses especially, there is a need to improve self-motivation and proactiveness in students. Compared to problem-based learning, students in PBL courses must manage their time, and resources and understand their role and task differentiation based on strong self-direction (Mills & Treagust, 2003).

To succeed in PBL, one critical skill students must develop is information literacy. In a study of an undergraduate course, Bankermans and Plotke (2018) found that it is necessary to incorporate activities in the course curriculum to actively support the development of skills that promote assessment and evaluation but curriculum designers for PBL courses face the challenge of embedding features that engage both the intrinsic and extrinsic motivations of students. An emerging trend in this regard is the application of gamification to promote student engagement (O'Donovan et al., 2013; Tan & Hew, 2016). Deterding et al. (2011) define gamification as the use of game-design elements in a non-game context such as embedding of intrinsic motivation in meaningful activities and self-learning checkpoints. In a review of Performance-based assessment for Machine Learning at the K-12 level, Rauber and Gresse von Wangenheim (2022) found that gamification activties were used in several contexts to increase student learning. However, only a few empirical studies have examined the effects of gamification in universities and higher education (e.g., Dicheva et al., 2015; Hanus & Fox, 2015; Souza et al., 2019).

Antonaci et al. (2018) reflect that while gamification models have been applied in schools and online classes, the problem of implementing gamification techniques into less game-oriented project units is still under-discussed in the literature (Laskowski, 2015; Tan & Hew, 2016). There have been efforts to incorporate gamification at the undergraduate level with a particular focus on skills-specific areas such as data science and machine learning (Durán-Rosal et al., 2023), which are areas where structured activities and competition can clearly link to knowledge gains and assessment performance. Thus, there remains a clear need to explore gamification in courses with a broader skillset or multidisciplinary focus, particularly where there are not only diverse students but diverse learning needs. Researchers such as Smiderle et al. (2019) and Hanus and Fox (2015) also highlighted the need to map the success of gamification models with different student diversities and student cohorts.

This research used a quantitative survey design to understand the possibility of applying gamification to a post-graduate STEM course in an Australian institution. The aim was to gather insights into how curriculum designers can embed meaningful gamification into a project-based learning course to support intrinsic and extrinsic motivation. We use the theory of self-determination to analyze the survey data as there is strong evidence that this theory can understand motivation through its

concepts of autonomy, competence and relatedness (Martin et al., 2018). To this end, our survey also aims to differentiate between self-directed and not self-directed learners, a necessary insight in a course that is highly dependent on self-motivation and self-directedness. The outcomes of our study are used to suggest the inclusion of certain gamification elements in STEM PBL courses. Our study is guided by the following two research questions:

Q1: What role does meaningful gamification play with respect to self-determination theory in the context of project-based learning courses?

Q2: How can gamification be used to foster relatedness, competence and autonomy in students from different demographics and diversified educational backgrounds enrolled in project units?

The next section provides a Literature Review of relevant studies to define our concepts and position our research contributions. Following this we present our Methodology, detailing the study setting and the data collection phase. Our Results analyze and interpret the data, with our findings the basis for a set of design recommendations for using gamification in a STEM PBL courses. The Conclusion summarizes the paper and indicates limitations and future research.

II. Literature Review

Student engagement is a broad concept, with faculties and administrators still struggling to effectively implement student engagement at both, the institutional and course levels (Mandernach, 2015). O'Donovan et al. (2013) found that student engagement is frequently neglected, which can lead to disengagement, cheating and learned helplessness. The literature reveals that there has also been a steady decline in the number of students who finish their studies on time, which highlights the importance of student engagement (Iosup, & Epema, 2014). Student engagement is often considered a product of student motivation and is presented by the self-determination theory of motivation, which has a strong foundation as a basis for fostering the intrinsic motivation of students (Martin et al., 2018).

A. Self-Determination and Student Engagement

Self-determination theory assumes that all individuals, regardless of gender, age, or culture, possess three fundamental psychological needs that move them to act or not to act: autonomy, relatedness, and competence (Tan & Hew, 2016; Gagné & Deci, 2005). Skinner (2008) suggests that autonomy, or having a sense of freedom to pursue choices based on interest, is expected to have an effect on higher levels of emotional engagement. Tan and Hew (2016) explain that competence or mastery of a topic being studied encourages the learner to further participate in project activities, and Furrer and Skinner (2003) find that relatedness or sense of belonging is linked to increased levels of behavioural and emotional engagement which is also identified as an effective component of student engagement by Mandernach (2015). In a study of MOOC students, Martin, Kelly and Terry (2018) found that self-determination is critical for designing frameworks for online courses as it engages learners more successfully than previous approaches, and when done effectively, contributes positive functional outcomes in terms of quality of motivation, self-regulation, learning, organization and integration, vitality, and well-being.

Project-based learning pedagogies are gaining attention, with more research exploring how classroom conditions and learning environments influence student choices, which in turn can inform practices and foster outcomes such as self-efficacy, metacognition, effort regulation and collaboration (Stefanou et al., 2013). Stewart (2007) explored the relationship between self-directed learning among students and project-based learning in post-graduate courses, with a key finding being that students with high self-management achieved higher learning outcomes in project-based learning courses.

In general, the self-determination theory has been applied to a wide range of educational contexts in previous studies and results indicate that the satisfaction of these basic psychological needs had a mediating effect on learning outcomes, by supporting intrinsic or other autonomous forms of motivation.

B. Meaningful Gamification in Education

Antonaci et al. (2018) define gamification as the application of game elements to a non-game scenario to create an effect on or change in user behavior. Laskowski (2015) describes the main goal of gamification as applying a specific structure of tasks based on game objectives and rules that are to be completed by users. Gamification can take a variety of forms, including the creation of social competition and the incentivizing of behavior through game-based mechanisms such as badge and reward systems, and the creation of challenges and leaderboards (Hanus & Fox, 2015; Souza et al., 2019).

At its core, gamification corresponds to extrinsic motivation and a variety of human desires, such as the need for reward, status, achievement, self-expression, competition, and belonging (Tan & Hew, 2016). Meaningful gamification not only uses game mechanics to provide extrinsic incentives but also applies student-centred activities to make a course meaningful to participants and provide intrinsic motivation. These activities are related to the self-determination theory of motivation and can be used along with game mechanics to boost student motivation. Fig. 1 shows a high-level gamification model summarizing common elements found across the literature.

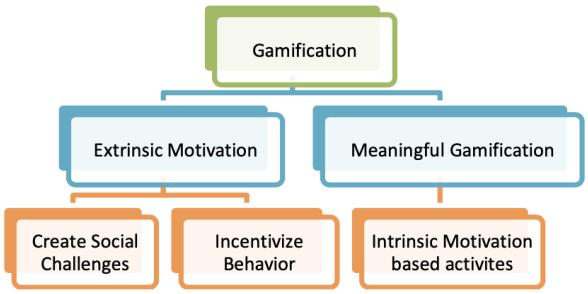

Fig. 1: Core elements of a gamification model.

O'Donovan et al. (2013) suggest that the gamification model has been successful in medical, social, lifestyle, business and educational contexts, and further applying gamification techniques in a university setting can improve students' understanding. From their study, Tan and Hew (2016) found that there was a difference in the uptake of gamification between high-performing students and students considered non-achievers. Laskowski (2015) used an experiment in a higher education setting where the author employed gamification techniques in two different courses during two academic years to demonstrate that the gamified group of students resulted in higher involvement, attendance levels and increased homework completion. Further, Iosup and Epema (2014) also applied gamification to undergraduate and graduate courses and found that gamification not only correlated with an increase in the percentage of passing students but also in participation in voluntary activities and challenging assignments. In the context of project work and development in high schools, Souza et al. (2019) developed a gamification-based assessment methodology called GAMED which is an assessment methodology that introduces systematic steps to improve student engagement through gamification by improving aspects such as motivation and teamwork. In a K-12 context where gamification was used to build teamwork among students learning machine learning and AI approaches, Sakulkueakulsuk et al. (2018) found that it was particularly useful for helping students adopt the futuristic and interdisciplinary thinking that is required in STEM courses.

C. *Self-determination Theory and Gamification*

According to Zichermann and Cunningham (2011), gamification includes a challenge-achievement-reward loop that promotes the production of dopamine which in turn, creates satisfaction and positively impacts student engagement. Iosup and Epema (2014) suggest that gamification gives the educator several powerful and predictable tools for influencing human motivation and behaviour. As student engagement is identified with self-determination theory that includes autonomy, relatedness and competence (Tan & Hew, 2016; Gagné & Deci, 2005), game mechanics can be used to cater to these needs. For example, using an 'early bird' badge for motivating students to download and read lecture material before class can fulfil the student's needs for autonomy while awarding a 'reply warrior' badge to motivate students to respond to each other queries' can be a way to boost relatedness. This is consistent with the research findings by O'Donovan et al. (2013) in two undergraduate courses that were gamified and encouraged students to remain more engaged in the coursework.

Various researchers such as Dicheva et al. (2015), Hanus and Fox (2015) and Souza et al. (2019) found that very few empirical studies have examined the effects of gamification, particularly in the context of universities and higher education. Hanus and Fox (2015) highlight that while gamification leads to engagement, future gamification research should investigate specific elements of gamification rather than as an overarching concept so that the effectiveness of different mechanics can be parsed out. Additionally, there has not been significant research in the past that maps any gamification model with different student diversities and cohorts based on demographics and previous education backgrounds (Marques et al., 2019; Hanus & Fox, 2015). The concept of meaningful gamification presented by Tan and Hew (2016) suggests that it has a positive impact on student engagement but there is a need to explore its effect in the context of project units with a mixed student cohort which is presented in this paper.

III. METHODOLOGY

We investigated student attitudes towards the possible inclusion of meaningful gamification elements in a PBL course through an exploratory survey administered to students enrolled in the course. The analysis is completed using descriptive statistics to identify different student cohorts and their preferences. Triangulation is used to compare these findings with the results presented by Tan and Hew (2016) in their research and validate if meaningful gamification can be used to promote student engagement in project-based learning.

A. *Study Setting*

The survey was administered in a post-graduate STEM course at a major Australian university in Semester 2, 2020. The course was 24 credit points and involved interdisciplinary ICT research projects that ran for the 13-week semester. The course was being delivered online due to the COVID-19 pandemic, which was a new experience for many of the students and may have influenced how they interpreted and responded to questions. The curriculum for the course is similar to the research context of Bankermans and Plotke (2018), in that students need to deliver key assignments based on their discipline and information literacy skills.

To complete all assessment items, students must apply for a research topic that is supervised by an academic staff member. Topics are diverse and can be interdisciplinary and span a range of different fields, such as data science, networking, security, machine learning and AI, information systems, engineering and social issues. The assessment items were iterative research-based written pieces, including a proposal, a journal article and a research seminar. While students were part of project groups,

each assessment was individual, with each student required to develop their research question and research design. Students attended lectures and tutorials throughout the semester, were expected to access libraries for help with information retrieval and engage with their project group in a way that doesn't compromise academic integrity. These students were selected through convenience sampling since the researchers were involved in the course. However, all surveys were anonymous and there was no coercion involved in their recruitment. Because students are expected to be self-directed, proactive and develop a broad array of skills, the study setting serves as an appropriate forum to gain new insights into possible gamification designs for a STEM PBL curriculum.

B. Data Collection

An online survey was approved by the university's Human Research Ethics committee. Out of 267 students enrolled in the course, 43 completed the survey. The instrument was structured into three themes (T1, T2, T3) mapped against self-determination theory to align the results with the research questions: T1: Relatedness (Sense of Belonging), T2: Competence and T3: Autonomy. There were six Likert scale questions developed by the researchers on a scale of strongly disagree to strongly agree (1-5) and three multiple choice questions (MCQ) that were gamification options for assessment that students either needed to select one or the other. The survey questions and responses from six Likert scale questions are presented in Table 1, where M=mean and SD=standard deviation to show students' levels of agreement.

