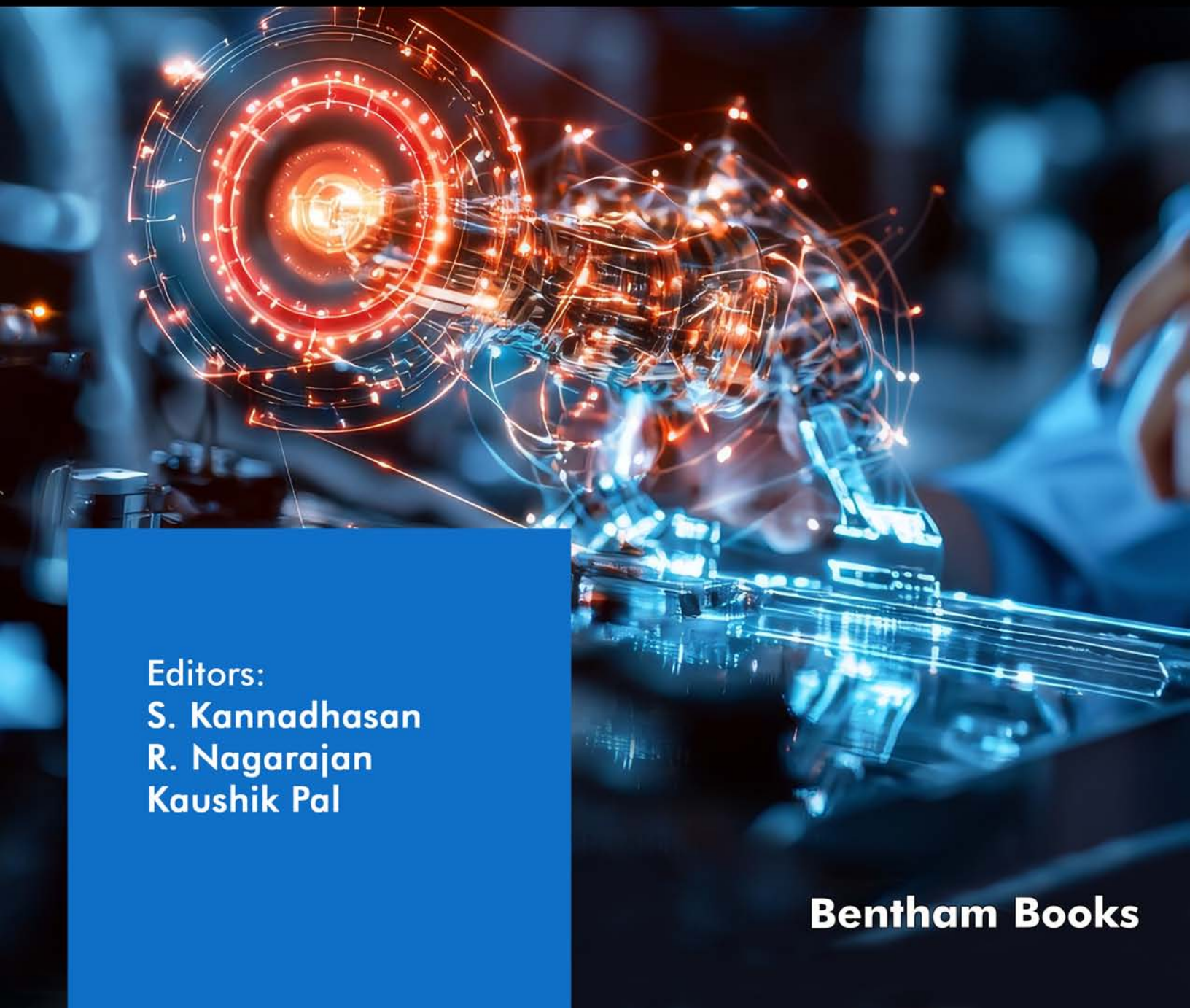
INTELLIGENT TECHNOLOGIES FOR SCIENTIFIC RESEARCH AND ENGINEERING



Editors:
S. Kannadhasan
R. Nagarajan
Kaushik Pal

Bentham Books

Advanced Technologies for Science and Engineering

(Volume 4)

Intelligent Technologies for Scientific Research and Engineering

Edited by

S. Kannadhasan
Department of Electronics and Communication Engineering
Study World College of Engineering
Coimbatore, Tamil Nadu,, India

R. Nagarajan
Department of Electrical and Electronics Engineering
Gnanamani College of Technology, Namakkal, Tamil Nadu, India

&

Kaushik Pal
Laboratório de Biopolímeros e Sensores
Instituto de Macromoléculas, Universidade Federal do Rio de Janeiro (LABIOS/IMA/UFRJ)
Rio de Janeiro
Brazil

Advanced Technologies for Science and Engineering

(Volume 4)

Intelligent Technologies for Scientific Research and Engineering

Editors: S. Kannadhasan, R. Nagarajan & Kaushik Pal

ISSN (Online): 3029–2859

ISSN (Print): 3029-2840

ISBN (Online): 979-8-89881-531-8

ISBN (Print): 979-8-89881-532-5

ISBN (Paperback): 979-8-89881-533-2

Published by Bentham Science Publishers Pte. Ltd. Singapore,

in collaboration with Eureka Conferences, USA.

First published in 2026.

need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Pte. Ltd.
No. 9 Raffles Place
Office No. 26-01
Singapore 048619
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

PREFACE

The book on Intelligent Technologies for Research and Engineering contains new research findings from academics. This edited anthology covers a range of research topics on science, engineering, and technology over forty six chapters. Discussion topics include artificial intelligence learning techniques, computerised medical image processing, human-computer interface for hand gesture recognition, community energy storage, e-learning, diabetes risk prediction, solar cells, hydrogen fuel cells for cars, and more. Many great engineering achievements have been made in the past century alone that we now generally take them for granted. For most of the globe, technology enables access to a plentiful supply of food and clean drinking water. For many of our everyday tasks, we depend on electricity. We can deliver products and services to any location in the world with great simplicity. Growing advances in communications and computer technology are creating new avenues for entertainment and information discovery. Even with these incredible technological accomplishments, there are undoubtedly still a ton of outstanding opportunities and challenges to be addressed. A lot of them are blurry, and many more are undoubtedly beyond most people's comprehension, even if some seem clear.

The latest advances in solidification research and the problems the community faces in the 21st century in processing and analysis are presented. On behalf of the editors, we would like to offer our appreciation to everyone who took part in this project. First and foremost, we offer all credit and respect to our almighty Lord for his bountiful grace, which enabled me to finish this book successfully. Moving forward, the authors, whose excellent work is at the core of the book, are acknowledged, and we gratefully congratulate all those involved and wish them great success. We would like to take this time to thank our family and friends for their support and encouragement while we worked on this book. We would also like to express our gratitude to the writers for their contributions to this edited book. In addition, our special thanks is extended to Bentham Science Publishers and its whole team for facilitating us in publishing this work and us providing the opportunity to present our work to the audience.

The content of this book is summarized as follows:

1. In Chapter 1, The practise of "lifelogging" involves documenting an increasing amount of one's everyday experience with the intention of using the recordings in the future as a memory aid or the foundation for data-driven self-development. Therefore, the usefulness of the generated lifelogs depends on the lifeloggers' ability to efficiently sift through them. The logs' intrinsic multi-modality and semi-structure allow them to combine data from a variety of sources, including cameras and other wearable physical and virtual sensors. As a result, expressing the data in a graph structure allows for the effective capturing of all created interrelations. Alternative methods must be developed to capture the higher-level semantics because it is impossible to manually or mechanically annotate each entry with a significant amount of semantic context. We describe Improved Life Graph (ILG), the first method for building a Knowledge Graph-based lifelog representation and retrieval solution, which can capture a lifelog in a graph structure and enhance it with external data to help with the connection of higher-level semantic information.

2. Chapter 2 presents that as a result of increased competition and a decline in new clients, the global telecommunications sector is suffering from a dramatic decline in revenue. Most operators first spend a significant portion of their income on expansion in order to maintain competitive advantages and attract a large user base. A company's ability to boost its selling,

marketing, and servicing operations across all client touchpoints is greatly aided by a well-developed client Relationship Management (CRM) strategy. Predicting customers' propensity to leave is a major challenge in CRM. The purpose is to identify consumers who could be at risk of leaving based on their historical data and actions. In the study under consideration, data mining methods were used for accurate churn forecasting. In this case, we preprocess the dataset using the normalised k-means approach. After the picture has been preprocessed, attributes are chosen using the minimal Redundancy & Maximum Relevance (mRMR) method. When making a decision, it favours qualities that have low connections among themselves and a strong relationship with the class (output). The ability to classify or forecast client turnover based on the provided attributes is explored using a Support Vector Machine with Photon Swarm Optimisation (SVM with PSO). To optimise the SVM's hyperparameters, PSO is used. Additionally, the problem of discovering a local optimal solution is avoided, and the accuracy of the classification is enhanced. The experimental results show that the proposed system is superior to the current one in terms of processing time.

3. Chapter 3 presents Multilayer perceptrons (MLPs) along with support vector machines (SVMs), which are examples of TMLTs that have been utilized effectively for churn prediction in the past, but only after considerable time and energy were spent configuring the training parameters. Choosing appropriate training settings for unsupervised learning is usually an ad hoc process that relies on experimentation. When it comes to churn forecasts, deep neural networks (DNNs) have demonstrated to be much more accurate than TMLTs. To set the instruction hyperparameters for DNNs throughout churn modelling, however, requires more time and effort because of DNNs' more complicated design and their ability to analyze vast volumes of non-linear input data. This creates extra difficulty for novice machine learning professionals and researchers. Few studies have been conducted to date to determine how various hyperparameters affect DNN performance when used for churn prediction. When it comes to churn modelling, DNNs aren't backed by much in the way of experimentally developed heuristics to help with hyperparameter selection. To better predict customer attrition in the banking sector, this work conducts an experimental analysis of the effect of adjusting DNN hyperparameters. The deep neural network (DNN) simulations beat the MLP across three separate trials, with the DNN models using a rectifier activation function in the hidden layers and the MLP using a sigmoid activation function in the output layer. Rems Prop training was more accurate than Adam, AdaGrad, Ad delta, and Adam ax, and it was also more effective than stochastic gradient descent (SGD). The DNN did best when the number of batches was smaller than the total number of data points in the test set. This study provides heuristic insights that may be useful to academics and practitioners when DNNs are used to predict churn from CRM table data in the financial services industry.

4. Chapter 4 suggests that event logs may be used for process mining. Event logs may include confidential data that cannot be analysed without agreement, preventing process mining. Anonymizing the event log prevents anybody from being identified by it. Differential privacy guarantees anonymity. Heterogeneous private event log anonymity aims to provide a log with high usefulness and a privacy guarantee. Event log anonymization methods introduce noise into traces by replicating, perturbing, or filtering them. Subsampling before noise injection improves the privacy-utility trade-off, according to research on differential privacy. Subsampling enhances privacy. Libra uses this observation to anonymize event logs. Libra takes numerous samples of trace from a log, separately enters noise, maintains statistically meaningful traces from every sample, then makes up the samples to create a uniquely private log. The suggested method yields far better utility for identical privacy assurances than baselines.

5. Chapter 5 tells us that banks and other financial organisations rely heavily on credit scoring

(CS) as a method of risk management since it is both effective and necessary. It reduces financial risks and gives sound advice on loan disbursement. As a result, businesses and financial institutions are exploring innovative automated solutions to the CS dilemma in an effort to safeguard their own resources and those of their clients. The use of various machine learning (ML) as well as data mining (DM) approaches has led to significant progress in CS prediction in recent years. The Deep Genetic Hierarchical Network of Learner (DGHNL) is a novel approach developed for this study. Support Vector Machines (SVMs), k-Nearest Neighbours (kNNs), Probabilistic Neural Networks (PNNs), and fuzzy structures are just some of the many types of learners that may be used in the suggested method. The Statlog German (1000 occurrences) approval of credit dataset from the UCI machine learning library was used to evaluate our model. We use a DGHNL model with five unique learner types, two feature extraction methods, three kernel functions, and three methods for optimising model parameters. In addition to conventional cross-validation (CV) and train-testing (stratified 10-fold) methods, this model employs a cutting-edge biological layered training (participant selection) approach. Because it makes use of coordinated and shared information (the DGHNL architecture and the optimisation of it), our approach is innovative. Using data on German credit approvals from Statlog, we show that the suggested DGHNL model can obtain a prediction accuracy of 94.60% (54 errors per 1000 classifications) with its 29-layer architecture.

6. Chapter 6 suggests that the primary use of forecasting short-term loads is in control centres, where it is used to investigate shifting consumer load patterns and estimate the value of the load at a future time. It is a crucial piece of equipment for building a smart grid. Several dimensions of influence impact the load parameters. In this research, we present a Residual Neural Network (ResNet)/Long Short-Term Memory (LSTM) hybrid model for load forecasting, which can better take advantage of the time series properties of load data and lead to more reliable predictions. Before feeding the data into the ResNeT network for feature extraction, it is first rebuilt using numerous feature parameters. The second step in short-term load forecasting using LSTM is to feed the feature that was extracted vector into the network. Finally, the technique is compared with other models using a real example, demonstrating that the proposed combination method has better prediction accuracy and confirming the practicality and superiority of input-parameter feature extraction. Additionally, this study examines weather prediction based on a variety of elements and characteristics.

7. Chapter 7 presents a deep learning method, named Convolutional Neural Network (CNN). Natural language processing problems, such as text classification, are simplified using this approach. In this study, we use a deep learning strategy, namely the CNN method, to address the issue of text classification. CNNs, which require a large amount of time as well as finances to train and use, have been greatly impeded by the rise of Big Data and the increased complexity of tasks. To get around these problems, we introduce a MapReduce-based CNN that rethinks what a CNN has learnt by breaking it down into a series of smaller networks and training them in parallel. Subsets of incoming text are analysed by many autonomous networks.

8. Chapter 8, it is well recognised that data preparation is necessary to provide reliable data with which mining tools may derive useful insights. Preprocessing methods need to be modified when dealing with multiple sources of continuous data. This study suggests using active rule-based systems, and more particularly, complex event processing (CEP) systems and engines, to aid domain experts in the definition and execution of pretreatment tasks for data streams. Our method's key innovation is the way it allows domain experts to easily manage temporal data by formulating preprocessing methods as identifying events rules stated in a SQL-like language. This concept is implemented in a freely accessible software

package that combines a CEP processor with libraries for web-based data mining. To test the efficacy of our method, we provide three real-world examples of applications in which CEP rules preprocess data streams by incorporating new temporal information, modifying features, and handling missing values. Preprocessing activities may be expressed in a flexible and high-level fashion using CEP rules, as shown by the experiments, without incurring excessive memory and time overheads. The generated streams of data not only enhance classification algorithms' predictive accuracy but also simplify decision models and shorten the time required to learn.

9. Chapter 9 offers a fresh theoretical perspective on the issue of interpretive topic modelling. Instead of using words or n-grams as the fundamental units of analysis, as in more traditional methods, this method employs whole sentences. Specifics of the proposed method include clustering of phrase embeddings and probabilistic sentence assessments within the text corpus. Sentence frequency distributions within subjects and topic frequency distributions throughout the text are both estimated using the topic model. Since sentences, unlike words, are more meaningful and include entire grammatical and semantic structures, our method allows for explicit understanding of themes. The process for doing this automatically is also given. Sentence embeddings are obtained via the use of context embeddings built on the BERT paradigm. Our method also demonstrates the feasibility of integrating both internal and external information sources in the subject modelling process, allowing for big data processing. In conventional topic modelling methods, the text corpus itself stands in for the internal knowledge source, and this source is often a single one. BERT, a machine learning model first trained on a massive amount of textual data, serves as the external knowledge source and produces context-dependent sentence embeddings.

10. Chapter 10 shows that by analyzing the geographical, chronological, and semantic components of geographic data, it is possible to reconstruct users' real route itineraries and get insights into their preferences and behavior. In order to analyse tourist traffic patterns in A-level scenic spots in Jiangsu and Zhejiang across time and space, this research collects and preprocesses data from Weibo check-ins at these locations. The author used a temporal perspective, looking at how the check-in data fluctuated between 2016 and 2018, as well as how it differed on weekends, holidays, and weekdays. The acquired data were subjected to a spatial kernel density analysis, which revealed the most active regions. Lastly, the vacation travel mode and characteristics were identified through an examination of spatial and locational flows and their orientations. The results of this study provide the groundwork for the growth of wisdom tourism.

11. Chapter 11 presents the amount of data available online is growing at an exponential rate. The Internet has a wealth of information that can be mined for specific details on any niche subject, depending on the user's needs. When faced with a bewildering array of options, users often find it difficult to decide which product to purchase online. In this case, it is crucial to review the available data to advise consumers on products and learn from other customers. The proposed approach allows us to efficiently filter, prioritize, and convey very important information, therefore mitigating the problem of information overload. To narrow down a list of options based on your individual tastes is the job of a recommendation system. The approach relies heavily on several different types of similarity measurements. It is generally agreed that collaborative filtering is the most effective method for making specific recommendations to users or providers. Since there are limitations to the user-based collaborative filtering strategy, the item-based strategy is considered an alternative. To address this shortcoming, we analyzed the effectiveness of several similarity calculation methods by comparing correlation-based and distance-based similarity measures, aiming to improve recommendation performance. The study's findings were utilized to design a better

technique, which uses statistical accuracy measures to provide the best informed suggestion possible. This study's findings provide a benchmark for evaluating and comparing similarity measurements. This work aims to help readers select suitable distance measures for datasets, and to make it easier to compare and evaluate new similarity measures against established ones.

12. Chapter 12 claims that recommender systems have quickly become an integral part of people's everyday digital lives as it is present on virtually every online service today. Modern deep learning-based models can only function at their peak when fed with a massive amount of data. Multiple domains, including Amazon, restaurants, and breweries, have offered datasets that meet this condition. The hotel industry has seen relatively few advancements and databases, with even the largest review dataset being in the hundreds of thousands rather than millions. Traditional collaborative-filtering methods are also inapplicable to the hotel domain due to its increased data sparsity compared to standard recommendation datasets. In this research, we present HotelRec, a TripAdvisor-derived, massively scaled hotel recommendation dataset comprising 50 million reviews. To the best of our knowledge, HotelRec is the largest recommendation dataset in a single domain, including textual reviews (50M vs 22M) in the hotel domain (50M vs 0.9M).

13. Chapter 13 suggests that, the field of Recommender Systems (RS) research has expanded to include a broad range of AI methods, from simpler ones like Matrix Factorization (MF) to more advanced ones like Deep Neural Networks (DNN) in recent years. Because they only consider a linear combination of user and item vectors, traditional Collaborative Filtering (CF) recommendation algorithms like MF have limited learning capacity. Neural collaborative filtering (NCF) uses deep neural networks (DNN) in combination with collaborative filtering (CF) to learn non-linear correlations. CF approaches still have issues with cold start and data sparsity, though. To improve recommendation accuracy, address cold starts, and address data sparsity, this research proposes a new hybrid RS, Neural Matrix Factorization++ (NeuMF++). We propose NeuMF++, which improves upon NeuMF by adding Stacked Denoising Autoencoders (SDAE) for a more accurate latent representation. GMF++ and MLP++ can be combined to form NeuMF++. By combining Generalized Matrix Factorization (GMF) with Multilayer Perceptrons (MLP), NeuMF provides a robust NCF architecture. By combining the linearity of GMF with the nonlinearity of MLP, NeuMF achieves state-of-the-art results. At the same time, GMF++ and MLP++ have been developed due to the successful integration of latent representations into the original GMF and MLP. NeuMF++'s learning capacity is greatly improved by the latent representation it obtains from the SDAEs' latent space, which enables it to learn user and item characteristics accurately. However, NeuMF++'s performance may suffer if GMF++ and MLP++ are forced to share feature extractions. Consequently, enabling GMF++ and MLP++ to learn features independently increases their adaptability and dramatically boosts their performance. The experimental root-mean-square error of 0.8681 obtained by NeuMF++ in a real-world dataset shows that it performs exceptionally well. Additional data, such as text or photos, can be added to NeuMF++ in later development. NeuMF++ allows for the incorporation of several neural network building elements to create a more robust recommendation model.

14. Chapter 14 shows that finding relevant information online has gotten increasingly difficult as the amount of data available on the internet has grown exponentially. In high-data-density, complex-domain settings, the recommendation system may be a big aid to users in making decisions. In the recommender system, several approaches have been presented. In the recommendation system, collaborative filtering is a common practice. The cold-start problem is one of the remaining issues with collaborative filtering approaches. To address this issue, we offer a movie recommendation system that uses social network analysis and collaborative

filtering. We used user preferences like age, gender, and profession to generate a connection matrix, and then used that matrix to use community identification based on edge betweenness centrality to cluster people. The suggested system will then propose movies to new members based on the preferences of the existing users in the group. Utilizing MAE, we demonstrate the superiority of the suggested technique.

15. Chapter 15 shows that the financial and emotional toll of cardiovascular disease is growing. As a result, we created a model for predicting comprehensive healthcare resource use (Adherence Score for Healthcare Resource Outcome, ASHRO), which includes patient health behaviors, and investigated its relationship with clinical outcomes, with the goal of improving the economy as a whole and the quality of the healthcare system. Data from a massive database of health insurance claims, long-term care insurance, and health checkups were used in this investigation. Patients admitted to hospitals with cardiovascular conditions (ICD-10 I00-I99) constituted the study population. The objective variable was medical and long-term care expenses, while the explanatory variable was a broadly defined composite adherence measure. Multiple regression analysis and random forest learning (AI) were utilized to calibrate predictive models, which were then used to generate ASHRO ratings. The prediction model's discriminatory and evaluative abilities were measured using the area under the curve and the Hosmer-Lemeshow test, respectively. Over a 48-month follow-up, we used propensity score matching to examine the overall mortality of the two ASHRO 50% cut-off groups after adjusting for clinical risk variables. Out of a total sample size of 48,456 patients, 61.9% were men, and the mean age at hospital release was 68.3 9.9 years for those with cardiovascular disease. Machine learning was used to adjust eight factors (secondary mitigation, rehabilitation intensity, direction, proportion of days addressed, overlapping outpatient visits/clinical laboratory and physiological tests, medical attendance, and generic drug rate) into a single index that served as the adherence score classification. The total coefficient of determination from the multiple regression analysis was 0.313 (p 0.001). The total coefficient of determination in a logistic regression analysis with 50% and 25%/75% cut-off values for medical and long-term care expenses was statistically significant (p 0.001). There was a statistically significant correlation between ASHRO score and death rate at the 50% cutoff (2% vs. 7%; p 0.001).

16. Chapter 16 shows the goal of this research is to examine how e-commerce and web-based businesses might benefit from the use of Big Data Analytics in managerial decision making. The data used in this analysis comes from a single e-commerce website's database. User interactions with the website, such as page views, product additions, and online purchases, would be recorded. Association Rule Mining Algorithm (APRIORI), K-Means Clustering, and Pearson's Correlation Coefficient are only a few of the algorithms that will be utilized to evaluate the dataset. The information will be analyzed and used to provide insights into users' interactions with the website, enabling the identification of patterns that may inform future actions. For instance, which product receives the most attention and sales, how many pages are viewed before a purchase is made, what percentage of customers buy the product again, and so on. This study would also determine whether the company's current Big Data implementation can be enhanced and whether doing so would be a smart use of resources.

17. In Chapter 17, with the help of a deep artificial neural network (ANN; i.e. deep learning), a new method is provided to simulate and reconstruct yearly surface mass balance (SMB) data across glaciers. An open-source regional glacier evolution model now includes this technique as its SMB component. While conventional glacier models increasingly include physical processes, we instead use data science to create a parameterized model. Deep learning or Lasso (least absolute shrinkage and selection operator; regularized multilinear regression) can be used to model annual glacier-wide SMBs from topo-climatic variables,

while a glacier-specific parameterization is used to update the glacier's shape. On a dataset of 32 French Alpine glaciers, we evaluate and cross-validate our nonlinear deep learning SMB model against other typical linear statistical techniques. Results show that deep learning is superior to linear approaches, with an estimated r 2 of 0.77 and a root-mean-square error (RMSE) of 0.51 m w.e., thanks to increased accuracy (up to +47% in space and +58% in time) and explained variance (up to +64% in space and +108% in time). The temporal dimension accounts for around 35% of the nonlinear behavior recorded by deep learning. The key unknowns in the evolution of glacier geometry stem from the initial ice thickness measurements. These findings support the application of deep learning in glacier modeling as a potent nonlinear tool for reconstructing or simulating SMB time series for individual glaciers across a region for past and future climates.

18. Chapter 18 shows the rate at which remote sensing technology is advancing, which means that our access to remote sensing data is better than before. The age of big data is here. Data collected using remote sensing exhibit hallmarks of Big Data, including hyperspectral features, high spatial resolution, and high temporal resolution. Using geographical feature and remote sensing data, this work provides a feature-supporting, marketable, and efficient data cube for time-series analysis and conducts a comparative assessment of water cover and vegetation change. This study defines remote sensing data cube (SRSDC) with a focus on spatial features. The purpose of this data cube is to offer a fast, flexible, and scalable way to analyze massive amounts of RS information using spatial features. It gives a general summary of the SRSDC's structure. The SRSDC provides feature translation to transform spatial feature information into query operations and spatial feature repositories to store and manage vector feature data. This article explains how a feature data cube and distributed execution engine were developed for use in the SRSDC. The evaluation of a feature data cube and a distributed execution engine is carried out using the production process and long-term remote sensing analysis as examples. As a new strategic resource for humanity, big data has risen to the top of the knowledge economy's mountain range. Data analysis methods, including supervised, unsupervised, and hybrid approaches, are the backbone of knowledge discovery techniques.

19. Chapter 19 provides suggestions for integrated, cutting-edge, and efficient tools, methodologies, and technologies for accessing and processing increasingly growing volumes of data in diverse fields. Personalizing a patient's care is a challenging task that requires the doctor to sift through and make sense of massive volumes of data. The scientific community behind precision medicine might benefit greatly from a unified system that facilitates data discovery, integration, preprocessing, model construction, storage, analysis, and visualization. The software package provides researchers with a simple, quick, and adaptable method for processing data, with the ultimate goal of enabling intelligent management, analysis, and visualization of massive genomic data. Services, data sets, and databases are at their disposal, or they can supply their own information for processing.

20. Chapter 20 shows that as a result of increased competition and a decline in new clients, the global telecommunications sector is suffering from a dramatic decline in revenue. Most operators first spend a significant portion of their income on expansion in order to maintain competitive advantages and attract a large user base. A company's ability to enhance its selling, marketing, and servicing operations across all client touchpoints is greatly aided by a well-developed client Relationship Management (CRM) strategy. Predicting customers' propensity to leave is a major challenge in CRM. The purpose is to identify consumers who could be at risk of leaving based on their historical data and actions. In the study under consideration, data mining methods were used for accurate churn forecasting. In this case, we preprocess the dataset using the normalized k-means approach. After the picture has been preprocessed, attributes are chosen using the minimal Redundancy and Maximum Relevance

(mRMR) method. When making a decision, it favours qualities that have low connections among themselves and a strong relationship with the class (output). The ability to classify or forecast client turnover according to the provided attributes is explored by using a Support Vector Machine and Photon Swarm Optimization (SVM with PSO). To optimise the SVM's hyperparameters, PSO is used. Additionally, the problem of discovering a local optimal solution is avoided, and the accuracy of the classification is enhanced. The experimental results show that the proposed system is superior to the current one, and the processing time.

21. Chapter 21, Multilayer perceptrons (MLPs), along with support vector machines (SVMs), are examples of TMLTs that have been utilized effectively for churn prediction in the past, but only after considerable time and energy were spent configuring the training parameters. Choosing appropriate training settings for unsupervised learning is usually an ad hoc process that relies on experimentation. When it comes to churn forecasts, deep neural networks (DNNs) have demonstrated to be much more accurate than TMLTs. Setting the instruction hyperparameters for DNNs throughout churn modelling, however, requires more time and effort because of DNNs' more complicated design and their ability to analyse vast volumes of non-linear input data. This creates extra difficulty for novice machine learning professionals and researchers. Few studies have been conducted so far to determine how various hyperparameters affect the DNN performance when used for churn prediction. When it comes to churn modelling, DNNs are not backed by much in the way of experimentally developed heuristics to help with hyperparameter selection. To better predict customer attrition in the banking sector, this work conducts an experimental analysis of the effect of adjusting DNN hyperparameters. The deep neural network (DNN) simulations beat the MLP across three separate trials, with the DNN models utilizing a rectifier functional for activation in the hidden layers and the MLP using a sigmoid in the output layer. Rems Prop training was more accurate than Adam, AdaGrad, Ad delta, and Adam ax, and it was also more effective than stochastic gradient descent (SGD). The DNN did the best when the number of batches was smaller than the total number of data points in the test set. This study provides heuristic information that may be useful to academics and practitioners when DNNs are used for predicting churn from table data for CRM in the financial services industry.

22. In chapter 22, this paper offers a fresh theoretical perspective on the issue of interpretive topic modelling. Instead of using words or n-grams as the fundamental units of analysis, as in more traditional methods, this method employs whole sentences. Specifics of the proposed method include clustering of phrase embeddings and probabilistic sentence assessments within the text corpus. Sentence frequency distributions within subjects and topic frequency distributions throughout the text are both estimated using the topic model. Since sentences, unlike words, are more meaningful and include entire grammatical and semantic structures, our method allows for explicit understanding of themes. The process for doing this automatically is also given. Sentence embeddings are obtained via the use of context embeddings built on the BERT paradigm. Our method also demonstrates the feasibility of integrating both internal and external information sources into the subject modelling process, enabling big data processing. In conventional topic modelling methods, the text corpus itself stands in for the internal knowledge source, and this source is often a single one. The BERT, a machine learning model that was first trained on a massive quantity of textual data, stands in for the external knowledge source and is responsible for producing the context-dependent sentence embeddings.

23. Chapter 23 presents that the day-to-day activities of millions of people all over the globe have been significantly altered as a direct result of the meteoric rise in the popularity of social media and online social networks such as Facebook, Twitter, Instagram, and TikTok. The ease with which data may be gathered, collected, and analyzed, in addition to the high degree

of social and financial interest in doing so, has sparked the interest of a broad range of research sectors. This has resulted in increased emphasis being paid to the research being conducted in these areas. Each agent is given a decentralized control mechanism that enables them to communicate, draw conclusions from their discussions, and learn from one another. Utilizing a network topology, this method makes it easier for dynamic agent organizations to alter the geometry of agent interactions in order to meet the particulars of the situation at hand.

24. Chapter 24 shows that the number of individuals using social networking and microblogging sites has increased dramatically in recent years, providing an interesting window into the perceptions of this age. People's opinions may be gauged in large part by looking at user reviews of various products, companies, brands, individuals, forums, films, etc. Analysts saw a need to automate the categorisation of evaluations into positive and negative categories; therefore, they developed algorithms to do so. Sentiment analysis is the name given to the automatic categorization process it enables. The fundamental objective of this study is to apply the Support Vector Machine (SVM) artificial intelligence algorithm to the task of classifying feelings and texts for product evaluations. This paper will do so by doing an in-depth analysis of several datasets utilised for this purpose. The Support Vector Machine (SM) learning technique has been trained, tested, and implemented across a variety of data sets to determine the polarity of ambiguous feelings. The major goal of this work is to apply the Support Vector Machine (SVM) artificial intelligence technique to analyse several datasets for sentiment and text classification, with the end result being improved categorization of product reviews. A support vector machine training system is trained, tested, and simulated on many datasets in this work to determine the polarity of ambiguous feelings or reviews. We found that among the available classification algorithms, Support Vector Machine (SVM) produces the highest accuracy (89.98%) right off the bat. Including additional sentence forms would further improve the achieved accuracy. As a final result, it establishes that the SVM method is reliable. Application/Improvements: Models generated by the use of the Support Vector Machine algorithm for learning are evaluated for their performance. Finally, a highly accurate and powerful classification method, the Support Vector Machine (SVM), has been developed.

25. In Chapter 25, it is well recognised that data preparation is necessary to provide reliable data with which mining tools may derive useful insights. Preprocessing methods need to be modified when dealing with multiple sources of continuous data. This study suggests using active rule-based systems, particularly complex event processing (CEP) systems and engines, to aid domain experts in defining and executing pretreatment tasks for data streams. Our method's key innovation is the way it allows domain experts to easily manage temporal data by formulating preprocessing methods as identifying events rules stated in a SQL-like language. This concept is implemented in a freely accessible software package that combines a CEP processor with libraries for web-based data mining. To test the efficacy of our method, we provide three real-world examples of applications in which CEP rules preprocess data streams by incorporating new temporal information, modifying features, and handling missing values. Preprocessing activities can be expressed in a flexible, high-level manner using CEP rules, as shown by the experiments, without incurring excessive memory or time overhead. The generated streams of data not only enhance classification algorithms' predictive accuracy but also simplify decision models and shorten the time required to learn.

