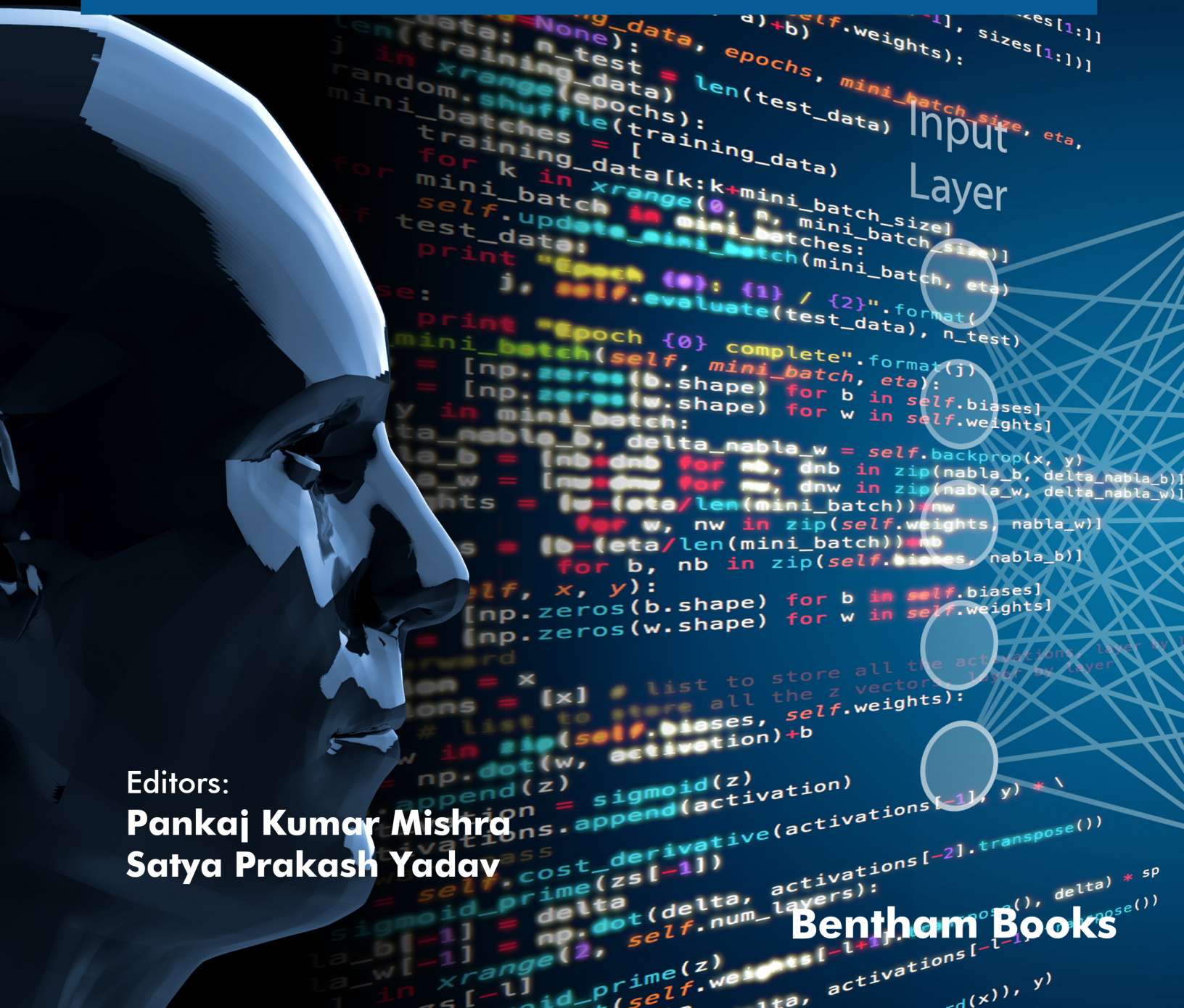
DEMYSTIFYING EMERGING TRENDS IN MACHINE LEARNING



Editors:
Pankaj Kumar Mishra
Satya Prakash Yadav

Bentham Books

Emerging Trends in Computation Intelligence and Disruptive Technologies

(Volume 2)

Demystifying Emerging Trends in Machine Learning

Edited by

Pankaj Kumar Mishra
Hi-Tech Institute of Engineering and Technology
Ghaziabad, U.P.
India

&

Satya Prakash Yadav
School of Computer Science
Engineering and Technology (SCSET)
Bennett University, Greater Noida, U.P.
India

Emerging Trends in Computation Intelligence and Disruptive Technologies

(Volume 2)

Demystifying Emerging Trends in Machine Learning

Editors: Pankaj Kumar Mishra and Satya Prakash Yadav

ISBN (Online): 978-981-5305-39-5

ISBN (Print): 978-981-5305-40-1

ISBN (Paperback): 978-981-5305-41-8

First published in 2025.