TABLE I: Likert scale questions structured by themes.

T1: Relatedness	**M**	**SD**
Q1: I feel engaged if my peers and supervisor in my project group know me on a personal level	3.63	0.14
Q2: I will feel more engaged in group work if there's an online forum where all the students of my project group can ask questions and interact with each other	4.14	0.15
T2: Competence questions	**M**	**SD**
Q3: I would find the course motivating if I earned imaginary points and badges for any accomplishment or task completion	3.46	0.17
Q4: I will be more competitive and perform better in my assignments if grades are released in the form of 'leader board' rankings	3.00	0.19
T3: Autonomy Questions	**M**	**SD**
Q5: I would be happy if I could earn some extra marks through bonus readings and mini-tasks for the course apart from assignments	3.95	0.13
Q6: I would feel more engaged in my studies if I could participate in the design and development of a project unit	3.74	0.12

The three MCQ questions are also segregated based on the above themes and were used to understand student perceptions using post hoc analysis. The questions and options associated with them are shown in Table 2.

TABLE II: MCQ questions structured by themes.

T1: Relatedness
Q7: Which of the following would make you feel more engaged with a project? *A dedicated platform to interact with team members and supervisor OR a general communication tool for interaction*
T2: Competence
Q8: Which of the following would make you participate in tutorials more? *An option to earn recognition in the form of points/badges for small tasks and activities I complete during the tutorials OR A standard guideline sheet issued by my tutor for the tutorial tasks*
T3: Autonomy
Q9: What do you think would increase your motivation to be engaged in a project unit? *The flexibility to choose and complete from multiple assignments and earn marks based on difficulty OR A single assessment option with standard marking for all the students.*

These questions came from conversations between the researchers about possible gamification applications that would work in the study setting. For example, for Q7 in Table 2, the dedicated platform referred to communication systems such as Slack. This question was asked due to some project groups using this platform, while other groups relied on general communication tools such as email.

Demographic data and background information was also gathered through the survey. This included prior education and industry experience. This information enabled us to explore diversity among respondents and how students from different backgrounds feel towards certain gamification themes. We also asked the students to consider to what extent they identified as self-directed learners on a scale of 1-5 (Strongly Disagree to Strongly Agree). For analysis, we classified students who selected agreed or strongly agreed as self-directed (SD) learner, and students who chose neutral, disagreed or strongly disagreed as not self-directed learners (NSD). The diversifications of our respondents with numeric results are shown in Table 3.

TABLE III: Diversifications of the survey respondents.

Diversification	**Type**	**%**	**N = 43**
Enrolment	International	65%	28
	Domestic	35%	15
Gender	Male	70%	30
	Female	30%	13

Background	No Experience	25%	11
	With Experience	75%	32
	• IT Background	28%	9
	• Non-IT	72%	23
Learner style	Self-directed	69.7%	30
	Not-self-directed	30.3%	13

For our analysis, the main diversifications we considered are international, domestic, SD and NSD learners since these are common student types in a PBL course in Australia.

IV. Results

The survey responses were analyzed using post hoc analysis. Student responses to the questions were filtered and students were clustered into different diversifications as defined in Table 3. Due to the small sample size, we used simple descriptive statistics to gain insight into the attitudes of each diversification. Because diversifications stem from the same dataset, there is overlap between students – for example, a domestic student could also fall into SD or NSD learner. The relationship of student choices with different parameters was observed and profiling was done to enhance the interpretation of results. This approach was chosen because a similar study in gamification by Tsay et al. (2018) used this form of triangulation to measure the effectiveness of a gamified curriculum, which led to increased engagement in online learning in an undergraduate course.

We use stacked aggregated bar charts to show the results for survey questions across the four groups. For this analysis, we reduce the 5-point survey scale to three categories – Disagree/Strongly disagree, Neutral and Agree/Strongly agree to cluster responses.

A. *Relatedness (Sense of belonging)*

The theme of Relatedness describes the students' sense of belonging to a project within the course in particular. In Figures 2, 3 and 4, we present the results of the students' responses to the questions in this theme.

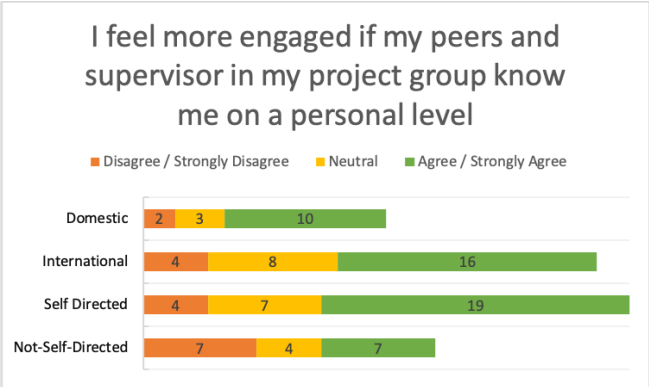

Fig. 2: Being known on a personal level.

Fig. 2 shows that students had reasonable agreement that being known on a personal level would increase engagement.

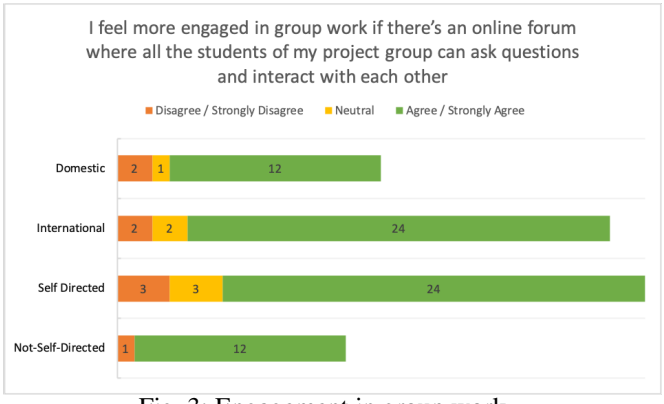

Fig. 3: Engagement in group work.

In Fig. 3, it is observed that an online forum was very important to all types of students.

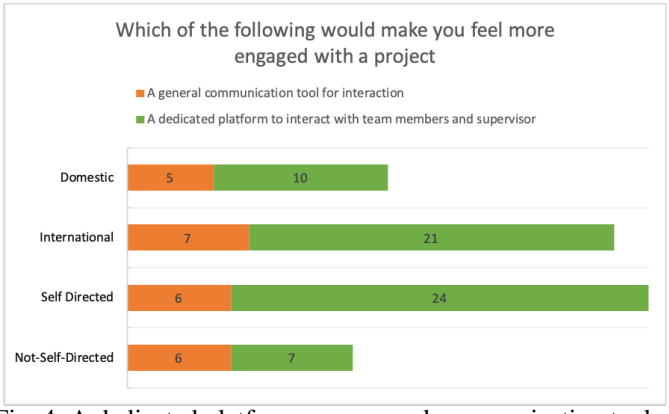

Fig. 4: A dedicated platform vs a general communication tool.

In Fig. 4, apart from mixed responses from NSD learners, all other groups preferred a dedicated communication channel to interact with peers and supervisors instead of legacy options like email.

B. *Competence*

Competence relates to intangible gamification rewards, with responses shown in Figures 5, 6 and 7.

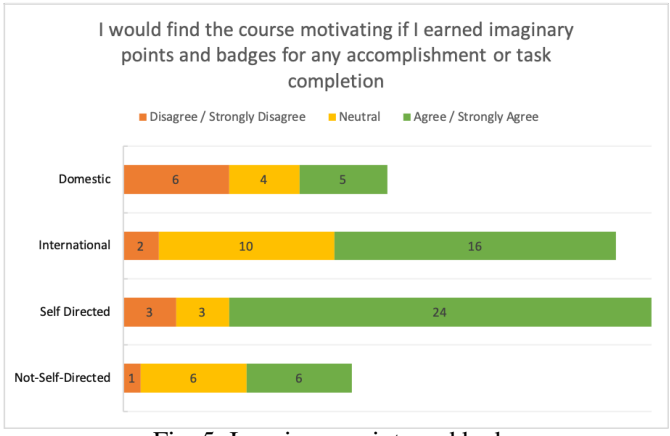

Fig. 5: Imaginary points and badges.

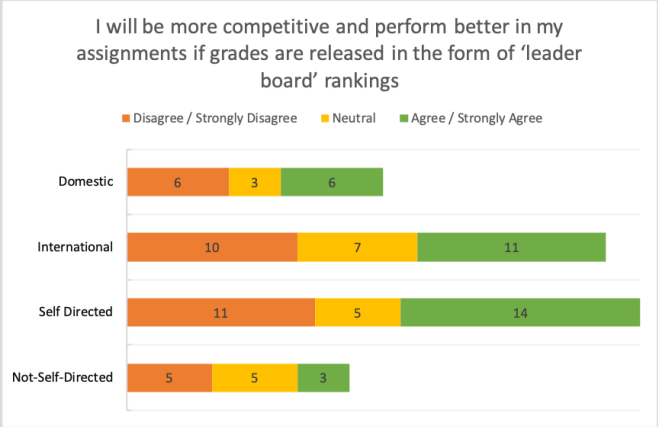

Fig. 6: 'Leader board' rankings.

In Figures 5 and 6, we can observe that there is a split between students in terms of intangible gamification rewards. Imaginary points and badges seemed to be preferred by SD learners as expected. However, for all other student groups, there was less enthusiasm, with domestic students and NSD learners far less interested in this element. Fig. 6 shows a clear disinterest in a 'leader board' ranking system, indicating that most students in the PBL were not motivated by this particular intangible reward.

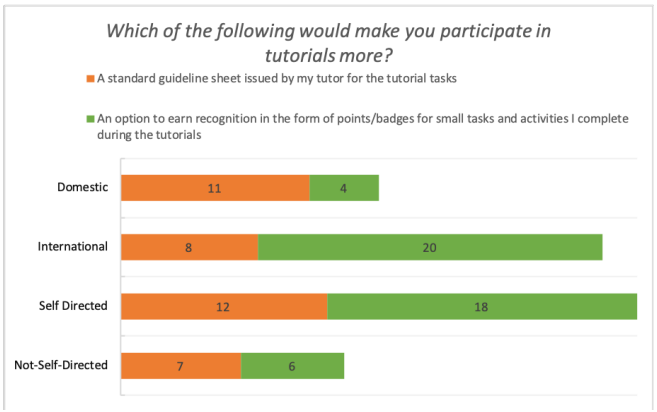

Fig. 7: Standard guidelines vs earning recognition in tutorials.

In Fig. 7, while international students prefer extrinsic forms of gamification elements and prefer to earn points and get recognised for the activities they complete, domestic students showed no such interest. This is also the case with NSD learners who do not like the idea of earning recognition through points and badges as shown in Fig 7.

C. *Autonomy*

In Figures 8, 9 and 10, we compare the responses relating to autonomy, which are the tangible and physical investments students can make into a PBL course. Students in Fig. 8 showed strong support for receiving extra marks based on performing additional functions to their standard assignments. However, in Fig. 9, we observe split between groups towards participating in developing and designing PBL assignments.

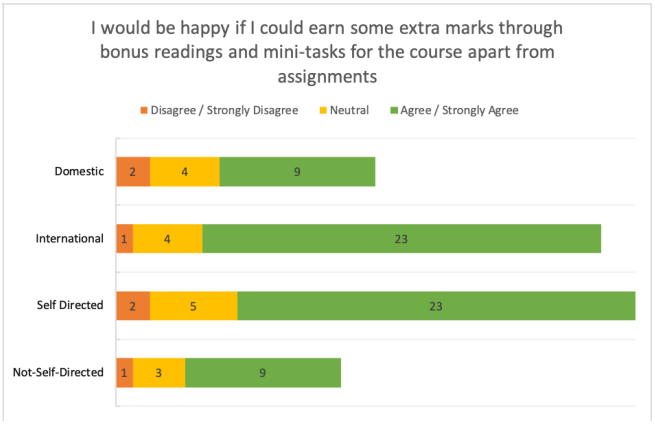

Fig. 8: Bonus readings and mini-tasks for extra marks.