26. In chapter 26, call records demonstrate client interest in various businesses. Multi-dimensional attribute dependence with communication day and time may help targeted advertising. Frequent and extensive inter-service links show that consumers of one service may have opportunities in the other. Multi-granulation rough sets address prospect discovery

from call record interest characteristics. Conventional intra and inter-pattern mining methods have increased the amount of processing and the vast space of statistically irrelevant patterns. This solution fixes these difficulties. The method produces food and restaurant target audiences using one month of anonymised Thai telecom service supplier call data and confirms some fascinating mathematical properties of knowledge systems.

27. In chapter 27, the rate at which remote sensing technology is advancing means that our access to remote sensing data is better than before. The age of big data is here. Data collected using remote sensing exhibits hallmarks of Big Data, including hyper spectral features, a high spatial resolution, and a high time resolution. Using geographical feature data and remote sensing data, this work offers a feature-supporting, marketable, and efficient data cube for time-series analytic application, and conducts a comparative assessment of water cover and vegetation change.This study defines the remote sensing data cube (SRSDC) with a focus on spatial features. The purpose of this data cube is to offer a fast, flexible, and scalable way to analyze massive amounts of RS information using spatial features. It gives a general summary of the SRSDC's structure.The SRSDC provides feature translation to transform spatial feature information into query operations and spatial feature repositories to store and manage vector feature data.This article explains how a feature data cube and distributed execution engine were developed for use in the SRSDC. Evaluation of a feature data cube and distributed execution engine is carried out using the production process and analysis of long-term remote sensing as examples. As a new strategic resource for humanity, big data has risen to the top of the knowledge economy's mountain range. Data analysis supervised learning methods, unsupervised learning methods, and their mixtures and modifications are the backbone of knowledge discovery techniques.

28. In chapter 28, with the help of a deep artificial neural network (ANN; i.e. deep learning), we provide a new method to simulate and reconstruct yearly surface mass balance (SMB) data across glaciers. An open-source regional glacier evolution model now includes this technique as its SMB component. While conventional glacier models increasingly include physical processes, we instead use data science to create a parameterized model. Deep learning or Lasso (least absolute shrinkage and selection operator; regularized multilinear regression) can be used to model annual glacier-wide SMBs from topoclimatic variables, while a glacier-specific parameterization updates the glacier's shape. On a dataset of 32 French Alpine glaciers, we evaluate and cross-validate our nonlinear deep learning SMB model against other typical linear statistical techniques. Results show that deep learning is superior to linear approaches, with an estimated r 2 of 0.77 and a root-mean-square error (RMSE) of 0.51 m w.e., thanks to increased accuracy (up to +47% in space and +58% in time) and explained variance (up to +64% in space and +108% in time). The temporal dimension accounts for around 35% of the nonlinear behavior recorded by deep learning. The key unknowns in the evolution of glacier geometry stem from the initial ice thickness measurements. These findings support the application of deep learning in glacier modeling as a potent nonlinear tool for reconstructing or simulating SMB time series for individual glaciers across a region for past and future climates.

29. Chapter 29 discusses that providing suggestions for integrated, cutting-edge, and efficient tools, methodologies, and technologies for accessing and processing increasingly growing volumes of data in diverse forms is a major problem in clinical data analysis and knowledge discovery. Personalizing a patient's care is a challenging task that requires the doctor to sift through and make sense of massive volumes of data. The scientific community behind precision medicine might benefit greatly from a unified system that facilitates data discovery, integration, preprocessing, model construction, storage, analysis, and visualization. The software package provides researchers with a simple, quick, and adaptable method for

processing data, with the ultimate goal of enabling intelligent management, analysis, and visualization of massive genomic data. Services, data sets, and databases are at their disposal, or they can supply their own information for processing.

30. Chapter 30 shows that tourism destinations and their online and social media information have made selecting and visiting them difficult. Tourists find tourism suggestion systems attractive, but designers must be able to deliver personalised services. This study proposes a personalised tourist system of recommendations that extracts user preferences. For this, tourist social network user reviews are a rich resource of preferences. To identify visitor preferences, remarks are preprocessed, semantically grouped, and sentimentally analysed. The characteristics of attractions are extracted from all user evaluations. Finally, the proposed suggestion system semantically matches user preferences with attraction attributes to suggest the most relevant attractions. The technology also filters undesirable goods and improves recommendations based on time, location, and weather. The Python-based recommendation algorithm is tested using TripAdvisor data. The suggested system improves the f-measure.

31. In chapter 31, The primary use of forecasting short-term loads is in control centres, where it is used to investigate shifting consumer load patterns and estimate the value of the load at a future time. It's a crucial piece of equipment for building a smart grid. Several dimensions of influence impact the load parameters. In this research, we present a Residual Neural Network (ResNet)/Long Short-Term Memory (LSTM) hybrid model for load forecasting, which can better take advantage of the time series properties of load data and lead to more reliable predictions. Before feeding the data into the ResNeT network for the extraction of features, it is first rebuilt using numerous feature parameters. The second step in short-term load forecasting using LSTM is to feed the feature that was extracted vector into the network. Finally, the technique is compared to other models using a real example, demonstrating that the proposed combination method has better prediction accuracy and confirming the practicality & superiority of input parameters feature extraction. Additionally, this study conducts studies in weather prediction based on a variety of elements and characteristics.

32. Chapter 32 suggest that data mining has become an increasingly significant method for conducting data analysis as a result of the rapid increase of databases used by a large number of contemporary businesses. The community of people who study operations research has made major contributions to this discipline, particularly by formulating and solving a large number of data mining problems as optimization problems. Additionally, data mining techniques may be used to solve a number of applications that study operations research. The purpose of this study is to offer an overview of the relationship between operations research and data mining. The basic objectives of the study are to highlight the spectrum of interactions between the two areas, present specific instances of significant research effort, and provide extensive references to additional significant work in the area. The purpose of this study is to examine not only the many optimization techniques that may be used for data mining, but also the process of data mining itself, as well as the ways in which operations research techniques can be utilized at almost every stage of this process. The report also identifies many potentially fruitful avenues for further investigation throughout its body. In the last part of the study, many applications connected to the administration of electronic services, including customer relationship management and customization, are investigated.

33. Chapter 33 shows that the field of Recommender Systems (RS) research has expanded to include a broad range of AI methods, from simpler ones like Matrix Factorization (MF) to more advanced ones like Deep Neural Networks (DNN) in recent years. Because it only takes into account a linear combination of user and item vectors, traditional Collaborative Filtering (CF) recommendation algorithms like MF have limited learning potential. Neural

collaborative filtering (NCF) uses deep neural networks (DNN) in combination with collaborative filtering (CF) to learn non-linear correlations. CF approaches still have issues with cold start and data sparsity, though. In order to increase recommendation accuracy, deal with cold starts, and fill in gaps in data sparsity, this research offers a new hybrid-based RS called Neural Matrix Factorization++ (NeuMF++). We propose NeuMF++, which improves upon NeuMF by adding Stacked Denoising Autoencoders (SDAE) for more accurate latent representation. and MLP++ can be combined to form NeuMF++. By combining Generalized Matrix Factorization (GMF) with Multilayer Perceptrons (MLP), NeuMF provides a robust NCF architecture. By combining the linearity of GMF with the non-linearity of MLP, NeuMF is able to produce state-of-the-art results. At the same time, GMF++ and MLP++ have been developed due to the successful integration of latent representations into the original GMF and MLP. NeuMF++'s learning capacity is greatly improved by the latent representation it obtains from the SDAEs' latent space, which enables it to accurately learn user and item characteristics. However, NeuMF++'s performance may suffer if GMF++ and MLP++ are forced to share feature extractions. Consequently, enabling GMF++ and MLP++ to independently learn features increases its adaptability and dramatically boosts its performance. The experimental root-mean-square error of 0.8681 obtained by NeuMF++ in a real-world dataset shows that it performs exceptionally well. Additional data, such as text or photos, can be added to NeuMF++ in later development. NeuMF++ allows for the incorporation of several neural network building elements to create a more robust recommendation model.

34. In chapter 34, The goal of this research is to examine how e-commerce and web-based businesses might benefit from the use of Big Data Analytics in managerial decision making. The data used in this analysis comes from a single e-commerce website's database. User interactions with the website, such as page views, product additions, and online purchases, would be recorded. Association Rule Mining Algorithm (APRIORI), K-Means Clustering, and Pearson's Correlation Coefficient are only few of the algorithms that will be utilized to evaluate the dataset. The information will be analyzed and utilized to provide insights into users' interactions with the website, allowing for the identification of patterns that might inform future actions. For instance, which product receives the most attention and sales, how many pages are viewed before a purchase is made, what percentage of customers buy the product again, and so on. This study would also determine whether or not the company's present Big Data implementation can be enhanced, as well as whether or not doing so would be a smart use of resources.

35. Chapter 35 presents that different e-learning recommendation strategies that benefit both students and teachers have been developed in recent years. In these cases, it is necessary to provide students and teachers with individualized instruction through the use of online learning systems tailored to their specific needs. In this study, we employ a split-and-conquer strategy-based clustering technique to design a smart recommender that can automatically adjust to the needs, preferences, and skill levels of each individual learner. The recommender is self-learning and does automated analyses of learner preferences and features. Using a divide-and-conquer approach, several learning modalities are grouped together for analysis. To extract the learners' functional patterns, the suggested cluster-based linear pattern mining approach is used. The machine then makes insightful suggestions by considering the ratings of common patterns. The proposed model offered critical learning tasks to learners based on their learning style, interest categorization, and talent traits, and it was tested on a variety of learner groups and datasets. The suggested cluster-based recommender was found to increase recommendation performance in experiments by leading to more lessons being finished by learners compared to those in the no-recommender cluster group. It was determined that over 65% of the students used all evaluation criteria while assessing the suggested

recommendation tool. The simulation results showed that the suggested recommender achieved higher metric values for larger learners. There were statistically significant variations in the assessed measures when the number of students was more than 1000. Using a computational framework that varied with |L| (the size of the suggestion list) and the characteristics of the students, we were able to identify the reasons for the observed discrepancies. The students had similar positive reactions to the recommender's precision and quickness. Standard deviation and mean of parameters for Recall (List, User) and Ranking Score (User) measurements differ significantly from other approaches for the sample dataset studied. The developed strategy outperformed competing strategies on all relevant criteria. In contrast to many well-known approaches, our recommender achieves the lowest mean absolute error across all clusters.

36. Chapter 36 shows that by analysing the geographical, chronological, and semantic components of geographic data, it is possible to reconstruct users' real route itineraries and get insights into their preferences and behavior. To analyse tourist traffic patterns at A-level scenic spots in Jiangsu and Zhejiang across time and space, this research collects and preprocesses Weibo check-in data for these locations. The author used a temporal perspective, examining how check-in data fluctuated between 2016 and 2018 and how it differed on weekends, holidays, and weekdays. The acquired data were subjected to a spatial kernel density analysis, which revealed the most active regions. Lastly, the vacation travel mode and characteristics were uncovered by an examination of spatial and locational flows and flow orientations. The results of this study provide the groundwork for the growth of wisdom tourism.

37. Chapter 37 shows that a user's interests and goals can be deduced with the use of a Recommendation System (RS), a new form of technology that employs knowledge discovery methods. User intent analysis and corresponding suggestion triggering become increasingly challenging in the face of exponential data growth. This work proposes a unique Deep Knowledge Graph (DKG) to do the necessary data analysis and construct the RS. DKG employs a Deep Convolutional Neural Network (DCNN). Our proposed DKG explicitly models the KG's end-to-end high-order connectivities. It recursively propagates the embeddings from a node's neighbors, which can be people, things, or traits, using an attention approach to assess the relative significance of the neighbours. In terms of theory, our DKG outperforms existing KG-based recommendation methods since it does not rely on regularization or an explicit representation of high-order relations. Empirical results on public benchmarks show that KGAT outperforms state-of-the-art methods like RippleNet and Neural FM. The benefits of the attention mechanism for interpretability, as well as the efficacy of embedding propagation for high-order relation modeling, have been shown in subsequent studies.

S. Kannadhasan
Head of the Departmnet
Department of Electronics and Communication Engineering
Study World College of Engineering
Coimbatore, Tamilnadu- 641105

R. Nagarajan
Department of Electrical and Electronics Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

&

Kaushik Pal
Laboratório de Biopolímeros e Sensores
Instituto de Macromoléculas
Universidade Federal do Rio de Janeiro (LABIOS/IMA/UFRJ)
Rio de Janeiro, RJ- 21941-901
Brazil

List of Contributors

Ashish Kumar Srivastava	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Avadhesh Kumar	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Alisha Sikri	Department of AIML, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India
A. Alli	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
Ajay Chakravarty	College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradaubad, Uttar Pradesh, India
Avadhesh Kumar	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Ajay Rastogi	Department of College of Computing Sciences & I.T, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Amit Singh	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Abhilash Kumar Saxena	College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Anurag Singh	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Ajay Shanker Singh	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Biswajeet Kumar Pandey	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Bhawna Wadhwa	Department of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India
C. Kalaiarasan	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
Chandrasekar	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
C.R. Manjunath	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Dhruv Galgotia	Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India
F. Rosita Kamala	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
Gopal K. Shyam	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
G. Geetha	School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India

G. Sindhu Madhuri	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Gaurav Kumar Rajput	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Gaurav Londhe	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
H.K. Shashikala	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
H.S. Shreenidhi	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Jagdish Chandra Patni	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Jayanthi Kannan	Department of Information Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
J. Vijay Fidelis	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
Kuldeep Singh Kaswan	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
K. Vanitha	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Krishnan Batri	School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
M. Veera Nagaiah	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
M.S. Sowmya	Department of Information Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Merin Thomas	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Mohammed Zabeeulla	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Meenakshi Sharma	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
M. Chandra Sekhar	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
Meenakshi Sharma	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Nitin Gaur	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
Priyanka Chandani	Department of DS, CSBS, AL-CSBS, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India
Prabha Nair	Department of IT and M.Tech Integrated, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India

Philomine Roseline	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
R. Pallavi	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
Rajat Bhardwaj	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Ranjana Sharma	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Ramesh Chandra Tripathi	College of Computing Science & IT, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Rajesh Pandian	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
R. Mahalakshmi	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
R. Pachayappan	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
Ramesh Sengodan	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
S.H. Shruthishree	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Sahana Shetty	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
S. Santosh	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
S. Kannadhasan	Department of Electronics and Communication Engineering, Study World College of Engineering, Coimbatore, Tamil Nadu, India
Sunanda Das	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Shambhu Bhardwaj	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Saira Banu Atham	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
S. Gadug	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
Sheetal	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
S. Senthilkumar	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
Somashekhara Reddy	Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India
T. Harish Naik	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India

Tushar Mehrotra	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Vineet Saxena	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Vetrimani Elangovan	Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
Veena S. Badiger	Department of Computer Applications, Presidency College, Bangalore, Karnataka, India
V. Gokul Rajan	Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

CHAPTER 1

A Study of Big Data Techniques for Extracting Valuable Information

Ashish Kumar Srivastava[1,*] and **Rajat Bhardwaj**[2]

[1] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: When dealing with unstructured data, information extraction is crucial for identifying named entities and events, which is essential for natural language processing. The exponential growth of information in the agricultural sector has made data extraction increasingly challenging. Despite the widespread use of deep learning-based approaches in smart farming for tasks such as crop cultivation, identifying diseases, weed removal, and yield production, the persistent impacts of meteorological, soil, pest, and fertilizer data make it challenging to discover the semantic relationships among the extracted data. The paper is divided into two sections. First, we present a data preprocessing method for cleaning up input corpora of any ambiguity; Furthermore, we provide a novel method for agricultural named entity identification, events, and relations that is based on deep learning techniques. This method utilizes multilayer perceptrons and the Adam optimizer with corrective capabilities. Agricultural, meteorological, soil, along with insect and fertilizer corpora, were used to train and assess the proposed algorithm. On common measures such as accuracy, sensitivity, and specificity, the experimental findings show that the proposed algorithm beats current approaches like Weighted-SOM, LSTM+RAO, PLR-DBN, the KNN, as well as Naïve Bayes.

Keywords: Adam optimizer, Agricultural sector, Algorithm, Big data techniques, Crop cultivation, Farming, Identifying diseases, Information extraction, Natural language processing, Weed removal, Yield production.

INTRODUCTION

A large portion of India's GDP comes from agriculture, which is particularly vulnerable to the effects of climate change. For example, the Indian agriculture sector is particularly susceptible to risks due to major characteristics such as tiny

[*] **Corresponding author Ashish Kumar Srivastava:** Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: dsrivastava@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

land holdings, over-reliance on fertilizers, and the monsoons [1 - 3]. The absence of processing algorithms results in a significant quantity of unorganized agricultural data being underused. The major variables for making choices in emerging nations like India continue to be human specialists and government regulations.

From a policymaker's point of view, factual confirmation according to current information is still missing [4]. Over the last several decades, climatic variability has had a significant impact on agricultural water resources, crop development and growth, as well as crop output over wide areas [5 - 8]. Crop production models for both rice and wheat in the Indian peninsula are based on simulation approaches, and researchers analyze the climate-crop connection using long-term fertility, regional data, and other predictable field studies [9]. The majority of Uttarakhand's land is arable, but farmers there see farming as an unrealistic way to ensure their families' nutritional needs are met. Maize and rice, sometimes called Kharif or monsoon crops, are Uttarakhand's main cash crops. Due to natural factors such as the continual danger of avalanches, substantial amounts of erosion, and collapses during rains, the Kharif crop yield in the Uttarakhand area is much lower than in other areas. Agricultural land that relies on rainfall for crop production is essential. The use of rain-fed techniques accounts for about 80% of Uttarakhand's agricultural output [10]. Diverse agroecosystems, each with its own unique hydro-geological setting, crop types, and harvesting methods, define a high-resilience system. Factors such as direction and amount of slope, height, soil type, irrigation conditions, and local knowledge might affect the variety [11]. However, traditional agricultural rotations as well as techniques can help keep this diversity intact. If you want your crop to thrive, it's not only the environment that matters; soil and fertilizers play a just as significant part. Nevertheless, present machine learning approaches like ANN, Bayes networks, and Gaussian kernel-based Support Vector Machine (SVM) can't determine which pests and fertilizers are best suited to a given soil type [12, 13]. Factors such as electrical conductivity, pH, and the macro- and micronutrients of the chosen crop determine the state of the soil [14]. In order to increase the crop's production, farmers use these soil quality indexes to determine which pesticides and fertilizers will work best. Data sources might include domain-dependent information, domain-independent or independent unstructured/semi-structured corpora, as well as extraction methods defined by the user [15]. Structured databases, such as relational and graph databases, store the knowledge extracted from the input data by the data extract (IE) engine. In order to estimate rainfall for agricultural production in the Dehradun area, researchers have suggested relatively restricted empirical as well as computational methodologies, combined with the relevant land information. To address this need, this study proposes to use unstructured agricultural literature

from the Uttarakhand area to extract Named Entities (NER) and event relationships (or event meaning).

RELATED WORKS

The last two decades have seen tremendous advancements in the use of machine learning and deep learning to address data extraction problems in many different industries, including agriculture, biometric identification, medical imaging analysis and retrieval, illness diagnosis, and many more. This literature review outlines previous efforts to apply machine learning to the agriculture industry. The use of ANN in the India-specific GCM. Using precipitation yields from GCM, the suggested method aimed to forecast the Indian Summer Maximum Rain values. In order to get scale predictions for India as well as its sub-divisional areas on a monthly basis, the ANN technique was linked to several ensemble organizations from the GCMs. In order to reduce the occurrence of over-fitting, the present research trained the ANN method utilizing the double-fold approval technique and straightforward randomization. The ANN-derived rainfall prediction is based on GCMs and is established by examining the percentage of linear errors in the probability time, contrast, box plots, and absolute error. After implementing the ANN systems used by these GCM experts in forecasting, the results of the experiment indicated the crucial alterations. Although logical considerations may impact the framework/variable, the data sets are dependent on previous estimates of the main factor. To estimate crop yields, Satir *et al*. [30] suggested using vegetation indices in conjunction with Stepwise Linear Regression. Accurate mapping of an area's crop patterns was achieved by combining object-based categorization with a multi-temporal Landsat data collection. Here, we approximated the forecast using real-time measuring techniques such as Mean Percent Error (MPE). Cotton, maize, and wheat had their MPE evaluated using a combination of soil salt levels and other factors. A single variable was used to lower the accuracy of weather forecasts based on the data collected. Using long-term weather data, hybrid algorithms like LASSO, ANN, and penalized regression techniques with elastic net (ENET), PCA, and SMLR to forecast rice yield. Based on the experimental data, LASSO-ENET performed well since these approaches simplified the model as well as avoided overfitting with amplitude coefficients. The hybrid models performed well in the West Coast of India crop forecast task, according to the pairwise numerous comparisons test. However, poor performance is produced by combining feature selection techniques with feature extraction using neural networks, such as PCA-SMLR. This is due to the fact that PCA failed to include the variable that is dependent while modifying the input variables. According to the study, a Hybrid Wavelet-based artificial neural network (HWNN) that included ANN, particle-swarm optimization (PSO), Social Knowledge, as well as Multi-Resolution Analysis was

CHAPTER 2

Adaptive Recommendation Systems for Improved Information Analysis

M.S. Sowmya[1,*] and **Nitin Gaur**[2]

[1] *Department of Information Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: There has to be a better design of screening devices to meet the rising need for health screening. We develop health screening machine learning models and suggest screening elements to provide individualized health care. For each screening item, a machine learning algorithm is created while the system is offline using guidelines and clinical findings from institutions to construct a synthetic data set. A customer's health states, as well as machine learning algorithms, are used by the proposed server online to provide a real-time inventory of screening items. The performance investigation showed that for 1,000 users online at once, the server responded in under a second, and the learning model's accuracy was close to 100%. This work introduces a recommendation system that can automatically adjust to changes in the screening environment.

Keywords: Machine Learning, Text Mining, Adaptive Recommendation Systems, Improved Information Analysis, Health Screening.

INTRODUCTION

Finding a decent hotel and making a reservation are travelers' first priorities. It now takes more time than before to look for a hotel online due to the sheer amount of data. The Recommendation Systems (RS) sector is booming due to the usefulness of these systems in facilitating decision-making and providing accurate information about the sought-after product or service. Finding the right hotel might be a challenge when using text reviews, rankings, votes, evaluations, and video views [1]. Many researchers are interested in Heterogeneous neural networks with graphs (GNNs) because of their strong performance as a neural

[*] **Corresponding author M.S. Sowmya:** Department of Information Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: ms.sowmya@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

network-based graph representation method. It handles complex networks with numerous links and nodes well, but massive amounts of semantic data and heterogeneity pose major challenges. The attention mechanism is a fascinating new concept in deep learning that has tremendous promise in many domains [2].

The importance of choosing a good hotel location and making a reservation for lodging has grown in recent years. The availability of a vast quantity of information online has led to a dramatic rise in the frequency and complexity of online hotel searches. The rise of Recommender Systems (RSs) may be attributed to their practicality in decision-making and their ability to provide accurate details on the desired good or service. It is becoming increasingly difficult to get hotel recommendations due to the proliferation of written reviews, rankings, votes, scores, and video views [3]. More and more individuals are logging into social media; therefore, it's crucial that they get exposed to the people they know with the things they're interested in. One function of recommendations is to increase users' social networks, which in turn encourages them to fill out their profiles. Using social interaction graphs, we can see how Twitter users engage with one another and who they follow [4]. One out of ten patients in developed nations experiences some kind of damage while hospitalized. By reviewing previous errors and identifying potential threats, Critical Incident Reporting Systems (CIRS) aim to lessen this. But incompatibility limits the effectiveness of Germany's 16 CIRS [5]. In order to explore the semantic aspects of users' preferences, review-based recommender systems combine user-generated reviews with rating-based models. A number of recent studies have shown that review information may enhance recommendation ability. Nevertheless, the majority of previous research has focused on improving the review-based representation learning component, which overlooks the underlying connection between ratings and reviews and fails to capture the fine-grained semantic elements [6]. A major obstacle to achieving the promised advantages of eHealth is the absence of integration and interoperability across diverse health systems. The easiest way to transition from isolated apps to eHealth solutions that work together seamlessly is to set up policies and standards for Health Information Exchange (HIE). On the other hand, data on the present state of HIE policies and standards throughout Africa is lacking [7].

RELATED WORKS

By engaging with vast quantities of diverse data and the desires of potential customers, researchers presented an intelligent method for providing real suggestions in [8]. Collaborative Filtering (CF) is a common recommendation system method. We presented a new method for CF recommendation that uses polarity detection and user sentiment analysis to build a feature matrix for hotels.

Hotel guests' peculiar habits may be better understood with the use of a multidisciplinary approach that incorporates lexical, syntactic, and semantic research. When we were making our recommendations, we looked at the hotels' customer service records and the quality of the firms they were affiliated with. The user's travel preferences will inform the system's hotel choices. More than one study has made use of data collected from hotel websites. The calculations were finished after this. This, therefore, ought to serve as the most important benefit of the description.

By using GNNs, the suggested architecture accomplishes multi-level semantic attention [9]. We used a combined dataset that included movies and TV shows from both IMDB and Netflix for further study of the findings. Utilizing heterogeneous graphs with multi-level semantics, this research primarily presents a method for movie recommendation. The spectator and the director are treated as one entity in our suggested paradigm. We also used the suggested framework to merge two datasets throughout the study. The next step was to test the graph neural network system's performance on the varied graph. We found that the suggested model performed better than the prior techniques by using the strategy. When our model's framework is used, the results are effective in terms of multidimensional anti-Semitism.

To help patients make educated judgments on the drugs that are most suited to their individual medical circumstances, the authors of the research suggest a customized medicine recommendation strategy in [10]. Research to Come: Adding patient medicine reviews to the suggested method of medication suggestion is one way to make it better. Medication prediction accuracy may be further enhanced by analyzing MC-based ratings with an improved aggregation algorithm.

In an effort to meet the demands of prospective clients, researchers presented an intelligent method for dealing with large-scale heterogeneous data in [11]. One of the most popular RS methods for making suggestions is the collaborative filtering, or CF, approach. We provide a new method for CF recommendation that uses polarity identification to create a hotel feature matrix from opinion-based sentiment assessment. To learn about guests' feelings about the hotel's amenities and their types of stays (solo, family, couple, *etc*.), our method integrates lexical, syntactic, and semantic analysis. The suggested algorithm tailors hotel suggestions to individual guests by considering their preferences and the hotel's amenities. With the help of the big data Hadoop platform, the created system can manage diverse data sets, and it can also use fuzzy rules to suggest hotel classes depending on the kind of guests. Two hotel websites provide real-world datasets that are used to conduct different studies. Additionally, the F-measure, recall, and

CHAPTER 3

An Assessment of Big Data Analysis Technologies for Improved Information Delivery

G. Geetha[1,*] and **Kuldeep Singh Kaswan**[2]

[1] *School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: Academic and industry researchers have found sentiment categorization to be an intriguing subject due to the exponential expansion of online reviews. Collecting annotated training data is challenging, but the reviews help with several areas. Despite the development of several sentiment categorization algorithms, information retrieval remains inaccurate, inefficient, and sluggish when it comes to convergence speed. The authors of this article propose using the Spider Monkey Crow Optimization approach (SMCA) as a sentiment classification paradigm for training Deep Recurrent Artificial Neural Networks (DeepRNN). In order to remove inappropriate information and save the user's search time, the telecom assessment is used to eliminate stemming and stop words. In contrast, review sentiment is extracted using SentiWordNet for feature extraction. In addition to the extracted SentiWordNet features, DeepRNN also takes into account supplementary data like hashtags, numerical values, expanded words, and punctuation when classifying emotions. In order to get the required review, we use the distance metric Fuzzy K-Nearest Neighbor (FuzzyKNN). The proposed SMCA-based DeepRNN has been evaluated and tested extensively, and its results in terms of precision (95.5%), accuracy (97.7%), recall (94.6%), and F1-score (96.7%) are superior to those of its competitive models.

Keywords: Big Data Analysis, Improved Information Delivery, Fuzzy K-Nearest Neighbor (FuzzyKNN), DeepRNN, and Spider Monkey Crow Optimization algorithm.

INTRODUCTION

Researchers have started using a movie recommendation system based on sentiment analysis, and recommender systems in general have recently become

[*] **Corresponding author G. Geetha:** School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: geetha.g@jainuniversity.ac.in

popular apps for dealing with information overload *via* tailored suggestions. A new method for emotive categorization is developed in this work, which aids in the recommendation of positively rated films [1]. By sifting through vast amounts of data, recommender systems assist people on the Internet in identifying content that could pique their interest. The addition of social trust connection data to the recommendations system enhances recommendation accuracy, according to recent research. Suboptimal recommendation performance is the outcome of the majority of current research, which focuses on explicit trust connections [2]. Over the last 10 years, several scholars have focused on the subjectivity of social media microblogs, sentiment analysis, and opinion mining. One useful tool that consumers have at their disposal is movie recommendation systems. The data accessible online is steadily rising as the number of users or watchers continues to rise daily. As a result, concerns about computing, analytics, and big data have emerged [3]. In order for online stores to make a profit, recommendation algorithms are crucial. Many other fields may make use of it. Classification of recommendation systems is dependent on whether they are contfor Emotion Labeling.ent-based, collaborative, or hybrid. In the absence of a sufficient quantity of data to generate suggestions, these systems fail [4]. Users trust the opinions of complete strangers while making purchases on a variety of websites. Users seeking to make judgments might benefit to a certain extent from this shared online information since it gives insight into how people view things like goods, services, or events. Outcomes for movies as well as product reviews, blogs, and social media posts have been achieved *via* the continuous study of sentiment identification [5].

Product suggestions, assessments of movies, political campaigns, game forecasts, and many other areas of our lives have been impacted by social media platforms like Twitter, thanks to its ever-increasing pleasant features and the simplicity with which one may share opinions with the whole globe. In most cases, just a small number of people have the clout to sway a whole organization or network to adopt a certain viewpoint. Among the many ideas utilized and researched in the field of social network analysis, “network centrality” ranks high [6]. Blogs, opinions, comments, and postings across many social media platforms populate today's online world. The film business has found sentiment analysis to be a useful technique for automatically classifying reviewers' polarized opinions. Sentiment analysis may determine the quality of a movie review based on linguistic trends [7].

RELATED WORKS

Give an example of a trust model that uses social network activity to determine how trustworthy a user is in [8]. By taking into account both open and closed

social trust ties, the suggested trust model not only represents a more realistic and fine-grained degree of user trust, but it also enhances the number of trust relationships. The TrustSVD social recommendation algorithm, which uses matrix factorization, is then updated to include this enhanced social trust data. By comparing our suggested technique to SVD++ and conventional TrustSVD, we discovered that it produces more accurate predictions, leading to better user experiences. This was confirmed by analyzing the prediction results using the Douban-600k dataset, which is available on the Douban Movie website.

In order to make the system of suggestions meaningful and efficient, it is necessary to enhance the recommendation services over the old one, as stated in [9]. To address these problems, this article lays out a distributed Apache Hadoop framework-based solution for producing large-scale, effective recommendation services based on movie scores, votes, Twitter likes, as well as reviews pulled from a variety of external sources by a web bot. A deep semantics analyzer that combines user movie interest (UMA) with recurrent neural networks (RNN/LSTM attention) is used to generate emotions from reviews. A more meaningful movie suggestion list tailored to the user's tastes on a mobile app is efficiently generated by the suggested recommender after evaluating multivariate.

This study aims to apply a deep learning strategy that clusters comparable users based on demographic factors to overcome the initial user cold start issue in recommendations for movies [10]. The suggestions are generated by a deep neural network using this encoding. Our method's efficacy is confirmed by the results of the analysis.

The best match is located by researchers referencing the criteria established in a study [11]. Text categorization is a subfield of NLP that aims to categorize texts according to the meaning and context of a given phrase. Many sentiment analysis text classification methods have been developed utilizing neural networks. The objective of this research is to use the IMDb dataset for movie reviews and apply neural network emotion assessment to them. Here, we focus on CNN-GRU and CNN-bidirectional GRU, two separate multi-branch models. Despite CNN-bidirectional GRU's somewhat better accuracy, the findings demonstrate that CNN-GRU achieved equivalent outcomes using much less training time.

For monotonous submodular maximizing under single as well as multiple knapsack constraints, including scalable implementations in networked and streaming environments, the first adversarially resilient technique was described in [12]. Our technique produces a strong summary of nearly optimum (up to polylogarithmic times) size for a single knapsack limitation, which can be used to generate a constant-factor approximation of the ideal solution. Our estimate gets

CHAPTER 4

Analyzing the Effectiveness of Big Data Tools for Analyzing Complex Data Structures

Avadhesh Kumar[1,*] and **S. Santosh**[2]

[1] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: A lot of individuals have been devastated by the COVID-19 epidemic, which has caused a lot of worry, dread, and confusing emotions. A wide range of complicated emotions has surfaced after coronavirus immunizations were introduced. With the use of deep learning algorithms, we hope to decipher their feelings in this study. Right now, there's no better place to let your emotions and thoughts out than on social media. Twitter, in particular, can give you a good sense of what's popular and what others are thinking. To better comprehend the wide range of opinions on vaccination, we set out to conduct this study. From December 21 to July 21, the collected tweets were used in this study. The most popular vaccinations that have lately been accessible worldwide were detailed in the tweets. The opinions of individuals on different immunizations were assessed using the VADER tool, which is a Natural Language Processing (NLP) application. After dividing the responses into positive and negative categories, we had a clearer understanding of the whole issue. According to our numbers, 33.96 percent were in favor, 17.55 percent were against, and 48.49 percent were unsure. While doing our study, we also took into account the timestamps of the tweets because attitudes changed over time. To evaluate the prediction models, the RNN-oriented architecture—composed of Bi-LSTM and LSTM—was used. When it came to accuracy, LSTM achieved 90.59% and Bi-LSTM 90.83%. To further confirm our models and conclusions, we used other performance indicators, including accuracy and F1-score, as well as a confusion matrix. This research lends credence to the goal of global coronavirus eradication and sheds light on public sentiment about COVID-19 vaccinations.