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.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

PREFACE

Artificial intelligence (AI) and machine learning (ML) are revolutionizing the way we approach healthcare and various industries. AI and ML are being used to improve patient outcomes, reduce costs, and increase efficiency in the healthcare industry. AI is also used in medical devices to predict and identify diseases, classify data for disease outbreaks, and optimize medical therapy.

In this book, we explore the role of neural networks in AI and ML in the medical and health sector. Neural networks are being used in oncology to train algorithms that can identify cancerous tissues at the microscopic level with the same accuracy as trained physicians. Various rare diseases may manifest in physical characteristics that can be identified in their premature stages by using facial analysis on patient photos.

The book also explores the role of AI and ML in various industries such as finance, retail, manufacturing, and more. AI is being used to improve customer experience by providing personalized recommendations based on customer data. In manufacturing, AI is being used to optimize supply chain management by predicting demand and reducing waste.

This book is a comprehensive guide for anyone interested in learning about the role of AI and ML in medical, health sectors, and various industries.

Pankaj Kumar Mishra
Hi-Tech Institute of Engineering and Technology
Ghaziabad, U.P.
India

&

Satya Prakash Yadav
School of Computer Science
Engineering and Technology (SCSET)
Bennett University, Greater Noida, U.P.
India

List of Contributors

Ayush Gandhi Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Ansh Kataria Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Akhilesh Kalia Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Abhinav Mishra Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Abhishek Singla Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Amarpal Yadav Department of AI, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India

Dikshit Sharma Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Deepak Minhas Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Darleen Grover Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Dhiraj Singh Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Himanshu Makhija Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Jaspreet Sidhu Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Jagmeet Sohal Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Punjab, India

Jatin Khurana Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Jagtej Singh Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Jaskirat Singh Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Mridula Gupta Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Madhur Taneja Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Manpreet Singh Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Manish Nagpal Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Madhur Grover Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

N. Sengottaiyan School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, India

Pratibha Sharma Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Pavas Saini Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Prabhjot Kaur Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Prakriti Kapoor Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Preetjot Singh Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Prateek Garg Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Pratik Mahajan Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Rohaila Naaz	College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India
Rahul Mishra	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Rajat Saini	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Siddharth Sriram	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sidhant Das	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sorabh Sharma	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sukhman Ghumman	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sulabh Mahajan	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sakshi Pandey	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Simran Kalra	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Shubhansh Bansal	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sover Singh Bisht	Department of DS, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India
Saniya Khurana	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Saket Mishra	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
Sahil Suri	Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Sourav Rampal Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

T. R. Mahesh Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, India
Department of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

Tarang Bhatnagar Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Tarun Kapoor Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Vaibhav Kaushik Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

CHAPTER 1

A Method Based on Machine Learning to Classify Text for the Field of Cybersecurity

Siddharth Sriram[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Rapid advancements in networks and computer systems have opened a new door for immoral acts like cybercrime, which threaten public safety, and security, as well as the global economy. The purpose of this proposal is to analyse IP fraud and cyberbullying as two distinct types of cybercrime. The primary goals of this study are to use instances of cybercrime to provide a short examination of cybercrime activities, and the family member principles, and propose a pairing schema. Using the Naive Bayes (NB) & Support Vector Machine (SVM) artificial intelligence techniques, cybercrime instances are categorised according to their ideal qualities. The Twitter data in the Kaggle database has been clustered using K-means. User ID, sign-up date, referral, browser, gender, and age as well as IP address are just a few of the most useful information used to educate the computer. Total 151,113 datasets were used for experimental analysis of the suggested algorithm's performance. The accuracy of the suggested approach, 97%, is higher than that of the current method (NB). The challenge of regression may be easily surmounted with the use of the random forest method for the categorization of the resultant cybercrimes. The planned study uses age categories as the foundation for identifying the different offenses.

Keywords: Cyber security, Machine learning techniques, Text classification.