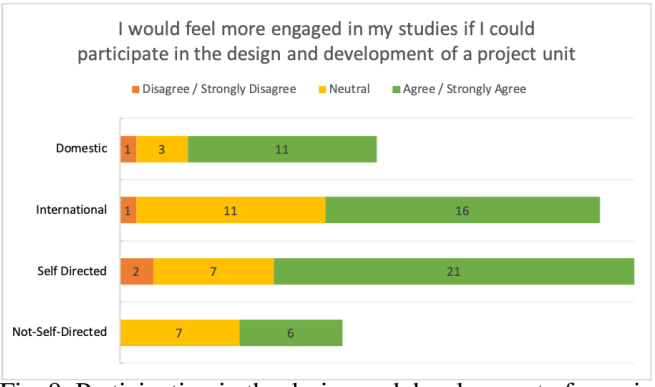

Fig. 9: Participating in the design and development of a project unit.

Domestic and SD learners showed generally strong enthusiasm, but international and NSD learners were less interested in participating in the development of research topics. Fig. 10 shows the results of the MCQ for this theme.

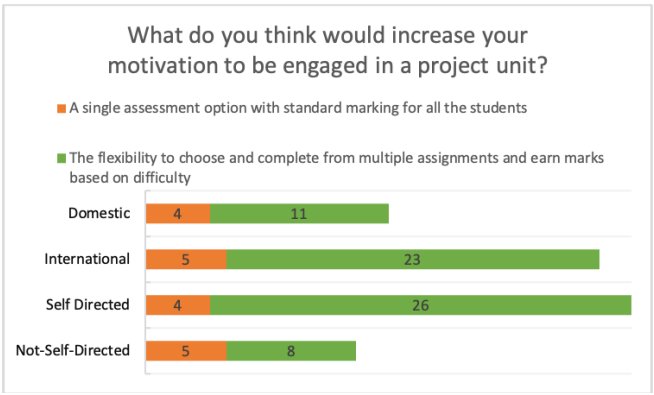

Fig. 10: Flexibility of choice vs a single assessment option.

Fig. 10 depicts that all students prefer having the flexibility to choose assignment levels and earn marks based on difficulty. This is also particularly favoured by SD learners who generally remain proactive and like the idea of earning marks based on the difficulty of the problem. Similarly, international students prefer having the option to choose the difficulty level of the assignments because of their diversity and technical

experiences.

Following the descriptive statistics comparisons, we utilize a Mann-Whitney test to compare the groups of students who consider themselves self-directed (SD) and not self-directed (NSD). The Likert scale questions were tested across the themes of Relatedness (T1), Competence (T2) and Autonomy (T3), as seen in Table 4.

TABLE IV: Mann-Whitney test results comparing self-directed and not self-directed learners.

Item	Type	Mean Rank	Sum of Ranks	U	P
Q1. I feel engaged if my peers and supervisor in my project group know me on a personal level	SD	22.53	676	179	0.68
	NSD	20.77	270		
Q2. I will feel more engaged in group work if there's an online forum where all the students of my project group can ask questions and interact with each other	SD	21.9	657	192	0.94
	NSD	22.23	289		
Q3. I would find the course motivating if I earned imaginary points and badges for any accomplishment or task completion	SD	21.5	645	180	0.7
	NSD	23.15	301		
Q4. I will be more competitive and perform better in my assignments if grades are released in the form of 'leader board' rankings	SD	23.3	699	156	0.3
	NSD	19	247		
Q5. I would be happy if I could earn some extra marks through bonus readings and mini-tasks for the course apart from assignments	SD	21.6	648	183	0.76
	NSD	22.92	298		
Q6. I would feel more engaged in my studies if I could participate in the design and development of a project course	SD	23.12	693	161	0.38
	NSD	19.42	252		

The results of the Mann-Whitney test showed no meaningful comparisons across the diversifications. The most noticeable differences were in questions 4 and 6. SD learners favoured leader board rankings (23.3) as compared to NSD learners (19), while SD learners indicated they would feel more engaged if they could participate in the design of a course (23.12) compared to NSD learners (19.42). There was no significance across any of the items however, with $p>0.05$ for each item.

V. Meaningful Gamification in STEM PBL

Our results reiterate that individuals have distinctive needs, which is a common theme in student engagement research. For gamification to work in a PBL course, supervisors must understand the characteristics of their cohort, and students need to feel connected to the topic and their group. While the assessments in this course were individual, socialization was extremely important for students to succeed, as peers and extracurricular services such as library liaisons are critical to students building relevant skills.

For this study, the first research question was "*What role does meaningful gamification play with respect to self-determination theory in the context of project-based learning courses?"* In terms of Relatedness, most students indicated they wanted an online space for the project group but there was a split between needing to be known at a personal level. This is understandable in an ICT course since students are expected to be self-directed and approach people on their own. This could be a barrier, particularly for introverted students. Students who make no attempt to learn from peers or approach extracurricular services may develop an over-reliance on project supervisors or develop learned helplessness (O'Donovan et al., 2013). Rewards and gamification elements that encourage students to access library support to develop information literacy and academic writing need to be embedded into a STEM PBL. This is in line with the findings by Tan and Hew (2016), where a Mann-Whitney test revealed that the students who accessed a gamified interaction forum were more engaged in the course than the students enrolled in a traditional course.

Comparing the Competence and Autonomy results, it is clear that actual rewards like bonus marks are preferable to the student groups rather than virtual or intangible rewards. All student groups strongly favored this element as in Figure 8, compared to the badges and leader boards in Figures 6 and 7. Leader boards were largely unfavored by students, indicating that competitiveness was not a desired element in this PBL course. This may have been inferred from interpreting the question to mean a system in which marks and names of students were exposed to the whole cohort, which was obviously not desirably to most of the students. This is a similar observation to Tan and Hew (2016) in their research in which they concluded that competitive activities may only be appealing to performance-oriented students (individuals who are interested in doing better than others). Adding tangible benefits such as extra marks for additional tasks could be a viable option for a PBL course of this nature, as it is structured around developing research skills. Students who do not access the library or do extra readings of their own volition may lose out on marks or develop deficiencies anyway, so these extra marks may incentivize their willingness to engage socially or put in extra effort, which should be the goal of a research course.

Autonomy also included a question about whether students would like to participate in the design of the PBL course research topics. As shown in our analysis, SD learners strongly favoured this, whereas NSD learners did not which again provides a key insight into the distinction in attitude between these groups. Ultimately the goal is to encourage NSD learners to become more active and self-directed in their approach to research-based assessment.

The results in Autonomy indicate that a majority of students would be open to the flexibility of choosing assignments based on difficulty, which is reflected in the research by Tan and Hew (2016) where they conclude that the use of game mechanics has a positive effect on motivating students to engage with more difficult tasks in the course. In addressing the first research question, we found that meaningful gamification can assist with student self-determination when there are tangible, beneficial rewards. Students also responded positively toward contributing toward their projects through giving themselves and classmates choices to help define their experience.

VI. Design Recommendations

The second research question was "*How can gamification be used to foster relatedness, competence and autonomy in students from different demographics and diversified educational backgrounds enrolled in project units?"* This question is answered by our gamification design recommendations, presented in Table 5. These are design for project supervisors and course designers who may consider

TABLE V: Design recommendations for STEM PBL course designers

Factor of Influence	**Autonomy**		
Guideline	***Diversification***	***Method***	***Frame of Reference***
Flexibility	All students	Provide options for students to choose from a list of questions based on level of difficulty instead of a fixed assessment structure with standard marking guidelines.	Provision of optional tasks or allowing students a method of self-reflection and perceived choice boosts intrinsic motivation (Martin et al., 2018).
Bonus Marks	All Students	Award points for recommended readings in the course, once the points reach a certain threshold, these can be converted to actual marks.	Martin et al. (2018) suggest that unexpected rewards may increase course enjoyment and motivation which also corresponds to the results of this research.
Student Inclusion	Students with academic experience	Involve students over a certain GPA in their previous studies in setting up the course structure and identify student expectations from the unit.	Tan and Hew (2016) suggest that students like to have control over their learning path which gives students interest-based preferences which is crucial for achieving autonomy.
Factor of Influence	**Competence**		
Rewards and Recognition-based tasks	Students with IT background	Inclusion of badges (reply warrior badge, high-achiever badge), progress bars, difficulty level-based tasks in the course and the use of motivational messaging.	A perceived sense of progression and recognition of success is a key design practice that has a strong influence on competence (Martin et al., 2018)
Refrain from trivial forms of gamification	All students	Use meaningful gamification to identify students' need for competence. If students in a course, do not support competitive environments, do not use extrinsic gamification elements such as leader board rankings	Tan and Hew (2016) concluded in their research that the creation of highly competitive environments or public recognition platforms is not appealing to all students. This also served as a hypothesis in the research by Martin et al., 2018 that suggests that such forms of contingencies potentially undermine intrinsic motivation.
Factor of Influence	**Relatedness (Sense of Belonging)**		
Communication Channel	International Students	A dedicated online forum where the students specific to a group can sign up and interact with the team members and the supervisor, make explicit the expectations of the supervisor's role and the level of socialization required to succeed.	Co-construction of knowledge and sharing of ideas is a crucial factor of motivation as witnessed in findings of the Mann-Whitney test by Tan and Hew (2016) which revealed that access to a gamified interaction forum led to an increase in engagement.
Cohort Considerations	All Students	Creation of rich profiles with avatars, varying backgrounds, interests and language preferences to be present in the ICT tool implementation.	Martin et al., (2018) reflected that including a frame of reference of the learners and creation of personas help in gauging the perspective of possible course participants when there is no direct access beforehand.

implementing gamification to enhance student engagement in a STEM PBL course. Generally, stakeholders need to:

- Understand the needs of the enrolled students early in the course. Providing an option for students to make decisions related to their course structure and giving them the flexibility to engage with the course coordinator as much or as little as they want would foster a sense of presence, also highlighted by Martin et al. (2018) in their design recommendations for improving autonomy.
- Prepare for student cohorts and use gamification elements relevant to the students and their needs. Tan and Hew (2016) found that public recognition and ranking systems may only be appealing to specific students, and in our study, we also found that there was little interest in imaginary or virtual rewards.
- Project supervisors should encourage socialization and create dedicated platforms where the students within a project group can interact and initiate conversations, which is highlighted by Martin et al. (2018) in their research results as students wish to interact with peers more often in the course.
- Gain student choices and preferences for marked assessments and tutorials and provide assessment flexibility to students. This is similar to the creation of an optimal challenge (Martin et al., 2018) for students and allows them to set their own goals based on their personas.
- Provide options to students to earn bonus marks apart from regular assessments. The results indicate that a majority of students would like to have the option to earn bonus marks through task completion which can be implemented as unexpected rewards that may enhance enjoyment and have an effect on intrinsic motivation.

VII. Conclusion

This study analyzed the survey results from 43 students to understand how meaningful gamification through self-determination theory can be embedded in a research based STEM PBL course. We divided our respondents into different diversifications to gain insight into their attitudes across three key themes: Relatedness, Competence and Autonomy. Importantly, in our classification of self-directed (SD) and not-self-directed (NSD) learners, we found differences in attitudes toward virtual and tangible rewards. Previous studies in gamification were mostly restricted to students in high school or undergraduate courses. This study extends research into PBL courses at the tertiary level, which has broader demographics and diverse student backgrounds, providing challenges for the academic staff.