Keywords: Big Data Tools, Complex Data Structures, COVID-19 pandemic, Natural language processing.

[*] **Corresponding author Avadhesh Kumar:** Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: avadheshkumar@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

INTRODUCTION

Improving national strength with competitiveness in the context of big data is a crucial problem for future international competition, given the rising popularity of big data applications and the fiercening of global rivalry [1].

There have been a variety of contexts, techniques, and methodologies used to study medication adherence. Still, there are a lot of moving parts when it comes to managing drug adherence regimens as a complex phenomenon, and our understanding and ability to measure them all are limited [2]. In recent times, Deep Neural Networks (DNNs), transformers, and ensemble models have shown exceptional SOTA performance in a variety of disciplines, including Computer Vision (CV), machine learning, and Natural Language Processing (NLP), among many more. Along with these benefits come drawbacks, such as models that are computationally too costly, include millions (or perhaps billions) of parameters, and are too heavy to be implemented on the client side of web applications, mobile devices, or embedded systems. To get around these problems (large size & inference computation time), researchers are looking at active model compression approaches like distillation [3]. Constant innovation in financial goods in the age of big data shows high innovation value and immediately drives a considerable revolution in the conventional financial business. Using big data analysis to study and improve the effectiveness of financial product innovation, along with resource allocation, is now crucial [4]. Modern culture jumped headfirst into the information technology age. Big Data (BD) and the Internet of Things (IoT) both came out around this time and will be very important going forward. With the advent of the IoT as well as BD analytics, we are entering an intelligence age, and these technologies are crucial to the success of traditional companies in an era of tremendous integration across many sectors. Modern business management relies heavily on performance management, particularly for SMEs [5]. The spiritual need to fully immerse oneself in the distinctive rural culture and to consistently raise the profile of rural tourism has grown in tandem with the modernization of rural tourism. Nevertheless, the outcome is sensitive to seemingly small changes in components because of the intricate nature of the tourist sector. Hence, conventional wisdom is that the tourist industry's future is completely unpredictable [6]. Improving smart transportation resource management and passenger coverage has become more dependent on driver behavior analysis in recent years. Information such as driving actions, acceleration, velocity, and fuel consumption is contained in the real-world environment by the driver's principles. Because feature assessments and classification do not take advantage of mining information, driving pattern analyses in big data are complicated [7].

RELATED WORKS

A study focuses on researching how China may use big data to bolster its cultural soft power [8]. The paper's authors speculate that this influence stems from Confucian society, which has had a profound impact on China for over two millennia. Confucian principles boost China's soft power culturally in the digital era, which might lead to better living conditions and higher incomes for the general populace.

The study delves deeper into the impact of environmental restrictions on this connection *via* these technologies in its discussion of the link between R&D expenditure and business financial success [9]. From 2007 to 2016, a sample of listed businesses' imbalanced panel data was used, and appropriate regression models were created using logical reasoning. Based on empirical research, we know that R&D spending correlates positively with financial success and has an inverted U-shaped connection with environmental regulatory severity. As a result, the maximum efficiency curve shows that stricter rules lead to a lower level. As a result, companies will be less likely to invest in manufacturing, research and development, and other high-value areas of financial performance as a result of stricter environmental restrictions.

As part of their quality control efforts, the authors of [10] examined the papers that used 91 transcriptomics datasets that were stored in the Gene Expression Omnibus database. Transcriptomics studies include more flaws in their reporting and analysis than one may think, according to this study. In order to improve the basic criteria for generating, analyzing, and reporting gene expression data, this paper concludes with a few recommendations for researchers and reviewers.

Our study in [11] suggests a way to use Patch-Based Convolutional Neural Network models (PBCNNs) and large data analysis to segment brain tumors early on. We use the BraTS datasets from 2012 to 2018. Profiling, cleaning, transformation, and enrichment are some of the preprocessing procedures used to improve the data quality. To measure how well the proposed model works, many indicators are used. A few examples include the precision, sensitivity, accuracy, Dice similarity coefficient, and rates of false positives and true positives. Our research shows that the suggested technique beats state-of-the-art methods for the first brain tumor segmentation. This is particularly true in cases of brain tumors, but the suggested strategy may also help doctors make faster and more precise diagnoses, which improves patient outcomes overall. In addition to demonstrating the promise of PBCNN models in medical imaging research, this study stresses the significance of large amounts of data in this area.

CHAPTER 5

Automated Reasoning and Topic Detection in Text Clustering

Ranjana Sharma[1,*] and **K. Vanitha**[2]

[1] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: This work gives a new theoretical approach to interpretative topic modelling. Instead of words or n-grams, this technique analyses complete phrases. Clustering phrase embeddings and probabilistic sentence evaluations in the text corpus is the suggested technique. The topic model estimates sentence frequency distributions within topics and text-wide topic frequency distributions. Our technique provides explicit topic interpretation since sentences are more significant than words and encompass full grammatical and semantic frameworks. The automated process is also provided. Sentence embeddings are produced using BERT-based context embeddings. Our technique also shows that topic modelling may integrate internal and external information sources for large data processing. Conventional topic modelling uses a single text corpus as the internal knowledge source. The external knowledge source is the BERT, a machine learning model trained on huge textual data, which generates context-dependent phrase embeddings.

Keywords: Text Clustering, Topic Detection, and Automatic Reasoning.

INTRODUCTION

The number of Indonesian internet users is expanding quickly. This is supported by data showing that there are 55 million users of the internet in Indonesia as of June 30, 2012, up from only 2 million in the year 2000.

In 2014, there was a 6% rise in internet use data compared to 2013. The number of internet users in Indonesia in 2014 was 88.1 million, as reported by the APJII. This is an increase over last year's total of 71.2 million. The present media

[*] **Corresponding author Ranjana Sharma:** College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; E-mail: sharmaranjana04@gmail.com

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

landscape has been influenced by the ever-increasing number of internet users. One of the key requirements for delivering information to convey the newest news for those who use the internet was identified by the mass media as a change or resolution towards a more sophisticated approach, namely social media, such as news portals. As of June 12, 2012, there were over 2000 news items published daily [1 - 5]. Some news stories with identical titles or content are loaded on many portals, resulting in a proliferation of news stories with identical material. This puts viewers in a difficult position, as they are forced to choose between seemingly similar but ultimately unrepresentative news stories.

The purpose of topic modelling, a kind of statistical analysis, is to unearth these underlying semantic patterns in text retrieval, classification, record annotation, social significance information classification, propaganda recognition, media outlets evaluation, and analyzing social networks, *etc*., are just some of the many uses for topic modelling in the NLP field. The topic model determines the texts' connected subjects and the words that make up each topic. Unobservable model parameters are estimated in the topic model in the form of two matrices, one of which determines the discrete range of probabilities of term occurrence throughout each topic and the other—the discrete probability of topic occurrences within each text [6 - 10]. Probabilistic Latent Semantic Analysis (PLSA) by Thomas Hoffman was the first topic model. In addition, Latent Dirichlet Allocation (LDA) was introduced by David Blei. This framework is a twist on the PLSA formula. As an a posteriori allocation of unknown model parameters, Dirichlet distributions are suggested for use in LDA, specifically for the distribution of words within topics as well as the distribution of topics within texts. Furthermore, the observed text data is utilised in conjunction with Bayesian inference to determine its posterior distribution of parameters. This model's popularity in the academic community has led to the creation of other variants [11, 12].

Additive Regularisation of Topic Models (ARTM) was a method for topic modelling introduced in 2014 by Konstantin Vorontsov. In this method, a particular topic model is decided by selecting a regularizer that imposes constraints on the unidentified numbers of the members of two matrices. In contrast to LDA variants, here only point estimations of unidentified variables are assessed, rather than their distributions estimated using Bayes's method or Gibbs sampling. The ARTM-based topic model now dominates LDA-improved topic models [13, 14]. The ARTM method does this by combining the smoothing of backdrop topics with the decorrelation and sparsity of domain-dependent topics provided by a variety of regularizers. Terms, which might be words, n-grams, or phrases, are the fundamental analytical units for the vast majority of topic models. As a consequence, we may think of the distribution of opinions on any given

subject as a series of these atomic building blocks. Since there is no logical relationship between the concepts, this format makes it more difficult to comprehend the results.

Because of this, determining the semantic relationship between them is challenging, and words lose their original context. The meaning of a phrase is usually better conveyed *via* its surrounding context than through the term itself. For this reason, it is crucial to create topic modelling techniques that completely reflect the context in the topics they produce during analysis. The approaches discussed below are meant to address this issue. Some work has been done in topic modelling to take into account the cohesiveness of text parts, such as word order in a phrase [15]. SentLDA, a model that takes into account how often words appear together in sentences, was suggested. The words, however, continue to serve as the primary analytical building blocks here.

RELATED WORKS

In natural language processing, one of the most intriguing and promising avenues for resolving a wide range of practical issues is to make use of the semantic information already present in the texts that are being analysed. In order to get higher-quality outcomes, it is essential to operate with word meanings rather than word labels. In text analysis, for instance, the precise meaning of a word may be gleaned from its surrounding context. When terms are used in comparable contexts, it's likely that their meanings are interchangeable. Using a specialised function that can quantitatively generalise contextual information for particular words in the text's embeddings, it is feasible to compare the senses of other words by examining the connections between their embeddings. With this kind of function that maps keywords in the content to embedded contexts, it's feasible to do semantic evaluation by comparing the meanings of various sub-samples of text. Based on neural networks, Google created the BERT model, a generative transformer, in 2018. For each word or symbol in the text, the BERT model that had been trained produces a contextual embedding. Words or tokens may have their embeddings continuously updated based on their context, according to the generative nature of the BERT model. The semantic closeness of the matching words may be inferred by comparing the analogy of the produced embeddings. It provides several options for integrating a cognitive dimension into various NLP activities [16 - 18]. The text's nuanced meaning may be stored in BERT and then used to address a wide range of practical issues. The BERT is useful since it can adapt to different settings.

In contrast to word2vec, we are not vectorizing word labels but rather their situation-dependent implementations, which may have quite varied meanings

CHAPTER 6

Automated Reasoning Tools and their Application in Text Mining

G. Sindhu Madhuri[1,*], Tushar Mehrotra[2] and **S. Kannadhasan[3]**

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[3] *Department of Electronics and Communication Engineering, Study World College of Engineering, Coimbatore, Tamil Nadu, India*

Abstract: Data mining is an application that searches for information in various databases in order to discover new facts and patterns. Conventional data processing technologies and applications are inadequate for handling the massive amounts of data available in SMEs today. Important for SMEs' decision-making is the adoption of efficient and effective data mining tools. This study describes and analyzes seven popular open-source data mining tools: KEEL, KNIME, Orange, RapidMiner, R Project, Tanagra, and WEKA.

Keywords: Automated Reasoning Tools, Application in Text Mining, SMEs' decision-making, data mining scans data sources, Data processing tools.

INTRODUCTION

With the proliferation of both software and computer networks, the potential harm that may be inflicted by such assaults has skyrocketed. Achieving low-effort, high-reward traffic monitoring and finding requires efficient identification of anomalous network activity. When providing automated services for a variety of applications, it is crucial to have reliable methods for detecting suspicious Internet traffic [1]. It is becoming more critical to manage data processing activities as well as the resources used by them as the volume of data continues to expand. Users are increasingly executing their work in the cloud since maintaining an individual architecture is either infeasible or unprofitable. Several approaches

[*] **Corresponding author G. Sindhu Madhuri:** Department of Computer Science and Engineering, JAIN (Deemed-t--be University), Bangalore, Karnataka, India; E-mail: g.sindhumadhuri@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

have been suggested to swiftly profile towards an optimal configuration or to find one using information from prior runs, as configuring workloads as well as resources is often difficult. However, training such techniques using performance data is often difficult and expensive [2]. The use of geothermal energy is on the rise due to North China's booming economy. Locating promising areas in North China for deep geothermal resource extraction is an immediate priority. Despite the widespread distribution of geothermal assets in the northern China Plain, choosing targets and assessing thermal resource management are crucial for the safe, stable, and efficient development of deep geothermal assets in North China [3]. Healthcare organizations rely on engineering project management services for a variety of reasons, including the development, implementation, and oversight of healthcare-related infrastructure as well as technological initiatives. Data mining methods used in engineering project management might maximize allocation of resources, decision-making, and project planning and execution. Data mining is the process of discovering hidden patterns, correlations, and trends in massive datasets by the use of automated extraction of information rather than human study [4].

The old advertising paradigm of businesses is being tested by the ever-increasing amounts of consumer data. Businesses are gradually shifting from passive to active marketing as a result of the fast development of data mining technologies, which has sparked novel concepts in marketing methods. The primary goal of technology for data mining is to examine data, gain an understanding of the relationship between variables, and ultimately make predictions about the future. Consequently, businesses integrate technology with their advertising campaigns and then use big data to construct an accurate advertising model [5]. The proliferation of online communities not only produces an enormous amount of data but also causes dramatic changes in our everyday lives. We can learn a lot about people's online communication, interaction, and working styles by analyzing massive data from social media. In addition, designers might benefit from this information as they work to enhance social media platforms to better meet the demands of users. In light of this, research into mining and analyzing large amounts of social media information has become one of the most active and promising fields in recent years [6]. There will be significant shifts in managerial responsibilities and business procedures brought about by the advent of artificial intelligence (AI). Artificial intelligence is reshaping organizational leadership and decision-making. Key competences and business processes, like managing knowledge, are showing the impacts of AI. Consumer outcomes like perceptions of service quality as well as satisfaction are also showing the effects of AI [7].

RELATED WORKS

This work presents ODFM, a new method for detecting anomalous traffic that relies on data feature optimization and mining. To cut down on data detection and make anomalous traffic detection easier, we create a feature selection approach to lower the feature research dimensions and set up a P2P (peer-to-peer) traffic identifying module to filter as well as mine the associated service traffic. The suggested method has proven to be successful and competitive in the broad tasks of anomalous network traffic identification [8], thanks to experimental findings that show a significant improvement in detection accuracy.

We provide a cooperative method in this study for users to share anonymised execution records of workloads, mine them for patterns, and leverage clusters of past workloads to optimize future optimizations. Using a publicly accessible trace dataset, we assess our initial solution for mining activity implementation graphs, as well as show that work groups generated from traces alone have predictive power [9].

From an extensive data mining vantage point, this work sheds light on the restricted hydrocarbons manufacturing procedure and provides a practical and precise way to forecast production as well as assess the economic viability of projects in the Cardium deposit. More than fifty factors related to Cardium formation are included, and for the first time, several artificial intelligence techniques are described and evaluated using information from technology, petrophysics, and geology. The approach presented in this research may be readily used to reliably anticipate output from additional unusual fields and formations, including the Montney, Eagle Ford, Fuling, and Bakken Shale plays, provided that the necessary data is available [10].

Using synthetic EM data, the suggested approach was thoroughly evaluated. A CMD2 (GF Instrument) EM equipment with GPS was used to regulate its efficiency, which was based on minimal actual data. When utilizing the Advanced Mode, the automated identification of unmineable inclusions became much more accurate. However, the Simple Mode computing method has the dual benefit of allowing real-time examination of the dug mine slope and being independent of equipment location. The approach for rock inclusions identification is introduced in this paper, which may help optimize mining operations by making them safer, more efficient, saving money, and less harmful to the natural world [11].

In order to improve the company's precise advertising model's clustering performance, as well as the K-means algorithm's sensitivity to cluster center setting up, our work optimizes and extends the K-means algorithm in conjunction with the artificial colonies of bees algorithm. Implementing evaluation and

CHAPTER 7

Hybrid Intelligence for Information Analysis for Machine Learning and Automated Reasoning

Merin Thomas[1,*] and **Ramesh Chandra Tripathi**[2]

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *College of Computing Science & IT, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

Abstract: Social networking and microblogging have grown, reflecting this generation's ideas. User reviews of products, companies, individuals, forums, films, *etc*., may gauge people's opinions. Analysts automated positive and negative judgment categorization using algorithms. Its automatic categorization is sentiment analysis. This study uses SVM to classify product reviews' feelings and phrases. The study will critically assess several datasets utilised for this aim. The Support Vector Machine (SVM) learning approach has been trained, verified, and used on many data sets to assess ambiguous sensory polarity. To improve product review categorization, this research employs SVM artificial intelligence to examine several datasets for sentiment and language classification. This work trains, evaluates, and simulates a support vector machine training system to recognise ambiguous emotions or review polarity on numerous datasets. SVM has the highest starting accuracy (89.98%) among classification methods. Adding sentence forms improves accuracy. It concludes that SVM is reliable. Application/Improvements: Support Vector Machine models are evaluated. Finally, the Support Vector Machine (SVM) classifies better and more accurately.

Keywords: Information Analysis, Machine Learning, Automated Reasoning, and SVM.

INTRODUCTION

The process of machine learning was a line of reasoning that looks at how computations may learn from data.

Such computations function by constructing a model based on data sources and using it to make predictions or decisions rather than only adhering to updated

[*] **Corresponding author Merin Thomas** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: merin.thomas@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

standards. Computational statistics and machine learning are very connected. The algorithms used for machine learning may be broken down into the following categories from a technical standpoint: There are many types of learning, including Supervised, Unsupervised, Semi-Supervised, and Reinforcement Learning. Supervised neural networks, in opposition to unsupervised training, occur when various input items are provided with a labelled result (also called a supervisory signal). In semi-supervised learning, there is a large number of input items, but only a subset of these objects is labelled with an output value. This kind of learning falls between supervised and unsupervised learning. However, with reinforcement learning, a system is trained by the environment to choose the optimum course of action depending on the rewards it receives.

Do people's ideas and beliefs have different degrees of detail? This refers to the overarching feelings expressed in a single statement or document, such as a revision document. Classification problems have often been used to describe the work of analysing reader perspectives, ideas, and sentiments. Component-based sentiment analysis often serves two major purposes: identifying positive feelings conveyed in texts, and detecting the hidden semantic component in writings written by cretins. The application of machine learning techniques for sentiment analysis and opinion mining presents a promising avenue for automating the information-gathering, processing, and interpretation processes. To determine the polarity of views, a technique known as “Sentiment Analysis” is used. Positive, negative, and neutral are all ways to describe polarity. Similar concepts are referred to by different names, but they all mean the same thing. Documents, web pages, social media feeds, and so on may all be analysed using sentiment analysis to reveal the underlying tone of the feedback they contain. Emotion analysis is a sort of categorization analysis in which data is broken down into distinct groups. These groups may be dichotomous (positive/negative) or multi-dimensional (happy/sad/angry/*etc*). Sentiment analysis is a text categorization approach that uses the words inside a document to discover the underlying attitude (positive or negative) of the author. Using sentiment analysis, we can quantify the underlying feeling of a product review, which in turn helps forecast the customer's future purchase behaviour. The reviews simplify the process of evaluating the merchandise. Sentiment analysis has become an increasingly popular and important method for making sense of the vast amounts of data we generate every day. Companies may now automatically elicit critical insights and streamline all background operations thanks to this. Extracting, identifying, and characterising the emotional content of a text unit may be done with the use of machine learning and natural language processing. When a supervisor uses this approach, the general tone of a review document or phrase is conveyed with a single phone call. Classification problems that involve analysing overarching sentiments in text are often created, such as dividing a review article into positive and negative

categories. Opinion mining, analysis of sentiment, sentiment extraction, and efficient rating are just a few of the many names for sentiment categorization.

RELATED WORKS

Here, you'll find a summary of the many methods proposed for doing aspect-based sentiment analysis on product reviews by various writers. Information from online user reviews is analysed using machine learning methods. Labelling a feature-wise rating for each product based on the person being studied is the main emphasis [1]. A system that uses the Internet to recommend and evaluate products is being developed. They used NLP to scan reviews and Naive Bayes classification. Polarity information and product feature comments were also gathered. Star ratings, evaluation date, helpfulness ranking, and review polarity are only a few of the elements that may be used to graphically inform the customer which of two things is superior [2]. Opinion Mining, formerly referred to as emotion analysis, is a kind of Natural Language Processing and Information Extraction that identifies the user's ideas or perspectives based on whether they are represented as either favourable or unfavourable comment and quotation clusters in the text. Methods for Sentiment Analysis that are either supervised or data-driven include Naive Bayes, Maximum Entropy, and Support Vector Machines. A Support Vector Machine (SVM) is used for classification, taking into account both the accuracy of the classification and the sentiment of the categorization [3]. The software is able to deduce that the opinion mining method did not make direct use of social network data. Both online and offline advertising may benefit from the methods suggested in this research. Due to its potential use, the computational approach to opinion, emotion, and subjectivity has received a lot of recent attention [4]. They take a vocabulary of representative words for each category and use those words to generate a representative article for each class. To better classify texts using supervised learning, they create a system that takes lexical information into account. Next, a composite growth Naive Bayes classifier is constructed by adaptively pooling the data sets from both of these models [5]. The meaning of a statement might change depending on the surrounding context. Since “contextual dependency” and “label redundancy” are two distinguishing features of sentence sentiment classification, they suggest a new approach for categorizing sentiment based on CRFs to account for both factors. Using CRF, they attempt to capture the restrictions of context on the mood of the statement. Many applications, such as electronic commerce, *etc*., rely heavily on the ability to extract these subjective messages and analyse their orientations [6]. Next, we look at how a polarity classification method makes use of four distinct kinds of fundamental linguistic information sources. However, our suggested method does not increase efficiency further by including dependency-based information or by excluding objective elements from feedback. The purpose of polarity analysis is

CHAPTER 8

Performance Analysis of Rule-Based Reasoning in Complex Data Environments

Sunanda Das[1,*] and **Gaurav Kumar Rajput**[2]

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

Abstract: It is well recognised that data preparation is necessary to provide reliable data with which mining tools may derive useful insights. Preprocessing methods need modification when dealing with several sources of continuous data inflow. This study suggests using active rule-based systems, particularly complex event processing (CEP) systems and engines, to aid domain experts in defining and executing pretreatment tasks for data streams. Our method's key innovation is how it enables domain experts to easily manage temporal data by formulating preprocessing methods as event identification rules expressed in a SQL-like language. This concept is implemented in a freely accessible software package that combines a CEP processor with libraries for web-based data mining. In order to test the efficacy of our method, we provide three examples of real-world applications in which CEP rules are used to preprocess data streams by incorporating new temporal information, modifying features, and dealing with missing values. Preprocessing activities may be expressed in a flexible and high-level fashion using CEP rules, as shown by the experiments, without incurring excessive memory and time overheads. The generated streams of data not only aid in enhancing classification algorithms' predicted accuracy, but also permit simplifying decision models and shortening the time required to learn.

Keywords: Complex Data, Automated Reasoning, Rule Mining and Complex Event Prediction.

INTRODUCTION

The advent of an Internet of Things (IoT) era has increased the significance of data analysis and machine learning approaches [1].

These two fields of study are anticipated to play a significant role in the development of appealing ideas like Industry 4.0 and Society 5.0 [2]. So, there are

* **Corresponding author Sunanda Das:** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: sunanda.das@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

always fresh proposals for uses of data mining software. Among the most significant difficulties linked to these principles is the automated identification of a human state or a response at a specific time or period [3]. Such issues may be articulated for a wide range of complicated systems that rely on human operators; for instance, in the medical field, operators may include the elderly in nursing homes or the mentally ill. Before an operator's status is automatically detected, the operator themselves needs to be constantly assessed. Commonplace sensors are enough for this task. Non-invasive vital sensing using Doppler sensors was presented in [4] and has since been used for human condition monitoring to improve efficiency. In the IoT context, this approach is considered important because its low computational requirements allow it to be implemented on small-scale processors, enabling longer battery life. This research investigates the effects of music listening and non-listening on the human state. Each participant's non-contact vital sensing measurements were standardised and pre-processed before the tests began. Based on this, we formulated two separate categorization problems to address. Fuzzy methods [6] and artificial neural networks [5] were employed to handle the aforementioned categorization challenges. For the purpose of automating the construction of ANNs, this research proposes two variants of the Differential Evolution technique (DE) [7], which was originally created for tackling multiobjective optimisation issues. The objective was to develop an ANN that could solve a given classification problem effectively using a simple architecture [8]. For the purpose of automatically generating fuzzy rule-based classifiers, COBRA, an optimization method based on population dynamics, was also used [9]. Data mining techniques such as support vector machines (SVM) [10], k-nearest neighbours (kNN) [11], decision tree (DT) [12], the Hybrid Evolutionary Fuzzy Classification method (HEFCA) [13], and regular artificial neural networks have all been employed to address classification issues pertaining to illness detection in humans. The findings from each of the classifiers are compared and contrasted. Knowledge discovery relies heavily on data pretreatment (Garcia *et al*., 2014). Data preparation entails a number of time-consuming but crucial steps, including cleaning, transformation, filtering, mapping, and integration (Zhang *et al*., 2003). In addition, using an appropriate preprocessing approach may improve mining outcomes (Crone *et al*., 2006; Uysal & Gunal, 2014). Unfortunately, data preparation is often accomplished using inelegant and error-prone techniques that need considerable familiarity with statistics and advanced programming abilities. Preprocessing activities may be quite challenging for domain specialists, but they might benefit enormously from high-level solutions & help (Bilalli *et al*., 2019). Common preprocessing tasks may be aided by tools like Weka & software packages accessible in languages like R and Python; however, these programmes have a steep learning curve.

It is challenging to reuse procedures and incorporate them into professional systems because domain specialists are often required to customize features or adapt generic approaches to suit their specific application area. (Huang *et al*., 2016). With more and more people wanting to learn from data streams, new computational methods for real-time data mining are required (Gama, 2010; Gaber, 2012; Ghomeshi *et al*., 2019). Data pretreatment is crucial for stream data mining, yet there is a dearth of literature on the topic (Ramrez-Gallego *et al*., 2017). The majority of current methods either focus on static data or attempt to alter existing data mining techniques. Dynamic feature selection and separation methods are still uncommon in software libraries, even though there is an increasing amount of work on the subject (Barddal *et al*., 2019; Ramrez-Gallego *et al*., 2018). Domain experts may lack a thorough understanding of data mining techniques and toolsthus, it is important to provide them with options to describe and perform preprocessing activities when data transformation with filtering is required. Active rule-driven linguistics might provide online data mining systems with a more expressive and efficient way of expressing and enacting data preparation operations (Corporate Act-Net Consortium, 1996). At the same time, rules' excellent understanding has made them popular in expert systems (Grosan & Abraham, 2011); this may make it simpler for experts to articulate the necessary preprocessing processes. On the other side, event-driven systems benefit greatly from the sophisticated data manipulation capabilities offered by these languages. For instance, a quickly-processing stream engine (Affetti *et al*., 2017) may be an ideal platform to carry out preprocessing methods, especially if the incoming data includes temporal context. Complex event processing (CEP) platforms are one such tool, letting users spot relevant events and share their verdicts with others (Cugola & Margara, 2012). These rule-based systems are distinguished by the extensive language support they provide for specifying rules and patterns in a SQL-like syntax. Several recent studies have shown the increasing prevalence of IoT applications in fields as diverse as monitoring the environment (Gomes *et al*., 2020; Sun *et al*., 2019), hospitals (Loreti *et al*., 2019), connected transport (Kousiouris *et al*., 2018), and and trash management (Pardini *et al*., 2020). Among others, it has sparked renewed interest in CEP in recent years. Our proposal is based on the idea that CEP languages and engines, with their fast processing streams, rule-based engine, and SQL-like syntax, are ideally suited for the description as well as execution of stream preparation tasks within online mining data processes. In this study, we use CEP to establish pretreatment rules for data streams, which is the first time this has been done. As a first step, we design several kinds of ECA (event-condition-action) procedures for preparing data streams with different goals in mind, such as data modification and combination. Following this, we demonstrate that CEP language ideas, including spatial and temporal doors, pattern matching as well as filtering, temporal

CHAPTER 9

Evaluating the Usefulness of Big Data in Decision Making

Dhruv Galgotia[1,*] and **Mohammed Zabeeulla**[2]

[1] *Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Connected to the web of things, "Big data," an increasingly popular phrase, describes the massive amounts of data frequently produced by modern healthcare systems. The sheer volume and complexity of the information make it difficult to extract actionable insights for forecasting and policymaking. Big data technologies provide significant infrastructure support to enable critical decision-making and better achieve the objectives of IoT systems. Privacy protection, data integrity, and verification of identity are three tenets that must be strictly upheld in big data for healthcare service administration. To solve these problems, this study suggests a big data ecosystem-enabled, scalable computer system that allows authenticated data access methods for health data analytics enabled by the Internet of Things. A large-scale data analytics tracking system, as well as a data storage/access system generated from blockchain, make up the two primary sub-architectures of the architecture that is suggested. In order to analyze, store, and authorize verified use of information collected by IoT-enabled devices, this solution utilizes big data systems as well as blockchain architecture. Data linkability may be avoided, and unauthorized users can have no access to private data by using the zero-knowledge protocol. The results demonstrate the efficacy of our method in resolving healthcare privacy and big data analytics issues.

Keywords: Big data analysis, decision making, blockchain, iot, healthcare.

INTRODUCTION

In cognitive production that is powered by big data, techniques for situational awareness, neural networks with convolution, as well as cutting-edge computing, construct event modeling as well as prediction. Data mining, predictive analytics, and sensor-actuator networks are the building blocks of automated control systems.

[*] **Corresponding author Dhruv Galgotia:** Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: ceo@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

Object recognition methods, deep learning computations, and visual perception tools all contribute to better robotic systems for navigation. Fusion of data technologies, cooperative operating methods, and context recognition tools enables configurable manufacturing procedures. Contributing technologies include methods for control and decision-making, edge and cloud computing, as well as remote intelligent devices, image detection, identity verification, and recognition. Technologies for monitoring contextual data, cloud computing, vision and navigation systems, and intelligent simulation environments are employed in their development. Visual cognitive algorithms, context-aware tools, and artificial neural networks are used to optimize robotic manufacturing processes.

Innovation and enhanced decision-making in service industries depend on accurate and up-to-date knowledge. This is why data is so important for service industries. Data that has been collected and evaluated thoroughly is used to predict what will happen and how specific parameters are trending in the future [1]. As we begin to fully grasp the advantages of this position, we realize that new technical breakthroughs have enabled us to generate and gather more data in almost every aspect of our lives, particularly in our social connections, research, jobs, and health. Consequently, “big data” has become a common phrase. Particularly challenging is the fact that healthcare businesses are producing data at a quick pace [2]. All sorts of diverse, multi-spectral, incomplete, and ambiguous primary source observations are a part of healthcare big data, including unstructured, semi-structured, and structured information (e.g., on demographics, illness, injury, therapy, problems with mental and physical health, as well as illness management). Structured data includes information such as ICD codes, traits, genotypes, and genomic data, while unstructured information includes medical imaging, observations, environmental variables, clinical notes, lifestyle choices, medications, and health economic data [3]. Additionally, advancements in the Internet of Things (IoT) have sped up data-driven applications in healthcare, smart cities, transportation, and networking [4]. Particularly, the medical field has been instrumental in the use of various sensors and technological devices to monitor patients' vital signs and other health indicators [5]. Information in excess of ever-increasing amounts has resulted from the expansion of omics fields, including genomics, proteomics, and metabolomics [6]. Electronic health records (EHRs) are replacing traditional medical records, further facilitating data dissemination. Medical professionals, epidemiologists, and healthcare policy analysts want to improve patient care and community health by using such massive amounts of data [7]. Setting up the required infrastructure, techniques, and tools is critical for making good use of the generated big data [8].