INTRODUCTION

In order to collect energy from the wake created by oscillating foils, a machine-learning model is constructed. For array deployments of oscillating foils, the function of the wake structure is crucial because the unstable wake significantly affects the performance of downstream foils. User sentiment is gleaned from social media using sentiment analysis [1, 2]. It is a way of organizing the text's ideas into positive, negative, and neutral categories. The outcomes of training and classifying the Twitter dataset have varied depending on the strategy utilized by

[*] **Corresponding author Siddharth Sriram:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: siddharth.sriram.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

the researchers. The algorithm is crucial, especially in time-sensitive industries like airlines and retail [3, 4]. Smartphones, tablets, and other Internet of Things (IoT) devices are frequently used in settings as diverse as the household and the factory. Bluetooth Low Energy (BLE) is used by many of these gadgets as a control or data transmission mechanism [5, 6]. These devices are susceptible to simple attacks because of their lack of robust security mechanisms and the inherent weaknesses in their software as well as communication components. Machine learning is often used for classification purposes [7]. In order to classify images and other remote sensing data, GIS experts often use deep neural network-based classification methods. Graph neural networks (GNNs) can be utilized for identifying geographical features by taking their topology into account, which is useful for data represented as a graph, such as line or polygonal spatial data [8, 9]. Using GNNs to group spatial objects into several categories is suggested in this article. Three alternative methods were tested, two of which depended only on the classification of text and one which combined text classification with a matrix of adjacency. The suggested method's application case was the categorization of planning zones in LSPs [9]. The outcomes of the trials demonstrated the importance of object topological information in enhancing GNNs' classification accuracy. Input characteristics including document length, training data representation format, and network architecture all need to be considered for optimal model performance [10].

RELATED WORK

In this study, we look at 46 different kinematics of oscillation foils to find vortex wakes in pictures of vorticity fields and to adjust the wake parameters according to the input kinematic elements. There are three types of wakes that are classified using a network of convolutional neural networks (CNN) that has lengthy short-term memory units. Utilizing an unsupervised convolutional auto-encoder with [Formula: see text]-means++ clustering, four separate wake patterns are identified, which corroborate the differences in foil kinematics. Future research might use these patterns to predict how foils positioned in the wake would behave and to build optimal foil combinations for tidal energy collecting [11].

In this paper, we provide an optimisation-based machine-learning method for tweet classification. The process consisted of three distinct steps. The process begins with gathering and organising data, then moves on to improving it *via* feature extraction, and then concludes with reclassifying the revised training set using machine learning methods. Different algorithms provide different results. Sequential minimal optimising with decision trees has been proven to have a high degree of accuracy (89.47%) when compared to other machine learning techniques [12].

In the first part of this essay, we saw how to get unprocessed information on network traffic while simultaneously launching a MitM assault on BLE devices. Second, we investigate the possibility of using machine learning—and specifically a combination of unsupervised and supervised methods—to identify this kind of assault. We reconstructed the model of BLE interactions using two unsupervised approaches and utilised those models to spot suspicious data batches. The packets within each batch were then categorised as either normal or attack using a classification approach based on the Text-CNN technology. The results of our model reconstruction demonstrate that our classification approach is very accurate (0.99), with a low false positive rate (0.03) [13].

We used LIME to explain the model's predictions and a conglomeration of deep learning algorithms (CNN-GRU) to categorize four distinct cardiac arrhythmia kinds as part of this study. A 1D convolutional neural network (CNN) served as the basis for training the model. A well-liked local explanation method, LIME can simulate the behavior of every machine learning system and provide an explanation for it. Unfortunately, LIME is limited to explaining data in tabular, textual, and visual formats. In order to better illustrate LIME on the signal dataset, we advocated for a heat map to be used to draw attention to relevant regions of the heartbeat signals. Our approach also allows for more accurate heartbeat segmentation by accurately extracting characteristics such as the QRS Complex, P Wave, and T Wave from electrocardiogram, or ECG, records. Tests for recall, accuracy, precision, f1 score, as well as area under the receiver operator curve of characteristic (AUC-ROC) were conducted using ECG lead II from the MIT-BIH datasets to evaluate the proposed hybrid model. We directly compare the suggested model to the independent CNN and GRU algorithms to show that it is more accurate and has a better ROC [14].

Corona Virus and Conspiracies Multimedia Analysis Task of the MediaEval 2021 Challenge is dedicated to investigating claims of wrongdoing in relation to the COVID-19 pandemic. Our HCMUS group takes several methods based on various pre-trained models to handle 2 separate jobs. We provide 5 iterations of Task 1 and 1 of Task 2 based on our experiments. While the BERT [5] trained model is included in both runs, the first run also incorporates a subjective assessment for acquiring semantic features prior to training. Our third and fourth runs will be more method-diversified with the use of naïve Bayes classifiers [4] as well as a long short-term memory framework [8]. By combining many ML and DL models, Run 5 is able to perform a multimodal analysis of textual data [3]. A single-run Bayesian categorization approach is used in the final phase of subtask 2. In the end, our approach yields scores of 0.5987 on task 1 as well as 0.3136 on task 2. The author's copyright for this article is valid until 2021 [15].