One limitation of the survey was the lack of specifics around the questions, such as under what conditions the students will be awarded badges. For example, in this course, it may be feasible to provide virtual rewards to them for completing a library module on information searching. If such conditions had been outlined, the levels of agreement may have been different. In addition, the study only presents a post hoc analysis of the survey results. The study did not explore why students from different cohorts prefer certain options as it does not capture qualitative data from students. The small sample size of students filtered into these sub-groups further limits the generalizability of the results.

Future research should focus on the implementation of these guidelines to identify if the recommendations work in actual study settings. Further investigation should use qualitative approaches to understand student decisions and behaviours; the use of design-based research by Anderson and Shattuck (2012) is suggested. Design-based research would enable the researchers to iteratively improve the gamified project unit over time while identifying the obstacles. This approach could potentially yield more generalizable practical design principles for using gamification in STEM PBL courses as opposed to a one-off theoretical study.

References

[1] Abdul Jabbar, A. I., & Felicia, P. (2015). Gameplay engagement and learning in game- based learning: A systematic review. *Review of educational research*, 85(4), 740-779. https://doi.org/10.3102/0034654315577210

[2] Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research?. *Educational researcher*, *41*(1), 16-25. https://doi.org/10.3102/0013189X11428813

[3] Antonaci, A., Klemke, R., Kreijns, K., & Specht, M. (2018). Get Gamification of MOOC right! How to Embed the Individual and Social Aspects of MOOCs in Gamification Design. *International Journal of Serious Games*, 5(3), 61 - 78. https://doi.org/10.17083/ijsg.v5i3.255

[4] Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in higher education*, 47(1), 1-32. https://doi.org/10.1007/s11162-005-8150-9

[5] Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011, September). From game design elements to gamefulness: defining gamification". In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments (pp. 9-15). https://doi.org/10.1145/2181037.2181040

[6] Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society*, 18(3). https://www.jstor.org/stable/jeductechsoci.18.3.75

[7] Durán-Rosal, A.M., Guijo-Rubio, D., Vargas, V.M., Gómez-Orellana, A.M., Gutiérrez, P.A., & Fernández, J.C. (2023). Gamifying the Classroom for the Acquisition of Skills Associated with Machine Learning: A Two-Year Case Study. In *International Joint Conference 15th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2022) 13th International Conference on EUropean Transnational Education (ICEUTE 2022)*. CISIS ICEUTE 2022 2022. Lecture Notes in Networks and Systems, vol 532. Springer, Cham. https://doi.org/10.1007/978-3-031-18409-3_22

[8] Furrer, C., & Skinner, C. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95(1), 148–162. https://doi.org/10.1037/0022-0663.95.1.148

[9] Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational behavior*, 26(4), 331-362. https://doi.org/10.1002/job.322

[10] Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research*, 98(3), 184- 192. https://doi.org/10.3200/JOER.98.3.184-192

[11] Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & education*, 80, 152-161. https://doi.org/10.1016/j.compedu.2014.08.019

[12] Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review.

Computers & Education, 90, 36-53. https://doi.org/10.1016/j.compedu.2015.09.005

[13] Iosup, A., & Epema, D. (2014, March). An experience report on using gamification in technical higher education. *In Proceedings of the 45th ACM technical symposium on Computer science education* (pp. 27-32). https://doi.org/10.1145/2538862.2538899

[14] Laskowski, M. (2015, March). Implementing gamification techniques into university study path-A case study. *In 2015 IEEE Global Engineering Education Conference (EDUCON)* (pp. 582-586). IEEE. https://doi.org/10.1109/EDUCON.2015.7096028

[15] Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, 16(2), 193-205. http://www.iier.org.au/iier16/mackenzie.html

[16] Mandernach, B. J. (2015). Assessment of student engagement in higher education: A synthesis of literature and assessment tools. *International Journal of Learning, Teaching and Educational Research*, 12(2). https://www.ijlter.org/index.php/ijlter/article/view/367

[17] Martin, N., Kelly, N., & Terry, P. (2018). A framework for self-determination in massive open online courses: Design for autonomy, competence, and relatedness. *Australasian Journal of Educational Technology*, *34*(2). https://doi.org/10.14742/ajet.3722

[18] O'Donovan, S., Gain, J., & Marais, P. (2013, October). A case study in the gamification of a university-level games development course. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference* (pp. 242- 251). https://doi.org/10.1145/2513456.2513469

[19] Rauber, M. F., & Gresse von Wangenheim, C. (2022). Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping. *Informatics in Education*. https://doi.org/10.15388/infedu.2023.11

[20] Sababha, B., Alqudah, Y., Abualbasal, A., & AlQaralleh, E. (2016). Project-based learning to enhance teaching embedded systems. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(9), 2575-2585. https://doi.org/10.12973/eurasia.2016.1267a

[21] Sakulkueakulsuk, B., Witoon, S., Ngarmkajornwiwat, P., Pataranutapom, P., Surareungchai, W., Pataranutaporn, P., & Subsoontorn, P. (2019). Kids making AI: Integrating Machine Learning, Gamification, and Social Context in STEM Education. In Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018 (pp. 1005–1010). IEEE. https://doi.org/10.1109/TALE.2018.8615249

[22] Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, 100(4),765–781. https://doi.org/10.1037/a0012840

[23] Smiderle, R., Marques, L., Coelho, J. A. P. D. M., Rigo, S. J., & Jaques, P. A. (2019, July). Studying the Impact of Gamification on Learning and Engagement of Introverted and Extroverted Students. *In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (Vol. 2161, pp. 71-75). IEEE. https://doi.org/10.1109/ICALT.2019.00023

[24] Souza, P., Mombach, J., Rossi, F., & Ferreto, T. (2019, July). Gamed: Gamification- based assessment methodology for final project development. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (Vol. 2161, pp. 359-361). IEEE. https://doi.org/10.1109/ICALT.2019.00114

[25] Stefanou, C., Stolk, J.D., Prince, M., Chen, J.C. & Lord, S.M. (2013). Self-regulation and autonomy in problem- and project-based learning environments. Active Learning in Higher Educationy, 14(2), 109-122. https://doi.org/10.1177/1469787413481132

[26] Tan, M., & Hew, K. F. (2016). Incorporating meaningful gamification in a blended learning research methods class: Examining student learning, engagement, and affective outcomes. *Australasian Journal of Educational Technology*, 32(5). https://doi.org/10.14742/ajet.2232

[27] Tsay, C. H. H., Kofinas, A., & Luo, J. (2018). Enhancing student learning experience with technology-mediated gamification: An empirical study. *Computers & Education*, *121*, 1-17. https://doi.org/10.1016/j.compedu.2018.01.009

[28] Wu, H. P., Garza, E., & Guzman, N. (2015). International student's challenge and adjustment to college. *Education Research International*, 2015. https://doi.org/10.1155/2015/202753

Selected Ph.D. Thesis Abstracts

This Ph.D thesis abstracts section presents theses defended in 2021 and 2022. These submissions cover a range of research topics and themes under intelligent informatics, such as prophylactic treatments to misinformation and disinformation, fatal disease detection using hybrid deep learning, geriatric care monitoring, graph model, interactive visualization, machine learning-assisted corpus exploration, large scale data exploration, remote patient monitoring, representation learning, secure content delivery in IoT, soft biometrics-based person retrieval, and safety solutions in smart cities.

AN INTERDISCIPLINARY ASSESSMENT OF THE PROPHYLACTIC EDUCATIONAL TREATMENTS TO MISINFORMATION AND DISINFORMATION

Kevin Matthe Caramancion
kcaramancion@albany.edu
University at Albany, State University of New York, USA

MISINFORMATION and Disinformation, both types of Information Disorder in the cyber world, operate on a systemic level. Several factors enabling their persistence, including laws, policies, and technological mediators, have been investigated in the literature. Cybersecurity frameworks and guidelines specify that the target victims, as part of the human factors, hold a degree of responsibility for the persistence of any threat. This dissertation re-shifts the lens to the intended targets of the falsehood attacks, the information consumers. This factor of consideration was investigated through its three-part, multi-method research phases.

The first phase, an interdisciplinary qualitative exploration of the different styles and techniques of preventive/prophylactic educational campaigns and theories that can be effectively used in raising awareness of the information consumers by employing a focus group of experts, revealed that ALA's CRAAP (Currency, Relevance, Authority, Accuracy, and Purpose) has the advantage and sustained its classical value over its counterpart prebunking (from the theory of inoculation). On the other hand, the second phase, where an outright comparison of the existing and suggested preventive treatments from the previous phase through quantifications of theories and actual user experimentations, revealed that users under the CRAAP treatment displayed greater detection accuracy (DV1) than the users on prebunking treatment. However, users on prebunking treatment resulted in a faster assessment time (DV2) compared to the users under the CRAAP group. Finally, in the third and final phase, this dissertation redesigned and improved the prevailing mis/disinformation predictive models by including the educational conditioning of users, as a factor, in forecasting, in multivariate format, their cyber risk from these attacks of deception.

Beyond policy implications, this project's significance includes contributions to the improvement of methods in scientific inquiry within the domain of risk modeling by combining self-reported data with digital trace data. Furthermore, this dissertation takes pride in its interdisciplinary design, considering the best practices of the granular fields involved to enable a more thorough investigation and probe through integration.

FATAL DISEASES DETECTION FROM ECG SIGNALS AND MRI IMAGES USING HYBRID DEEP LEARNING MODELS

Hari Mohan Rai
harimohanrai@gmail.com
Department of Electronics & Communication Engineering
Dronacharya Group of Institutions, Greater Noida, Uttar Pradesh, India

AUTOMATED and computerized detection and diagnosis of critical diseases or abnormalities from the human body with the highest accuracy is a very vital task for medical experts, researchers, and scientists. The work is divided into two segments: arrhythmias detection from Electrocardiogram (ECG) signals and tumor detection from MRI images. The ECG signal is obtained by recording the electrical activity of the human heart, and it is a noninvasive method employed as a principal diagnostic tool for the detection of cardiac arrhythmias. We know that the major disadvantage of the first system is that it is not very effective and is not suited for big datasets. Furthermore, the performance of the first system is dependent on the manual feature extraction method.

The six types of arrhythmias detection from ECG dataset has been employed using hybrid deep learning models (CNN-LSTM). To verify the model's performance, 123,998 ECG beats from the MIT-BIH arrhythmias database (MITDB) and the PTB diagnostic database (PTBDB) have been utilized. Because both datasets are extremely unbalanced, three data balancing methods have been employed to resample and balance the dataset in order to enhance the accuracy of the minority class classifications in both datasets. Using the SMOTE-Tomek link sampled dataset, the ensemble method is able to achieve an overall accuracy of 99.10 percent on a test dataset of 24,800 observations. The proposed hybrid deep learning models on big datasets offer very high accuracy for the detection of cardiac arrhythmias from ECG signals. Another hybrid deep learning model, UnetResNext-50, has been proposed for the detection of brain tumors from magnetic resonance imaging (MRI) images, which is one of the most deadly diseases. It is primarily the merging of the features of two distinct deep learning models, UNet and ResNext-50, with some modifications to the layers in the hybrid deep learning model that has been proposed.

The proposed model with the ResNext-50 backbone omits associations that take care of the degradation issue of deep

CNN models with more layers while improving pixel quality toward the vanilla Unet decoder. Total magnetic resonance imaging (MRI) dataset consisting of 3929 MR images, comprising 1373 images with tumors and 2556 images of non-tumorous kind (without tumors). The MRI dataset is primarily preprocessed by resizing, cropping, pixel normalization, and data splitting techniques before applying for segmentation and detection tasks. Also, the dataset used to train the proposed model is insufficient, therefore 12 different data augmentation (DA) functions are employed to generate a big training dataset (42,432) that can be used to train the DL models effectively. The six types of performance measurement metrics used to evaluate model proficiency are the Intersection over union or Jaccard index, F1-score, DICE score, precision, accuracy, and recall. The proposed model (UnetResNext-50) performance has been compared to two other deep learning models, Vanilla U-Net and UnetResNet-50, utilizing six types of performance assessment metrics to validate the system efficiency and accuracy. Post-processing techniques that use DICE and intersection over union (IoU) values are now being used to enhance tumor segmentation visibility. The developed hybrid deep learning model has great accuracy and accuracy for tumor identification, and it has 99.7score.