RELATED WORKS

Its stated goal is to perform the aforementioned tasks using the NEWS data. To create complex queries, it is necessary to first study the NEWS data and build it. A significant data challenge arises from the complexity and volume of the data in relation to data analytics. Also, bear in mind that the healthcare business is not making the most of its resources because data confidentiality and integrity issues are widespread, even though this data is believed to be crucial for improving health outcomes and reducing costs [9]. Data privacy and integrity verification using efficient methods and methodologies is of the utmost importance for addressing the limitations of these challenges. Two North East hospitals in the United Kingdom are overseen by South Tees Institutions NHS Foundation Trust1, which is responsible for patient data. These hospitals store 5 million records, each containing 20 tables with around 50 attributes related to patients. Executing queries on large datasets, however, can take weeks or even months owing to the data's massive volume and complexity, as well as user-related concerns during UDF construction [10]. In healthcare, managing heterogeneous data is a key challenge in big data analytics. However, this data presents an opportunity for better healthcare for countless people. Big data platforms help address issues in the healthcare delivery system by analyzing large, complex healthcare data. The use of highly parallel, multi-tenant environments is common in big data processing systems such as Hadoop2 and Spark3, but these environments are vulnerable to software and hardware errors that can reduce performance [11]. A complex query is intentionally designed to make it more difficult to obtain responses when dealing with challenging data. A sizable internal healthcare Internet of Things surveillance network is established by transferring data from various wearables and sensors to servers. Problems with data confidentiality (the protection, accessibility, storage, and longevity of personally identifiable information) and data integrity (the assurance that data is accurate, comprehensive, consistent, and securely stored in compliance with the General Data Protection Regulation, or GDPR)) arise organically as the [12]. With the help of extensive analytical systems, the Internet of Things (IoT) ecosystem for healthcare generates the most sensitive large data. Processing and storing such massive amounts of data on state-of-the-art computers raises serious concerns about data integrity, as erroneous diagnoses might put lives in jeopardy. Security and privacy are additional issues when working with large amounts of sensitive health data. Public cloud computing, in particular, makes cloud-hosted big data systems more vulnerable to hacker attacks. According to McAfee [13], 3.1 million users' data that was stored in the cloud has allegedly been hacked in the last year. Big data systems, which often rely on centralized datacenters, pose a danger to health data, as this instance shows. To avoid these kinds of incidents, healthcare institutions should have strict protocols to ensure the confidentiality and integrity

CHAPTER 10

Evaluating the Role of Data Mining in Enhancing Decision-Making

Alisha Sikri[1,*]

[1] *Department of AIML, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India*

Abstract: An obstacle for sentiment-based recommendation systems is the abundance of information available online. Because of flaws in their data-cleaning and analysis processes, existing machine learning (ML) methods fail to provide trustworthy recommendations. To get over this issue, this research proposed a recommender system that uses Twitter reviews and is based on Lite deep learning (LightDL). The information retrieved from Twitter is cleaned using data-cleaning processes. After that, the LightDL algorithm is trained to use this preprocessed data to identify the relevant characteristics (hashtags, unigrams, bigrams, etc.) in each dataset. We have learned four types of information from this: semantic, grammatical, figurative, and aspects based on tweets. Finally, the data is rated as positive, negative, or neutral based on the learned attributes. A battery of tests is run in MATLAB to evaluate the model's performance on many metrics, including recall, accuracy, precision, f-measure, and error rate. Specifically, the proposed LightDL model outperforms all other models across all metrics and achieves 95% accuracy on the Twitter dataset.

Keywords: Lightweight Dl, Sentiment Analysis, Recommender System, Twitter Data.

INTRODUCTION

Many years have passed since businesses began using transactional system data recorded in their day-to-day operations to help in making choices. This goal has been achieved via the use of various business intelligence tactics and data analysis methods. Novel processing processes, able to handle such massive amounts of data, have been developed in recent years; these mechanisms are known as Big Data [1]. Data quantity, information diversity, and data velocity have all increased. CPSS creates complex heterogeneity and large amounts of data from

[*] **Corresponding author Alisha Sikri:** Department of AIML, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India; E-mail: alisha.sikri@niet.co.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

many sources by integrating humans, machines, and data into large-scale computerized systems. Academics are very interested in graphs of knowledge because of the critical role they play in empowering large-volume, poor-quality data sets to support CPSS's clever services and applications. Knowledge is described in the Resource Description Framework (RDF) as directed labeled graphs in the shape of subject-predicate-object triples [2]. Despite the widespread use of artificial intelligence (AI) and big data to automate network administration tasks, managers still sometimes resort to heuristic analysis of system data in the event of an unexpected cybersecurity issue or fault. Though existing methods focus on reducing the pool of potential data, a diverse range of information is necessary to tackle complex cybersecurity problems [3]. During the big data age, technology for data mining is mostly used to assess the collected information. Afterwards, all inductive reasoning regarding the information is computerized in order to find potential patterns. With recent advances in data mining technologies, the national psychological wellness dataset can be used to efficiently address a range of early warning indicators related to mental wellness. Approaches: One such use case is the extraction of key characteristics and details from psychological data using data mining [4]. There is a pressing need for innovative technical solutions to address the growing complexity of city infrastructure, the abundance of urban data, and the need for interdisciplinary research and decision-making for long-term city sustainability. Data and process connectivity, automated reasoning over knowledge, and large-scale data handling are just a few areas where ontologies have shown to be useful tools for practitioners [5].

Scholars have noted that findings from conventional mature organizations cannot be fully transferred to the study of current online businesses due to numerous distinctions between conventional and contemporary enterprises. Consequently, the study of digital transformation has seen a proliferation of novel ideas throughout the last few decades. More research on its definition, conceptual dimension, and measurement has been conducted during the span of its long history of growth. Despite the fact that effect thinking will be important for the success of corporate digital transformations, very little research has examined its effects in this context [6]. Big data has become an essential part of many industries, helping businesses gain valuable insights and make informed decisions. Nevertheless, to make good use of big data, it is essential to ensure data quality. Therefore, practitioners and academics have been increasingly focused on big data quality over the past few years because of its substantial influence on decision-making. While there have been studies on the accuracy of data deviations, they tend to focus on narrow issues, such as outliers or inconsistencies [7].

RELATED WORKS

In this chapter, we put forth a tensor-based paradigm for analyzing knowledge. This framework helps with knowledge graph modeling, fusion, and argumentation. We start by fully representing heterogeneous knowledge networks using Boolean tensors. Afterwards, we detail a set of graph tensor procedures for extracting, modifying, and aggregating high-order graphs of knowledge. To further infer the link among any two entities, we compute the relationship path tensor by performing tensor 1-mode combination operations on the information network representations tensor, as well as the entity's representations tensor. Lastly, we present a case study demonstrating that the proposed paradigm is successful and practicable [8].

In this research, we describe MADPM, a system in which robots use data-processing technologies to store and assess data from networks. In MADPM, data-processing sequences are formed by interactions between agents. As part of this procedure, we plan a sequence's structure based on needs, as well as a way to augment it for an in-depth examination that supports intuitive reasoning. During testing, we put the initial system through its paces with five case studies. Based on the findings, administrators may benefit more from a complex data display than from the chosen representation, which was selected for its single-faceted nature. Providing a multi-system presentation of information for heuristic reasoning in network administration tasks is the end product of our suggested technique [9].

Preparing data, mining it for insights, analyzing the findings, and using a decision tree algorithm are all parts of the data mining early warning systems for problems with mental health that this study outlines. Results from the experiments show that the early-warning technique presented in this article can reach an excellent precision of over 96%—second tier of proof. Clinical trials examine the efficacy of a therapy [10].

We take a look at the ways in which ontologies have bolstered smart city services, and we compile an in-depth overview of the issues tackled and the results obtained so far through the use of ontology-based apps. To achieve this goal, we systematically reviewed the literature on ontology and its influence on the development of modern intelligent cities. Our review findings additionally influence our proposal for a taxonomy of the city's sub-domains covered by the taxonomies we uncovered, as well as a list of the scientific field's current priorities for study. Finally, we delve further into a few unanswered questions and focus on the areas where semantic technology has already been shown to be beneficial in improving the smart city idea [11].

CHAPTER 11

Harnessing Big Data for Advanced Business Intelligence

V. Gokul Rajan[1,*] and **H.S. Shreenidhi**[2]

[1] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Thanks to the rapid development of remote sensing technologies, we now have greater availability of remote sensing data than in the past. Big data has arrived. Hyperspectral features, high geographical resolution, and high temporal resolution are characteristics of Big Data that are present in data acquired by remote sensing. By combining geographical feature data with remote sensing data, this study compares changes in water cover and vegetation and provides a feature-supporting, useful, and efficient data cube for use in time-series analytics. This research provides a spatially-focused definition of the SRSDC, or remote sensing data cube. This information cube provides an efficient, adaptable, and scalable method for analyzing large volumes of RS data using spatial attributes. It provides an overview of the SRSDC's organizational structure. For query operations, the SRSDC offers feature translation, and for storing and managing vector feature data, it offers spatial feature repositories. The article details the process of creating a distributed execution engine and a feature data cube for the SRSDC. The manufacturing process and long-term remote sensing evaluation are used as examples to assess a feature data cube and a distributed execution engine. To the pinnacle of the information economy now stands big data, a fresh strategic asset for humanity. Methods for discovering new information rely on data analysis, supervised learning, unsupervised learning, and combinations of the two.

Keywords: Big Data, knowledge discovery, business intelligence, data analysis (DA),.

INTRODUCTION

Throughout our research and professional experience, we have noticed that many professionals in related fields have trouble making heads or tails of the various methods for analyzing large data sets because of the complexity of the terminology and the fuzziness of the concepts involved.

[*] **Corresponding author V. Gokul Rajan:** Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: gokulranjan@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

While developing and testing a BI system, for example, we were asked why concepts such as big data (BD), data analysis (DA), and Knowledge Discovery (KD) were not part of it. These questions are valid since we classify BI, BD, DA, and KD as a set of ideas related to big data analysis. However, there is no universally accepted single standard or framework that encompasses all of these domains. While these and related inquiries may lie beyond the purview of a given research project, answering them may necessitate additional time and effort, particularly in a research environment where debate must be grounded in an evaluation of the relevant literature.

Each of the papers reviewed in this article takes a somewhat different tack on the question of how best to analyze large amounts of data, but they all cover the same ground. Rather than devoting considerable time and energy to clarifying the differences between BD and BI ideas from a data structure perspective, the literature pays little attention to issues like conceptual ambiguity, misunderstanding, and confusion, as well as the interrelationships across approaches.

Because of this, we decided that the area of big data analysis needed a thorough evaluation. To analyze and evaluate BI, DA, BD, and KD critically so that their similarities, distinctions, and relationships can be identified; (ii) to provide other researchers with a visual representation that can help them respond quickly to questions, concerns, and problems that arise from not understanding the concepts mentioned in this paper, and (iii) to lay the groundwork for future discussions between researchers and practitioners.

Business Intelligence

One way to look at business intelligence (BI) is as an umbrella term for all the things that help a firm collect, analyze, display, and disseminate information: strategies, procedures, programs, data, items, technology, and technical architectures [1]. Businesses may gain an advantage over rivals [2] and even forge deeper connections with consumers and increase sales [3] by adopting a more nuanced perspective of their clientele. It is a crucial aspect of organizational growth because it gives companies a leg up in situations with positive information asymmetry [1, 4, 5], enabling them to maximize earnings, optimize resources and processes, and make more proactive [6] and strategic [7] choices. Business Intelligence is used not only at the strategic and tactical levels but also at the operational level due to its many benefits.

Users of business intelligence tools may be better equipped to anticipate and respond to changes in the business environment [8]. Its goal is to improve

operational performance management by providing stakeholders with a deeper understanding of the business.

Businesses may benefit from business intelligence (BI) by extracting actionable insights and previously unknown knowledge from their operational data, which can then be utilized to make more informed predictions, calculations, and assessments [10]. Traditional business intelligence [9] centered on extract, transform, and load (ETL), data warehousing (DW), and reporting (R). Data exploration and visualization, however, have become significant research foci for the next generation of BI [11, 12]. There seems to be a shift in focus from static reporting to interactive visualisations, as research themes have shifted from a "ometrics summary" to "odiscovering the causes and consequences of the events the metrics describe" [12]. Data mining, text analytics, near real-time BI, self-service BI, and BI in the cloud are only a few examples of the new business intelligence advancements driven by corporate competitiveness [13].

Data Analytics & Big Data Analytics

The goal of data analytics is to improve decision-making by applying computer analysis to existing datasets [14]. According to Ridge, "oany activity involves the integration of statistical techniques to data for the goal of obtaining insight from the information" is what the word "odata analysis" (DA) really means [15].

This multidisciplinary field encompasses a wide range of disciplines, including statistics, machine learning, pattern recognition, applied research, data mining, artificial intelligence, business intelligence, prescriptive analytics, and descriptive analytics, among many others. Thus, study themes and concerns pertinent to DA ideas include those already defined as research fields in BI, including visualisation, cloud computing, and data exploration [16 - 21].

RELATED WORKS

Big data has become an invaluable strategic asset for humans in this era of the information economy. Relying primarily on correlational data instead of traditional causality, it is representative of a fresh data-intensive paradigm in science that has arisen as a result of theory, computer models, and experience. It is reshaping the way scientists perform their work and will soon become a vital force in scientific advancement. Data science and AI depend significantly on insights from large-scale remote sensing data, which lies at the intersection of statistics, computer science, and earth science. To find a predictive, functional, or practical data structure, it must be able to explore all of Earth's big data without being constrained by the many data types. Important methods for discovering new information include both supervised and unsupervised educational approaches, as

CHAPTER 12

Examining the Role of Big Data in Advanced Statistical Modeling

Meenakshi Sharma[1,*] and **Chandrasekar**[2]

[1] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Here, we provide a novel approach to simulating and reconstructing annual surface mass balance (SMB) data across glaciers using a deep artificial neural network (ANN; *i.e.*, deep learning). This method is currently included in the SMB component of an open-source regional glacier development model. Instead of using physical processes, which are becoming more common in traditional glacier models, we use data science to build a parameterized model. A annual glacier-wide SMB may be modelled using topoclimate data with deep learning or Lasso (least absolute shrinkage and selection operator; regularized multilinear regression), and the glacier's form can be updated using a parameterization specific to the glacier. Our nonlinear deep learning SMB approach is tested and validated on a dataset of 32 French Alpine glaciers, compared with other conventional linear statistical methods. Compared with linear methods, deep learning outperforms them with a root-mean-squared error (RMSE) of 0.51 m w.e. and an estimated R^2 of 0.77. This is due to the fact that deep learning can explain up to 108% more variation and up to 47% more space. Nearly one-third of deep learning's observed nonlinear behavior is attributed to the temporal dimension. The original measurements of ice thickness are the source of significant uncertainties in the development of glacier geometry. These results provide evidence that deep learning may be a powerful nonlinear tool for glacier modeling, enabling us to reconstruct or simulate SMB time-series data for specific glaciers within a given area to predict past and future environments.

Keywords: Big Data, Advanced Statistical Modeling, AI, nonlinear behavior, linear statistical methods.

INTRODUCTION

Significant changes are occurring in glaciers worldwide due to human-caused climate change [1].

* **Corresponding author Meenakshi Sharma:** Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail meenakshi.sharma@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

Major environmental and societal effects are predicted as a result of the worldwide retreat of glaciers, despite significant regional variations. About 10% of the global population lives in alpine areas and the neighboring plains; glaciers are crucial to their existence because they provide water for farming, hydropower generation, industry, and household consumption. Numerous terrestrial and aquatic ecosystems rely on these water sources for their existence because they continue to produce runoff even during the warmest and driest periods of the year. Correctly anticipating and mitigating the associated environmental and socioeconomic repercussions requires an accurate prediction of future glacier development. Estimates of glaciers' future contribution to sea level have been calculated using a variety of global glacier development models throughout the past decade. Irrespective of methodology, all glacier models must address the following two primary processes that dictate the course of glacier evolution: (1) Glacier mass equilibrium, which is the difference among the mass gained *via* development (*e.g*., avalanches, snowfall, and refreezing) as well as the mass lost *via* ablation (*e.g*., calving, melt, sublimation of ice, firn, and snow), and (2) ice flow factors, that is the downward movement of ice due to gravity. It is difficult to develop a model that accurately simulates these processes on a global scale without making significant assumptions or simplifications [1 - 5]. Recent work has enhanced these models' depiction of ice flow dynamics by replacing empirical parametrizations with simpler physical models. However, most large-scale models of glacier development rely on temperature-index models to depict the glacier mass balance. A linear connection between positive degree-days (PDDs) and ice/snow melt is used in such a model. The fact that they work well in large-scale studies with a sparse number of data is a major factor in their popularity and has led to their outperforming more complicated models. Although these models may approximate the climate-glacier communication, it should be remembered that both systems exhibit non-linear behavior, especially when subjected to pre-processed forcings such as PDDs13. A nonlinear model, such as a deep ANN, may replace these more traditional methods. The use of artificial neural networks (ANNs) for regression problems in glaciology has been mostly uncharted; for example, only a small number of researchers have utilized shallow ANNs to predict a single glacier's ice thickness14 or mass distribution.

As climate proxies, glaciers can illustrate the changing climate to an international audience (IPCC, 2018), making them one of the most iconic representations of climate change. In the coming decades, mountain glaciers may significantly affect the water quality of glaciated catchments, according to several studies (*e.g*., Hock *et al*., 2019; Beniston *et al*., 2018). It is important to accurately forecast the range of hydrological, ecological, and economic effects that the melting of mountain glaciers might have. The change in hydrological regimes, including their timing and amplitude, is highly dependent on future climate scenarios (Huss and Hock,

2018). In order for society to adapt to changing hydrological and climatic conditions, understanding these changes is essential. These problems can be partially addressed by using glacier and hydro-glaciological models, which provide a range of outcomes under varying climate change scenarios. Understanding the evolution of glaciers at sub-regional and regional scales requires defining (a) the surface mass balance (SMB) and (b) glacier dynamics. To model these procedures, researchers have used a range of granularities. Different approaches may be used to replicate these processes on a large scale, that is, over many glaciers at a watershed size [6 - 10].

Half a century ago, the first mathematical models employed in glaciology were simply multiple linear regressions based on a limited set of meteorological parameters (Hoinkes, 1968; Martin, 1974). The advent of artificial intelligence has played a crucial role in the remarkable advances in statistical modeling over the last few decades. In contrast to other geosciences domains such as oceanography, climatology, and hydrology, the glaciological domain has not made full use of these methods, according to our contention (see, for example, Ducournau and Lguensat *et al*., 2018; Fablet, 2016; Jiang *et al*., 2018; Marçais, Rasp *et al*., 2018; and de Dreuzy, 2017; Shen, 2018). However, investigations have begun to rely more on statistical methods. Steiner *et al*. (2005) used ANNs in the first study to investigate mass balance of the Great Aletsch Glacier in Switzerland. Compared with the standard stepwise multiple linear regression, they found that a nonlinear model performed better at representing mass balances in glaciers. The link between climate and glacier mass balance was also discovered to include a sizeable nonlinear component [11 - 15].

RELATED WORKS

In spite of Lewis Fry Richardson's 1922 proposal for NWP, the advent of programmable computers in 1955 marked the beginning of its practical implementation [1]. The past 30 years have seen tremendous advancements in weather forecasting using neural networks. Model output statistics (MOS), a method for comparing model results with observational data [12 - 14], was the most common technique for improving numerical models' predictive capabilities before 2000. In 1983, [15] a weather-forecasting method was developed that combined statistical and dynamic elements. A study [16] in 1991 marked a turning point in dynamic modeling. The growing gap between the current and forecast timeframes is accompanied by a lack of trustworthiness due to high computational demands, long prediction horizons, and the absence of design approaches for selecting model structure and parameters [13, 16, 17]. Other limitations and challenges include these approaches. A method for predicting the lowest possible temperature using artificial neural networks and backpropagation

CHAPTER 13

Exploiting User Preferences through Content-based Recommendation Systems

Shambhu Bhardwaj[1,*] and **Rajesh Pandian**[2]

[1] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Present on virtually every online service today, recommender systems have quickly become an integral part of people's everyday digital lives. Modern deep learning-based models can only perform at their peak when fed massive amounts of data. Multiple domains, including Amazon items, restaurants, and breweries, have offered several datasets meeting this condition. The hotel industry has seen relatively few advancements and databases, with even the largest review dataset being in the hundreds of thousands rather than millions. Traditional collaborative filtering methods are also inapplicable to the hotel domain due to its higher data sparsity compared to standard recommendation datasets. In this research, we present HotelRec, a TripAdvisor-derived, massively scaled hotel recommendation dataset comprising 50 million reviews. To the best of our knowledge, HotelRec is the largest recommendation dataset in a single domain and includes textual reviews (50M vs 22M) in the hotel domain (50M versus 0.9M).

Keywords: User preferences, content-based recommendation systems, HotelRec, hotel industry, textual reviews.

INTRODUCTION

In this age of rapidly evolving internet technologies, recommendation systems (RSs) have piqued the interest of both businesses and consumers alike because of their practicality and significance in online shopping and in improving customer acceptance.

It is widely held that the success of an online business is directly proportional to its customers' loyalty. Online booking and reservation systems are also an integral

* **Corresponding author Shambhu Bhardwaj:** College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; E-mail: shambhu.bharadwaj@gmail.com

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

part of the travel market. According to a study by Mariani *et al*. [1], tourism is the most influential sector of the global economy. To avoid any potential issues with their hotel accommodations, travelers from all over the world rely on recommender systems to find the best available hotels and make reservations online well in advance of their arrival dates.

Some recommender systems have recently been developed to help travellers obtain a list of hotel recommendations before booking, as reported by Liu *et al*. [2]. Conventional recommender systems handle only homogeneous data; hence, the performance of hotel recommendation systems suffers when dealing with the heterogeneous nature of most Internet and online data. Particularly when dealing with complex data in its many numerical, textual, and visual formats, developers must pay close attention to data heterogeneity when building recommender systems. According to Li *et al*. [3], there aren't many recommenders out there that can deal with heterogeneous data. The ones that do employ customer ratings, but don't take into account other forms of user feedback, like votes, reviews, rankings, or video views, that are accessible on social media. We have integrated many user feedback mechanisms into the proposed approach, such as YouTube video plays and votes. The novel and advantageous recommender system we propose does two things. The first step is to utilize a hotel characteristic matrices that takes both numerical and textual information into account to provide accurate recommendations; the second step is to analyze user-contextual information, such as reviews, ranks, votes, and YouTube video views, in order to extract emotions from reviews. The second point is that different user types have different requirements and interests; thus, a fuzzy module may categorize them into different categories and provide hotel suggestions based on them. Similar to how “room,” “food,” with “cleanliness” are the most important amenities for a family, “pool,” “spa,” with “gym” may be more desired by each individual traveler. It is not only “computer” and “WiFi” that can be useful for a business traveler.

To get accurate and useful suggestions, recommender systems were proposed by Zhang and Mao [4]. If a recommendation is relevant, it implies that it takes into account the customer's preferences and interests. Customers' past experiences are included in a recommender system alongside the hotel's qualities and amenities. The primary objective of this paper is to provide a method that can effectively manage and analyze varied online datasets in order to make precise hotel recommendations according to the customer's preferences.

Sentiment analysis and opinion mining of user reviews to produce a polarity rating reflecting the extent to which a person loves or hates a hotel was the most demanding component of managing the data for our multifeature hotel recommendations system. Our proposed recommender considers customer

feedback, numerical ranking votes, and video views to deliver more accurate results than existing recommenders. Existing recommenders depend on user assessments of a hotel's qualities and facilities. User sentiment about a hotel is expressed as a polarity score in text. We selected a big data solution based on Hadoop because it easily deals with data heterogeneity and variety, which is especially important given that the provided method employs both numerical and textual data. The primary focus of this study is categorizing visitors into five types (individual, family, company, friend, and couple). We will take into account a variety of rating criteria and employ a feature-based approach to sentiment analysis on user feedback. The hotel's location, room, cleanliness, service, and personnel are just a few of the categories on which TripAdvisor's users can provide feedback. Sentiment analysis, also known as opinion mining, is the practice of extracting insights from written feedback. With the proliferation of online resources comes the inevitable necessity to sift through material in light of individual tastes. Every digital activity, from online buying and social networking to music streaming and hotel reservations, now includes the use of a recommender system. Researchers have explored recommender systems for over 30 years (Bobadilla *et al*., 2013). The many models and datasets generated over the years include movie datasets (Harper and Konstan, 2016), Amazon product datasets (McAuley *et al*., 2015; He and McAuley, 2016), and music datasets (Celma, 2010). Covington *et al*. (2016) found that large-scale recommender systems based on deep learning led to improved suggestion quality. Although deep learning-based models enhance recommendation performance, massive data sets are required to achieve this. For example, state-of-the-art models shown in recent years need enormous datasets numbering in the millions for optimal performance (Wang *et al*., 2019; Liang *et al*., 2018; He *et al*., 2017). A few articles have addressed hotel recommendation, such as Zhang *et al*. (2015) and Wang *et al*. (2011). Additionally, the largest publicly available collection of hotel reviews has 870,000 records (Li *et al*., 2016). Traditional collaborative filtering techniques are not well-suited to the hotel domain due to the sparsity of the data (Wang *et al*., 2015; Khaleghi *et al*., 2018; Musat and Faltings, 2015). In addition, hotel ratings are unique due to the longer duration and increased number of factors to examine (Khaleghi *et al*., 2018).

LITERATURE SURVEY

Various studies have examined the issue of suggestion, which is as ancient as the subject itself. These studies have focused on various contexts, including Amazon products, breweries, restaurants, pictures, music, movies, and more. While other attributes, such as text, sub-ratings, date, and helpfulness, may also be included in datasets, ratings are the sole constant. The number of user-item interactions may range from many thousands to several hundred thousand. Our best research

CHAPTER 14

Exploring the Application of Data Mining in the Detection of Fraudulent Transactions

Priyanka Chandani[1,*]

[1] *Department of DS, CSBS, AL-CSBS, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India*

Abstract: Many businesses, including those in the banking sector, have begun offering their services online in response to the meteoric rise in internet use. Worldwide, financial fraud is on the rise, leading to enormous losses. From now on, threats like unlawful transactions and unusual assaults may be actively detected by developing sophisticated financial fraud detection systems. Data mining and ML have become more important tools for addressing these problems in recent years. However, in order to speed up processing, analyze massive amounts of data, and detect new attack patterns, these systems need a lot of improvements. According to this study, a Deep Convolution Neural Network (DCNN) can be used to detect financial fraud utilizing smart algorithms. When dealing with large datasets, this approach enhances detection accuracy. The performance of the proposed model is assessed by comparing it to other models that use auto-encoder methods, deep learning, and current machine learning models on a revised credit card fraud database. The results of the experiments demonstrate that the proposed model attained a 99% detection accuracy in only 45 seconds.

Keywords: Artificial neural network, deep learning, financial fraud, convolution neural network, data mining.

INTRODUCTION

Big data, characterized by its heterogeneity, is created every day in massive quantities as a result of the exponential expansion of technological advances. A stumbling block continues to be the integration of heterogeneous data. The integration and completion of the business information required have become arduous processes [1]. Technology, big data, as well as open source material, have all contributed to the meteoric rise of the company intelligence (BI) industry over the last decade. Despite this expansion, private-sector use of open government

* **Corresponding author Priyanka Chandani:** Department of DS, CSBS, AL-CSBS, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India; E-mail: priyanka.chandani@niet.co.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

information (OGD) remains low due to a lack of awareness of its benefits. Private enterprises' limited use of OGD raises hopes that it might inspire novel concepts and support data-driven decision-making [2]. Due in large part to the coronavirus epidemic, which has increased companies' online presence, social networking has emerged as a major player in modern business and organizational operations. Researchers and company owners alike are increasingly intrigued by the prospect of using social network data to power company analytics. Social media has given company owners a new tool to better understand their customers and adapt their operations to their ever-changing demands. Twitter makes it simple to keep tabs on what's trending, what's set to gain traction, and important information like the source of the trend and the people involved [3].

A number of recent technical developments—such as the Internet of Things, machine learning, and big data analytics—have already had an effect on business operations. Among the next technologies, artificial intelligence (AI) might radically alter the way marketers approach their craft. Artificial intelligence has the potential to be useful in many areas of today's business world. The fate of human civilization will be determined by artificial intelligence, according to the philosophical as well as scientific consensus [4]. To prevent the coronavirus from spreading rapidly, many entities monitor the number of confirmed cases daily. Big Data technologies handle very large, diverse, and dynamic datasets that are inherently unmanageable with current methods. There are four steps for handling Big Data before it generates usable information: acquisition, access, analytics, and application. The purpose of business intelligence is to make sense of a company's operational and transactional information. Using graphical representations of data, Big Data Analytics examines both current and historical company choices [5]. Thanks to advancements in IT, BI, and information technologies, data is now both a storage medium for massive amounts of information as well as an analytical method in and of itself. Businesses may better adapt their operations to achieve successful production if they have a proper grasp of the benefits of analyzing company data and information from business intelligence [6]. People in every walk of life are becoming increasingly dependent on information technology as the global economy expands and technological and scientific progress advance. A comparable data system, known as intelligence for business (BI), is likewise under continual development in the corporate sector. Additionally, the use of IT has increased rivalry in the corporate world. In addition to the commercial acumen of corporate decision-makers, a scientific, precise collection of data, made possible by rapid BI tools, is essential for a company to uncover new markets and seize opportunities in the cutthroat business world [7].

RELATED WORKS

Presently underway projects, patterns in development throughout history, and potential future directions were the foci of this study. This article employs a CiteSpace-based bibliographic analysis, using 681 unique papers from 2000–2021, culled from the WoSCC and Scopus indexes. We found the most influential nations, universities, authors, publications, and references in the field of academia. The study delves into international social networks and academic partnerships. The degree of cross-disciplinarity is quantified. Investigation methodologies, business intelligence (BI) applications, and obstacles are addressed independently as we delve into the historical patterns of burst keywords and hotspot dispersion. Possible causes of 2021 hotspot explosions are investigated. Lastly, recommendations are made for future scholars, and the course of research is forecasted. The findings indicate that in the following years, research on COVID-19, healthcare, 5G, as well as big data/AI driven business intelligence techniques will be highly sought. Consequently, the findings of this study may be used to influence future studies, especially those pertaining to the direct selection and method application domains [8].

Innovation and knowledge management's (KIM) impact on sustainability and innovation initiatives is expanded upon in this essay by Danny Sampson, as well as Marianne Gloet. This article describes the results of a longitudinal study that included a number of medium- and small-sized food and drink exporters from Australia. Dilek Cetindamar, Baraah Shdifat, as well as Eila Erfani's "Understanding the Big Data Analytical Capability as well as Sustainability Supply Chains" is the SI's concluding article. Sustainable supply chain efficiency (SSCP) and big data analytics capacity (BDAC) are reviewed in this literature review. After that, they provide some suggestions for future research and highlight deficiencies in the current literature. We trust that you will find this SI's extensive coverage of current and relevant subjects, organized under the theme of "Business Analytics & Big Data as Tools for Innovation as well as Sustainable Development of Companies" [9], to be both interesting and useful.

We have created an original structure for Heterogeneous Integration of Information and Analysis to tackle the challenges of handling different big data. Companies often employ big data analysis, a method for extracting knowledge, for company intelligence purposes. Unfortunately, data mining struggles with issues like limited memory and high processing costs, thus it can't handle really huge data sets well. Here, we provide a Convolutional Neural Network (CNN) design for analyzing massive data sets that are diverse in nature. Last but not least, the suggested approach is the quickest to integrate data structure, according to the trial findings [10]. It is also a great framework for business evaluation.

CHAPTER 15

Leveraging Text Mining Techniques for Efficient Information Discovery

M. Chandra Sekhar[1,*] and **R. Pachayappan**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: Researchers in the field of information systems often talk about human capital. Organizations need trained personnel to create and implement IT objects. This is particularly true when dealing with sophisticated technologies like AI. Natural language processing (NLP) and computer vision (CV) are two important subfields of AI. From a business perspective, this article compares and contrasts the skills required of CV and NLP experts. To achieve this goal, we used a text mining-based analytical pipeline to analyze AI job ads. Named entity recognition and word vectors were used to assess actual job ads from both sub-disciplines that were scraped from a big multinational online employment portal. We suggested an improved machine learning method, the Artificial Neural Network (ANN), for the last skill analysis. The requisite skill sets for the two job descriptions may be different. It calls for individualized thought as there is no one-size-fits-all profile of an AI expert.

Keywords: Text Mining, Information Discovery, Employee Skill Analysis, Computer Vision.

INTRODUCTION

Ads have evolved into an essential tool for companies in today's business world to boost their profile as well as expand their operations. Businesses across many industries use customized ads to increase their ROI and leverage consumer demographic information. This is especially true in the e-commerce, social media, recreation, and television industries. Digital billboards in shopping centers display the same ad layout every time, regardless of targeting, proving that many businesses still cannot get their demographics quite right [1]. There is growing concern about the surveillance nature of targeted online marketing systems, which

* **Corresponding author M. Chandra Sekhar:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: p.biswajeet@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

gather personal data from individuals and then use it to automatically classify and infer things. Examining the implications of advertising targeting may be enriched by examining the lives of LGBTQ+ persons, whose identities challenge conventional ideas about who is considered "normal" and entitled to privacy, autonomy, as well as a right to autonomy [2]. The use of machine learning has been enabled by data, which has accelerated data-driven material design. It remains a challenge to collect data mechanically from scientific research, which produces a vast volume of high-quality, trustworthy data [3]. Customers express their interest in using certain services by contacting numerous firms listed in their telecom call history. Time and the day of such messages, together with multifaceted characteristic dependency, provide valuable information for targeted advertising. Customers who use a particular service might also be potential customers for another service if there is a strong correlation between the two [4]. The use of digital health data is on the rise to identify individuals with comparable histories of illness, diagnoses, and outcomes, and to suggest multidisciplinary therapies for them. Expert physicians and novices alike work together to produce these files. A practical and transferable approach to finding individuals who have very similar clinical traits is using data mining in large, unorganized datasets. The Imagine Institute conducted a bioinformatics evaluation of the exome and focused on next-generation sequencing [5]. One of the main reasons internet advertising is growing is the use of programmatic advertising, which employs big data to deliver targeted, tailored marketing materials. One significant component of programmatic advertising is in-app adverts, sometimes known as in-app ads. Customers' individual requirements determine the timing and location of ad delivery in in-app marketing, an income stream closely tied to app development. Text messages on social media platforms are emerging as a powerful new kind of advertising that may influence customers' purchasing decisions, all thanks to electronic word-of-mouth (e-WOM) [6]. Surprising trends in educational institutions can be discovered through the rapidly developing interdisciplinary field of educational data mining. Various automated learning systems have been investigated in light of the difficulties educational institutions face in assessing learners' performance and enhancing instructional management [7].