CHAPTER 2

A Practicable E-commerce-Based Text-Classification System

Sidhant Das[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: This article examines the features of the dealer's brush list evaluation material in light of research findings on misleading assessment and identification of online purchasing. A Gated Recurrent Unit (GRU) model using keyword weighting is presented as a solution to the issue that it is challenging for the DL model to collect the feature data of the whole assessment text in a false evaluation identifying job. The TF-IDF technique is first used to generate the list of keywords, and then that list's weight is applied to the word vector. Finally, a weighted vector of words is categorised using this method of the model to finish the recognition job of erroneous evaluation, replacing the pooling component of the GRU model with a constrained Boltzmann machine. By using a variety of text categorization algorithms and comparing their results in terms of correctness and performance, this research aspires to represent the practical benefits of applications that use machine learning in the real world. We built a system that can run several text classification algorithms, and we used that system to create models that were educated using actual data taken from E-Commerce, a virtual fashion e-commerce platform. The Convolutional Neural Network technique achieved the greatest mean accuracy of 96.08% (with a range of 85.44% to 99.99%) with an average deviation of 5.65%.

Keywords: Feature extraction with machine learning, Text classification, Text mining.

INTRODUCTION

E-commerce platforms should prioritize pre-sale customer support since it helps improve the purchase experience for potential consumers. We propose AliMe KG, a domain knowledge graph in E-commerce that contains user issues, POIs, item information, and linkages there between [1, 2] to better serve consumers. It is

[*] **Corresponding author Sidhant Das:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sidhant.das.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

useful for gauging user interest, preparing for sales conversations, and writing up guides. Trading online is not a novel idea [3]. By removing the need for physically stored papers and replacing them with digital ones, businesses may save time and money [4, 5]. E-commerce, or electronic commerce, is a kind of business transaction that takes place *via* the open Internet. It has become a major part of people's daily routine. Therefore, businesses of all sizes must work quickly to create a reliable E-commerce system [6, 7]. The primary components analysis (PCA) is a statistical method used to examine the impact of key variables and attributes on the growth of electronic commerce [8]. The findings are calculated using the PCA model and expert ratings. Primary components and factor analysis (PCFA) are used to examine the effectiveness of the security mechanism used in E-Commerce transactions. The analysis findings represent the major function indexes of assessment in security methods in E-Commerce affairs [9], and are based on the PCFA model and the ratings by experts. Based on the results of the case study, it is clear that some indices play a crucial part in the analysis and assessment of the primary function indices, and that all evaluations should be considered. Using logic and reasoning, we may deduce that using PCFA to evaluate security techniques in E-Commerce business is feasible, given the goal of reducing data loss [10].

The following is the outline for this paper. The experimental setup is described briefly in Section 2, our study is described in depth in Section 3, along with the datasets, evaluation metrics, and features used in our analysis, and the results are evaluated, a conclusion is drawn, and future work is discussed in Section 4.

RELATED WORK

The goal of this research is to help AirAsia better serve its online shoppers by analysing the influence of e-commerce sites on the airline's customers' purchasing habits. This study used a qualitative approach based on in-depth interviews. E-commerce websites have an influence on consumers' purchasing habits, and this report will help the organisation better understand this effect and implement successful ideas. AirAsia has to learn more about its customers' purchasing habits if it wants to succeed. As a result, AirAsia has to invest time and energy into studying customer preferences to better serve them [11].

We used AliMe KG in a number of real-world use cases, including shopping guides, property inquiry answering, and the creation of selling points, and saw significant financial returns. In this work, we systematically explain our method for building a domain knowledge graph out of unstructured text, and we show its commercial usefulness *via* a number of examples. Our results demonstrate the

viability and potential benefit of extracting structured information from unstructured text in the vertical domain [12].

This research investigates the ways in which women have established themselves in various sectors of the online economy. How has the rise of online shopping affected the status of women in society? And how can governments encourage e-commerce growth in low-income regions? Most examples of businesses run by women that have found success in business-to-consumer (B2C) e-commerce focus on selling niche goods to affluent customers. The government plays a pivotal role in stimulating the growth of e-commerce by actual realistic actions [13], even if it is widely recognised that the private sector should take the main role in the creation and usage of e-commerce.