FREELANCING GERIATRIC CARE MONITORING SYSTEM IN AUSTRALIA

Hamid Ali
hamydaly@gmail.com
University of Southern Queensland, Australia

Due to the advancement in medical science, extreme decline has been seen in the death rate. Life expectancy has increased considerably, and people are enjoying a long life. Because of the reduction in the mortality rate, 20 percent of the world's population is expected to be 60 or older by 2050. It is quite common to transfer elderly people to Residential Aged Care Facilities (RACFs) from in-home care in developed countries like Australia. But due to the increasing elderly people and shortage of Personal Care Assistants (PCAs), RACFs are not enough to provide care and services. With the RACFs, elderly people lose their autonomy, independence and social interaction, to name a few things. That's why most Australians prefer to live in their homes and are frequently visited by the PCAs. AI-enabled freelancing strategy can help the geriatric care monitoring system in managing the demand and supply of PCAs in Australia.

In this paper, we propose AI-enabled freelancing strategy to help the geriatric care monitoring system in managing the PCAs. The goal is to propose a cheaper, more inclusive, and novice idea of integrating AI with freelancing strategy. Traditional research has been unable to solve this disproportionate issue of increasing elderly people and decreasing PCAs. Traditional research talks about the remote patient monitoring but that still have so many issues up to date. This novice idea will not only track but also predict the vital signs (body temperature, pulse rate, respiration rate, blood pressure etc) and send the signals to PCAs when help is needed. Vital signs are useful in detecting or monitoring medical problems at home. In the previous models, PCAs are bound in their duties, and they pay a visit to the elderly people without even being asked for. Since the Australian Government uses pay-as-you-go strategy, which means younger people are paying for the aged-care through taxation, it's going to be difficult for the government to manage in the future with less young generation and an increasing older generation.

There are many technologies that can be used but they are too expensive to implement, or the solutions are not concrete enough to solve the issue of aging people. Different solutions like ambient and wearable sensing, deploying of VR and implementation of 'smart shoe insoles' are too costly to implement. Previous research has given qualitative solutions while this research, which is of an exploratory nature, proposed a concrete solution to tackle the current problems and add to previous literature. With this freelance work, it will take the burden off full-time PCAs in the future. Similarly the intake of aged-care workers has been increasing, which will help to meet the growing demand of PCAs, as attracting and retaining workers is a big problem. The main purpose of clustered and in-home care is to support elderly people and giving the residents more autonomy in their routines and more flexibility when it comes to activities and outdoor access. So, this is a more flexible model, and it is going to help the Australian aged-care sector in the coming decades.

GRAPH MODEL FOR SCHEMA AND DATA MAPPING

Sonal Tuteja
sonalt9@gmail.com
Jawaharlal Nehru University
New Delhi, India

Graph model has emerged as a data modeling technique for large-scale applications with heterogeneous, schemaless, and relationship-centric data. It can help unveil relationships otherwise hidden in heterogeneous data sources. The flexibility and schemaless nature of graph models have led to various modeling techniques to map data from several data models into a graph model. However, the unification of data from heterogeneous source models into a graph model has not received much attention. Furthermore, designing graph models based on queries to be addressed has not been explored much. In addition, the role of data mapping techniques in a graph model for query performance has not been considered.

Addressing these research gaps, we propose a framework for unifying heterogeneous data sources into a graph model. We also analyze and compare the unified graph's query performance, scalability, and database size with heterogeneous source data models. We design various graph models for an e-commerce application incorporating queries to be executed and compare their performance with the baseline graph model. We define different data mapping techniques for graph models and verify their equivalence with the source graph model.

We observed that the graph model outperformed the relational and ontology models in all performance measures, except for aggregation queries. In query-driven graph models,

incorporating new nodes and edges improved the query performance of selection, projection, path-traversal operations, and their combinations. The designed data mapping techniques can be used in graph models for creating different relationships. Thus, our thesis designs and develops frameworks and techniques for improving a graph model's performance.

INTERACTIVE VISUALIZATION FOR INTERPRETABLE MACHINE LEARNING

Dennis Collaris
d.a.c.collaris@tue.nl
Eindhoven University of Technology Netherlands

MACHINE learning has firmly established itself as a valuable and ubiquitous technique in commercial applications. But these models are often complex and difficult to understand. Understanding models is particularly important in high-impact domains such as credit, employment, and housing, where the decisions made using machine learning impact the lives of real people. The field of eXplainable Artificial Intelligence (XAI) aims to help experts understand complex machine learning models. In recent years, various techniques have been proposed to open up the black box of machine learning. However, because interpretability is an inherently subjective concept it remains challenging to define what a good explanation is. We argue we should actively involve data scientists in the process of generating explanations and leverage their expertise in the domain and machine learning. Interactive visualization provides an excellent opportunity to both involve and empower experts.

In this dissertation, we explore interactive visualization for machine learning interpretation from different perspectives, ranging from local explanation of single predictions to global explanation of the entire model. We first introduce ExplainExplore: an interactive explanation system to explore explanations of individual predictions (i.e., local). For each explanation, it provides context by presenting similar predictions, and showing the impact of small input perturbations. We recognize many different explanations may exist that are all equally valid and useful using traditional evaluation methods. Hence, we leverage the domain knowledge of the data scientist to determine which of these fit their preference. To ensure these contributions can be broadly applied, we introduce a software library that enables interoperability with a wide range of different languages, toolkits, and enterprise software. Next, we propose the Contribution-Value plot as a new elementary building block for interpretability visualization, showing how feature contribution changes for different feature values. It provides a perspective in between local and global, as the model behavior is shown for all instances, but visualized on a per-feature basis. In a quantitative online survey with 22 participants, we show our visualization increases correctness, confidence, and reduces the time needed to obtain an insight compared to previous techniques.

This work highlighted that a small difference in feature importance techniques can result in a large difference in interpretation, and warranted a follow-up human computer interaction contribution to characterize the data scientists' mental model of explanations to explore the differences between existing techniques. Finally, we introduce StrategyAtlas: a visual analytics approach to enable a global understanding of complex machine learning models through the identification and interpretation of different model strategies. These model strategies are identified in our projection-based StrategyMap visualization. Data scientists can ascertain the validity of these strategies through analyzing feature values and contributions using heat maps, density plots, and decision tree abstractions. As computing the local feature importance values for an entire dataset is computationally expensive, we complement this work with an algorithmic contribution called LEMON to improve the faithfulness of explanation results, enabling significantly sped up computations of StrategyMap projections.

MACHINE-LEARNING-ASSISTED CORPUS EXPLORATION AND VISUALISATION

Tim Repke
repke@mcc-berlin.net
Information Systems Group, Hasso Plattner Institute at the University of Potsdam, Germany
https://hpi.de/naumann/people/tim-repke.html

TEXT collections, such as corpora of books, research articles, news, or business documents are an important resource for knowledge discovery. Exploring large document collections by hand is a cumbersome but necessary task to gain new insights and relevant information. Our digitized society allows us to utilize algorithms to support the information seeking process, for example, with the help of retrieval or recommender systems. However, these systems only provide selective views of the data and require prior knowledge to issue meaningful queries and asses a system's response. The advancements of machine learning allow us to reduce this gap and better assist the information seeking process. For example, instead of sighting countless business documents by hand, journalists and investigators can employ natural language processing techniques such as named entity recognition.

Although this greatly improves the capabilities of a data exploration platform, the wealth of information is still overwhelming. An overview of the entirety of a dataset in the form of a two-dimensional map-like visualization may help to circumvent this issue. Such overviews enable novel interaction paradigms for users, which are similar to the exploration of digital geographical maps. In particular, they can provide valuable context by indicating how a piece of information fits into the bigger picture. This thesis proposes algorithms that appropriately preprocess heterogeneous documents and compute the layout for datasets of all kinds. Traditionally, given high-dimensional semantic representations of the data, so-called dimensionality reduction algorithms are used to compute a layout of the data on a two-dimensional canvas. In this thesis, we focus on text corpora and go beyond only projecting the inherent semantic structure itself. Therefore, we propose three dimensionality reduction approaches that incorporate additional information into the layout process: (1) a multi-objective dimensionality reduction algorithm to jointly

visualize semantic information with inherent network information derived from the underlying data; (2) a comparison of initialization strategies for different dimensionality reduction algorithms to generate a series of layouts for corpora that grow and evolve over time; (3) and an algorithm that updates existing layouts by incorporating user feedback provided by pointwise drag-and-drop edits. In the scope of this thesis, we also developed system prototypes to demonstrate the proposed technologies, including pre-processing and layout of the data and presentation in interactive user interfaces.

MINIMIZING USER EFFORT IN LARGE SCALE EXAMPLE-DRIVEN DATA EXPLORATION

Xiaoyu Ge
xiaoyu@cs.pitt.edu
Department of Computer Science University of Pittsburgh, USA
https://d-scholarship.pitt.edu/41742

THE ever-increasing supply of data is bringing renewed attention to data exploration, a technique that serves as a key ingredient in a widely diverse set of discovery-oriented applications, including scientific computing, financial analysis, and evidence-based medicine. One major challenge for those discovery-oriented applications is the need to extract useful pieces of knowledge from data while requiring little to no specification of the information that is being searched for. As the traditional searching and data mining techniques fall short in meeting such challenging demands, data exploration techniques aimed at intelligently assisting users in constructing precise exploratory queries have recently generated a lot of interest in both the academic and industrial communities and have led to the development of a variety of semi-automatic data exploration approaches. Among such approaches, Example-driven Exploration is rapidly becoming an attractive choice for exploratory query formulation since it attempts to minimize the amount of prior knowledge required from the user to form an accurate exploratory query.

This dissertation focuses on interactive Example-driven Exploration, which steers the user toward discovering all data objects relevant to the users' exploration based on their feedback on a small set of examples (i.e., data objects selected from the underlying dataset). Interactive Example-driven Exploration is especially beneficial for non-expert users as it leverages human-in-the-loop paradigms and enables them to circumvent query languages by assigning relevancy to the presented examples and leveraging them as proxies for the intended exploratory analysis. However, existing interactive Example-driven Exploration systems fall short of supporting the need to perform complex explorations for data that are large, unstructured, or high-dimensional. To overcome these challenges, in this dissertation, we have developed novel methods that facilitate the Example-driven Exploration paradigm in four different areas: data reduction, example selection, data indexing, and result refinement, which help to support largescale complex discovery-oriented applications.

The novelty of our approach is anchored on leveraging active learning and query optimization techniques. The prior enables semi-automatic exploration and reduces manual user effort. The latter helps to reduce the potential exploration space for the exploration system. Together they strike a balance between maximizing accuracy and minimizing user effort in providing feedback while enabling interactive performance on the system level for exploration tasks with arbitrary, large sized datasets. Furthermore, our proposed approach extends the exploration beyond the traditional structured data by supporting a variety of high-dimensional unstructured data and enables the refinement of results, which prevents the results from being overwhelming to the user when the exploration task is associated with too many relevant data objects. To affirm the effectiveness of our proposed models, techniques, and algorithms, we implemented multiple prototype systems and evaluated them using real-world datasets. Some of them have also been incorporated into domain-specific analytic tools. Our comprehensive evaluations have shown that our exploration methods help to reduce users' manual effort by up to 9x while achieving the same accuracy as the state-of-the-art alternatives. Furthermore, our data reduction and result refinement methods significantly reduced the system run-time and achieved a speedup of up to 159x when exploring large and complex datasets.