RELATED WORKS

This study introduces a fundamental component extractor, an easy-to-use MS/MS data analysis tool capable of extracting characteristics specified by the consumer. This software incorporates, for the first time, the sequential neutral loss characteristics and the abundance of finished ions as fundamental components, in addition to the typical neutral losses and product ions. One example of the instrument's usefulness is the identification of nine sesquiterpene dimers in Artemisia heptaphylla that have not yet been reported. Artemiheptolide I (9),

among these dimers, showed an IC50 of 8.01 ± 6.19 μM when tested against influenza A/Hongkong/8/68 (H3N2) *in vitro*. In addition, IC50 values ranging from 3.46 to 11.77 I1/4M were observed for two recognized guaianolide derivatives (16 and 17) when tested against hepatitis A/Puerto Rico/8/1934 H1N1, H3N2, and influenza B/Lee/40. This method has many broad applications, such as improving the annotation effectiveness of LC-MS/MS analyses and effectively discovering new natural compounds. It may also be used to capture derivatives with specified segments [8].

To fill gaps in our understanding, provide new experimental data, and construct new qualitative Abnormal Outcome Pathways (qAOPs) for UVB, this research revisited the deadly consequences of UVB on crustaceans. The objectives were accomplished *via* the use of both computational and experimental methods. A varying amount of synthetic UVB was administered to Daphnia magna in a targeted testing experiment in the laboratory. To measure the impact of UVB on various levels of biological organization, the company used targeted bioassays. The NIVA Risk Evaluation Databases (NIVA RAdb) is a new computer tool used to compile fresh experimental findings and data from prior publications into a toxicity pathway network. Current and previous information were used to create a network of AOPs and to evaluate the weight of evidence. We also fitted the D to the AOPs to find quantitative connections between significant events. Convert the magna data into pre-made models. Potentially useful for future de novo qAOP creation for chemical and non-chemical stresses, the provided approach encompasses all stages of qAOP construction and assessment [9].

The primary focus of this study is the development of a gender and age-based customized advertising system for use in retail centers. The "I-Advert" system was developed by using machine learning and image processing approaches. The "I-Advert" solves the problems with current methods by providing malls with online marketing that is both automated and targeted. Indoors and in a real-life small business, we tested the equipment to gain a full understanding of it [10].

In order to get a better understanding of the LGBTQ+ community's (N=18) experience with internet advertising, thoughts on ad targeting, and how these systems interact with their homosexuality as well as other identities, we conducted semi-structured. Participants characterized the lack of variety and complexity in online ad content as stereotyped and tokenizing, a characterization reflected in our findings. However, their underlying mistrust and disapproval of the extractive, non-consensual nature of ad targeting also clashed with their demands for improved ad content. They wanted more say over what happens to their information and focus, wanted the option to opt out completely, and were worried about privacy issues related to persistent data aggregation, including

CHAPTER 16

Exploring the Potential of Big Data for Monetization of Information

Meenakshi Sharma[1,*] and **Somashekhara Reddy**[2]

[1] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: The public's swift adoption of wearable sensors and their rapid worldwide spread, together with the rapid advancement of the Internet of Things (IoT), have resulted in most companies and organizations being overwhelmed with massive amounts of data these days. Finding ways to use the data deluge could provide companies with an edge in the market, as data can enhance several aspects of a company. There has been some monetization in the industry across verticals in terms of stacking connected devices with various SaaS choices, such as subscription plans or smart device insights. A significant advantage in today's cutthroat digital marketplace will go to those who can properly monetize data rather than just hoard it, says the "machine economy" sprouting up in this sector. Innovations in big data, analytics, and AI have created new opportunities for competitiveness, with data strategically employed as an asset that can generate new revenue streams due to its ever-changing nature. Because of this surge, a plethora of new resources—including systems, platforms, tools, and markets—have emerged to help organizations make effective use of data. Actually, the goal of new business models is to shift the balance of power away from data-harvesting companies and toward customers. The selling of user data to companies like data analytics corporations is being promoted by startups and NGOs. Data monetization encompasses more than just data sales. It is also possible to implement measures that enhance data. Companies may generate money from data in three ways: 1) by improving internal processes or making more informed business decisions; 2) by integrating data into their main offerings; or 3) by selling data to current or potential customers. This article will address the key aspects of Internet of Things data monetization and the challenges that accompany it, with a focus on the healthcare industry. The topics covered will include data management, scalability, regulations, interoperability, security, and privacy. Not only that, but a detailed reference architecture for the healthcare data economy and a case study on the detection and prediction of cardiac abnormalities using privacy-preserving machine learning (PPML) and multiparty computing (MPC) methods are also provided.

[*] **Corresponding author Meenakshi Sharma:** Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, Karnataka, India;
E-mail: meenakshi.sharma@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

Keywords: Big Data Analysis, Decision Making, Blockchain, IoT, Healthcare.

INTRODUCTION

It is hard to imagine life in the early 21st century without data visualization. Data visualisations are useful in many different settings since they help with knowledge and decision-making. People may find themselves in space with the use of GIS and environmental maps, and in time with the help of schedules, timetables, and timelines.

Visualization tools such as infographics, dashboards, and interactive charts make complex processes, subjects, and systems easier to grasp. For both comparative and relational understanding of the world, people use a variety of visualisation methods. These include, but are not limited to, bar charts, pie charts, line graphs, and tree diagrams. Crucially, in order to make sense of the apparently infinite quantities of biometric and behavioural data that modern activities consistently produce, visualisations are relied upon by individuals, organizations, and governments. Two prime examples of this are the quantified self movement and the pervasiveness of linked wearable gadgets such as smart watches and fitness trackers. This research argues that the accompanying visualisations of these devices stand for a decentralised, networked management of bodies, promoting individual self-regulation as part of a broader neoliberal power structure dependent on biopower [1 - 5]. Data visualizations, such as those on wearables, are seen daily by many people and are usually powered by big data. Despite its profound effect on consumer culture, interpersonal dynamics, and public policy, this vague and ominous concept comes out as somewhat abstract when brought up in casual conversation. The term “big data” refers to three separate ideas: one way of thinking about data and computers, data with certain properties, and the analytical techniques that go along with it. The term describes “a computational turn in thought and research” (Boyd and Crawford 2012, 665) that puts an emphasis on quantitative computing tools and massive volumes of easily available data, such as metadata generated by website users' actions, over any other approach. Data is characterized as “big” when it is rapidly collected, processed in real-time, has low accuracy, is correlated across sources, and is significantly more voluminous than what was previously available in a particular domain (Ekbia et al. 2015, 1525-6; Leszczynski 2015, 967). Methods for discovering patterns inside and across massive datasets are associated with big data. These methods include automated data purification, neural networks, machine learning, and the predictive generation of new data (Salvo 2012, 37). Inaccurate data is a big issue in visualizing huge datasets. Oftentimes, big data mixes and confuses metadata with content, algorithmically produced “activity-based intelligence” with genuine human behavior, and digital user profiles created from aggregated data sets with

real, live humans. Actual people might be harmed by decisions derived from big data visualizations that rely on collected digital profiles (Crampton 2015, 521; Poster 1990, 126).

Data visualisation is the glue that holds big data together and makes it understandable to humans. It supposedly improves understanding and, ideally, leads to rational, ethical human decision-making by modifying and filtering information. The focus here is on human decision-making due to the fact that visualization is an essential human capability [6 - 10]. In the many rapidly automated and algorithm-dependent industries and professions, computers and robots do not need a visual representation of data to guide their decisions (Valle 2013, 2040). Decisions based on big data (by assessing visualisations) are considered too important to entrust to computers in many professions, despite our society's increasing dependence on big data. Consequently, there is a growing collection of big data-driven visualizations available to the public. Although data visualization is necessary in most areas of knowledge in the modern era, it is especially important and relied upon in research, commerce, and government sectors that are undergoing big data transformations (Ali et al. 2016, 656; Cook, Lee, and Majumder 2016, 135; Crampton 2015, 520). Hepworth and Canon (2018), 53, and Shoresh and Wong (2012) 5, both agree that visualization is an important tool for academics to employ for analyzing study results and for communicating those findings to a wider audience. Managing supply chains, finding monetizable behaviors, and determining acceptable pricing points are all possible *via* the visualization of aggregated data on content creation and behavior (Salvo 2012, 39). Whether it is for environmental protection, national security, political representation, or health care, governments rely on visualization to inform big data-based public policy decisions at the local, state, and federal levels (Salvo 2012, 39).

Organizations seeking to develop a fruitful data monetization strategy should have an in-depth understanding of the many data monetization strategies, including their implications, opportunities, and limitations. There has been an attempt to provide academic research in a number of areas, but there is still a lot of ground to cover since data monetization is still a relatively new subject. To inform the academic community about the potential of data monetization research, Liu and Chen conducted the first Systematic Literature Review (SLR), which was published in 2015. The writers aided readers in understanding data monetization by providing an approach that combines Analytics 3.0 (advanced analytics) with BI&A 3.0 (mobile and sensor-based analytics), as well as application scenarios and guiding principles. By combining data intelligence with traditional analytics, big data, and efficient methods of data collecting and processing, companies may achieve measurable business gains in Analytics 3.0. This research just included

CHAPTER 17

Exploring the Potential of Data Visualization to Enhance Data Analysis

Dhruv Galgotia[1,*] and **Biswajeet Kumar Pandey**[2]

[1] *Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Providing suggestions for integrated, cutting-edge, and efficient tools, methodologies, and technologies for accessing and processing increasingly growing volumes of data in diverse forms is a major problem in clinical data analysis and knowledge discovery. Personalizing a patient's care is a challenging task that requires the doctor to sift through and make sense of massive volumes of data. The scientific community behind precision medicine might benefit greatly from a unified system that facilitates data discovery, integration, preprocessing, model construction, storage, analysis, and visualization. The software package provides researchers with a simple, quick, and adaptable method for processing data, with the ultimate goal of enabling intelligent management, analysis, and visualization of massive genomic data. Services, data sets, and databases are at their disposal, or they can supply their own information for processing.

Keywords: Clinical data analysis, Knowledge discovery, Medicine, Model construction, Patient's care, Storage, Visualization.

INTRODUCTION

In this study, we developed a theory of Big Data visualization. Rendering models that aid Knowledge Discovery is a high-level goal of big data processing. This necessitates research into the features we call large Data, as well as thoughts on the many kinds of visual representations for large data.

Using what may be called (Synthetic) Metapictorial renderings, I suggest, is a potential modeling alternative toward attaining knowledge discovery from huge data. Massive volumes of experimental data have been produced as a direct outcome of the computer simulations. As a result, new research paradigms

* **Corresponding author Dhruv Galgotia:** Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: ceo@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

emerged, such as data-intensive science, and methods for processing large amounts of data were developed. Because of this shift, new research methods and data-driven discoveries will be required. During this process, we actively look for meaningful connections and employ cutting-edge techniques for knowledge acquisition. Knowledge discovery is founded on "data-intensive decision making" [1] and is made possible by the methodologies and technology of the new paradigm. Big Data refers to the ever-increasing volume, veracity, and freshness of datasets used in virtually every field of study, including healthcare [2]. Scientists can now create, store, and analyze data at breakneck speeds. Big data today encompasses not just massive data storage but also improved data analysis and interpretation. There is constant research and development into novel approaches to bettering data acquisition, storage, cleansing, processing, and interpretation. Precision medicine relies on a thorough understanding of human genetic variation, including both common and unusual mutations linked with disease susceptibility [3]. To provide the best possible care for each patient, precision medicine advocates personalized approaches to diagnosis, medical decision-making, treatment, and therapy [4]. Personalized treatment, intelligent medication design, population screening, and mining electronic health records might all benefit from the use of novel methods for extracting meaning from massive volumes of data. There is a dire need for collaborative networks that share data and knowledge, and there is a pressing need for standardization of data content, format, and clinical criteria. Using a text mining framework, a data mining strategy is designed for SVM-based information extraction from the biomedical domain [5]. New methods for analyzing and visualizing large amounts of data have contributed to the explosive expansion of precision medicine [6]. Large, detailed databases of genomic, transcriptomics, proteomics, or metabolomics variables, as well as conventional clinical patients' characteristics and treatment records, have emerged at an ever-increasing rate in the era of big data and the advent of electronic healthcare records [7]. However, in the context of their direct use in a clinical setting, such data are generally highly diverse, high-dimensional, noisy, and poorly interpretable [8]. Small sample sizes, imprecise technology, variations in clinical trials, a wide variety of patient groups, and an unstandardized health care system model can all reduce the reliability of the results. The term "big data analytics" refers to the practice of analyzing massive datasets that often include disparate data types. The goal of big data analytics is to unearth previously hidden relationships, complex patterns, imbalanced data sets, consumer and clinical preferences, and other valuable insights [9]. New methods for early prognosis of cancer therapy results have been discovered by scientists [10]. Because of the rapid development of new medical technology, researchers now have access to a wealth of biomedical data. Predicting how an illness will progress, however, is one of the most fascinating

and difficult issues facing clinicians today. Medical researchers have found great success with artificial intelligence methods because they can be used to forecast future outcomes of a disease type by discovering and identifying models from complex datasets and their relationships. The foregoing suggests that tailoring medical care to an individual patient is a challenging task that requires doctors to sift through copious amounts of data. The examination of medical records can benefit greatly from the use of big data technology. Researchers in the field of precision medicine might benefit greatly from a unified system that facilitates every step of the process, from finding relevant data to integrating, preprocessing, constructing models, storing, analyzing, and visualizing it.

RELATED WORKS

In the foreword of Fayyad and Grinstein's Information Visualization in Data Mining and Knowledge Discovery, Stephen G. Eick writes, "Visualization is the link between the two most powerful information processing systems: humans and the modern computer." It's "easy to be overwhelmed by the volumes of data that are now routinely connected," the author writes. Data mining is a reduction method that works in harmony with human talents. While I disagree that we are "easily overwhelmed," I do think that those whose jobs need them to work with large datasets are under so much pressure that they would benefit greatly from any way to speed up and simplify the discovery of new information [11-15]. This is especially true in processes whose goal is to do what is now called anticipatory analysis on the data by visualizing it in a new light. The first steps in learning came through observing the world around us, and this practice continues to this day. The scale of big data is growing to the point that it is starting to resemble the bigger systems seen in nature. A new naturalesque/synthetic imagery, called metapictorial, may thus be the best approach for depicting massive data, especially when used to facilitate the discovery of new information. These three components would make up even the most basic diagram contrasting actual with metapictorial imagery: Three contrasts are shown in Fig. (**1**): 1) pictures of reality vs synthetically manufactured "reality-styled" imagery (metapictorial); 2) derived, continuous mathematics versus applied formulation (algorithms and computing efforts); and 3) invisible physical models versus invisible relational models.

PROPOSED WORK

This publication is part of a project that provides a scientific framework for intelligent management and analysis of huge data streams for biomedical research and precision medicine.The main benefit is that hypotheses and decision options are generated automatically and can then be verified and validated using biological datasets and the knowledge of scientists. The purpose of this work is to

CHAPTER 18

Exploring the Potential of Recommender Systems for Semantic Analysis

Vineet Saxena[1,*] and **Gaurav Londhe**[2]

[1] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Tourism destinations and their online and social media information have made choosing and visiting them difficult. Tourists find tourism suggestion systems attractive, but designers must be able to deliver personalised services. This study proposes a personalised tourist recommendation system that extracts user preferences. For this, tourist social network user reviews are a rich resource of preferences. To identify visitor preferences, remarks are preprocessed, semantically grouped, and sentimentally analysed. The characteristics of attractions are extracted from all user evaluations. Finally, the proposed suggestion system semantically matches user preferences with attraction attributes to suggest the most relevant attractions. The technology also filters undesirable goods and improves recommendations based on time, location, and weather. The Python-based recommendation algorithm is tested using TripAdvisor data. The suggested system improves the f-measure.

Keywords: Sentiment analysis, recommendation systems, big data, tourism.

INTRODUCTION

Lifelong educational platforms like MOOCs use AI to make online learning intuitive. AI-powered learning analytics and

Adaptive learning toolkits help the university 4.0 assess learning needs. Open learning platforms and learner-specific customization are essential to their success.

Thus, recommender systems enhance learning and facilitate MOOC search across platforms. A recommendation system predicts a user's opinion by comparing their profile to reference data [1]. Distance learners differ in background, history,

[*] **Corresponding author Vineet Saxena:** College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; E-mail: tmmit_cool@yahoo.co.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

competence, learning style, and academic activity [2]. “Recommending materials for learning to a specific learner” is tougher [3]. It is challenging to predict students' interests given the scope and variety of MOOCs. Employing information structures such as ontologies to personalize student profiles to meet their needs and attributes might simplify the selection of learning resources [4].

Tourism websites and social media generate massive amounts of data and comments. This data and visitor reviews help travelers choose places and attractions. Tourists struggle to analyse massive data sets manually. Personalized recommendation systems for tourism have been proposed. They extract user preferences and provide more personalized suggestions. Some recommendation algorithms group users by similarity in previously visited areas and provide them with the same suggestions. Users' visits to tourism attractions are not enough, thus their reviews are crucial. Another set of recommendation algorithms analyses comments to determine consumer tastes. User evaluations are matched to information regarding attractions to find the best matches [5].

This form of tourist recommendation algorithm exploits standard terms in comments, independent of user mood. Thus, negative terms emphasized in user input may be misinterpreted as desires. Tourism sentiment analysis is often neglected. Tourism recommendation systems should incorporate the following: determine preferences by looking for ideas rather than keywords, use sentiment analysis of user comments to identify positive and negative preferences, and provide background-aware recommendations. According to the authors, no tourist recommendation system has all of the following criteria. This research presents a sentiment-based tourist recommendation system. This system extracts user preferences using word processing and sentiment analysis. The preference extraction component extends the first study by extracting and preprocessing user ratings regarding attractions. Semantic clustering and sentiment analysis extract preferences. This study proposes a personalized recommendation system based on aggregated user evaluations of attractions [6 - 10].

The algorithm also takes into account the user's location (to find nearby attractions), the time available (to see whether those attractions are open), and the weather (to make suggestions suitable for the current climate). Python is used to create the suggested system. To test the effectiveness of the suggested approach, an experiment is carried out on TripAdvisor1, a popular travel website. To that end, have collected data from 2018, including the comments and visits of 100 individuals. Precision, recall, and f-measure all point to the suggested recommendation system's superior performance in the assessment [11 - 15].

RELATED WORKS

It is stated [4] that recommender systems may be categorized in accordance with the information learned about their users, the relationships between those users, the things to recommend, and, lastly, the classifications of objects that their users might determine. Most studies on recommenders in e-learning "focus on these traditional recommendation techniques" [3, emphasis added]; hence, we were able to categorize three different RS types for the filtering processes used in distant learning: content-based, collaborative-based, and hybrid-based. However, a recommendation strategy that gives weight to semantic links between ideas, is sensitive to learner preferences, and is regularly updated, is essential in the e-learning setting. When making recommendations, knowledge-based systems take into account the user's wants and requirements as well as the information they already possess. Knowing how to "reason about the connection between a need & a possible recommendation" [5] is the ultimate objective of any knowledge-based RS. Furthermore, it encourages information exchange and reuse, which is fundamental to the concept of online education.

Knowledge-based (KB) recommenders, first and foremost, establish a harmony between user requirements and product characteristics. Therefore, it recommends products based on deductions about the user's tastes and requirements [6, 7]. The system's acquired domain knowledge is then compared with the user's needs to determine which items are most appropriate and helpful for that user [1, 8]. "Knowledge-based systems depend on the concept 'Tell me whatever fits my needs'" [8]. Therefore, "this information will occasionally include explicit operational knowledge about how specific item features meet user needs" [7], such as when a platform for e-commerce asks its consumers to pick certain characteristics of offered items. KB recommenders depend heavily on a repository of data containing user, object, and inference rules that determine which items are most likely to appeal to which users. Unlike other recommendation systems that rely on users' evaluations of products, they incorporate a semantic component into the suggestion process and thus require knowledge engineering approaches [3] for their creation.

In addition to "the most used concept in the field of computer science research: Ontology is a formal, explicit description of a shared conceptualization...", ontologies provide the semantic web technologies and serve as a knowledge-based foundation for recommendations [9]. Indeed, by leveraging semantic and interoperability concepts across numerous platforms, it enables the integration of an inference framework for individualized suggestions. Therefore, it allows for "reusing as well as sharing information across a broad spectrum of systems" [10]. Collectively, "ontology to represent understanding of the items and users in a

CHAPTER 19

Text Mining for Intelligent Information Analysis for Opportunities and Challenges

Saira Banu Atham[1,*] and **A. Alli**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: One noticeable development of the expanding internet age is text mining. Accurate exploration of its potential will enable various real-time applications to leverage text mining's unique characteristics. Conversely, social media platforms like Facebook and Twitter provide vast amounts of data that must be processed to derive actionable insights. Therefore, the purpose of this work is to explore the feasibility of using text-mining algorithms to analyze Twitter data. We want to use this knowledge to improve disaster management by applying the suggested technique to datasets pertaining to natural disasters. The three steps that make up the suggested approach are as follows. The subject of the database is first determined using a Latent Dirichlet Allocation (LDA) model. An SA is then suggested as a second step. The themes found in the sample are categorized into three feelings using this SA, which is applied to the LDA findings. We suggested using an ANN model for sentiment analysis. The third step is to apply data-mining algorithms to the themes in each sentiment using textual analysis (TA). At last, we find out how the supplied tweet feels. Anybody looking to enhance business intelligence analysis (BIA) procedures across industries may benefit from the suggested methodology's significant improvements in data text mining, including correctness, reliability, and the discovery of new insights.

Keywords: Data Mining, Text Mining, Automated Information Analysis, and Latent Dirichlet Allocation (LDA).

INTRODUCTION

Because of the large amount of data and the highly dimensional structure of text, it is difficult to summarize data succinctly; hence, robust representation of complex data is crucial for efficient natural language processing for event categorization. Machine learning technologies derive insights from textual linkages; visualization helps clarify these patterns, especially in event extraction

[*] **Corresponding author Saira Banu Atham:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail sairabanuatham@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

tasks [1]. Industry, education, and research all rely heavily on visualization systems for the insights they provide and for how they improve decision-making. These technologies enable the graphical depiction of information and complex procedures in an intelligible way, thereby improving comprehension, analysis, and data transmission [2].

The global electrical system might be affected by solar protons and other high-energy charged particles, as well as by phenomena such as Forbush reductions. In good weather places, these impacts have been examined by looking at disruptions to a chance gradient in ground-based measurements. The potential gradient nocturnal curve is examined in this work as it deviates from the average values observed in good weather situations between solar proton events as well as Forbush reductions [3]. Essential components of woodland fire prevention include risk assessment, suppression strategies, and damage evaluation. When it comes to weather prediction, topomorphology, and socio-economics, the data that is involved is dispersed and diverse. It also has a spatiotemporal analysis component with many scales. All of these things highlight how important it is to improve at making choices using the right tools and techniques [4]. Students will learn how to analyze data to identify potential policy problems that state legislatures face through this case study that uses IRS SOI data on migration. Learn how to use Alteryx for the extraction, transformation, and loading (ETL) process, as well as Tableau to create data visualizations; these are data skills that entry-level accountants will need in any area. First, to improve students' ETL abilities; second, to help them become better data visualizers; third, to help them become better critical thinkers; and fourth, to help them become better communicators both verbally and in writing. According to open-ended responses and pre- and post-learning assessments, the case accomplishes these educational objectives [5]. Adenomatous polyps can be detected and removed before they progress to colorectal cancer (CRC), thereby greatly improving patient outcomes. During a white-light colonoscopy, polyps are usually found and removed. Regrettably, there is still a substantial proportion of interim tumors that develop between CRC screening sessions. This is associated with inadequate polyp removal as well as poor screen visibility of polyps [6]. Adenomatous polyps may be detected and removed before they advance to colorectal cancer (CRC), which greatly improves patient outcomes. During white-light colonoscopies, tumours are typically identified and removed. Due to issues with polyp visibility as well as partial removal, the time frame for cancer rate among CRC screening appointments is still somewhat high [7].

RELATED WORKS

Using natural language processing (NLP), we show how visualization may help at every step of our example, which aims to find news stories that might trigger state-led mass executions [8]. This includes exploratory data analysis, neural network training evaluation, and post-inference validation.

Graphs were used to represent network traffic data, with nodes representing devices and edges representing interaction instances. Afterwards, we trained our artificial intelligence algorithms using these graph properties. Our results show that the evaluated machine learning models—including logistic regression, support vector regression, and K-means clustering—perform somewhat better when graph theory is applied to network data. These findings highlight the importance of graph-theoretic representations for improving machine learning algorithms' discriminative performance on network data [9].

Afterwards, we trained our artificial intelligence algorithms using these graph properties. Our results show that the evaluated machine learning models—including logistic regression, support vector machines, and K-means clustering—perform somewhat better when graph theory is applied to network data. These findings highlight the importance of graph-theoretic representations for improving machine learning algorithms' discriminative performance on network data [10].

A knowledge-based maintenance system (PMS) backed by RAM assessment is developed in this work. By analyzing data, the system can determine current RAM levels and forecast future trends, allowing for adjustments to the maintenance schedule. Decisions are supported at both the practical and managerial levels by monitoring the depicted information. Periodic data from the main engine on a cargo ship are used to demonstrate the suggested method. Fuel injection valves are the primary objects in the data display. The findings demonstrate that the improved PMS system enables the efficient, timely scheduling of both regular and unexpected ship repairs. The RAM study offers valuable insights to technical management and ship engine room personnel regarding the current and future possibilities of PMS systems aboard. To digitalize the process and enhance compliance with the requirements of ISM Code 10 Clause and TMSA 4A.4, the additional research focuses on developing an upkeep plan app [11].

This research aims to fill that void by analyzing scientific papers in the Scopus database through a bibliometric lens. The examined data spans the years 1980–2022, and its primary goal is to trace the origins, development, and rise of space programs through reference, events, cooperation, and grouping. The UAE

CHAPTER 20

Bridging Logic and Data to Automated Reasoning for Big Data Analysis

Krishnan Batri[1,*] and **Ajay Chakravarty**[2]

[1] *School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradaubad, Uttar Pradesh, India*

Abstract: The primary use of forecasting short-term loads is in control centres, where it is used to investigate shifting consumer load patterns and estimate the value of the load at a future time. It's a crucial piece of equipment for building a smart grid. Several dimensions of influence impact the load parameters. In this research, we present a Residual Neural Network (ResNet)/Long Short-Term Memory (LSTM) hybrid model for load forecasting, which can better take advantage of the time series properties of load data and lead to more reliable predictions. Before feeding the data into the ResNet network for the extraction of features, it is first rebuilt using numerous feature parameters. The second step in short-term load forecasting using LSTM is to feed the extractedfeaturevector into the network. Finally, the technique is compared to other models using a real example, demonstrating that the proposed combination method has better prediction accuracy and confirming the practicality and superiority of input featureextractionparameters. Additionally, this study conducts studies in weather prediction based on a variety of elements and characteristics

Keywords: Big Data Analysis, Automated Reasoning, LSTM, ResNet, smart grid.

INTRODUCTION

Big data analytics seeks to aid organizations in making better decisions by unearthing previously unseen trends, patterns, and insights within large datasets. Thanks to the rapid distribution of this data, businesses can maintain their competitive advantage more quickly and with greater agility.

Tools and systems such as business intelligence (BI) technology enable businesses to combine unstructured and structured data from many sources. In

* **Corresponding author Krishnan Batri:** School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail krishnan.batri@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

order to get insight into the inner workings and performance of a company, users (usually workers) submit queries to these tools. The four methods of data analysis are used in big data analytics to uncover useful information and arrive at workable conclusions. The practice of predicting the weather dates back to the eighteenth century [1, 2]. Weather forecasting is the study of atmospheric data such as temperature, irradiance, air pressure, wind speed and direction, humidity, and precipitation. Large amounts of information are needed for weather forecasting. Additionally, these records lack any kind of structure. Consequently, using meteorological data for weather forecasting is difficult due to the large number of variables involved. The often shifting weather conditions determine the range of these variables. It is important to take into account the unique properties of weather when proposing an algorithm for weather prediction, including its continuity, data intensity, multidimensionality, and chaotic behaviour [3, 4]. Originally performed by hand, weather forecasting is now a computer-based process that necessitates high-tech equipment [5, 6]. Many factors outside our control might influence our predictions. Season, location, quantity input information, weather classifications, lead time, and validity time are all factors that may affect the forecast accuracy [7, 8]. The term "big data" refers to large quantities of unstructured data in digital form [9]. Traditional methods of data management are neither practical nor easy for processing enormous data sets [10, 11]. To get meaningful insights from enormous data, we need an outstanding performance infrastructure and an applicable big data mining approach [9]. Exploring enormous data sets to uncover hidden patterns, new connections, and other useful information to aid decision-making is known as "big data analytics" [12]. The use of big data in weather forecasting has the potential to improve forecast quality [13]. Since precise predictions are essential, we may use big data analysis to improve our weather forecasts. Human forecasters were crucial to the process in the past. But in the modern age of information, we utilise data and technology to make predictions [14]. Atmospheric datasets include, but are not limited to, precipitation, humidity, air pressure, radiation, solar intensity, and data collection. We also need a large volume of data collected from many different sources (big data). To deal with this mountain of information, sophisticated hardware and software are needed [15, 16].

This investigation expands upon presentations made at the 9th International Symposium on Computation in the Cloud, Data Science, & Engineering (Confluence) [1], which laid the groundwork by applying Machine Learning Algorithms to Big Data Analytics. Predicting the weather is one of the most valuable applications of technology, allowing people to better prepare for everything from air travel and satellite launching to agricultural harvests and natural calamities. Four sorts of weather forecasts may be created, depending on how long they are expected to last: As Soon As Possible, a) Predictions are long-

range forecasts often issued many days in advance. Long-term predictions go out months or even years, whereas short-term forecasts span only a few days or weeks [17 - 20]. In the beginning, forecasters compiled data from several weather stations and used statistical analysis to predict the weather. As ML methods gain in popularity and scope of application, faster processors become more widely available, and improved tools and methodology are introduced, researchers are starting to experiment with machine learning (ML) techniques in the field of meteorology, but it has stalled for a long time due to the descent of gradients problem and computer slowness. Technological advancements such as parallel processing, high-speed computer clusters, and graphics processors have enabled significant performance gains when applied to NNs. Recurrent neural networks (RNNs) are vulnerable to the Vanishing Gradient Problem due to their feedback connections. Gated RNNs, such as LSTM (Long Short-Term Memory neural networks, a term initially suggested by Schmidhuber and Hochreiter in 1997), produce gradients that neither grow nor shrink over time, therefore mitigating the aforementioned issue. They consist of one memory unit, a trio of gates (input, output, & forget), and a function that decides which data is saved, sent, and forgotten. Gradient-flowing paths may be built using forget-gate-controlled conditional weight self-loops in a long short-term memory network [2]. Its proficiency in learning causal relationships has made it a favourite in the text-sequencing industry [21 - 23].