In this paper, a.NET-based e-commerce framework is proposed. It provides examples of the most frequent E-commerce system functions, a system enclosure, designing concepts for the system's levels, and the most pressing technical issues associated with the system. The standard features of an E-commerce platform have been implemented on this one. As has been shown in the real world, this kind of E-commerce platform may fulfill the requirements of several typical businesses such as Electronic market, .NET framework, purchases, temporary storage, and subscriptions [14].

Evidence from case studies demonstrates that some features and qualities have a significant impact on the growth of the e-commerce industry; nevertheless, all of these elements must be considered in order to achieve a successful outcome. Analysis and debate lead to the conclusion that using PCA to analyse the impact of E-Commerce development is feasible [15], predicated on the assumption that data loss can be kept to a minimum.

PROPOSED WORK

Problem Formulation

Instead of basing their decisions on the semantic resemblance of words, text classification algorithms take into account the closeness of term vectors in the vector space using different approaches (such as cosine similarity). However, the similarity score suffers when words with distinct forms owing to suffixes, prefixes, or tenses are evaluated as separate vectors. We devised a procedure called “Stemming” to remove suffixes, prefixes, and tenses from words and return them to their root forms so as to circumvent this issue. Next, we use a process called “Vectorization” to convert every sentence in our data set into a numerical value. Words are converted into machine-readable values (such as integers or floating-point numbers) inside a limited vector space. Vectorization assigns a

CHAPTER 3

AI Model for Text Classification Using FastText

Sorabh Sharma[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The purpose of text categorization, a machine learning technique, is to automatically assign tags or categories to texts. Natural language processing (NLP)-based text classifiers can quickly analyse vast volumes of text and classify it based on emotions, themes, and human intent. FastText was created by Facebook's AI Research team and is available to the public as a free library. Its primary goal is the efficient and accurate processing of big datasets in order to provide scalable remedies for the problems of text categorization and representation. Traditional machine learning techniques used in most text categorization models suffer from issues including the curse of dimensionality and subpar performance. This research offers a fastText-based AI text classification model to address the aforementioned issues. The fastText approach allows our model to create a low-dimensional, continuous, and high-quality representation of text by mining the text for relevant information through feature engineering. The experiment uses Python to define the text dataset, and the results demonstrate that our model outperforms the baseline model trained using classic ML methods in terms of accuracy, recall, and F values.

Keywords: Emotional polarity judgment, Feature engineering, Machine learning, Text classification.

INTRODUCTION

As a result of the COVID-19 pandemic, all global activity ceased. Taking care of one's mental and emotional well-being is as crucial as taking care of one's physical well-being under such circumstances. Since telecommuting has become the norm, users now make use of Twitter to express how they're feeling about the situation, and their tweets may be analysed in this way [1, 2]. Sentiment analysis of text is a necessary part of many NLP tasks. The use of sentiment analysis to sift through Internet data for useful insights is becoming more important as social media platforms grow in popularity [3]. In light of recent developments, we're considering using models based on deep learning to manage sentiment categori-

[*] **Correspondending author Sorabh Sharma:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sorabh.sharma.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

zation. As a result of the service malfunction that happened on July 20, 2019, Bank Mandiri's share price plummeted 0.95 percent at the start of trade on July 22, 2019, compared to its opening price [4, 5, 6]. Looking for a connection between Twitter or online media coverage attitudes or views and the price of the stock of Bank Mandiri (BMRI) [7], the fall proceeded to reach 1.27 percent from the previous price. The use of text mining to acquire data, optimisation of the model using Logistic Regression, as well as feature expansion with FastText are used for predicting sentiment by decreasing lexical discrepancies [8]. Then, at the year's conclusion, a Pearson Correlation Coefficient test is conducted to see whether sentiment forecasts are correlated with stock prices from May through September. Connectivity to Bank Mandiri's stock price data in order to learn how satisfied customers are with a certain service offered by Bank Mandiri [9]. Automated text classification tasks have been proven to improve to a fair degree when combining efficient word representation methods (word embeddings) with contemporary machine learning models. However, in terms of inadequate word vector representations for training, the efficacy of such methods has not yet been assessed. When it comes to pattern identification, picture analysis, and text categorization, Convolutional Neural Networks have been shown to be very effective [10].

RELATED WORK

The goal of our research system is to analyse user emotions over time by collecting data from Twitter and processing it based on keywords. Fasttext, a text categorization tool developed by Facebook, does this by categorising tweets into the following states of mind as quickly and accurately as possible: angry, relieved, bored, happy, hateful, fun, loving, surprised, enthusiastic, sad, and empty. FastText is a popular Natural Language Processing (NLP) Library for representing and categorising text. Classifying emotions across time sheds light on the state of mind of the general public and how it has evolved [11].