REMOTE PATIENT MONITORING SYSTEM USING ARTIFICIAL INTELLIGENCE

Thanveer Shaik
Thanveer.Shaik@usq.edu.au
University of Southern Queensland, Toowoomba, Australia

HEALTHCARE applications are vastly dependent on artificial intelligence (AI) methodologies to enable forecasting capability and provide utmost care to patients in hospital as well as remotely. Especially in psychiatric care with acute mental illness and depressed suicidal tendency patients, the goal is to provide a safe and therapeutic environment to both patients and medical staff by avoiding the physical violence caused by aggressive and agitated patients. This could be achieved by monitoring patients continuously to detect their movements and vital signs such as heart rate, respiratory rate, and breathing. The frequency of manually recorded patient reports is limited due to staff availability and number of patients in a hospital. Remote patient monitoring (RPM) could enable the continuous monitoring of acutely ill patients in psychiatric care. Empowering the RPM strategy with AI methods could transform healthcare monitoring by predicting patients' future vital signs and also classify their body movements.

In this study, the aim is to propose an AI enabled RPM system with non-invasive technology. Traditional RPM systems with invasive technology such as electrocardiography (ECG) and photoplethysmography (PSG) touch patients' skin to record their data and could cause inconvenience, limiting their daily actions. We propose a non-invasive technology radio frequency identification (RFID) based on near-field coherent sensing (NCS) without touching patients' body and allow their daily activities. RFID passive tags will need to be arranged at different areas of body such as chest area, abdomen, and limbs to record the vital signs and their body

motion. The passive tags data will be retrieved via RFID reader-antennas to a computer to process and retrieve patients' vital signs. Advanced AI methodologies such as reinforcement learning, explainable AI (ExAI) and federated learning will be adopted for adaptive learning of patients' behavior, to enhance interpretability and transparency in AI model results for decision-making, and enable personalized monitoring with patient privacy, respectively. The adaptive learning will consider each patient as an individual learning agent in a hospital environment, attempting to achieve maximum rewards by following the designed policy of staying safe clinically.

In ExAI, AI models will be interpreted based on their weights and also estimate Shapley values to extract feature importance, and model behavior. Personalized monitoring will be achieved by adopting a federated learning approach in which each patient data will be monitored individually using a local AI model and pass only the model parameters or predictions to build a robust global model. The proposed AI enabled RPM system would monitor patients' behavior adaptively while forecasting their future vital signs and classifying their physical activities. This research contributes a novel patient monitoring system that learns patient behavior, and assists clinicians with a decision support system that makes timely interventions and avoids acute disturbances in psychiatric care.

REPRESENTATION LEARNING FOR TEXTS AND GRAPHS: A UNIFIED PERSPECTIVE ON EFFICIENCY, MULTIMODALITY, AND ADAPTABILITY

Lukas Galke
Lukas.Galke@mpi.nl
Kiel University, Germany Max Planck Institute for Psycholinguistics, Netherlands

FUELED by deep learning, natural language processing is becoming increasingly influential. Meanwhile, graph representation learning shows how to process graph data effectively. However, the size of language models and large, evolving graphs becomes increasingly challenging. The immense computing power and GPU memory requirements make it difficult for small companies and research labs to participate. Thus, this thesis aims to find efficient text and graph representation learning methods that can adapt to new data. This thesis is situated between text and graph representation learning and investigates selected connections.

First, we introduce matrix embeddings as an efficient text representation sensitive to word order. After self-supervised pretraining, the matrix product acts as sentence encoding for downstream tasks. Experiments with ten linguistic probing tasks, 11 supervised, and five unsupervised downstream tasks reveal that vector and matrix embeddings have complementary strengths and that a jointly trained hybrid model outperforms both. Second, a popular pre-trained language model, BERT, is distilled into matrix embeddings. To this end, we extend matrix embeddings with a bidirectional component and equip them with a strategy to encode sentence pairs. The results on the GLUE benchmark show that these models are competitive with other recent contextualized language models while being more efficient in time and space. Third, we compare three model types for text classification: bag-of-words, sequence-, and graph-based models. Experiments on five datasets show that, surprisingly, a wide multilayer perceptron on top of a bag-of-words representation is competitive with recent graph-based approaches, questioning the necessity of graphs synthesized from the text. Pretrained Transformer-based sequence models perform best but come with high computational costs.

Fourth, we investigate the connection between text and graph data in document-based recommendation systems for citations and subject labels. Experiments on six datasets show that the title as side information improves the performance of autoencoder models. We confirm this result under different experimental conditions: the number of all possible items and the fraction of already-present items per document. We find that the meaning of item co-occurrence is crucial for the choice of input modalities and an appropriate model. Fifth, we introduce a generic framework for lifelong learning on evolving graphs in which new nodes, edges, and classes appear over time. The task is to classify nodes and detect new classes based on textual and graph information. We experiment with five representative graph neural network models and three datasets based on scholarly articles: two citation graphs and one collaboration graph.

The results show that by reusing previous parameters in incremental training, it is possible to employ smaller history sizes with only a slight decrease in accuracy compared to training with complete history. Furthermore, weighting the binary cross-entropy loss function is essential for automatically detecting newly emerging classes. This work opens up new opportunities for efficient text and graph representation learning. It shows how recommender systems can exploit textual side information and lays the foundation for lifelong and open-world learning in evolving graphs with text-attributed nodes.

SECURE CONTENT DELIVERY IN TWO-TIER CACHE-AIDED SATELLITE INTERNET OF THINGS NETWORKS

Quynh Tu Ngo
T.Ngo@latrobe.edu.au
La Trobe University, Melbourne, Australia

INTEGRATING satellites into terrestrial communication systems has been identified as a critical solution in the new era of 6G to enable global connectivity for the Internet of Things (IoT). With better service reliability and coverage, satellite communications have been an effective complementary component for IoT in addition to cellular networks. However, embedding satellites into IoT networks poses challenges in excessive service delays, expensive cost, and security issues when relating space communications. A content delivered to users is forwarded from an Internet-connected gateway through satellites and ground stations, which extends the serving time in addition to very pricey and often limited satellite bandwidth. A large-scale wireless network created when embedding satellites into IoT is exposed to more security risks due to satellite broadcast nature and wide coverage.

To tackle these challenges, this thesis first studies edge caching techniques in satellite-terrestrial network (STN) to guarantee latency and cost of services, and then physical layer security (PLS) techniques are exploited in cache-enabled STN to ensure data confidentiality. We begin by examining a two-tier cache-enabled STN model with a focus on the content delivery latency analysis in terms of the successful delivery probability (SDP). We consider enabling caching capacity in combination with fullduplex transmission at both satellite and ground station to shorten service delivery time and reduce in-network traffic. We derive a closed-form expression for the SDP given consideration to the requested content distributions, realistic channel statistics, and three commonly studied caching configurations.

Based on the derived results, we investigate the system SDP performance under different transmission modes and network settings. The network capacity in terms of maximum number of supportable users, satellite bandwidth and energy consumption are also studied. Then we design an SDP maximization-based cache placement strategy subjected to caching capacity constraints at satellite and ground station. We then secure the transmission of the two-tier cacheenabled STN by exploiting the intelligent reflecting surfaces (IRS). A novel two-hop secure content delivery scheme is proposed, and the PLS performance is investigated in terms of the secure transmission probability. Closed-form expressions for the connection probability and secrecy probability over two cascaded fading channels, i.e., the Rayleigh–Rayleigh and the Rayleigh–Shadowed-Rician fading channels are derived. Based upon the derived results, we form the system secure transmission probability, which we aim to maximize when jointly designing the transmission rates and caching probability at satellite and ground station.

To further explore the potential of IRS enhancing PLS systems, we study a hybrid IRS-assisted secure multiuser multiple-input single-output STN and analyze system performance in terms of worst-case secrecy sum-rate. A robust and secure beamforming design problem is formulated for satellite as well as hybrid IRS under practical outdated channel state information and power consumption models. We leverage deep reinforcement learning (DRL) to solve the problem by proposing a fast DRL algorithm, namely deep post-decision state–deterministic policy gradient (DPDS-DPG). DPDS-DPG exploits prior known system dynamics by integrating the PDS concept into the traditional deep DPG (DDPG) algorithm, resulting in faster learning convergence. Simulation results show better learning efficiency of DPDSDPG than DDPG with comparable achievable system secrecy rate, demonstrating the performance gains of employing hybrid IRS over conventional passive IRS to support secure communications.

SOFT BIOMETRICS-BASED PERSON RETRIEVAL FROM UNCONSTRAINED SURVEILLANCE VIDEO

Hirenkumar Jagdishchandra Galiyawala
hirenkumar.g@ahduni.edu
Ahmedabad University, Gujarat, India
http://hdl.handle.net/10603/429586

In today's security world, one interesting direction is to search and locate a person of interest from unconstrained surveillance videos using a textual description. An automated computer vision system that translates a description and retrieves the person would greatly assist. A description, e.g., a tall female with a grey short sleeve t-shirt and white jeans with a handbag, is a prime entity for retrieval. It is because the description has personal attributes (soft biometrics) like height, gender, clothing color, and clothing type. This thesis considerably contributes to developing state-of-the-art approaches for retrieving person(s) from videos using a soft biometrics-based textual description. The contributions are as follows; (1) The development of cascade filter-based approaches incorporates adaptive torso patch extraction, better height estimation, and Intersection-over-Union (IoU)-based bounding box regression, (2) The design and implementation of an attributes' weighting module for a person ranking-based approach, (3) The creation of a dataset for person attribute recognition, (4) The development of a multi-attribute learning-based model with fewer parameters which predicts all attributes using a single model, (5) The design and development of a state of the art end-to-end person retrieval approach.

The initial approach uses semantic segmentation for precise feet and head point extraction for better height estimation using the camera calibration approach. It uses only three attributes, i.e., height, gender, and torso color, where gender and color models are trained using AlexNet. This approach achieves an average IoU of 0.363, a performance of 0.522 with IoU $\geq$ 0.4, and a True Positive Rate (TPR) of 54.12%. It was improved by introducing adaptive torso patch extraction, bounding box regression, better data augmentation for color and gender models, and DenseNet-169 classification models. Person retrieval is done using height, torso type, torso color I, torso color II, and gender. This approach performs well for the person with torso type "no sleeve" as chances of unwanted pixels from the human body are minimized to improve the classification accuracy. It achieves an average IoU of 0.569, 0.746 with IoU $\geq$ 0.4, and a TPR of 76.21%.

Further improvement is made by utilizing the contribution of each attribute and a ranking strategy. This strategy performs well for a person with partial occlusion, and the resulting approach outperforms previous approaches and achieves state-of-the-art performance. This approach achieves an average IoU of 0.602, 0.808 with IoU $\geq$ 0.4, and TPR of 82.14%. Introducing a single attribute-recognizable model enables a solution toward an end-to-end person retrieval system. Considering features' correlation, a better dataset and attention mechanism with focal loss proved best for attribute recognition. It outperforms existing techniques by utilizing only five attributes. This approach achieves an average IoU of 0.667, 0.856 with IoU $\geq$ 0.4, and a TPR of 85.30%. Findings from this dissertation contribute to achieving the best performance in person retrieval from unconstrained surveillance video using soft biometrics. The work is supported by the BRNS, Govt. of India grant to supervisor Mehul S Raval.

TOWARDS SUSTAINABILITY AND SAFETY SOLUTIONS IN A SMART CITY

Federica Rollo
federica.rollo@unimore.it
Enzo Ferrari Engineering Department University of Modena and Reggio Emilia Modena, Italy

A smart city is a place where technology is exploited to help Public Administrations make decisions. Technology can contribute to managing multiple aspects of everyday life, offering more reliable services to citizens and improving the quality of life. However, technology alone is not enough to make a Smart City; suitable methods are needed to collect data generated by technology, analyze and manage them in such a way as to produce useful information. The thesis focuses on two aspects of Smart Cities: sustainability and safety. The first aspect is addressed by studying the impact of vehicular traffic on air quality through the development of traffic and air quality sensor networks and the implementation of a chain of simulation models.