RELATED WORKS

In this section, we review previous work on applying NN to the problem of weather prediction. Studies of the various Data Mining algorithms currently used for weather forecasting have identified potential pitfalls in applying more traditional data mining techniques [4]. Using a five-layer Deep NN and a four-layer Stacked Auto-Encoding mechanism to learn features from raw data, with SVR for prediction, the authors [5] examined 30 years of Hong Kong meteorological data. Several measures have been used in the review process, including Normalized Mean Squared Error (NMSE), R-squared, and Directional Symmetry (DS). Following the steps outlined in a study [6], we used a three-layer BPNN model to detect and analyze non-linear correlations in the normalized weather data that we obtained from www.weatherunderground.com. To develop a practical inference procedure, the authors studied IGRA data from 60 locations. The highest-quality validation results were achieved with two-layer RBMs containing either 50 or 150 hidden neurons. Using CNNs for the first time, the authors of [8] were able to better predict Australia's monthly rainfall than both the Bureau of Meteorology's model and an MLP on days with above-average annual rainfall. Researchers have also proposed using hybrid genetic algorithms (CFPS-HGA) based on climate factors and neural network parameters of choice to fine-

CHAPTER 21

Investigating the Benefits of Text Mining for Information Analysis

S. Senthilkumar[1,*] and **M. Veera Nagaiah**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: Thanks to advancements in educational technology, new approaches to controlling and understanding the teaching-learning processes have arisen, such as instructional analytics (TA). Instructional analytics is a new yet exciting approach to learning that has the potential to dramatically improve teaching quality and student outcomes by helping educational institutions better use the massive volumes of information generated from student data. This study offered a framework for educational process data mining with artificial intelligence (EPDM+ML) that uses student evaluation of teaching (SET) data to contextually evaluate teachers' performances and provide recommendations. The EPDM+ML model was created and implemented using a mix of text mining and artificial intelligence (AI) technology. It is based on descriptive decision theory, which examines why learners tend to make choices through quantitative analysis of textual data. To this end, the study considers instructors' gender while analyzing the pedagogical elements that influence students' suggestions for their professors and the effect of students' subjective experiences on their assessments of instructors. Predicting a student's potential recommendation for a teacher is conceivable with SET data such as students' gender, average mood, and emotional valence. In actuality, we mined text for the different sentiments and emotions (comment intensities) that students conveyed in the SET. Then, we used a Kruskal-Wallis test to determine which elements were the most important by analyzing the covariance using the quantitative data (average emotions and emotional valence). Furthermore, we examined the gender conceptions to see whether there were any differences in the suggestions students made for instructors. Taking into account both the mood and the sentiments expressed in the comments, we find that most (n=85,378) were positive in tone, while a sizeable portion was neutral. Female students were more likely than male students to propose that teachers take into account students' emotions in statements with an emotional valence (11.8%) and opinions (23.6% positive and negative) (p=.000). Men, on the other hand, exhibit borderline behavior in terms of sentiment (p=.077) and emotions (p=.056). The EPDM+ML model was also validated using the k-fold cross-validation method, with 63.6% of the optimal k-values observed, and it proved to be an efficient and effective predictor of students' recommendation scores for teachers, with high and acceptable values of precision (1.00), recall (1.00),

* **Corresponding author S. Senthilkumar:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: senthilkumars@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

specifcity (1.00), accuracy (1.00), F1-score (1.00), and zero error-rate (0.00). Theoretically, we find that the suggested approach (EPDM+ML) is helpful for effective examination of SET and its consequences in the field of education. There is a growing need to improve teaching and learning processes and students' learning experiences in an ever-evolving educational ecosystem. However, this can be done by identifying the most important factors that students use to evaluate and recommend teachers.

Keywords: Text mining, information analysis, teaching analytics, students' evaluation.

INTRODUCTION

Methods based on artificial intelligence (AI) have recently emerged as potent instruments for revolutionizing healthcare. Despite the impressive effectiveness of neural network classifiers (MLCs) in image-based diagnosis, analyzing vast amounts of EHR data is still a challenge [1]. Algorithmic law is a relatively new field of law that focuses on the automation of legal reasoning. New approaches based on artificial intelligence (AI) have the potential to revolutionize healthcare. Analysis of varied as well as enormous EHR data is still difficult, even if neural network classifiers have been effective in image-based diagnosis [2]. The study and documentation of computational legal argumentation is a relatively new area of study within the larger field of law.

An important aspect of creating Machine Learning algorithms is the input pipeline, which typically takes in information and processes it in some way. Thoughts on synchronicity, heterogeneity in fine-grained profiling data, and parallelism make effective input pipeline implementation difficult [3]. With applications such as reasoning monitors, decision summarizers, contention recommendations, and semantic viewers, mining trends in reasoning from evidence-intensive legal decisions can enhance the efficiency of legal services and broaden public access to justice. Automating the classification of phrases stating whether the prerequisites of relevant law have been met in a given legal case is a significant task for these use cases [4]. Text analysis, particularly in the malware realm, is a difficult task. Even when using a natural language processing technique to extract malware-related entities from unstructured information, such as text, there are limitations due to the lack of a specific Named Entity recognizer. We must automate the extraction of data from text, including Ransomware entities. This data could then be used for knowledge reasoning purposes, such as profiling Ransomware behavior using publicly accessible online information. The data analysis and extraction procedures are not without their difficulties when dealing with informal, unstructured content, such as that found in online forums [5]. Creating systems of control for autonomous manufacturing operations that can handle data integration, semantic interoperability, monitoring, and decision-

making is a major challenge for the development of Industry 4.0 applications [6]. Automatic language analysis has recently shown promise as a tool for the diagnosis of moderate cognitive impairment (MCI). The majority of research on MCI prediction using a combination of linguistic processing and artificial intelligence has examined only one language task [7].

RELATED WORKS

For technological forensic systems that involve collaborative criminal analysis and prediction, we have developed an Automatic Learning Framework ontology. Execution of Automated Machine Learning (AutoML) is also performed based on the minimal viable ontology. This implementation will be evaluated both quantitatively and qualitatively to assess its effectiveness in helping investigative agencies represent, reason, and extract useful information from the diverse and extensive datasets they collect. By comparing our proposed generic Smart Forensic Framework with existing systems using qualitative and quantitative metrics, we can determine its performance in Digital Forensics applications. To encourage forensic intelligence organizations to utilize the features and capabilities of the AutoML Smart Forensic Framework, we will present insights along with performance indicators gathered in our study [8].

Here, we examine how using one's own memories can enhance facial behavior analysis. Automated facial recognition and two forms of contextual information are combined in a set of multisensory neural network trials we run to make emotional predictions about video viewers: Start counting using Arabic symbols. List the audiovisual content of a video and the reported free-text description of memories that are generated by it. Our findings show that, in addition to face analysis, both types of background offer a model for understanding the variability in viewers' emotional reactions [9].

Using a methodology known as computational grounded theory, we rigorously assessed students' written claims on the likelihood of opposing chemical reactions in this research. Our goal when applying an unsupervised clustering approach was to thoroughly assess the logic patterns as well as the granularity levels used by students in their written reports by evaluating their argumentation patterns. We combined data-driven clustering with a theory-driven architecture to automate the analysis of observed argument patterns, and a comprehensive 20-category rubric was developed based on this research. The use of state-of-the-art deep learning methods to assess the complexity of students' arguments was supported by almost perfect agreement between machine and human scores and by easily interpretable findings from pre-trained big language models coupled with deep neural

CHAPTER 22

Data-driven Strategies for Resource Optimization Using Data Mining

Bhawna Wadhwa[1,*]

[1] *Department of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India*

Abstract: Data mining has become an increasingly significant method for conducting data analysis as a result of the rapid increase in databases used by a large number of contemporary businesses. The community of people who study operations research has made major contributions to this discipline, particularly by formulating and solving a large number of data mining problems as optimization problems. Additionally, data mining techniques may be used to address a number of applications in operations research. The purpose of this study is to offer an overview of the relationship between operations research and data mining. The basic objectives of the study are to highlight the spectrum of interactions between the two areas, present specific instances of significant research effort, and provide extensive references to additional significant work in the area. The purpose of this study is to examine not only the many optimization techniques that may be used for data mining, but also the process of data mining itself, as well as the ways in which operations research techniques can be utilized at almost every stage of this process. The report also identifies many potentially fruitful avenues for further investigation throughout its body. In the last part of the study, many applications connected to the administration of electronic services, including customer relationship management and customization, are investigated.

Keywords: Data mining, resource optimization, goods, services, optimization, customization.

INTRODUCTION

Human resource management is the theoretical basis for and practical application of the HRM system. Data gathering and analysis are used to compile all HR-related material, including modules for hiring, approving employees, calculating salaries, and more [1].

Human resource managers will find this helpful in their day-to-day operations. The adoption of the human resources system is driven primarily by businesses'

[*] **Corresponding author Bhawna Wadhwa:** Department of Computer Science and Engineering, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India; E-mail: bhawna.wadhwa@niet.co.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

desire to achieve the greatest possible financial returns from their investment in human capital. Human capital, a term coined in response to the rise of the information economy, is increasingly seen as on par with, or even more crucial than, physical assets like buildings and machinery. Furthermore, individuals are the repository of information [1]. The full potential of human resources can only be realised through careful human resource management, which is essential for making the most of one's expertise.

From its inception to the present, the HRM system has gone through three distinct phases:

1. The origins of the modern human resource management system may be traced back to the late 1960s, coinciding closely with the emergence of computing itself [2]. Nonetheless, the HRM system in its early stages can only realize extremely rudimentary computing capabilities owing to the embryonic development of computer technology.
2. Since the advent of the personal computer in the 1970s, significant progress has been achieved in the areas of database and application [3]; this includes the creation of effective tools for human resource management. Third-generation HRM is known as "Information Human Resource (Electronic Human Resource, E-HR)," in contrast to the second-generation system, which was developed primarily by computer scientists but failed to achieve both technical and practical improvement. It makes extensive use of computer and network technologies and takes into account the real requirements of human resource managers [4]. Unfortunately, the majority of HRM tools that companies use today are still in the early stages of data collection, storage, and querying. What's more, the data analysis operations that do exist can only extract superficial data [5], which makes it impossible to delve deeply into dynamic data, such as the company's changing development capabilities or workforce quality. It is challenging to provide scientific support for sustained business growth if the HRM system's central data value is not fully exploited [6].

As Internet technology has become more widely used, the amount of data available online has increased massively [7]. People are increasingly interested in learning how to use technology to unlock the hidden value in large data sets. Under these conditions, data mining technology has emerged at a massive scale. Data mining has a significant effect in numerous areas, including sports, transportation, and scientific research, in recent years [8]. Organizations need to leverage HR's resources to grow sustainably. The concept of human capital [9] proposes that a company's personnel are its most significant asset because they enable it to endure and thrive in the face of market competition. Accurate and

comprehensive HR planning has to be the number one objective for any enterprise that strives to achieve sustained, long-term success as well as formulate a more permanent growth path in light of the present situation of increasing human resource expenditures [2]. The only way to bring down employee costs is to adopt a more effective way of distributing spending [3]. Suppose HR cost planning is thorough and accurate in advance. The best workers in a certain job should not be used as a benchmark for staffing [4], even if their labor productivity is much better than that of average or mediocre workers in the same position. We can only identify the most appropriate individuals for the role by carefully broadcasting the position and measuring the talents, knowledge, and level of various workers, as well as the demands of the position [5, 6].

It is important to remember, however, that the HR portion is not a straightforward selection procedure, but rather one that exclusively uses scientific methodologies. HR computing intelligence enables businesses to access all HR-related materials [10]. To make appropriate management decisions, businesses may rely on data and information that are both applicable in practice and easy to understand in terms of the enterprise's growth. Human resource management uses data mining technologies, and the results may be broken down into three broad classes of information: The first kind is time-sensitive information. Information of this sort is primarily reflected in people's rosters, both at the individual and organizational levels [14]. The former includes details such as headcount, job titles, years of experience, educational attainment, certifications, and even employees' family information. [11].

RELATED WORKS

Most people mean by “computational intelligence” the laborious task of discovering fresh, potentially valuable, and legitimate connections in data [12].

Data mining has been used in a variety of contexts and disciplines during the last few years. One such context is business management, which includes well-established specializations like customer management, manufacturing management, and finance management.

Human resource management (HRM) seems to have recently joined these corporate application fields [15 - 20]. More and more studies have been conducted over the past few decades to pave the way for broader deployments of HRM data mining. Human resource management (HRM) includes a wide range of tasks, from recruiting and hiring new employees to tracking turnover rates, planning training and development programs, and assessing performance. A cursory literature review reveals a rapidly expanding subfield of data mining study tailored to HR needs and, thus, of great practical use to HR practices.

CHAPTER 23

Investigating the Role of Data Mining in Enhancing Business Performance

Prabha Nair[1,*]

[1] *Department of IT and M.Tech Integrated, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India*

Abstract: While data mining finds extensive usage in the scientific and technical communities, it also finds beneficial uses in the financial and marketing spheres, where it helps address specific problems. By analyzing the actual situation, a data mining-based decision support system improves the organization's performance. Due to factors such as competition, costs, tax pressures, and other factors, every firm experiences economic volatility. A more competitive environment is dragged into the organization *via* privatization, globalization, and liberalization. To stay ahead of the competition and accomplish your goals, you need a well-planned and implemented marketing strategy. By using an effective data mining strategy, marketing decision support systems help organizations reduce the responsibilities associated with analysis and strategic planning. This study offered a data mining-based decision support system for estimating a company's marketing plans by combining decision trees and artificial neural networks: data mining, decision trees, artificial neural networks, and marketing decision support systems.

Keywords: Data mining, business performance, knowledge-driven decision support system, data-driven decision support system, communication-driven decision support system.

INTRODUCTION

The safety, efficiency, and integrity of building initiatives may be guaranteed by taking measures to prevent work-related crimes early on. With the use of sophisticated big data analytics and predictive models, crime prevention efforts can be taken to the next level. Construction projects begin with the following steps: bidding, acquiring land, demolishing, obtaining permission, and procurement. During the first phases of building projects, the most common types of work-related crimes are bribery, corruption, and dereliction of duty [1]. The

[*] **Corresponding author Prabha Nair:** Department of IT and M.Tech Integrated, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India; E-mail: prabha.snair@niet.co.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

need for solutions to analyze large amounts of data is growing in tandem with the importance of information and communication technology (ICT) in facilitating and bolstering smart cities. A number of solutions based on AI, data mining, ML, and statistical analysis have been successfully implemented in areas such as climate research, energy management, transportation, air quality, and climate change analysis [2].

One of the most significant areas of application for big data analytics is predictive analysis, which uses a combination of complex analytical functions, primarily statistical analysis, along with other methods to forecast future events. The primary areas of use for large-scale data analytics technologies include corporate intelligence, public services, government decision-making, and related fields. To address the statistical analysis issues associated with time-series big data, one can use factor analysis, a method from classical statistics. However, factor analysis provides only information about the degree and variation of the data distribution around the median, and it does not account for how data object characteristics change over time [3]. Because of the exponential growth of AI and big data, which are grounded in the experiences of college students and made possible by cutting-edge technology, university curricula have also grown in many other ways. The design of electrical technology is based on a rational and scientific basis. Resulting from the optimization of derivative tools used in university courses as well as the use of digital information as a means of transport to enhance performance in other domains, including education and sports [4]. A Brief Overview: With the expected acceleration of global population growth in the coming decades, there will likely be a corresponding surge in demand for eggs and chicken meat. Pollution, land erosion, competition for limited resources, animal welfare, restrictions on growth promoters, antimicrobial agents, increasing dangers from animal infectious illnesses, and zoonoses are some of the many challenges that this expansion hides, despite the fact that it clearly represents an outstanding chance for the sector. Optimizing and increasing efficiency are the primary ways to boost chicken output. There is a once-in-a-lifetime chance to create tools that optimize farm profitability, reduce socio-environmental impacts, and increase human and animal health and welfare [5]. This opportunity is made possible by the growing capacity to generate massive amounts of data, also known as “big data,” as well as by the availability of resources to organize, distribute, integrate, and interpret this data using effortless as well as flexible algorithms. If a company wants to improve its decision-making or reduce risks, it needs a data-driven strategy plan that uses analytics. In this case, company data is vital. This study presents a statistically informed strategy for LendingClub, a financial services provider, that includes investigating the feasibility of integrating a big data platform with sophisticated feature selection capabilities [6]. There are new possibilities and challenges for businesses in this age of unparalleled information

accessibility. Consumer habits and market trends are both covered by the ever-increasing dataset. To make informed economic decisions, nevertheless, the issue of how to sift through such a mountain of data becomes paramount [7].

RELATED WORKS

In this post, we looked at how construction projects can leverage big data to improve planning and execution, making them safer for employees. The article began by examining the trends, patterns, and kinds of work-related offences that occur early in construction projects. Afterwards, we learned that big data technologies may help prevent such crimes. As a last step, we analyzed survey data and proposed using big data technologies to curb construction-related crimes at the preliminary phases of a project. Ten Chinese provinces rank first to 10th in the frequency of crimes committed on the job: Zhejiang, Beijing, China, Shandong, Guangzhou, Henan, Jiangsu, Hebei, Hubei, Fujian, and Liaoning. This study selected a number of construction projects across various locations using basic random sampling and quantitative research methods. The intended subjects of this study were workers in the construction sector, construction firms, key stakeholders, and associated professionals. To gather data, participants were asked to fill out an online questionnaire. The gathered data were examined using descriptive analysis methods in SPSS. This study's findings highlighted the value of big data technologies for preventing construction-related crimes at the project's first phases [8].

A comprehensive literature analysis of smart city big data analytics is presented in this work. For this research, we used a structured data mining process that included searching many sources for relevant materials using precise keywords. We have also reported the findings of a technical and thematic analysis of the selected literature, which included identifying different data mining and machine learning approaches. We also provide a categorization model that examines four facets of this field's research [9]. This includes models for data and computation, considerations for privacy and security, and the most important factors propelling the smart city industry. In addition, we highlight potential areas for further study and provide a gap analysis. We conducted a thematic analysis and identified the following themes: smart city governance, energy, transportation, economics, and environment. We lay out the most pressing issues in these areas, the most important data analytics studies that have attempted to address these problems, and the most promising avenues for further study [10].

Using an object-oriented programming approach, this research introduced a novel model that applies grey factor analysis to the large data analysis layer. At last, big data analytics uses this approach. Analyzing and comparing the findings yields a

CHAPTER 24

Learning Features and Preferences with Matrix Factorization Models for Recommendation Systems

C.R. Manjunath[1,*] and **Nitin Gaur**[2]

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: Over the last several years, research into recommender systems (RS) has grown to include a wide variety of artificial intelligence (AI) techniques, from more basic ones like Matrix Factorization (MF) to more complex ones such as Deep Neural Networks (DNN). Since they only examine a linear combination of consumer and item vectors, traditional CF recommendation algorithms like MF have little room for improvement in learning. Neuronal collaborative filtering (NCF) is a hybrid of deep neural networks (DNNs) and collaborative filtering (CF), enabling it to learn non-linear connections. But CF methods still have issues with cold beginnings and sparse data. This research presents a new hybrid RS, Neural Matrix Factorizing++ (NeuMF++), that can handle cold starts, boost recommendation accuracy, and address data sparsity. To improve NeuMF, we introduce NeuMF++, which uses Stacked Denoising Autoencoders (SDAEs) to better represent latent features. Merging GMF++ and MLP++ might yield NeuMF++. Mixed with Multilayer Perceptrons (MLP) and Generalized Matrix Factorization (GMF), NeuMF is a robust NCF architecture. By combining the linearity of GMF with the nonlinearity of MLP, NeuMF achieves state-of-the-art results. Simultaneously, the original GMF as well as MLP were enhanced with GMF++ and MLP++, respectively, by successfully including latent representations. Obtaining a latent representation from the SDAEs' hidden space significantly enhances NeuMF++'s learning power. This representation allows it to learn item and user attributes correctly. If feature extractions are shared across GMF++ and MLP++, however, NeuMF++'s performance could take a hit. As a result, its speed and flexibility are greatly enhanced when GMF++ and MLP++ are allowed to learn features separately. Using a real-world dataset, NeuMF++ achieved an experimental root-mean-square error of 0.8681, demonstrating its excellent performance. Additional data, such as text or photos, can be added to NeuMF++ in later development. NeuMF++ allows incorporating several neural network building blocks to create a more robust recommendation model.

[*] **Corresponding author C.R. Manjunath:** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail cr.manjunath@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

Keywords: Recommendation Systems, Matrix Factorization, Local Features, machine learning (ML), Generalized Matrix Factorization (GMF).

INTRODUCTION

In the last decade, machine learning (ML) has become increasingly popular due to advances in methodology enabled by the widespread availability of powerful computers and massive amounts of data. Recommender systems are one area where ML has been effectively used.

Recommender systems powered by machine learning are ubiquitous, influencing our choices across many domains, from entertainment to e-commerce. Companies increase revenue and customer loyalty while giving consumers a better, more individualized experience using recommender systems [1]. Despite the upsides, this achievement has spawned serious moral problems [2]. One of these is the rising number of privacy concerns arising from the proliferation of personal data collection for highly tailored suggestions. To train centralized models on the processors (servers or data processing organizations), conventional recommender ML approaches must collect sensitive user information, which might compromise their privacy. As consumers become more aware of the need to protect their personal information, they have higher expectations of the privacy protections offered by the services they use. For instance, while gathering and handling user data, systems of recommendations must comply with the GDPR (General Data Protection Regulation) of the European Union [3].

The information available to us today has grown at an exponential rate, making it more difficult than ever to sift through all of the available data and discover what we need. It's becoming increasingly esoteric to acquire the necessary info online. Therefore, word-of-mouth is quite important [4]. A recommender system is a type of knowledge filtering system that determines a user's preferences based on their past rating behavior [5] and the information at hand. News, movies, text mining, novels, *etc*., are only a few of the many applications of recommendation systems. Unfortunately, not every recommendation engine can deal with every conceivable scenario. There are two main approaches to designing a recommender system. In the first group, content-based filtering methods are used to define the characteristics of each individual or item. Think of a movie profile that details the film's genre, cast and crew, box-office performance, *etc*. Similarly, users' profiles may include numerical data or responses to an appropriate survey. The resulting profiles can be used in software to pair customers with items that meet their needs. A content-based recommendation system has the drawback of requiring external information, which can be difficult to obtain in practice. One alternative approach, and one of the most classic and eye-catching recommendation

algorithms [6], is collaborative filtering (CF), which plays a significant role in producing tailored suggestions [7]. For the purpose of discovering novel user-item correlations, CF analyzes user-dependencies and product-relationships [8]. Based on user history alone—which might include item ratings or transactions—very few CF systems can deduce non-existent links between individuals and products. This is because very few CF systems can identify sets of items with comparable ratings or users with complementary ratings and buying habits. CF is domain-free and superior to the content-based approach [9] because it can handle confusing and challenging information types. However, the sparse user-rating matrix problem is a limitation of the CF approach, leading to subpar suggestion precision. When this happens, it is normal practice to use an average rating for that individual or item to fill in the blanks. The level of imprecision can only be reduced slightly with this method. The reliability of the final suggestions is significantly affected by the accuracy with which missing values are imputed [10]. The benefits of matrix factorization hidden factor models, such as preserving accuracy when scaling data, minimizing estimation cost, and mitigating issues arising from high sparsity levels, have recently brought them under investigation [11]. Among collaborative filtering methods, this one is the most popular for revealing the hidden factors that impact a user's preferences. The memory-efficient, more specialized Matrix factorization-based technology outperforms the similarity-based recommendation technique, which considers only similarities between individuals and products to provide recommendations [12]. Because of its usefulness in collaborative filtering, matrix factorization lends itself well to the use of SGD and ALS as learning algorithms. Matrix factorization is a technique developed and popularized by the winners of the Netflix award, in which a huge matrix is broken down into smaller matrices. This strategy improved upon the efficiency of previous, primarily neighborhood-based, approaches of creating recommendations. Finding hidden components and decreasing the dimensions are the goals of this type of decomposition [13]. Modern matrix factorization techniques are built with exact feedback data, making data analysis straightforward [14]. However, service providers must build feedback sites for consumer evaluations and ratings, a challenging, time-consuming process due to user engagement. Inferring user preferences from a larger volume of implicit data enables recommenders to provide indirect suggestions based on user behavior [15]. Some examples of implicit feedback data include a user's online purchase history, page views, search terms, and even their mouse movements.

RELATED WORKS

Many academics have devoted significant time and energy to recommendation systems in recent years. Hongmei H. Li *et al*. present a more effective, optimal recommendation framework based on an all-weighted strategy. Several

CHAPTER 25

Leveraging Big Data to Enhance the Analysis and Use of Information

S. Gadug[1,*] and **Avadhesh Kumar**[2]

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: The goal of this research is to examine how e-commerce and web-based businesses might benefit from the use of Big Data Analytics in managerial decision-making. The data used in this analysis comes from a single e-commerce website's database. User interactions with the website, such as page views, product additions, and online purchases, would be recorded. Association Rule Mining Algorithm (APRIORI), K-Means Clustering, and Pearson's Correlation Coefficient are only a few of the algorithms that will be utilized to evaluate the dataset. The information will be analyzed and used to provide insights into users' interactions with the website, enabling the identification of patterns that may inform future actions. For instance, which product receives the most attention and sales, how many pages are viewed before a purchase is made, what percentage of customers buy the product again, and so on. This study would also determine whether the company's current Big Data implementation can be enhanced and whether doing so would be a clever use of resources.

Keywords: Big Data, Information Analysis, E-Commerce, ARM, Association Rule Mining Algorithm (APRIORI).

INTRODUCTION

Obtaining private space in the modern world is difficult. As more and more of our lives move online, so too does the risk of sensitive data leakage.

Data collection includes personal information from users of online services, including buying, trading, banking, and entertainment. While not all of these services are designed to abuse users' private information, some do. For instance, one method of Internet marketing is to use users' browser histories to provide

* **Corresponding author S. Gadug:** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: s.gadug@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

specific suggestions based on their interests and preferences. The diagnostic process can benefit from the study of sensitive data, such as medical data. Such application situations have wide-ranging applicability. Data transformation technologies are used to preserve personal information while still enabling valuable analytics. Given the inherent tension between protecting individual privacy and doing thorough research, this is easier said than done. The sheer volume of big data makes this a very difficult task from several angles. The ability to strike a compromise between privacy, usefulness, and performance is essential for any privacy-protecting analytic system. Several strategies for protecting privacy in big data analytics have been presented. Specifically, they focus on the privacy algorithms k-anonymity [1, 2], and l-diversity [3]. Methods based on k-anonymity use data transformation techniques to render personal information useless. In k-anonymity-based methods, the choice of k is crucial, as it determines the level of anonymity. Better privacy can be achieved by optimizing k, but at the expense of some of the data's usefulness.

Another restriction of optimization methods is the performance overhead they impose. Anonymization allows a middle-ground approach, where one goal is prioritized over the others, and the application's context explains this. The mathematical groundwork offered by the privacy algorithm known as differential privacy [3] has lately attracted more attention. Differential privacy is the central method, and lenient variants have been proposed [4]. To keep their users' data secure, tech giants like Google and Apple have implemented algorithms that show only a sanitized version of their data. The fundamental idea behind the algorithm is to safeguard users' personal information, whether they actively participate in the data analysis or not. Even while this method has been endorsed for privacy protection [5], useful analytical findings require a learning mechanism at least as powerful.

Companies that can efficiently collect data, analyze it for insights, and turn that insight into strategic action are likely to succeed in the market. Not only information about the market as a whole or how tough the competition is, but also how consumers are responding to our offerings. Understanding how people feel, speak about, and respond to our product might help us decide what steps to take next. The company requires genuine, unprompted, and honest customer feedback to achieve this. When people are asked questions or given forms to fill out by hand, the answers they provide may not accurately reflect their true thoughts and feelings, as they may be motivated to give an answer that benefits them personally. The result would be the collection of useless information, which might lead to disastrous consequences for businesses. Therefore, the “unconscious” response is what we seek, and Big Data can help us get there [6].

RELATED WORKS

The 5Vs of big data are volume, velocity, variety, veracity, and variability; they will be discussed in more detail below [5]. One, the quantity of storage space needed is directly proportional to its volume. Second, velocity is a measure of how quickly and accurately data is transmitted and processed in real time. Data comes in many formats, including text, images, audio, and video, and this diversity is a reflection of the data's structure or lack thereof. Fourth, data trustworthiness is a function of data veracity. Five, variation shows how varied the data points are from one another. One of the most crucial factors in developing a website is the user's behavior. Abraham Maslow, a humanist psychologist, proposed a "hierarchy of needs" as a theory of what people need to achieve their "pinnacle" (the point at which they decide to engage) in terms of behavior. The CEO of WebFx, William Craig, lists the many factors that determine a person's ability to reach "self-actualization" and explains their relative weight. 1) Ease of access. Everyone has access to the website and may utilize it. 2) Dependability: the site's reliability and consistency are both high; 3) Practicality: the interface is intuitive and straightforward; 4) Dependability: there is no downtime in the website's availability; 5) Usefulness: site visitors can access useful information, resources, and services, 6) Capacity to adapt: the website changes to suit the preferences of its visitors. The model suggests that a website's primary purpose should be to drive sales; thus, it's crucial that the website cater to and prioritize its customers' needs [6].

The main goal of analytics is to help businesses make better decisions and fix issues. Thus, analytics is a repository of information that comprises mathematical and statistical tools, machine learning algorithms, data management procedures such as extraction, transformation, and loading (ETL), and computer technologies such as Hadoop, which together extract useful, actionable information from data. When discussing analytics and business intelligence solutions that rely heavily on statistical analysis and data mining, the term "analytics" is often used. Data warehouses (DW), online analytical processing (OLAP), ETL, and relational database management systems (DBMS) constitute the backbone of most of these approaches. "Data analytics" refers to the process of collecting and analyzing data to draw conclusions. With the use of data analytics, businesses may get valuable insights that can improve decision-making and lessen the likelihood of fraud, mistakes, abuse, and danger [7]. The list also includes data visualization, AI, natural language processing, analytical database capabilities such as MapReduce, in-memory database analysis, and columnar data warehouses. All of these case studies illustrate the use of sophisticated analytics [8], which are applications of specialized analytics methods and associated tools. It is possible to perform big data analytics on incomplete or inconsistent data. Actually, they make no

CHAPTER 26

Leveraging Text Mining Techniques for Automated Information Analysis

Vetrimani Elangovan[1,*] and **F. Rosita Kamala**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: In recent years, the quantity of textual data carried *via* online digital files, e-mails, databases, and other formats has increased at an exponential rate due to the fast growth of Internet technology. Ontology and other contemporary models of knowledge representation rely heavily on attributes, which are formal descriptions of any real-world thing. Attributes may be either qualitative or quantitative. There is a wealth of literature attesting to the extensive research on mining non-qualitative attributes (such as part-of relations) from text and the web, but very little on mining qualitative attributes (such as size, color, taste, *etc*.). It became increasingly difficult for traditional intelligence data-processing technologies to meet the demands of these tasks. The accuracy of segmenting words in intelligence texts was significantly enhanced by our suggested intelligence dictionary-based word segmentation technique and the construction of an intelligence-specific dictionary.

Keywords: Text mining techniques, automated information analysis, ontology, word segmentation strategy, part-of relation.

INTRODUCTION

Cybercriminals target online social networks by creating fake identities and then using them to steal money, influence politics, or further their own agendas. OSNs often use machine learning (ML) abuse classifiers as a reaction. A practical and effective ML-based defense, however, requires the following: gathering sufficient ground-truth data with labels for model training; developing a scalable system to handle all current accounts on an OSN, which could number in the billions; and meticulously engineering features that are resistant to adversarial manipulation [1]. Finding social interactions is crucial for studying animal behavior, as it allows researchers to address several questions, such as how social hierarchies affect

* **Corresponding author Vetrimani Elangovan:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: vetrimani.elangovan@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

behavior and what causes agonistic behavioral disorders. The bulk of earlier research, however, relies on labor- and resource-intensive manual counting of social interaction kinds and numbers obtained from direct observation [2].

In practical contexts like computational finance, social networks, and recommender systems, graphs are an effective tool for depicting and evaluating intricate interactions. In complex systems, graph-based thinking is crucial for uncovering hidden trends and patterns and for forming conclusions about the interactions between elements. An understudied subject is reasoning about graphs using large language models (LLMs), despite the remarkable advances in automated reasoning from natural text [3]. In order to foster a more welcoming and courteous online environment, it is imperative to create automated systems that can recognize instances of sexism as well as other disrespectful and hateful actions. This is especially important given the growing importance of social media platforms. Still, given the variety of hatred categories and the author's goals to consider, these tasks are very difficult, particularly in the context of learning within a dispute regime [4]. To distribute spam and harmful content, violate users' privacy, or manipulate data to influence the stock market or electoral outcomes, social bots use social networks to create content and engage with network users, attempting to mimic or change user behavior. This can lead to numerous losses [5]. An automated method for identifying and making sense of the feelings expressed in text is known as sentiment analysis (SA). There has been a meteoric rise in SA's profile in the NLP community within the last decade. Social media analytics (SA) has grown in importance as a tool for businesses to gauge consumer sentiment and inform advertising campaigns, driven by the proliferation of online platforms. Furthermore, SA is used by academics to assess public opinion on many subjects [6]. Human values are qualities people hold in high regard, such as honesty, social responsibility, justice, privacy, and related concepts. There could be a wide range of negative consequences for people and society if software systems, particularly mobile app developers, disregard or violate such ideals. This mixed-methods research focuses on honesty as a crucial human value in software engineering, but previous studies have explored many other human values [7].

RELATED WORKS

The research referenced as [8] used a Real-Time Kinematic Global Navigation Satellite System (RTK-GNSS) localization device (RTK rover) affixed to the backs of sheep. The module was manufactured by u-blox of Thalwil, Switzerland. We tested the RTK rover's ability to continuously track the movement of seven sheep over four days, and the results showed an accuracy of around 20 cm. Using each sheep's geospatial data, we built social networks and tracked the first sheep

to move during a grazing period (the movement of the leadership) in the one-hectare test field. Finding the optimal location update rate with a 20 cm or 30 cm threshold distance revealed that social networks could be detected just as accurately with location sampling at 1 sample every second for 1 minute, then no samples for 4 or 9 minutes, as with ongoing measurements at 1 sample every 5 seconds. While operating in an outdoor field setting, the RTK rover collected accurate data on the social networks of a single flock of sheep using sampling techniques developed to prolong battery life.