A text emotion classification model using FastText & multi-scale DPCNN is presented to enhance the effectiveness of Chinese text emotion classification. The FastText model is first used to build a text vector matrix. After that, a multi-scale filter is used to extract numerous feature maps from the content vector matrix. At last, the DPCNN model is fed the combined feature maps for sentiment analysis. Tests are conducted using the Chinese sentiment mining corpus (ChnSentiCorp) data set, and the findings from various sets of tests that compare the accuracy of text sentiment classification using the proposed model with other techniques [12] demonstrate its efficacy.

In this research, we introduce the use of a Convolutional Neural Network (CNN) inside a framework we call fastText. To kick things off, we feed a CNN using word vector representations generated by fastText. The goal of FastText is to both create an automatic model of a single word and accurately represent word distance. This one allows the parameters to be set at a good CNN point, which improves the effectiveness of neural networks in this setting. Second, we develop a Convolutional Neural Network structure tailored to emotional analysis. In this architecture, we combine two groups of convolutional layers with pooling layers. To the best of our knowledge, this is the initial occasion that a 9-layer design and architecture model based on fastText and CNN have been utilized to evaluate the sentiment of sentences. To make our model more precise and adaptable, we employ the Normalisation and Dropout methods alongside Rectified Linear Unit (ReLU) regression. For this experiment, we used a publicly accessible dataset consisting of excerpts from movie reviews labelled as either "negative," "slightly negative," "neutral," "moderately positive," or "positive" to evaluate our methods. With a test precision of 96.4%, ourReLU paired network outperforms other neural network models on this dataset [13].

Applying Logistic Regression and including feature expansion in the FastText scenario with a split of 90% of training and 10% of testing yields 71.8% efficiency in the model improvement scenario. Using an optimisation model on both positive and negative sentiment, we find that negative sentiment correlates most strongly with changes in BMRI stock from May 2019 to the close of September 2019 [14].

In this research and experiments, we look at how the CNN model may be used to solve text classification issues. We used six publicly accessible benchmark datasets—Ag News, Amazon Full & Polarity, Yahoo Question choice Yelp Full, & Polarity—to train our classification model utilising the popular word embedding generation methodology, Fast Text. In addition, the suggested model has been evaluated using a non-benchmark dataset including tweets about U.S. airline companies. Fast Text as embedding words was shown to be a promising method in this study [15].

PROPOSED WORK

System Model

Fig. (**1**) illustrates the overarching concept of the approach we suggest in our study to categorise text as harmful, non-toxic, or ambiguous. To get started with this study, we integrated two separate datasets into one larger one. Following this first stage, we use ML classifiers and then DL-based algorithms for further breakdown.

CHAPTER 4

An Algorithm for Textual Classification of News Utilizing Artificial Intelligence Technology

Rahul Mishra[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The rate at which technology is improving is increasing all throughout the world. Every day, a tremendous amount of textual data is produced as a result of the Internet, websites, business data, medical information, and the media. Extraction of interesting patterns from text data with varied lengths such as views, summaries, and facts is a challenging issue. This work provides a deep learning (DL) algorithm-based approach to news text classification to address the issues of large amounts of text data and cumbersome features obtaining value in news. Although the relationship among words as well as categories has a significant impact on the categorization of news text, previous approaches to text classification relied solely on the knowledge of the connections between words to make their classification decisions. This research uses the idea of a tailored algorithm to provide a CNN, LSTM, and MLP-based customizable ensemble framework for categorising news text data. The proposed model is based on a parallel representation of word vectors and word dispersion. We feed the term vector to the CNN module to convey the relationship between words, as well as nourish the discrete vector corresponding to the relationship between words and categories into the MLP module to achieve deep learning of the spatial data on features, the time-series feature information, and the connection words and categories in news texts. Extensive experimental study confirmed the dependability and efficacy of the proposed approach. The experimental results demonstrated that the proposed method improved the most - in terms of precision, recall rate, and comprehensive value, while also addressing the problems of text length, extraction of features issues with the news text, and classification of news text.

Keywords: Artificial intelligence, CNN, LSTM and MLP, Text classification.