This work is part of the European Union co-financed TRAFAIR project, which aims to monitor in real-time and predict air quality on an urban scale in 6 cities, including the Italian city of Modena. The project requires the management of a large amount of heterogeneous data and its integration on a scalable platform. To manage spatio-temporal data, the Smart City data platform has been implemented as a PostgreSQL database (60+ tables and 435 GB of data in 2 years - only for Modena). The simulation models used in the project to reconstruct traffic flow and forecast air quality in the cities are based on sensor data. Since sensors are prone to errors, a data cleaning process is needed to ensure reliable results of the simulation. After studying the sensor data distribution and the correlation among sensors, several anomaly detection techniques have been implemented. A novel approach employing a flow-speed correlation filter, STL decomposition and IQR has been developed for traffic sensor data as well as an innovative ensemble method for air quality sensors considering the influence between pollutant measurements. The evaluation of real-world data of Modena demonstrated the efficiency and effectiveness of these techniques.

In the thesis, the safety aspect is examined by the development of a crime analysis project which aims at generating timely and pertinent information for crime reduction, prevention, and evaluation. Due to the lack of official data, this project exploits online news articles. The goal is to categorize news articles based on crime category, geolocate crime events, detect the date of the event, and identify other relevant features (e.g., what has been stolen during the theft). A novel framework has been developed for the analysis of news articles, the extraction of semantic information by using NLP techniques, and the connection of entities to Linked Data. The emerging technology of word embeddings has been employed for text categorization. News articles referring to the same event have been identified through the application of cosine similarity. Finally, a dashboard has been developed to show the geolocated events and provide statistics and annual reports. The framework allowed the production of the Italian Crime News dataset (available online), collecting 15,000+ news articles. The impact and scalability of such a framework has been evaluated on two online newspapers of Modena. This is the first framework that, starting from Italian news articles, provides analyses of crimes and makes them available through a visualization tool.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

WI-IAT 2023
The 2023 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology
TBA
https://www.wi-iat.com/wi-iat2023/index.html

The 2023 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'23) provides a premier international forum to bring together researchers and practitioners from diverse fields for presentation of original research results, as well as exchange and dissemination of innovative and practical development experiences on Web intelligence and intelligent agent technology research and applications.

Web has evolved as an omnipresent system which highly impacts science, education, industry and everyday life. Web is now a vast data production and consumption platform at which threads of data evolve from multiple devices, by different human interactions, over worldwide locations under divergent distributed settings. Such a dynamic complex system demands adaptive intelligent solutions, which will advance knowledge, human interactions and innovation. Web intelligence is now a cutting-edge area which must address all open issues towards deepening the understanding of all Web's entities, phenomena, and developments.

WI-IAT'23 aims to achieve a multi-disciplinary balance between research advances in theories and methods usually associated with collective intelligence, data science, human-centric computing, knowledge management, network science, autonomous agents and multi-agent systems. It is committed to addressing research that both deepen the understanding of computational, logical, cognitive, physical, and social foundations of the future Web, and enable the development and application of intelligent technologies. WI-IAT'23 provides a premier forum and features high-quality, original research papers and real-world applications in all theoretical and technology areas that make up the field of Web Intelligence and Intelligent Agent Technology. WI-IAT'23 welcomes research, application as well as Industry/Demo-Track paper submissions. Tutorial, Workshop and Special-Session proposals and papers are also welcome. This conference is officially sponsored by the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), ACM-SIGAI, Wilfrid Laurier University, and York University.

WI-IAT'23 will provide a broad forum that academia, professionals and industry people can exchange their ideas, findings and strategies in utilizing the power of human brains and man-made networks to create a better world. More specifically, the fields of how intelligence is impacting the Web of People, the Web of Data, the Web of Things, the Web of Trust, the Web of Agents, and emerging Web in health and smart living in the 5G Era. Therefore, the theme of WI-IAT'23 will be "Web Intelligence = AI in the Connected World".

Please refer to the conference website for further information about the venue and attendance dates.

IEEE ICDM 2023
The 23rd IEEE International Conference on Data Mining
Shanghai, China
December 1-4, 2023
http://www.cloud-conf.net/icdm2023/index.html

The IEEE International Conference on Data Mining (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative and practical development experiences. The conference covers all aspects of data mining, including algorithms, software, systems, and applications. ICDM draws researchers, application developers, and practitioners from a wide range of data mining related areas such as big data, deep learning, pattern recognition, statistical and machine learning，databases, data warehousing, data visualization, knowledge-based systems, and high-performance computing. By promoting novel, high-quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to advance the state-of-the-art in data mining.

Topics of interest at this year's conference include: Foundations, algorithms, models and theory of data mining, including big data mining. Deep learning and statistical methods for data mining; Mining from heterogeneous data sources, including text, semi-structured, spatio-temporal, streaming, graph, web, and multimedia data; Data mining systems and platforms, and their efficiency, scalability, security and privacy; Data mining for modelling, visualization, personalization, and recommendation; Data mining for cyber-physical systems and complex, time-evolving networks; and Applications of data mining in social sciences, physical sciences, engineering, life sciences, web, marketing, finance, precision medicine, health informatics, and other domains. There is also an encouragement of submissions for emerging topics of high importance, such as ethical data analytics, automated data analytics, data-driven reasoning, interpretable modeling, modeling with evolving environment, cyber-physical systems, multi-modality data mining, and heterogeneous data integration and mining.

Authors are invited to submit original papers, which have not been published elsewhere and which are not currently under consideration for another journal, conference or workshop. Paper submissions should be limited to a maximum of ten (10) pages, in the IEEE 2-column format (https://www.ieee.org/conferences/publishing/templates.html), including the bibliography and any possible appendices. Further information for preparing paper submissions are available on the conference website. There is a current call for papers, with the submission closing date on the 1st of July 2023 and authors notified of acceptance on September 1st. Papers must be submitted through the ICDM online submission system on the Call for Papers page on the

conference website.

IEEE ICDM 2023 will also host workshops and tutorials during the conference. Workshops are expected to be held on December 4th, with half- or full-day workshops complementing the main conference technical program through expanding new directions and applications of data mining for both practitioners and researchers in a particular field. ICDM also invites proposals for tutorials from active researchers and experienced tutors. Ideally, a tutorial will cover the state-of-the-art research, development and applications in a specific data mining direction, and stimulate and facilitate future work. Tutorials on interdisciplinary directions, novel and fast-growing directions, and significant applications are highly encouraged. Hands-on-tutorials are also encouraged, particularly ones that will allow attendees to learn a particular suite of tools, software and applications that are relevant to the data mining community. Software and tools should preferably be open-sourced and readily available to all participants.

Workshop proposals are due by March 5, 2023, while tutorial proposals are due by July 9, 2023. Both workshop and tutorial proposals must be submitted by email, with workshop proposals directed to the ICDM Workshops Chairs (emails available on the Call for Workshops page) and tutorial proposals sent to icdm2023chairs@gmail.com with the subject line ICDM2023_<tutorial_name>.

Registration for the conference closes on October 15, 2023. For further questions regarding the IEEE ICDM conference, please email ICDM2023Chairs@gmail.com.

ICHI 2023
The 11th IEEE International Conference on Healthcare Informatics
Houston, Texas, USA
June 26-29, 2023
https://ieeeichi.github.io/ICHI2023/

The IEEE International Conference on Healthcare Informatics (ICHI 2023) is a premier community forum concerned with the application of computer science, information science, data science, and informatics principles, as well as information technology, and communication science and technology to address problems and support research in healthcare, medicine, life science, public health, and everyday wellness. The conference highlights the most novel technical contributions to stakeholder-centered technology innovation for benefiting human health and the related social and ethical implications. ICHI 2023 will feature keynotes, a multi-track technical program including papers, posters, panels, workshops, tutorials, an industrial track, and a doctoral consortium.

Authors are strongly encouraged to submit their original contributions describing their algorithmic, methodological or empirical contributions, and theories relevant to the broader context of health informatics. Submissions can focus on one or more specific aspects of theory, design, development, evaluation, or deployment. The conference covers three tracks: Analytics, Human Factors and Systems. When submitting papers, the authors must select a track that is most appropriate for their submission. Further information about each track is available on the website.

Posters and demos are an excellent way to present innovative ideas, late-breaking work, concepts, work-in-progress, early stages of research, and preliminary results from implementation and validations to academic and industrial audience. These must also fit the three tracks and should be 1-2 pages (excluding references) in the same format as full and short papers. Workshops will be held on the day prior to the start of the main program (June 26th, 2023) and should address topics relevant to healthcare informatics. ICHI 2023 also offers a Doctoral Consortium that provides an outstanding opportunity for doctoral students working on health informatics problems to discuss their work in progress and receive feedback and guidance from Consortium mentors. There is also a call for papers for the Industry track where submissions are invited to describe practical implementations of healthcare informatics, and following the success in previous years, a call for tutorial proposals. Tutorials at ICHI 2023 will be presented by domain experts to cover current topics directly relevant to the conference theme of computing-oriented health informatics. A list of topics of interest to ICHI 2023 include health data science, health informatics, innovative technologies and implementation science. Tutorials can be half-day (3 hours) or full-day (6 hours) in length. Tutorial instructors must make a commitment to prepare tutorial materials (including slides and activities) that reflect the high-quality standard of ICHI.

All submissions for papers, posters and demos, workshops, the Doctoral Consortium, Industry track and tutorials must be made through the EasyChair system on the conference website. The deadline for all submissions is January 31, 2023.

IEEE BigData 2023
The 2023 IEEE International Conference on Big Data (IEEE BigData 2023)
TBA
https://bigdataieee.org/BigData2023/

In recent years, “Big Data” has become a new ubiquitous term. Big Data is transforming science, engineering, medicine, healthcare, finance, business, and ultimately our society itself. The IEEE Big Data conference series started in 2013 has established itself as the top tier research conference in Big Data. The 2023 IEEE International Conference on Big Data (IEEE BigData 2023) will continue the success of the previous IEEE Big Data conferences. It will provide a leading forum for disseminating the latest results in Big Data Research, Development, and Applications.

IEEE BigData 2023 will cover topics such as Big Data Infrastructure, Big Data Science and Foundations, Big Data Search and Mining, Big Data Management, Data Ecosystem, Big Data Learning and Analytics, and Big Data Applications. Each topic covers a number themes and sub-topics relevant to big data research.

The conference might be held in 2023. Please refer to the website for upcoming announcements and details about the date and venue for IEEE BigData 2023.

IEEE ICKG 2023
ICKG 2023: 17. IEEE International Conference on Knowledge Graphs (ICKG)
Amsterdam, Netherlands
November 4-5, 2023
https://
https://waset.org/knowledge-graphs-conference-in-november-2023-in-amsterdam

International Conference on Knowledge Graphs (ICKG) aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of Knowledge

Graphs. It also provides a premier interdisciplinary platform for researchers, practitioners and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of Knowledge Graphs.

Prospective authors are kindly encouraged to contribute to and help shape the conference through submissions of their research abstracts, papers and e-posters. Also, high quality research contributions describing original and unpublished results of conceptual, constructive, empirical, experimental, or theoretical work in all areas of Knowledge Graphs are cordially invited for presentation at the conference. The conference solicits contributions of abstracts, papers and e-posters that address themes and topics of the conference, including figures, tables and references of novel research materials. The call for papers cover themes such as: Automatic and semi-automatic creation of knowledge graphs; Data integration, disambiguation, schema alignment; Collaborative management of knowledge graphs; Quality control: noisy data, uncertainty, incomplete information; New kinds of knowledge graphs: common-sense, visual knowledge; Architectures for managing big graphs; Expressive query answering; Reasoning with large-scale, dynamic data; Data dynamics, update, and synchronization; Synthetic graphs and graph benchmarks; Innovative uses of knowledge graphs; Understanding and analyzing knowledge graphs; Semantic search; Question answering; and Combining knowledge graphs with other information resources.