To identify abusive identities in OSNs that have eluded more conventional abuse-detection methods, the authors of [9] introduce Deep Entity Categorization (DEC), a machine-learning framework. We take advantage of the fact that accounts may be hard to categorize on their own, but attackers have a far harder time replicating or manipulating at scale the social graph, which contains information about the accounts' and others' networks, features, and actions. The technology we use: • Extracts “deep features” by aggregating characteristics and behavioral information from the accounts' immediate and indirect social network neighbors using a “multi-stage multi-task learning” (MS-MTL) paradigm. Using a combination of just a few high-precision human-labeled samples and a large number of lower-precision computational labels in consecutive phases, this method leverages imperfect ground truth information. This design can handle billions of users with ease by using several sampling and reclassification procedures to reduce system load, and it produces a single model that accurately classifies a wide range of abusive accounts. Since DEC's deployment, Facebook has been able to continually classify all users, reducing the number of abusive accounts by an estimated 27%, even when compared to accounts found by more conventional approaches.

Using the audio sensor capabilities of a generic, unmodified wristwatch, researchers provide a workable method for automatically recognizing in-person conversations [10]. It demonstrates the advantages of feature fusion and includes feature representations obtained from various neural network configurations in its suggested architecture. An audio model trained to the voice of the person wearing the wristwatch or others close by is not necessary for the framework to function. In a semi-naturalistic investigation with 39 people across 18 households and 4 persons in free living, we assess our framework with F1 scores of 83.2% and 83.3% for recognizing user talk with the watch, respectively. In addition, we describe several methods to make our framework more practical in real-world settings and assess its real-time performance by running an algorithm on an actual wristwatch. We also make available our annotated database of discussions to encourage further study in this field.

CHAPTER 27

Optimizing Recommendations in Real-Time with Hybrid Recommendation Systems

Ajay Rastogi[1,*] and **H.K. Shashikala**[2]

[1] *Department of College of Computing Sciences & I.T, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: Recently developed e-learning recommendation systems aid students and instructors. These conditions need specialist online learning for students and educators. Split-and-conquer clustering based on a strategy creates a clever recommender for learners' requirements, preferences, and abilities. An automatic self-learning recommender evaluates learners using a separate-and-conquer analysis. Cluster-based linear pattern mining may reveal learner functions. The application suggests properly utilising common pattern scores. This technique was evaluated across different learner groups and datasets to provide relevant learning tasks based on learning style, interest classification, and competence. The recommended cluster-based recommender improved trial recommendation performance, helping learners complete more courses than the no-recommender group. Students scored the suggestion tool over 65% in all areas. The recommender increased metric values for bigger learners in the simulation. Statistics showed huge disparities in student measures exceeding 1000. A computational framework using |L| (recommendation list size) and student qualities explained the disparities. The kids liked the recommender's speed and accuracy.

Keywords: Hybrid Recommendation Systems, Real-Time Optimization, Text Mining, separate-and-conquer analysis, e-learning recommendation systems.

INTRODUCTION

There is a wealth of data accessible in today's globally linked environment. Sometimes consumers feel overwhelmed by the sheer volume

of material available online. A recommendation system aids consumers in making sense of this deluge of data. Therefore, e-commerce sites like Amazon and Netflix, as well as video-on-demand services like Hulu, utilize recommendation

* **Corresponding author Ajay Rastogi:** Department of College of Computing Sciences & I.T, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; E-mail: ajayrahi@gmail.com

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

systems. Both CBF and CF encounter difficulties when not enough people rate products, known as the cold-start problem [1 - 3] and the sparse rating problem [4]. As a result, there are two distinct varieties of cold-start issues: those associated with a brand-new user and those associated with a brand-new product. CBF, on the other hand, only has issues with new users for whom no prior preferences data exists. Also, determining nearest neighbors or nearby things might be difficult when user-item rating matrices are sparse [5]. Comments on many social media sites have been used to go around these restrictions [6]. The material has been evaluated based on the feelings expressed in the feedback. Most of this study has relied on analyzing social media comments about already released films to draw conclusions; however, there are still obstacles for forthcoming films. For upcoming movies, for instance, there is a dearth of metadata such as user reviews and ratings. It is difficult to build a recommendation system for a yet-to-be-released film due to limited data.

For the most part, traditional learning methods have been superseded by e-learning in recent years [7]. E-learning, which uses web innovations to provide students with access to learning online or offline at any time, is seen as more successful than traditional methods of education. Effective learning can occur anywhere, at any time, thanks to personalized e-learning [8]. Novel recommendation algorithms that should, in theory, exceed the state of the art have been the focus of e-learning research efforts as of late. Therefore, developing a better recommendation system is essential to raising the bar for student service. The suggested recommendation system relies heavily on the following subsystems: Learner, Domain, Application, Adaptation, and Session. This research provides a wealth of new strategies to enhance existing state-of-the-art recommendation systems. The proposed approach was used on a dataset including information on one thousand students. Based on our experiments, we found that students in the simulated cluster completed the course in less time and with fewer lessons than those in the no-recommender group. The suggested methodology was also shown to intelligently recommend learning materials based on learner profiles and preferences. The larger the gap between the performance measurements and the computational complexity required by well-known recommendation systems, the less accurate the recommendations will be. As a result, in order to address a wide range of practical concerns, it is necessary to create novel recommendation techniques that address the limitations of existing approaches.

RELATED WORKS

Some currently available e-learning systems were developed with a focus on certain aspects of instructional practices, data structures, and user profiles [9].

Some of the helpful tools and procedures that our team has developed include rules for content-based cooperative filtering and mining, a Bayesian model-based acquiring structure, a recommendation framework built around domain perspectives, a recommendation for mutual sorting for small files, an approach for filtering mutually *via* the estimated maximum likelihood, and a rules-based, tailored learning system through fuzzy theory. We presented ontology-based semantic recommendations and used artificial neural networks to construct an e-learning recommendation based on self-organized maps. To examine how students progress through courses, we built a machine learning and clustering-based recommendation model. Clustering algorithms and content filtering approaches were used to study and evaluate the various learning styles in the context of personalized object-based learning.

The following are some broad categories into which the newly developed recommender systems fall. Profile-Based Recommendation Methods: Recently, personalized recommendation systems have been created for use in a variety of contexts. Learner profiles may be examined using the User Profile-Oriented Diffusion (UPOD) technique. Based on the outcomes of the training phase, this system provides tailored suggestions. In the referral phase, suggestions are made based on characteristics of the student's profile. Systems for Suggesting Readable Material Content-based recommendations using convolutional neural networks (CNNs) were described. This technique is used to uncover the unseen aspects of various media uses. Textual data processing is used to suggest relevant materials in this approach. Mixed-Type Recommendation Frameworks: There have been some hybrid approaches to suggesting movie-based apps. Better movie suggestions were made using content-based filtering. The hybrid approach to personalized learning was created. The recommendation engine generates custom suggestions using data visualization. To address the ERP System and E-Agribusiness datasets, a feature-based recommendation system was also used. The processing speed, accuracy of suggestions, and average absolute error are just a few areas where reviewers have criticized prominent e-learning recommender systems. This paper proposes a recommender system that automatically adjusts to learners' needs, interests, and skill levels by fixing flaws in existing popular techniques. Students' tastes and habits may be automatically assessed and learned by the recommender. To accommodate these different learning styles, we use a split-and-conquer clustering strategy. The machine then makes insightful suggestions by considering the ratings of common patterns [9 - 12].

CF is a process that can hone in on what an individual user prefers by analyzing the preferences of people similar to that user. The inability to start up from cold is a major issue for people with CF. Yang *et al*. introduced a system that infers ratings based on user data. The greater the number of pages viewed by a user, the

CHAPTER 28

Proactive Automated Reasoning Systems for Spatial and Temporal Data Analysis

Amit Singh[1,*] and **Jagdish Chandra Patni**[2]

[1] *College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: By analysing the geographical, chronological, and semantic components of geographic data, it is possible to reconstruct users' real route itineraries and get insights into their preferences and behavior. To analyse tourist traffic patterns at A-level scenic spots in Jiangsu and Zhejiang across time and space, this research collects and preprocesses Weibo check-in data for these locations. The author used a temporal perspective, examining how check-in data fluctuated between 2016 and 2018 and how it differed on weekends, holidays, and weekdays. The acquired data were subjected to a spatial kernel density analysis, which revealed the most active regions. Lastly, the vacation travel mode and characteristics were identified through an examination of spatial and locational flows and their orientations. The results of this study provide the groundwork for the growth of tourism.

Keywords: Spatial and Temporal Data Analysis , Automated Reasoning Systems, Textual Analysis , tourism, tourist traffic patterns.

INTRODUCTION

In today's big data world, more and more individuals expect to access and share information whenever and wherever they need it, and they have become accustomed to doing so *via* their mobile intelligent terminals.

When it comes to mobile apps for gathering and disseminating data, Location-Based Service (LBS) has quickly become a standard. The adoption of these applications has created massive amounts of social media data that includes location information (*i.e.*, geo-tagged social media big data), which is continually expanding. This new kind of enormous social media data has introduced both

* **Corresponding author Amit Singh:** College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; E-mail: amit84376@gmail.com

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

opportunities and challenges to many disciplines of study, piquing the interest of academics in these areas. Similar to "checking in" on Facebook, users of several local life information services (such as Dianping.com, Google NearYou, Jiepang.com, *etc*.) may rate and review various businesses, such as restaurants, hotels, and attractions. Geo-tagged photos, like those shared on platforms like Flickr and Instagram, include not only a descriptive caption but also a geographic coordinate. Users may document their travels in detail, take pictures at any time, and make notes as they go, using numerous applications (such as Baidu Tourism, Bread Trip, and others). The aforementioned social media data (geo-tagged photographs, check-in data) contains not only the typical metadata elements such as title, tags, and author, but also time data (the time the picture was shot or the check-in was performed) and geographical location information. Finding well-known landmarks is a common use case for geo-tagged data mining. Photos of tourist destinations are shared on social media so regularly that studies aimed at identifying the most popular destinations and selecting the best ones based on user interests have become prominent [1 - 5]. Tourists' geo-tagged picture albums or check-in data are interpreted as chronological sequences of places, from which information about hotspots and visitors' routes may be gleaned. Hence, proof and assistance for applications such as tourism planning, tourist attraction development, and intelligent travel suggestions may be obtained by analysing the activities of visitors from social networking geo-tagged big data.

Technology advancements in GIS, GPS, and map visualisation have contributed to the rise in the popularity of mobile positioning devices and location services in recent years. Location-based social networks (LBSNs) [6] have emerged as a result of the convergence of these technologies with established ones. Since the LBSN integrates positioning and social interaction [7], its users may save and broadcast their whereabouts at their convenience. Users and locations serve as the LBSN's fundamental building blocks. The SMG data consists of the two types of units taken from social media.

Each user's SMG data depicts that person's path through time, while collectively they provide information on the modes of transportation and residential neighborhoods frequented by the population as a whole. The LBSN is an improvement over conventional social networks since it allows members to broadcast their precise locations to others. By doing so, online communities may more closely mimic their real-world counterparts [8 - 12]. Twitter and Weibo, two of the most popular social media applications, play a significant role in fostering the fast growth of the LBSN. These applications herald the beginning of the "user-generated content" age (CGM). At any time and from any location, everyone can share their thoughts and emotions. Users of social media platforms may be thought of as a global network of sensors that report on breaking events in

the physical world. As SMG matures, it will become more integrated with location services, generating massive amounts of SMG data. The popularity of reading travelogues online has increased as social media sites mature and web surfing becomes more mainstream. Researchers now have access to a wealth of information in the network environment, thanks to collected reputation data from visitors, allowing them to evaluate the image of tourist attractions and the spatial tourist pattern at individual locations. In particular, UGC from social platforms has become a trusted resource for travellers seeking information about potential destinations.

Furthermore, this data may include tourists' accidental thoughts and sentiments, allowing researchers to gather visitor evaluations and tourism destination images through opinion mining, as well as location data. The value of public opinion in networks has been established in several studies, with important implications for management. The study's goal is to shed light on critical challenges in tourism management by analysing the image and spatial pattern of popular tourist destinations through the lens of opinion mining.

RELATED WORKS

The first geographical information applications mainly catered to business requirements and specialised audiences. These services used one-on-one interviews to gauge consumer interest. Expert interviews alone are unlikely to be sufficient in this day and age, given the prevalence of location tracking and the growth in user numbers. The issue has been addressed by the introduction of network technology, which has enabled the development of effective solutions such as network observation [4] and network investigation [5]. The most up-t--date methods for gathering the locations of many users are summarised here. Harrower *et al*. [6] used the conventional assessment procedure to evaluate the knowledge, skills, and abilities of 16 geographers and lay users. Using an online poll, Huang *et al*. [7] gathered data on the accessibility of GIS applications from 385 legitimate respondents. Gareth *et al*. [8] used satellite imagery and Google Street View to determine camera tilt with latitude and longitude. Cao *et al*. [9] demonstrated the use of texture features as a visual component of images by retrieving them from GPS-tagged photos on photo-sharing websites and projecting land uses using supervised learning. Using GPS trajectories as a data source, Zheng *et al*. [10] retrieved user residence locations, produced the network structure by hierarchically clustering these sites, and used the Hyperlink-Induced Topic Search (HITS) method to identify interest points and hot scenic places. Interest points were extracted from GPS tracks by the authors Zhang *et al*. [11] and correlated with actions. As a first step in compensating for the dearth of location-activity labels, the activity-activity relationship was quantified using data

CHAPTER 29

Proving Knowledge Bases for Automated Reasoning for Information Analysis

G. Geetha[1,*] and **Abhilash Kumar Saxena**[2]

[1] *School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India*

Abstract: New avenues for intelligence collection have been made possible by the advent and proliferation of the internet from the perspective of military intelligence activities, but the ever-increasing amount of information stored on it has put a strain on traditional methods of intelligence analysis. The current state of intelligence analysis must be corrected by expanding and improving military intelligence analysis practices. The control node to intelligence agency analysis must be abandoned if the military is to obtain the depth of intelligence knowledge it requires, if the technology of data mining is to be introduced into the analysis conducted by intelligence agencies, and if a data mining model is to be used to construct the network of military and intelligence analysis. The model's algorithm for intelligence analysis, which is based on semantic analysis rather than traditional types of association analysis, may improve the efficiency and accuracy of military intelligence gathering as a result.

Keywords: Knowledge bases, information analysis, automated reasoning, military intelligence, Fractional calculus.

INTRODUCTION

Integer calculus has been around just as long as its fractional counterpart. Fractional calculus is an important branch of mathematics, but its early development was modest over the last three centuries because it received little support from the foundations of physics and mechanics.

Several complex algorithms in the abnormal dynamics of nature have been shown to defy traditional integer-order calculus models as research has progressed. Its performance is easier to model using a fractional-order system. Because of its

* **Corresponding author G. Geetha:** School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India: E-mail: geetha.g@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

non-locality and memory effects, a fractional calculus quasidifferential operator is ideal for modeling memory and inheritance systems as well as stochastic materials. Due to its lower performance, the second derivative is better suited to represent the historical dependence in the development of the system function than the integer derivative, which lacks localization. Fluid mechanics, electrochemistry, control theory, and viscoelasticity have all seen a rise in academic attention recently. The partial derivative operator method is gaining popularity and is being used more and more in many real-world contexts. Numerous sectors and organizations have embraced its tenets and methods, including engineering, economics, administration, the armed forces, and countless more. We may now build on the solid groundwork laid by these theoretical researches to address the theoretical challenges in systems engineering with greater knowledge and expertise.

What we call "Internet public information" is data that has been gathered and compiled from publicly available sources on the Internet. Timeliness, frugality, thoroughness, and a variety of credible sources define public information available online. On the other hand, the Internet makes knowledge accessible at any time and from anywhere, with great adaptability. The intelligence communities of many nations have come to recognise the significance of freely available public information online since the 1980s [1]. They have evolved into a crucial tool for spies to gather information. Because of its widespread use and ever-increasing bandwidth, the Internet is increasingly seen as a means of interaction with the wider world. The Internet has quickly become the "fourth medium," alongside print, broadcast, and visual forms of communication [2]. The term "military intelligence" refers to the collection of intelligence for military purposes. Intelligence gathered by the military covers a wide range of topics, from the quantity and variety of equipment and weapons to their deployment and operational plans and the technological and economic status of the nation as a whole. The goal of military intelligence gathering is to provide decision-makers with a wealth of valuable information and to analyse recommendations for use in formulating military strategies. The term "open-source intelligence" describes the process of gathering data from open sources, processing it, and delivering it to targeted users in a timely manner to meet their intelligence requirements. Research shows that anywhere from 40-95% of Western industrialised nations' national intelligence is gathered from open sources. There has been a plethora of research done in China on this topic in recent years. The growth of the Internet, and notably social media, has led to the network's progressive rise to prominence as people's primary information resource. The network has grown into the most significant open-source information resource, with vast databases covering commercial Industry, the military, consumer behaviour, and other related topics [3 - 5].

Now that data mining technology has advanced, users may extract meaningful insights from vast troves of network data that were previously inaccessible. In particular, research is now being conducted on the topic of "mining" social networks. Focus areas for studies include a wealth of information on military intelligence activities, such as weapons testing and the placement of weapons and equipment, among these networked resources. There is a vast amount of information available in the highly accessible online world, including news from all over the world, government regulations, research findings from academic institutions, and competitive economic intelligence, as well as ideas authored by individuals in the form of blogs, forums, and the like, and websites set up by terrorist organisations and anti-government groups for the purpose of promotion and liaision. Reportedly, the "Tibetan administration in exile" views the Internet as a "powerful instrument to recruit Tibetans opposing China," according to SIFY.COM. Study findings indicate that "Tibet independence" and the "Tibetan administration in exile" both maintain official websites [6]. Several Dalai Lama sites focus specifically on listing the online addresses of these "Tibet" groups.

Nowadays, all nations place a premium on gathering military intelligence online, with several having established whole departments devoted to the task. Although the proliferation of online data makes it easier to collect information, it also creates formidable challenges when analysing that data [7]. To some degree, military intelligence is affected by the fact that conventional methods of data analysis, although able to filter and anticipate data, are limited in their ability to perform complex statistical processing and analysis when data volumes are small [8 - 10]. Consequently, this paper's goal is to improve the efficiency of intelligence agency evaluation in a decentralized, open system that encompasses a plethora of established Chinese military intelligence analysis methods by introducing data mining methods into connectivity intelligence agency analysis and by building an intelligence agency conceptual framework based on data mining.

RELATED WORKS

Knowledge discovery *via* data mining is the process of uncovering hidden patterns and associations in large datasets. As of now, the technology has found widespread use across sectors such as banking, communications, retail, medicine, and even public security, intelligence, surveillance, and analysis. One way of looking at it is that its creation injects intelligence analysis into the realm of military action. Deeper intelligence agency analysis and so more effective intelligence decisions may be made with the use of modern data mining technologies, particularly in a network setting. The intelligence work of a data assimilation innovation goes even further by classifying the structures of smart

CHAPTER 30

Social Network Analysis for Movie Recommendations and Information Extraction

Jayanthi Kannan[1,*] and **Anurag Singh**[2]

[1] *Department of Information Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: Finding relevant information online has gotten increasingly difficult as the amount of data available on the internet has grown exponentially. In high-data-density, complex-domain settings, the recommendation system may be a big aid to users in making decisions. In the recommendation system, several approaches have been presented and collaborative filtering is a common practice. The cold-start problem is one of the remaining issues with collaborative filtering approaches. To address this issue, we offer a movie recommendation system that uses social network analysis and collaborative filtering. We use user preferences like age, gender, and profession to generate a connection matrix, and then that matrix is used to cluster people using community identification based on edge betweenness centrality. The suggested system then proposes movies to new members based on the preferences of the existing users in the group. Utilizing MAE, we demonstrate the superiority of the suggested technique.

Keywords: Sentiment analysis, movie recommendations, information extraction, social network analysis.

INTRODUCTION

The rate at which new knowledge is added to the world today far outstrips our capacity to absorb it. Online retailers frequently use the recommendation system. These recommendation systems emerged as a way to tailor content to individual consumers. Book, music, movie, news, CD, DVD, TV show, and product recommendation systems are all in use [1].

Providing a list of the best things, making recommendations based on demographics, and suggesting items based on users' past preferences are just a few of the numerous approaches to building a recommendation system [2].

* **Corresponding author Jayanthi Kannan:** Department of Information Science and Engineering, JAIN (Deemed-t--be University), Bangalore, Karnataka, India; E-mail: k.jayanthi@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

Recommender systems are often classified into three groups based on the criteria they use to make recommendations: rule-based, content-based, and collaborative filtering (CF) [3]. The foundation of the CF algorithm is the assumption that a user's latent interests can be inferred from their shared experiences or preferences. Finding users or items that are similar to the expected person or object is its primary function. One of the benefits of a collaborative recommendation system is that users are more likely to find anything of interest in the product, and no prerequisites are needed to use the recommended materials. There are still flaws in the CF algorithm, such as the cold start problem [4].

The “community detection in huge networks”, one of the techniques used to discover communities in social networks, has recently attracted the interest of researchers [5]. Researchers implemented this strategy in the movie recommendation system to improve the reliability of their recommendations. Many researchers focus on ranking measures, such as normalized discounted cumulative gain, which is a family of ranking measures widely used in practice and produces very efficient results alongside the method of community detection in social networks. The goal of this research is to provide a more effective approach than social network analysis. The proposed movie recommendation system uses social network analysis and the normalized discounted cumulative gain approach. The betweenness centrality approach was applied, which is important to social network research.

The cold-start problem occurs when the system has not had enough time to build up its user data to make any recommendations. An individual's tastes and interests must be included when creating a user profile for any recommender system. The user profile is created by considering how they interact with the system. The algorithm decides and offers recommendations based on the user's past history and actions [6]. The issue occurs when a new user or object is introduced into the system, for which there is insufficient data for the system to make a judgment. It can be challenging for the system to build a model from data for a new user who has not yet rated certain products or viewed some items. Because of this, similarity estimates between users may be off. There have been several attempts to find a solution to the issue. Embarak [7] proposed two methods to address the cold-start problem: node recommendation and batch recommendation. Then, he contrasted these two approaches with three others: the Triadic Aspect Technique, the Media Scout Stereotype Technique, and the Naive Filterbots Method. A new hybrid method was suggested by Basiri *et al*. [8] to address the cold start issue. With this method, you can get a sensible and useful recommendation.

RELATED WORKS

These days, you may find recommender systems employed everywhere from entertainment to information to product discovery [9]. Information about a user's interests can be automatically predicted or filtered using a technique called collaborative filtering (CF), which aggregates the preferences and tastes of multiple users. To better grasp what CF is, let us consider a common scenario: you want to see a movie right now, but you have no idea which one to watch. What do you do? In most cases, you ask your friends for recommendations. The CF approach centers on this principle [10].

Recently, recommender systems have gained much attention. It has found its way into literature, journalism, film, music, and even consumer goods. Additionally, there are expert recommenders [11], restaurant recommenders [12], collaborator recommenders [13], monetary service recommenders [14], romantic partner recommenders [15], and Twitter page recommenders. Users can shop for items that best suit their interests or needs with the help of these recommendation systems, which use filtering to predict ratings and customer satisfaction. Knowing the user's background can help predict how they'll respond in certain scenarios [15]. Because it may help users find things they might not otherwise, the recommender system is a valuable alternative to search algorithms. It is often done with a search engine that does not rely on conventional indexing methods. A recommender system is software that analyzes data, such as purchases or user preferences, to make educated guesses about what other people might like to buy. Some methods for building a recommender system are shown below [16 - 20].

Filtering based on content: Techniques for selecting content based on user input and object characteristics are constantly evolving. The information the user supplies is used by these techniques. Advice will be provided to users based on the information gleaned from their profiles. The accuracy of the engine improves as the user gives more information and follows the suggestions made by the engine. These formulas make an effort to suggest items that are quite similar to those previously enjoyed by the consumer. For instance, many of the goods that have been nominated are compared to things that the user has ranked in the past. Research into data retrieval and data filtering can build on this methodology.

Collaborative filtering, or CF, is a data-driven approach to user interest prediction and content filtering that relies on the aggregated preferences of many individuals. The CF approach is predicated on the idea that users who share each other's preferences are more likely to share their preferences on topics when those preferences differ from those of random users. Collaborative recommender systems get a list of products to recommend by looking at how similar users are

CHAPTER 31

Text Mining Approaches for Next-Level Information Analysis and Decision Support

R. Mahalakshmi[1,*] and **Veena S. Badiger**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: A company's postmanagement entails assessing, selecting, and employing individuals who logically and scientifically match the post requirements as well as are a good fit for the company's growth into their appropriate positions. The data that most accurately portrays the market need for data analysis skills is the job recruiting information published on the employment website. Nevertheless, this work uses text mining to analyze the information in online recruiting materials. This research serves as a valuable resource for analyzing job search data, since most of the material is presented in textual form. Workplace matching is the focus of this research, which draws on text mining as well as multi-criteria decision-making. With 66.45% of the candidates coming from the Internet business, technical positions within text mining are mostly concentrated there. Technical data analysis is not as in demand in other sectors; in fact, demand is below 15% across the board. Education continues to dominate as a demand supplier, accounting for 30.53% of the market, despite the dispersed and expansive nature of the business data analysis industry. But at 24.28% and 22.34%, real estate and media are in close proximity. The service sector will have a greater need for business data analysts than technical ones, in addition to the aforementioned three sectors. During the data-gathering step, the recruiting website's data is analyzed to extract job details, including pay, qualifications, required work experience, and the total number of recruits. ¬The data gathering step is carried out by this paper, which crawls the recruiting website for job information. Along with further details about the post, such as the number of recruits, the position's remuneration, the qualifications required, and the required work experience, this study may aid in analyzing recruitment data published by the website and swiftly reflect data on the market need for data analysis abilities.

Keywords: Text mining, advanced information analysis, decision making, text mining, job recruitment information.

[*] **Corresponding author R. Mahalakshmi:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: mahalakshmi@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

INTRODUCTION

To boost performance, organizations should consider improving their data collection processes. Process improvement strategies are becoming more complicated due to the large amount of datasets and the diversity of their attributes. For processing large datasets, previous strategies have not been successful in improving performance. When applied to large process datasets, data mining may reveal useful hidden patterns, which in turn can help improve strategies [1].

By giving workers more agency, Customer relationship management (CRM) boosts customer satisfaction. The customer-firm connection is complex, and CRM data is rarely evaluated across different markets or customer categories; these factors contribute to CRM's challenges in B2B e-commerce. For this reason, the CRM paradigm makes it hard to make good judgments [2]. Discovering new processes and verifying their conformity are two common uses of process mining in big data analytics. However, finding issues, researching remedies, and implementing them are usually the last steps, even though event log analysis may also improve business performance [3]. The ever-increasing data volumes have impacted conceptual modelling as a discipline. In this setting, process mining refers to a set of methods for deriving a process architecture from an execution-related event stream. Preliminary findings from mining processes and analysis algorithms may be useful for weeding out extraneous data, but accurate interpretation of those results requires human judgment, creativity, subject expertise, visual examination, and automation algorithms. In addition, business users often end up with complex process models that are difficult to understand after conducting a process research on a log of events [4]. Research in business process management (BPM) has led to a range of methods, techniques, and resources for creating, deploying, overseeing, and evaluating effective business processes. Process mining (PM) is a relatively new field that aims to enhance business process model analysis by extracting useful information from large volumes of event records [5]. Data and event logs are the building blocks of process mining, a relatively young field of research. As it aids in improving data-based knowledge of business processes, it is widely used in business organizations. When a novel method for analyzing data is founded on an integrated business process with information technology. The integration of process mining with deep learning algorithms facilitates a robust connection between BPM and BI strategies [6]. To analyze trends and identify new ways to reengineer corporate processes, social media has become the most critical data source. There are many different ways to use the data made available by social media. One example is for an entrepreneur to research the industry they want to join and determine what their customers need before releasing new products [7].

RELATED WORKS

To convey the improvement ideas, the authors [8] developed a framework that uses data mining methods to identify important hidden patterns in high-volume process datasets. A genuine set of processes and their characteristics was collected to assess the suggested framework. Next, techniques for classification, clustering, and feature selection were used to identify interesting patterns in the process dataset, which included a large amount of data. The findings then suggested ways to improve after being evaluated. Findings demonstrate that recognized patterns may back up efforts to enhance processes by suggesting ways to do so.

A study presented a hybrid methodology for business-to-business customer relationship management (B2B CRM) that makes use of genetic algorithms as well as data mining methods to enhance decision making [9]. The approach divides consumers into two groups: those who shop often and those who just stop by. The decision tree method's rules were optimized using a genetic algorithm, and an altered Data Mining - C5.0 was employed for consumer categorization. According to the findings, the suggested strategy successfully distributes funds to the most lucrative clientele. Compared to other algorithms, such as the conversional C5.0, k-means, and the Support Vector Machine (SVM) method, the suggested model performs better and has higher accuracy [9].

The study suggests a method for helping company decision-makers narrow down their alternatives and improve performance by analyzing process logs and using process mining [10]. A technique is presented and then illustrated through a case study, after describing how event log evaluation may be used to run performance checks and thereby limit business opportunities. A logistic case study with three trials was used for evaluation; in this research, many business situations were explored, and the worst-performing ones were excluded after a certain time. The studies' results show that the suggested method for reducing business scenarios helps decision-makers identify the best scenario.

In our paper [11], we provide a method for displaying the results of event log analysis to identify business process bottlenecks. Using visual analytics in a process mining setting allows us to provide business process insights that are easy to understand and analyze, rather than presenting users with overly complicated process models. After applying the suggested method to a manufacturing company's process, the findings demonstrate that visual analytics within the framework of process mining may detect performance concerns and bottlenecks and present them to the company's users in a straightforward and unobtrusive manner.

CHAPTER 32

Text Mining for Automated Data Analysis and Information Retrieval

Gopal K. Shyam[1,*] and **J. Vijay Fidelis**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: One new piece of software that's making waves is the recommendation system, or RS. It uses knowledge-discovery methods to infer the user's goals and interests. It is becoming increasingly difficult to decipher the user's intent and activate the suggestion appropriately as data volume grows exponentially. This research proposes a new Deep Knowledge Grid (DKG) for RS construction and huge data analysis. The DKG uses a Deep Convolutional Neural Network (DCNN). This DKG explicitly models the KG's end-to-end high-order connectivities. Using an attention approach, it iteratively propagates embedded information from a node's neighbors—persons, objects, or characteristics—while determining the relative significance of those neighbors. From a conceptual standpoint, our DKG outperforms state-of-the-art KG-based recommendation methods that either use regularization to model high-order relations or extract paths to exploit them explicitly. Results on publicly available benchmarks show that KGAT outperforms state-of-the-art methods like RippleNet with Neural FM. The attention mechanism's benefits for interpretability and the efficacy of embedding propagation in high-order relation modeling have been demonstrated in subsequent studies.

Keywords: Text mining, automated data analysis, information retrieval, Deep Convolutional Neural Network (DCNN), Novel Deep Knowledge Graph (DKG).

INTRODUCTION

The goal of topic identification is to identify and monitor trending subjects on social media. When it comes to individuals voicing their opinions on various topics, Twitter is perhaps the most popular site. Among them, the COVID-19 pandemic stands out. Governments and healthcare organizations might benefit from topic detection and monitoring in dealing with these phenomena [1]. A plethora of information on trending subjects and popular sentiment may be found

* **Corresponding author Gopal K. Shyam:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: gopalkirshna.shyam@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

in the messages of microblogs, such as Twitter. A deeper understanding of data flowing in from social networks can be achieved through topic identification and monitoring. Research in this area often sticks to a consistent set of predetermined themes (clusters) while detecting.

When data is cumulative and subject to change, these methods become inefficient. Furthermore, brief texts are not well-suited for nonparametric evolutionary subject models owing to data shortage [2]. Since popular microblog posts reflect broader public sentiment, subject identification on these sites is crucial for tracking and controlling public opinion. Traditional clustering methods, however, struggle to handle the massive amounts of microblogging data, which span a wide range of themes and contain substantial noise [3]. Plagiarism is a contested issue in many industries, but particularly in the music industry, where commercially produced music is worth a small fortune. Music plagiarism detection is a challenging issue since there are no objective measurements to determine whether a song is plagiarized. As a result, choices are often dependent on subjective arguments. A useful tool is automated music analysis algorithms that can detect musical similarities [4]. Finding the subjects of published papers online is known as topic detection. Getting the subject terms right and making them easy to grasp are two of the most crucial things for topic identification [5]. Books on finance reflect current developments in the industry, which, in turn, affect market movements and investor choices. Subject identification technology is employed to easily and swiftly extract subject information from the vast financial textual corpus [6]. Topic models that rely on clustering to generate topics claim to be more effective than generative probabilistic topic models. These models group high-quality phrase embeddings using the appropriate word-selection method. Nevertheless, these methods have limitations, such as inadequate parameter selection and models that fail to account for the quantitative relationships between words and themes, as well as between texts and subjects [7].