[*] **Corresponding author Rahul Mishra:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: rahul.mishra.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

INTRODUCTION

Cystoscopy clinical documentation consists of both written and visual information. However, inadequate data handling in ordinary clinical practice continues to restrict secondary applications of visual judgment data for educational as well as research reasons [1]. Methods: Three main components—data management, annotation management, and utilisation management—make up the conceptual framework developed for standardised cystoscopy documentation. For quality assurance and problem-solving, the Swiss-cheese paradigm was presented [2]. With FAIR (findable, readily available interoperable, and reusable) principles in mind, we outlined the supporting infrastructure that would be needed to put the framework into action. To guarantee conformity with FAIR principles [3, 4], we used two scenarios illustrative of data sharing for educational and research uses. Academic activity around adversarial instances in the picture area has exploded in recent years, after their discovery [5]. With the rise of AI in recent years, scientists have been more interested in studying rival samples in the written sector [6]. Given the rapidly growing amount of electronic resources [7], automated categorization of certain resources has become critically vital. Artificial intelligence (AI) was used to glean people's thoughts from social media. Despite this, most ongoing studies concentrate on extrapolating characteristics from texts. Due to the increased complexity of giving a single label to each text fragment and the sheer volume of accessible data, multi-label textual data categorization has emerged as a pressing issue. You may see this in the media and in emails [8]. Building a classification system that can categorise fresh data based on labelled online documents is a primary goal of automated web page classification research in web mining [9]. In order to solve text classification challenges, such as the categorization of online documents, machine learning techniques are adopted. Artificial immune systems of Computational intelligence use cues from biological immune systems to address a wide range of challenges in artificial intelligence, including categorization [10].

Nevertheless, the aforementioned combination of depth learning research only considers the relationship among words and ignores the relationship among words and categories, despite the fact that this is a crucial factor in the categorization of news text. The following are a few of the paper's most fundamental contributions: (i) This study chooses CNN, LSTM, and MLP models to present a custom MLP algorithm for news text categorization based on double input mixed depth learning, in line with the research concept of the combination of DL approaches based news text classification. (ii) Word vector and word dispersion are used in tandem to describe the suggested model. To achieve deep learning of news text's spatial data on features, time-series data feature information, as well as the

connection between words and categories, we feed a word vector representing the relationship between words into the CNN module, and a discrete vector representing the relationship between words as well as categories into the MLP module. (iii) Many experiments were performed to test the reliability and effectiveness of the strategy under consideration. The experimental findings demonstrate that the suggested technique significantly outperforms the alternatives in terms of the accuracy of predictions as well as other performance metrics, hence (iv) The method is a clear winner.

The remaining sections of the paper are laid out as follows. The relevant literature is presented in Section 2, and the suggested model for news material categorization is shown in Section 3. In Section 4, we provide the findings and analyses of our experiments, and in Section 5, we draw a conclusion.

RELATED WORK

In order to shed light on the causes of the observed fluctuations in the direct ownership and market liquidity in the United States, this article employs a text-based sentiment indicator. Market sentiment is taken from 66,070 news stories on the US real estate market in the research firm S&P Global Market database using an artificial neural network. The network is trained using a remote supervision technique and labelled data from Seeking Alpha, a crowdsourced financial advising portal, totaling 17,822 investment ideas. The derived textual emotions indicator is significantly associated with the depth as well as resilience aspects of liquidity in the market (proxied by Amihud's (2002) price impact measure) and the breadth dimensions (proxied by transaction volume), as determined by self-regressive distributed lag models that include contemporary as well as lagged sentiment as variables that are independent. These findings point to the existence of a long-term relationship between market mood and direct property market liquidity. This impact should be taken into consideration by market players when making investment choices and when pricing liquidity risk. Not only does this research add to the body of knowledge on text-based emotion indices in property, but it is additionally the first to use AI to extract sentiment from news stories in the context of market liquidity [11].

Successful implementation of the framework is carried out according to FAIR guidelines. The resulting cystoscopy atlas could be featured on a website dedicated to education; 68 full-length subjective videos and their annotation data were made available for use in AI projects addressing frame segmentation and classification issues at the case, lesion, and frame levels. Results from our research indicate that the suggested architecture makes it possible to save visual

CHAPTER 5

Analysis of the Sentiment of Tweets Regarding COVID-19 Vaccines Using Natural Language Processing and Machine Learning Sectionification Algorithms

Sukhman Ghumman[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The unique Coronavirus pandemic of 2019 (called COVID-19 by the globe Health Organisation) has exposed several governments throughout the globe to risk. The Covid-19 epidemic, which had previously only affected the Chinese population, is now a major worry for countries all over the globe. Additionally to the obvious health effects of COVID-19 epidemic, this study reveals its repercussions on the worldwide economy. The research went on to talk about how they analysed public opinion and learned new things about Covid-19 vaccinations by using content Analytics and sentiment evaluation in Natural Language Processing (NLP) using content from Twitter. To categorise and analyse the outcomes, researchers used two machine learning algorithms: logistic regression (LR), random forest, decision tree, and convolutional neural networks (CNNs). To better identify public opinion, several pre-processing methods were used and categorised responses into neutral, positive, and negative categories. The public's opinion on Covid-19 vaccinations is 31% favourable, 22% negative, and 47% neutral, according to the results of the emotion section distribution. CNN achieved 98% accuracy, according to the tested machine learning algorithms.