All submitted conference papers will be blind peer reviewed by three competent reviewers. ICKG has teamed up with the Special Journal Issue on Knowledge Graphs. A number of selected high-impact full text papers will also be considered for the special journal issues. All submitted papers will have the opportunity to be considered for this Special Journal Issue. The paper selection will be carried out during the peer review process as well as at the conference presentation stage. Submitted papers must not be under consideration by any other journal or publication. The final decision for paper selection will be made based on peer review reports by the Guest Editors and the Editor-in-Chief jointly. Selected full-text papers will be published online free of charge.

The deadline for Abstracts and Full-Text Paper submissions is January 31, 2023. All submissions should be through the conference website, where there are submission guides for Papers and Abstracts, and further Author Information. The Early bird registration deadline for the conference is October 4, 2023.

AAMAS 2023
The 22nd International Conference on Autonomous Agents and Multi-Agent Systems
London, England
May 29-June 2, 2023
https://aamas2023.soton.ac.uk/

Autonomous Agents and Multiagent Systems (AAMAS) is the largest and most influential conference in the area of agents and multiagent systems, bringing together researchers and practitioners in all areas of agent technology and providing and internationally renowned high-profile forum for publishing and finding out about the latest developments in the field. AAMAS is the flagship conference of the non-profit International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

After three years of virtual AAMAS conferences, many in the research community are eager for an in-person AAMAS conference once again. While virtual interactions are a reasonable substitute when used appropriately, they still lack key facets and benefits that in-person interactions offer. For those unable to travel to London due to personal health constraints or traveling restrictions, AAMAS will provide limited support to enable authors to present their work remotely for those who really cannot travel to the in-person event. The conference also plans to live stream the sessions using low-cost solutions. There will not be any support for other online infrastructure, such as a virtual platform or GatherTown. Should the COVID-19 situation worsen over the coming months to make a physical gathering in London unfeasible, AAMAS will transition the conference to either a virtual or hybrid setting.

Notifications were sent on 4 January. Out of 1015 submissions, 237 were accepted as full papers and a further 221 were accepted as extended abstracts. Therefore, the conference has an acceptance rate of 23.3% for full papers and 45.1% for full papers + extended abstracts. Papers are associated with 10 areas of interest, including: Coordination, Organisations, Institutions, and Norms; Markets, Auctions, and Non-Cooperative Game Theory; Social Choice and Cooperative Game Theory; Knowledge Representation, Reasoning, and Planning; Learning and Adaptation; Modelling and Simulation of Societies; Humans and AI / Human-Agent Interaction; Engineering Multiagent Systems; Robotics, and Innovative Applications. The list of accepted papers is available at: https://aamas2023.soton.ac.uk/program/accepted-papers/.

AAMAS 2023 will feature Tutorials, Demos, Competitions, the Doctoral Consortium, a JAAMAS track for presentation of articles that have appeared or been accepted for publication in the Journal of Autonomous Agents and Multi-Agent Systems, and the Blue Sky Ideas special track for visionary ideas, long-term challenges, new research opportunities, and controversial debate. The conference also features a workshop program, to be held on May 29-30, immediately prior to the main program of the AAMAS conference. The objective of the AAMAS 2023 workshop program is to stimulate and facilitate discussion, interaction, and comparison of approaches, methods, and ideas related to specific topics, both theoretical and applied, in the general area of Autonomous Agents and Multiagent Systems. The list of accepted workshops is available on the conference website. This year also features a workshop on diversity and inclusion, with the Multiagent Systems for Diversity and Inclusion (MSDI) workshop dedicated to discussing the conception and deployment of inclusive MAS.

Accommodation booking is now open, with TFI Lodestar the official accommodation partner of AAMAS and ICRA. There are plenty of options for discounted individual and group bookings. If you plan to use this service for booking your accommodation, please reserve your rooms no later than February 19, 2023.

AAAI 2023
The 37th AAAI Conference on Artificial Intelligence
Washington, DC, USA
February 7-14, 2023
https://aaai.org/Conferences/AAAI-23/

The purpose of the AAAI conference series is to promote research in Artificial Intelligence (AI) and foster scientific exchange between

researchers, practitioners, scientists, students, and engineers across the entirety of AI and its affiliated disciplines. AAAI-23 is the Thirty-Seventh AAAI Conference on Artificial Intelligence. The theme of this conference is to create collaborative bridges within and beyond AI. Like previous AAAI conferences, AAAI-23 will feature technical paper presentations, special tracks, invited speakers, workshops, tutorials, poster sessions, senior member presentations, competitions, and exhibit programs, and two new activities: a Bridge Program and a Lab Program. Many of these activities are tailored to the theme of bridges and all are selected according to the highest standards, with additional programs for students and young researchers.

The purpose of this year's Bridge Program is to tap into new sources of innovation by cultivating sustained collaboration between two or more communities, directed towards a common goal. AAAI-23's interpretation of bridge is broad and encompasses disciplines both within and outside of AI. Hence, the communities the Bridge Program is intended to bring together could be distinct subfields of AI, such as planning and learning, or different disciplines that contribute to and benefit from AI, such as AI and the humanities.

AAAI-23 welcomes submissions reporting research that advances artificial intelligence, broadly conceived. The conference scope includes machine learning (deep learning, statistical learning, etc), natural language processing, computer vision, data mining, multiagent systems, knowledge representation, human-in-the-loop AI, search, planning, reasoning, robotics and perception, and ethics. In addition to fundamental work focused on any one of these areas we expressly encourage work that cuts across technical areas of AI, (e.g., machine learning and computer vision; computer vision and natural language processing; or machine learning and planning), bridges between AI and a related research area (e.g., neuroscience; cognitive science) or develops AI techniques in the context of important application domains, such as healthcare, sustainability, transportation, and commerce.

Most papers in AAAI-23 will be part of the main track. This conference has two additional tracks, which focus on AI for Social Impact and Safe and Robust AI. As in past years, AAAI-23 will include a Track on AI for Social Impact (AISI). Submissions to this track will be reviewed according to a rubric that emphasizes the fit between the techniques used and a problem of social importance, rather than simply rewarding technical novelty. In particular, reviewers will assess significance of the problem being addressed; the paper's engagement with previous literature on the application problem (whether in the AI literature or elsewhere); both novelty of and justification for the proposed AI-based approach; quality of evaluation; facilitation of follow-up work; and overall scope and promise for social impact.

This year, AAAI-23 is introducing a new special track on AI systems that are safe and robust. AI is increasingly being deployed throughout society. To ensure that this technology is trustworthy, it needs to be robust to disturbance, failure and novel circumstances. Furthermore, the technology needs to offer assurance that it will reasonably avoid unsafe, irrecoverable situations. Submissions to this track will be reviewed according to a rubric that emphasizes the fit between the driving requirements of safety and robustness for AI systems and the methods and formalisms presented.

The conference this year features diversity and inclusion activities, with example topics and activities including new open research questions that affect diversity in AI, mentoring activities for students from underrepresented groups, community-building for young researchers (graduate students and postdocs), activities to expose undergraduates to research or K-12 students to AI, and reports on successes at increasing and sustaining diversity. These activities are sponsored by the Association for the Advancement of Artificial Intelligence. AAAI-23 also features a Demonstration program, New Faculty Highlight talks, a Senior Member Presentation Track (SMPT), a student abstract and poster program, a Tutorial and Lab Forum, the 28th AAAI/SIGAI Doctoral Consortium and an Undergraduate Consortium, which will provide students who are more than one year from graduation with significant enrichment opportunities in a professional setting.

The AAAI-23 workshop program includes 32 workshops covering a wide range of topics in artificial intelligence. Workshops are one day unless otherwise noted in the individual descriptions. Registration in each workshop is required by all active participants, and is also open to all interested individuals.

AAAI-23 plans to be an in-person conference and is exploring opportunities to complement this with remote participation. Given the volatile situation, however, the exact format of the conference can only be decided at a later stage.

SDM23
The 2023 SIAM International Conference on Data Mining
Minneapolis, Minnesota, USA
April 27-29, 2023
https://www.siam.org/conferences/cm/conference/sdm23

Data mining is the computational process for discovering valuable knowledge from data – the core of Data Science. It has enormous application in numerous fields, including science, engineering, healthcare, business, and medicine. Typical datasets in these fields are large, complex, and often noisy. Extracting knowledge from these datasets requires the use of sophisticated, high-performance, and principled analysis techniques and algorithms, which are based on sound theoretical and statistical foundations. These techniques in turn require implementations on high performance computational infrastructure that are carefully tuned for performance. Powerful visualization technologies along with effective user interfaces are also essential to make data mining tools appealing to researchers, analysts, data scientists and application developers from different disciplines, as well as usable by stakeholders.

The SDM conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students to network and get feedback for their work (as part of the doctoral forum). Everyone new to the field can also learn about cutting-edge research by hearing outstanding invited speakers and attending presentations and tutorials (included with conference registration). A set of focused workshops is also held on the last day of the conference. The proceedings of the conference are published in archival form and will also be made available online.

SDM23 has three main themes in Methods and Algorithms, Applications and Human Factors and Social Issues, with each having a broad number of relevant topics. This year's conference will have invited presentations and multiple two-hour minitutorials, as well as a half day tutorial on Data-Efficient Graph Learning. SDM23 features special events such as the IBM

Early Career Data Mining Research Award and the SDM Doctoral Forum, which provides a unique opportunity for PhD students in data science (including data mining, machine learning, databases, and pattern recognition) to present their doctoral dissertation in poster format and get feedback from SDM participants and senior leaders in the field. The conference also hosts two full day workshops: Algorithmic Fairness in Artificial intelligence, Machine learning and Decision Making, and Data Science for Smart Manufacturing and Healthcare. Please check the website for further details and updates about these events.

The pre-registration deadline for SDM23 is March 30, 2023. The Travel Fund application deadline has been extended to February 17, 2023. Refer to the website for further information.

IJCAI 2023
The 32nd International Joint Conference on Artificial Intelligence
Cape Town, South Africa
August 19-25, 2023
http://www.ijcai-23.org/

Starting from 1969, IJCAI has remained the premier conference bringing together the international AI community to communicate the advances and achievements of artificial intelligence research. Submissions to IJCAI-2023 will report on significant, original, and previously unpublished results on any aspect of artificial intelligence. Papers on novel AI research problems, on AI techniques for novel application domains, and papers that cross discipline boundaries within AI are especially encouraged.

In addition to the main track, authors will be able to submit papers to the two multiyear special tracks (AI for Good and AI, The Arts and Creativity), as well as the survey track; these tracks will post their own calls for papers later this year, and their deadlines, procedures and policies may differ from what is described on the IJCAI-2023 website. The special track on AI and Social Good is dedicated to research triggered by real-world key questions, is carried out in collaboration with civil society stakeholders, and uses AI to work towards the SDGs and LNOB. The track aims to encourage the application of AI to solve current global and local challenges and to strengthen the civil society-science-policy interface.

The conference is also calling for volunteers for Senior Program Committee Members and Program Committee Members, with the application forms available through the website. This year, IJCAI-2023 will feature tutorials, competition, workshop, demo and doctoral consortium programs, that will be held across the duration of the event. IJCAI-2023 also offers several attractive sponsorship packages to provide opportunities for smaller and larger companies and organizations to participate in the world's premiere AI research event, with more than 3000 expected attendees worldwide. The IJCAI-2023 Sponsorship and Exhibit Brochure is available on the website.

Registration opens on April 19, 2023. Please refer to the website for further details about the conference venue, including important local information and accommodation for specific IJCAI Conference hotels that have negotiated rates.