RELATED WORKS

In a study [8], a new method is proposed for grouping 001 texts across languages based on two established techniques: aligned embedding and community identification in graphs. Using clustering, our technique aims to uncover fundamental top005 ics in a multilingual dataset. 006 A quantitative assessment using silhouette as well as V-measure metrics is presented, along with a qualitative evaluation for which a new systematic method is proposed. An analyst's empirical evaluation finds that our algorithm has strong overall performance. Our 014 novel algorithms were deployed and utilized on a 015 every day for a major 013 multidisciplinary public consultation, which is the backdrop of the work we discuss.

A new communicative clustering method, called ComStreamClust, was proposed [9] for grouping related sub-topics (such as COVID-19) within a larger subject. The two datasets used to assess the suggested method were the FA CUP and the COVID-19. Compared with current approaches such as LDA, the results obtained with ComStreamClust confirm its usefulness.

This paper presents a method for discovering new topics by combining BERT and seed LDA clustering [10]. Building a seed LDA (sLDA) model is the first step in optimizing the LDA model. First, the seed words that the LDA model uses were analysed. Second, by fusing sLDA and BERT, the B-sLDA paradigm was constructed. The process of topic generation is guided by the seed phrase set, which is used to generate phrases with a certain likelihood. The sLDA model was used to acquire topic characteristics, while the BERT model was used to obtain text features. The output of the feature fusion process was fed into the method of K-means clustering, which yielded topic groupings. As a last step in the topic identification process, we sent the first 10 words of every cluster through the TF-IDF framework. The official dataset, which included 20 newsgroups, and the real public opinion information for "Shanghai COVID-19" that was acquired using a crawler, were both used in the experiment that was conducted in this study. It is clear from the experiments that the seed term set guides the LDA model, as sLDA outperforms the initial LDA model in complexity and coherence. The Silhouette Coefficient and the Calinski-Harabasz metrics show a considerable improvement in the B-sLDA clustering approach.

An evolutionary clustering approach that automatically determines the number of clusters is proposed [11] by researchers using the Bayesian nonparametric Dirichlet process mixture model. To make the algorithm better at handling short social media communications, they combined text similarity in topic recognition and temporal information gathered from tweets and other social network characteristics in a weighted amalgamation. Data acquired from Twitter over a 2.5-month period was used to evaluate the model. The results show that the topic identification performance may be improved using information from social networks. When it comes to clustering performance and topic coherence, the suggested approach is superior to earlier research, resulting in superior clustering on brief texts.

The authors [12] proposed a hierarchical Bayesian stochastic framework based on clustering to detect and monitor themes in online news articles. This technique would enable the sharing of topics across various stories in the corpus. The hierarchical Pitman-Yor mixing model, whose component densities are the inverted Beta-Liouville (IBL) distribution, has shown greater efficacy in text data modeling compared to the typically used Gaussian distribution; we built our

CHAPTER 33

Text Mining for Effective Information Extraction and Analysis

Ramesh Sengodan[1,*] and **T. Harish Naik**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: The study's overarching goal is to provide a methodology for employing domain ontologies to get health-related data from the Internet. The Internet has made patient-doctor dialogues commonplace. Solutions like HealthTap and AskTheDoctors make it possible for people to ask physicians questions about their health. The knowledge gap between experts and laypeople, as well as the language barrier, mean that most people seeking medical treatment online still struggle to ask effective questions. It is difficult to glean information from these non-expert accounts since they usually do not use technical language. The underlying applications, such as information retrieval, become less efficient as a result of this. In this paper, we provide an ontology-driven method for employing a meta-model to glean insights from such incomplete descriptions. When people seeking medical care don't share the same terminology as their doctors, a meta-model may help. To access the extensive medical vocabulary, the meta-model is mapped to SNOMED CT. To increase the coverage of layman's words during information extraction, it is mapped with WordNet. A data extraction prototype that relies on syntactical similarities is put into place in order to evaluate the approach's potential. Using the gold standard corpora established in Task 1 of ShARe CLEF 2013, the method demonstrated encouraging results in identifying medical ideas in actual medical papers, with an F-score of 0.79.

Keywords: Text mining, information extraction, ontology, WordNet, data-mining customer relationship management.

INTRODUCTION

With the growth of AI, data-mining customer relationship management systems may enhance Chinese businesses' management and decision-making capabilities while simultaneously boosting their bottom line [1]. Opportunities abound in big data analytics to support decision-making and expand our understanding of hospi-

* **Corresponding author Ramesh Sengodan:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: ramesh.sengodan@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

tality management. In marketing, consumer sentiment analysis, product and service enhancements, corporate valuation, and many other areas, hotel management stands to benefit greatly from user-generated content (UGC). What factors influence guests' happiness in the hotel industry has been the subject of much research [2]. CRM stands for customer relationship management and is a method for gathering information about customers, understanding their characteristics, and using that knowledge in targeted advertising. CRM has been around for a while. The truth is that CRM has always been around. But CRM has only lately become the center of interest. The following is some background: • Customers are now seen as a critical component of a company's competitive strength; • Businesses can better manage their customers thanks to data warehousing, mining data, and various other data technologies; and • E-Commerce has created a new platform for marketing and business, allowing us to convert online customer actions into data [3]. A large number of companies have made heavy use of data mining. The use of data mining is on the rise, if not already indispensable, in the healthcare industry. The healthcare industry as a whole stands to gain significantly from data mining technologies. Data mining has many applications in the healthcare industry, including fraud and abuse detection for insurers, decision-making for healthcare organizations on customer relationship management, treatment efficacy and best practices for clinicians, and improved, more cost-effective healthcare for patients [4]. If you want to make a smart choice about an RMS strategy that will increase efficiency and benefit, you need to know how service needs and RMS might interact. Mining efficiency is affected by the excessive number of candidate sets in classic association rule mining techniques, and the results are not straightforward for consumers to understand. In order to find service requests along with remanufacture services association rules, a mining approach based on the binary particles swarm optimization ant colony method is suggested [5]. Popular methods for discovering connections between things individuals buy in a database include market basket analysis (MBA). Mining association rules is a foundational data mining technique that extracts useful information from massive datasets. An ever-increasing amount of data is generated online as people use internet apps for activities such as online shopping and insurance. Large quantities are produced, and both people and society stand to gain substantially from efficient mining [6]. Numerical categorization has played a significant role in China's economic growth, driven by the country's rapidly expanding economy and society. The future of business is a customer-centric, complete operational service model. Improving the management strategy and the service quality procedure is the main concept. To manage the market in line with customers' genuine demand, the business economy takes appropriate steps to build a strong customer relationship, which involves understanding consumers' psychological consumption conditions [7].

RELATED WORKS

Classification of clients, cross-marketing, acquisition, and upkeep are the four data mining modalities discussed [8] under the framework of client relationship management (CRM). The data mining module's SPRINT method is used to classify customers. Utilizing FP-growth, an association rule technique without candidates, in cross-marketing further improves the system's practicability. Adopting the algorithm of the ideal customer retention plan within the context of digital intelligence technology helps businesses improve their operations and fine-tune their marketing tactics, while also compensating for the limitations of conventional CRM systems.

To find out what makes Greek hotel guests happy and what makes them unhappy [9], text analytics is used to dissect hotel reviews and identify how they correlated with guests' overall happiness. This study is helpful for hotel managers because it identifies the features and qualities of products and services that influence visitors' happiness or discontent, and it shows how the positioning and tactics of hotels influence customers' views about those features and qualities.

We set out to do two things with this essay: first, to define data analysis and its role in the sales and marketing domain; and second, to demonstrate the far-reaching implications of investigating and analyzing data in these areas. It also demonstrated that market operating companies are under pressure from the pervasive forces of competition to maximize profits for the benefit of shareholders by preserving and expanding their market share. The researcher conducted a critical literature assessment of relevant theoretical works in addition to the descriptive approach to accomplish the study's aims. All aspects of managing customer relationships rely on data analysis technologies, according to the report. The project can create a cohesive view of its customers from the mountains of data stored in its database by using data analysis tools [10]. Data mining methods are analytically powerful, which means they may help businesses transform their customers' data into valuable information for relationship management choices. The research also suggests that firms dealing with large volumes of data should adopt these approaches.

We demonstrate the use of web mining to automatically gather data from customers' websites for CRM systems in B2B settings in the case study [11]. In this case, we construct a data collection that satisfies the necessary high-quality requirements by extracting pertinent information using specialized local grammars. According to the examination, local grammars do a good job overall, though they may be too strict in some situations. To conclude, our case study shows that B2B CRM may benefit from web mining and local grammars,

CHAPTER 34

Enhancing Information Processing and Analysis with Text Mining

C. Kalaiarasan[1,*] and **Philomine Roseline**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: Many social media forums (blogs, Facebook discussions, *etc*.) have emerged as a result of the proliferation of the Internet, and users are able to access these sites from their mobile devices. These sites serve as venues for users to express and communicate their thoughts and experiences on current events and other global concerns. Before making a purchase, many customers check out review websites. In order to find useful product reviews and user experiences, people often explore popular websites. It is simple for us to gather massive volumes of product data, both structured and unstructured, and then analyze it to extract the specific product details you need. Researchers are increasingly turning to sentiment analysis—also called opinion exploration or opinion mining—to gather and assess customer sentiment and viewpoints. Emotional text mining is introduced in this paper, with a focus on reducing data dimensions *via* principal component analysis and singular value decomposition, and on improving accuracy and reducing implementation time through the use of an additional feature strategy. The research begins by proposing a method for preparing data to classify sentiment. Secondly, it leverages new features to improve emotion categorization precision. Singular value decomposition and principal component analysis are used to reduce data in the third step. Finally, it creates five modules with different features and analyzes their performance, both with and without stemming. When compared to the alternatives, the experimental results show that the proposed method is the most accurate and has the potential to drastically reduce implementation time.

Keywords: Text Mining, Intelligent Information Processing and Analysis, Principal Component Analysis, Data Dimension Reduction, Opinion Exploration (Opinion Mining).

[*] **Corresponding author C. Kalaiarasan:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: kalaiarasan@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

INTRODUCTION

In recent years, data cleansing, a sector of data mining, has seen tremendous growth. To provide the best service to consumers, it is crucial to ensure the reliability of data at every stage, from generation to reception. Data is complex, produced at tremendous velocity, and massive in volume, making the aforementioned job easier said than done. Data Cleaning has extensively used a wide range of approaches from several branches of computer science to achieve optimal performance [1]. Automatic fact-checking (AFC) evidence may take many forms, including text, tables, pictures, audio, and video. Although there is growing interest in using images for AFC, most prior research has focused on identifying altered or false images [2].

Recent advances in machine learning (ML) have enabled the training of high-accuracy predictive models even when dealing with complex multivariate relationships, such as interactions and heterogeneity. Knowledge exploration, accountability, and justice in domains such as healthcare require greater interpretability of ML models. Learning Classification Systems (LCSs) and other rule-based ML methods provide a happy medium between interpretability and prediction accuracy in complicated, noisy environments [3]. The capacity of transformer-based language models to reason in accordance with the logical principles stated in natural-language text has recently become the subject of study. Nevertheless, our current understanding of their thinking is limited because we cannot decipher the abstractions generated by the representations they use. Some argue that these models can't be applied to real-world scenarios because they rely too heavily on memorizing complex patterns in the data [4]. A novel context, unrelated to objects, as well as canonical concept patterns discovered in fact- or rule-based information sources, such as the UMLS ontology, may arise in complex querying scenarios, especially in the setting of EBM (evidence-based medicine) using biomedical literature. In addition, different lexical forms within the collection could conceal relationships between candidate ideas that are relevant to the present context, rather than in a single document [5]. Virtual reality (VR) provides a dynamic and immersive learning environment for construction workers to enhance their safety awareness. With the right VR peripherals, such as data-gloves and trackers that record events, users may experience a wide range of sophisticated interactions and realistic settings. Serious mishaps on construction sites may occur, for instance, when operators fail to properly interpret workers' hand signals [6]. The concept of the “smart home” is moving from the realm of fantasy to that of a practical, workable solution for assisted living, thanks to the proliferation and improvement of the Internet of Things. Recognizing activities is a crucial first step in providing prompt, correct help and service. Due to the prevalence of noise, similar activities, and missing data in everyday life, there are

several problems with inferring complex events using knowledge-driven reasoning algorithms. These include issues with scalability, computational complexity, generalization, and real-time segmentation of raw sensor data [7].

RELATED WORKS

To ensure data reliability, the authors [8] propose a rules-based data-cleaning system that leverages natural language processing. Using natural language processing (NLP), the mechanism outperformed comparable mechanisms that do not employ NLP in both effectiveness and efficiency. Although tested across a variety of healthcare datasets, the mechanism's applicability extends beyond healthcare and lends credence to the idea of generic data cleansing.

In a study [9], researchers propose a new activity they term chart-based fact-checking and present ChartBERT as the first attempt to compare AFC with chart evidence. To determine whether textual assertions are true, ChartBERT uses chart text, structure, and visual data. Our new dataset, ChartFC, has 15,886 charts, and we use it for assessment purposes. We demonstrate that ChartBERT achieves 63.8% accuracy, surpassing VL models, by methodically evaluating 75 distinct vision-language (VL) baselines. Based on our findings, this is a challenging but doable endeavor with several obstacles to overcome.

An automated process for interpreting LCS models for complex biological categorization is described [10] as LCS-DIVE, which stands for LCS Discovery and Visualization Environment. For biomedical data mining, LCS-DIVE employs a novel scikit-learn framework called ExSTraCS, an LCS designed to overcome scalability and noise issues. For each training set, it generates a model with human-readable IF-THEN rules and feature-tracking scores. The LCS-DIVE system combines feature-tracking ratings and rules to automatically define feature relevance and underlying additive, epileptic, or heterogeneous association tendencies in three ways: clustering, visualization creation, and cluster interrogation. To test LCS-DIVE's discriminatory power, we ran it through a battery of synthetic and benchmark datasets that encoded a wide range of complex multivariate relationships; thereafter, we used it to describe associations in a real-world pancreatic cancer investigation.

The authors [11] postulate that classifiers would be more reliable if given both factual and counterfactual explanations. To delve into the process of creating these kinds of explanations, we provide a novel approach to producing factual and counterfactual explanations for pretrained decision trees with fuzzy rule-based classifier scores. When it comes to understanding how classification algorithms work, experimental findings show that combining factual and counterfactual explanations within the framework of fuzzy inference systems works well.

CHAPTER 35

Text Mining Approaches to Enhance Knowledge Discovery and Information Analysis

R. Pallavi[1,*] and **Sheetal**[2]

[1] *Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India*

[2] *Department of Computer Applications, Presidency College, Bangalore, Karnataka, India*

Abstract: A user's interests and goals can be deduced with the use of a Recommendation System (RS), a new form of technology that employs knowledge discovery methods. User intent analysis and corresponding suggestion triggering become increasingly challenging in the face of exponential data growth. This work proposes a unique Deep Knowledge Graph (DKG) to do the necessary data analysis and construct the RS. DKG employs a Deep Convolutional Neural Network (DCNN). Our proposed DKG explicitly models the KG's end-to-end high-order connectivities. It recursively propagates the embeddings from a node's neighbors, which can be people, things, or traits, using an attention mechanism to assess the relative significance of the neighbours. In terms of theory, our DKG outperforms existing KG-based recommendation methods since it does not rely on regularization or an explicit representation of high-order relations. Empirical results on public benchmarks show that KGAT outperforms state-of-the-art methods like RippleNet and Neural FM. The benefits of the attention mechanism for interpretability, as well as the efficacy of embedding propagation for high-order relation modeling, have been shown in subsequent studies.

Keywords: Text mining, knowledge discovery, information analysis, neural networks.

INTRODUCTION

With the rapid expansion of the internet came a corresponding explosion in data. Because there is so much data to choose from, users often feel overwhelmed and end up doing nothing.

Several applications of recommender systems aim to enhance the user experience. Some examples include music recommendations [1], movie suggestions [2], and

* **Corresponding author R. Pallavi:** Department of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India; E-mail: Id-pallavi.r@presidencyuniversity.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

online purchasing [3]. The recommendation algorithm is the backbone of every recommender system. Recommender systems may be broken down into three distinct categories: those that rely on collaborative filtering (CF), those that rely on content, and hybrid systems that combine the two [4]. CF-based recommendations leverage similarities between people or items gleaned from interaction data to predict user preferences, whereas content-based recommendations leverage an item's content attributes. In recent years, it has been common practice to incorporate a knowledge graph (KG) into the recommendation engine. The supplementary data has aroused the curiosity of scholars. Nodes in a KG stand for entities, whereas edges show the relationships between them. It is possible to get insight into the relationships between items by mapping them and their attributes into the KG.

By including user and user-side data into the KG, it is possible to more accurately capture user preferences, object relationships, and user interactions [5]. Although KG has these benefits, it is not easy to use in RS because of its high dimensionality and unpredictability. The information graph embedding (KGE) technique is one way to preprocess the KG. It utilizes a vector of low-dimensional representations to map items and connections [6]. Classical KGE approaches, such as TransE and TransR, are better suited to in-graph applications, such as KG completion as well as link prediction, than to recommendation-making. These algorithms see semantic connections as head + relation = tail. Creating a chart is a process that comes more easily and intuitively.

As a rule, there are two main ways to represent RS systems [6],

- Path-based Methods - Path-based techniques provide predictive models with paths that include high-order information. To limit the routes, they have either constructed meta-path patterns or employed a route selection method to find relevant connections between the nodes. Such two-stage techniques have a considerable influence on overall performance; however, the first step of path selection is not optimal for the recommendation goal. It may be time-consuming to construct several meta-paths in a complicated KG with a wide range of interactions and entities, and this is because doing so effectively necessitates domain expertise.
- Finding new loss terms to replace the KG structure is one way regularization-based strategies help with recommender model learning. For instance, in order to train for both recommendation and KG completion, various studies make use of shared item embeddings. High-order relationships are only encoded indirectly in the recommendation-optimized model using these approaches. Lack of explicit modeling makes it impossible to reliably capture either high-order modeling

results or long-range connectivities.

- Modeling a novel RS for use in a number of contexts is important in light of the most pressing issues with existing approaches.
- In this study, we make the following contributions toward this goal:
- To simplify the suggested RS, we first collect data and do preliminary processing on that data.
- The Deep Knowledge Graph (DKG) is a knowledge graph constructed by a Deep Convolutional Neural Network (DCCN).
- The suggested DKG takes into account higher-order information when constructing DKG. In addition, the attention mechanism calculates the neighbour connectedness.
- Finally, the DKG-RS's performance is measured against the dataset.

RELATED WORKS

Yunkyoung Lee's System for Collective Filtering Suggestions

Collaborative filtering, one of the most well-known and thorough recommendation algorithms, considers the user's preferences while making product suggestions [7]. Because it is based on what people like, collaborative filtering-based recommendation systems may produce a reliable forecast when enough data is available. The recommendation system's user-based collaborative filtering has traditionally been its most critical component for predicting consumer behavior. Widespread usage, however, has exposed major issues, such as data sparsity and scalability, which are becoming more pressing as the population and diversity of commodities keep growing. Embedded representations, which may take up numerous terabytes of space in large-scale settings, are essential to machine intelligence in many applications, including recommendation systems. We take a look at mixed-dimension embeddings—a design for an embedding layer where the dimensionality of an embedded vector grows with the query frequency—as a potential remedy for this issue. Mixed dimensions has the potential to drastically decrease memory use without compromising or diminishing ML performance, according to our theoretical analysis and rigorous testing. Using the Criteo Kaggle dataset, our results demonstrate that the suggested mixed-dimensional layers may increase accuracy by 0.1 percent with half the amount of parameters and maintain accuracy with sixteen times fewer parameters.

In today's fast-paced world, recommendation algorithms are more useful than ever [8]. Because there are so many things that need to be done in a day, people are under continual pressure to get everything done. Since movie fans have limited mental capacity, movie recommendation systems are crucial for helping them find

CHAPTER 36

Advantages and Limitations of Big Data Analysis Tools

Dhruv Galgotia[1,*] and **S.H. Shruthishree**[2]

[1] *Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

Abstract: As the number and variety of data kept growing, conventional data processing technologies could not keep up with the demands of the big data setting. Because of this, new processing instruments tailored to specific needs have emerged. However, picking the ideal instrument could be difficult due to the abundance of options. To aid its mapping to different needs, this page provides a technical overview of the most common analytics tools, along with a literature review on the topic. Furthermore, we highlight the importance of choosing the correct tool for various uses, especially in smart city settings.

Keywords: Big Data, OpenSource Tools, Hadoop, MapReduce, Spark.

INTRODUCTION

From the first pictorial symbols to the most recent digital data, the evolution of data has been continuous and unfettered, expanding at an exponential pace. Properly structured and with a compact store design, a relational database may store little data. In contrast to large data sets, little data is built on a relational database design that allows for easy navigation across different subjects [1].

There is usually much familiarity with the little data stored in the data warehouse. In contrast, there is a deluge of complex, unstructured, non-relational data in our digital age that comes from all sorts of different fields. The major role of big data is to generate the ability to make sense of very small data sets; the term “Big Data” was not completely understood by science until after many others had done so. And materials, covering a wide range of fields such as academia, business, social media, and industry, Many fields have the potential to benefit greatly from

* **Corresponding author Dhruv Galgotia:** Department of Management, Galgotias University, Greater Noida, Uttar Pradesh, India; E-mail: ceo@galgotiasuniversity.edu.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

the opportunities presented by big data. Despite big data's continued prominence as a hot issue, most people agree that big data methods, in particular, are best used in conjunction with more conventional data tools [2]. Big Data appears to be both the biggest challenge and the most promising field for future research as we adapt to the massive data transformation [3].

Big data is often described as "high-volume, high-velocity, as well as high-variety data assets requiring cost-effective, creative forms of data processing for enhanced insight as well as decision making," as stated by Gartner [4].

Following Gartner's three V's (Volume, Velocity, and Variety), IBM data scientists developed the fourth V, "Veracity" [5], to address ambiguity and incompleteness in data, which adds another layer of complexity to big data organizations [6]. Using the four Vs of big data volume, velocity, variety, and veracity to discover and evaluate previously overlooked but critical data might lead to the fifth V, "Value," which is the most desired prospect for most firms.

Each of the five dimensions of big data volume, velocity, diversity, and value is included under its purview. These five characteristics, sometimes referred to as the "5Vs," define modern big data:

- Terabytes, Exabytes, and Zettabytes are the units of measurement for the dataset's size.
- Data Velocity is a real-time metric that displays the pace of data input and output for a process. Due to the time-sensitive nature of data processing, data velocity is defined as the rate of data creation per unit of time.
- Uncertain or missing data is characterized by data veracity.
- Data Variety is the variety of data types generated from diverse areas and resources. It includes structured, semi-structured, and unstructured data.
- Exploring, finding, and analyzing the most essential data in the dataset is the goal of Data Value.

RELATED WORKS

Various sectors, including healthcare, sports, markets, businesses, network security, and educational systems, have adopted Hadoop-based applications, and the authors have devoted most of their attention to describing the concept in [4]. According to Bajaber *et al*. [5], we are now in the "big data 2.0" age, when many processing technologies have developed to improve processing capabilities (*e.g*., MapReduce 2.0) and provide solutions tailored to businesses. For current distributed stream computing tools, there is a standard [6] that considers durability and fault recovery. Storm and Spark are pitted against one another in this article.

Open source real-time/near real-time processors have been studied by Liu *et al*. [7], who focused on their designs and platforms. In a research work [8], authors identified a few tools for real-time big data analysis and categorized the studies based on the tools used and the type of application for each. Their primary emphasis has been on applications in environmental protection, social media, healthcare, and surveillance.

The research paper [9] discusses big data analytics solutions. In addition, it has identified avenues for future study and open questions about platforms and techniques in this field. As a general, highly scalable Cloud-based solution, SMASH has been proposed by Gong *et al*. [10] for processing large-scale traffic data.

A different group of academics has already proposed a method for allocating resources in a shared cloud to handle streaming large data [11]. The goal is to maximize throughput utility with minimal variation across all topologies. It should be emphasized that Apache Spark is another appealing platform for large data processing, in addition to Hadoop. Although the two tools are frequently seen as rivals, it is well acknowledged that they are much more effective when used in tandem. Due to Apache Spark's growing profile as a premier big data analysis platform, many studies have focused on improving it.

BigDebug, a tool that provides dynamic, real-time debug primitives, was created by Gulzar *et al*. [12] for massive data processing in Spark. The spark long lineage issue may be solved with minimal impact on global performance using an automated checkpoint approach [13]. One may find an updated Spark framework, NetSpark [14]. Data serialization, network buffer management, and RDMA (hardware-supported Remote Direct Memory Access) improvements are part of this framework's plan to reduce Spark job execution times. To mitigate the effects of straggler machines, an approach known as Multiple Phases Time Estimation (MPTE) was proposed [15]. An improved task scheduler has also been designed, which allows for better scheduling of backup jobs. Additional research has developed spark-based frameworks to enhance the efficacy of big data analytics.

Actually, TR-Spark was developed by Yan *et al*. [16] to address the problems associated with transitory resources. This framework enables Spark-based apps to operate more efficiently by running as a secondary background activity on transitory resources. Two guiding ideas informed the development of this new framework: data size reduction, awareness scheduling, and resource stability. By combining these ideas, TR-Spark can adapt to the infrastructure's stability. A goal-oriented big data analytics paradigm is proposed by Park *et al*. [17] using Spark, with the aim of improving corporate decision-making. Decisions about

CHAPTER 37

Utilizing Contextual Recommendation Systems for Advanced Comprehension of Information

Sahana Shetty[1,*] and **Ajay Shanker Singh**[2]

[1] *Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India*

[2] *Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India*

Abstract: Creating a system to track the health of a machine is no easy feat. For this to work, a sufficiently large dataset of machine operation signals, operation-related background data, and diagnosis-related anecdotes is required. Nowadays, it is not difficult to address the two issues mentioned. Because it relies on non-numerical and imprecise information, the diagnostic experience is the most difficult to convey. However, in order to build a reliable monitoring system, the processing of the collected data is crucial. A system devoted to suggesting processing methods for condition monitoring is presented in this article's framework. A database and internal modules built using fuzzy logic are part of it. The technique proposes processing algorithms based on user-provided contextual information. The results of testing the suggested agent on two separate parallel gearboxes are also presented in this article. Processing algorithms using assigned model types are the outputs of the system. The findings demonstrate that the algorithms suggested by the system outperform those chosen at random. Accuracy gains ranging from 5% to 14.5% are achievable, according to the findings.

Keywords: Machine's condition monitoring system, user preferences, contextual recommendations systems, video recommendation system.

INTRODUCTION

Several practical uses of recommender systems include providing personalized product suggestions and aiding user decision-making. Since user preferences may vary depending on various circumstances (*e.g.*, time, place, and companion), it is possible to build context-aware recommender systems to tailor suggestions to different scenarios. Improved recommendation algorithms have recently benefited from deep learning and neural network approaches [1]. Online retail

* **Corresponding author Sahana Shetty:** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, Karnataka, India; E-mail: s.sahana@jainuniversity.ac.in

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

recommendation systems (like Amazon's) help consumers choose what to buy by suggesting related goods and services. For merchants to enhance revenue and maintain customer loyalty, it is essential to design a recommender system tailored to each client's needs. Having timestamps of user contact is crucial for learning sequential patterns from their interactions, understanding their short- and long-term preferences, and predicting what to propose next [2]. This is because users' hobbies and tastes change over time. Systems must be able to adapt to users' changing preferences across many online contexts in which they engage with end users. A user's preferences for an item can shift over time for a number of reasons, including changes in context or the nature of the work at hand, as well as external variables, both immediate and distant. To keep up with users' ever-changing interests, recommender systems must be able to record these changes in preferences [3]. Among the many applications of machine learning today, recommendation systems stand out. All parts of the contemporary user experience are affected by recommender systems, whether you're using Netflix to choose a new program to watch or Spotify to listen to an automated playlist. Matrix factorization is a popular technique for creating recommendation systems; it uses past ratings and user preferences to forecast how a user would rate a product [4]. One type of recommendation system is the video recommendation system found on e-learning platforms. This system uses algorithms to determine students' interests and preferences, then suggests instructional movies. Written evaluations and comments from students may give light on the instructional video's merits and shortcomings [5]. Combining recommendation algorithms with deep learning has become a research hotspot in the area of recommender systems, which are vital information technologies today. Numerous recommendation algorithms have been developed, each offering higher recommendation accuracy based on the LFM (Latent Factor Model) [6]. This model learns latent representations *via* matrix factorization with gradient descent to better match user preferences. In an effort to better match user ratings to specific items, recommendation systems inspired by cognitive science have gained significant interest in the last several years. Cognitive knowledge user-item evaluations are sparse and ambiguous, hindering the effectiveness of recommendation systems. Because of this, top-k recommendations are now among the most talked-about topics in the field. The purpose of top-k suggestions is to produce the best-ranked list of items based on user preferences [7].

RELATED WORKS

For context-aware suggestions, see reference [8] for an expanded, proposed generic framework for neural contextual matrix factorization, along with evaluation and comparison of a family of such models. Specifically, we consider two factors - the possible components into which contexts could be fused as well

as the embedding technique used to express context situations-and analyze how these affect the effectiveness of recommendations that are aware of context.

New recommender systems for single-player role-playing games based on latent-component models are proposed [9]. More specifically, we offer a tensor factorization approach for decomposing sets of bipartite matrices that capture the time-dependent interests and actions of the participants. Our algorithm outperforms prior handmade or collaborative filtering methods in recommending more interesting missions and retaining more players, according to comprehensive online bucket-type testing conducted while a commercial game was running.

With an eye on e-commerce product recommendation, this article presents a taxonomy of sequential recommendation systems (SRecSys). It classifies SRecSys into three broad categories: (i) methods based on neural networks, such as advanced models and matrix factorization; (ii) methods based on factorization and latent representation, such as Markov models; and (iii) methods based on sequence similarity, pattern mining, and sequential pattern mining. Products could be better understood using this classification, based on the current SRecSys literature in the e-commerce application field; the classification also provides the current status of the solutions and future research directions [10]. In addition, the approaches and their limitations are outlined, followed by a categorization of the assessed systems based on eight critical aspects. Based on studies conducted on e-commerce datasets (Amazon and Online Retail), integrating sequential purchasing patterns into the recommendation process and modeling consumers' sequential behavior enhances suggestion quality.

To extract aspects, sentiment, and semantic characteristics from user and item evaluations, the aspect-based neural network rating prediction approach (ADLRP) is proposed [11]. After that, latent variables for both people and objects are generated using the deep learning approach. Items' ratings are predicted using the matrix factorization approach based on these three characteristics. The experimental findings demonstrate that the proposed approach outperforms standard artificial neural networks and traditional rating prediction methods. The suggested technique improves rating prediction accuracy by accurately and effectively extracting aspect-specific sentiments and semantics from review texts.

The fact that people's preferences evolve over time is included in our recommendation algorithm [12]. To identify change-points in a series of user interactions that indicate significant changes in preference, we propose a Hidden Markov Model (HMM)- oriented approach that accounts for the sequential actions of all users in the data. To provide recommendations, the proposed approach employs a sequence-aware non-negative matrix factorization model with the

SUBJECT INDEX

S. Kannadhasan, R. Nagarajan, & Kaushik Pal (Eds.)

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y

Z

S. Kannadhasan

Prof. S.Kannadhasan is working as an associate professor in the Department of Electronics and Communication Engineering at Study World College of Engineering, Coimbatore, Tamilnadu, India. He has published around 50 papers in the reputed international journals indexed by SCI, Scopus, Web of Science, and major indexing, and more than 193 papers have been presented/published in national, international journals and conferences. Besides, he has also contributed a book chapter. He also serves as a board member, reviewer, speaker, session chair, and member of the advisory and technical committees of various colleges and conferences. He is a member of SMIEEE, ISTE, FIEI, FIETE, ACM, CSI, IAENG, SEEE, EAI Community etc.

R. Nagarajan

Prof. R.Nagarajan received B.E. in electrical and electronics engineering from Madurai Kamarajar University, Madurai, India, in 1997. He received M.E. in power electronics and drives from Anna University, Chennai, India, in 2008. He received Ph.D. in electrical engineering from Anna University, Chennai, India, in 2014. He has worked in the industry as an electrical engineer. He is currently working as professor of electrical and electronics engineering at Gnanamani College of Technology, Namakkal, Tamilnadu, India. His current research interests include power electronics, power systems, soft computing techniques and renewable energy sources.

Kaushik Pal

Prof. (Dr.) Kaushik Pal has vast academic experiences, skills and research background. He is on the editorial board of significant SCI or SCOPUS - indexed journals, and international publishers. He has reviewed around 150 articles. Prof. Pal is an expert group leader as well as the associate member of various scientific societies, reorganizations and professional bodies. He contributed around 10-plenary, 25- keynote and 30-invited lectures worldwide. As an expert group leader, he belongs to worldwide professional research collaboration and spans engaging him in many R & D Industrial organizations and national and international scientific societies.