Keywords: Covid-19 vaccine, Health informatics, Micro-blogging, Public opinions, Sentiment analysis, Social media, Twitter.

INTRODUCTION

It is no longer possible to overstate the importance of social media as a means of communication and expression for people, businesses, and governments all over the globe. The 2019 coronavirus disease (COVID-19) pandemic has made social

[*] **Corresponding author Sukhman Ghumman:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sukhman.ghumman.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

media platforms more important than ever for individuals to connect, share, and voice their opinions on a wide range of issues. Governments' and organisations' ability to react quickly to emerging threats may be enhanced by analysing this kind of textual data. In order to assist organisations in raising awareness, this project plans to conduct sentiment assessments on the issue of COVID-19 vaccine, as well as temporal and geographical analyses utilising textual data, and identify the most often discussed subjects. The effect of COVID-19 vaccination and how people react when exposed to the virus is now the most commonly contested subject in the medical community. Although several different vaccines have been created, it has never been quite clear whether or not they will provide total protection against the virus [1, 2]. Although many researchers have considered the topic of evaluating people's attitudes about vaccinations, few have actually used language processing approaches to the data they collected. Over the last decade, social media has seen rapid growth [3, 4]. This growth has been accompanied by both good and bad effects. People from all walks of life are able to contact each other directly across cultural and economic divides thanks to the exponential growth of networking *via* social platforms and websites. There are positive effects of social media on society, but there are also negative effects. In recent years, hate speech has emerged as a serious issue. Online forms of hate speech often include the use of threatening or insulting rhetoric [5, 6]. It might mean anybody or a specific group of individuals who have some common ground. The purpose of this automated essay assessment system is to score writings and deliver comments automatically. The usage of automated grading systems in the section room and on digital tests is on the rise. The purpose of this research is to build and assess machine learning models for automated essay assessment [7, 8]. In order to steal sensitive information from a person, such as login credentials or financial account numbers, phishing is a common method of using shared technologies. This occurs when the attacker poses as a reliable system in order to trick the victim into opening an email, clamorous material, or textual concerns [9]. The target is then tricked into opening a malicious attachment, which installs malware and prevents the machine from responding to user input until the ransom is paid. or the disclosure of confidential data. The leak of sensitive information is only one example of the devastating effects that might result from such assaults [10]. Theft of money or personal information, as well as other forms of illegal leverage, is considered a serious crimes by individuals.

Since feelings evolve over time, our study also examined how these tweets evolved over time, which is a unique and significant addition to the field. Furthermore, we demonstrated *via* study that our system is able to accurately discern the emotion of a phrase given text inputs. The purpose of this study and the value it adds are to better understand how people feel about becoming vaccinated against COVID-19. Researchers and policymakers in the health sector

will benefit from this information as they work to improve public trust in vaccinations and ensure the public's continued safety and awareness.

The analytical methods and results of the study's technical components are described in the following sections. In Section 2, we go through the processes and jargon specific to sentiment analysis. The functionality of effective visualisation tools, the explanation of how they represented emotions, and the results obtained are covered in Section 3. In Section 4, we explore the results of our study, its potential effect and benefits to humanity, and ways in which it might be enhanced to hasten the good of the planet. Section 5 contains the results of our research.

RELATED WORK

This research focuses on information collected from social networking sites and uses the NLTK (Natural Language Processing Toolkit) to improve sentiment analysis. The covid19 dataset was used to classify public tweet sentiments by polarity, recall, accuracy, and f1-value while encouraging the use of a combination of embedding words with the TFIDF vectorizer, data sheathed *via* fine-grained sentiment analysis, as well as machine learning methods like Linear SVC, SVM, as well as Nave bayes. In order to efficiently extract emotions from data, simulations utilise the features of the Python library VADER (Valence Aware Dictionary & sentiment Reasoner) [11].

We present in this study our approach to deal with and significantly reduce such hate speech. Many individuals harm others' sentiments by venting their hate and rage directly on social media. It would have serious consequences for them, regardless of their caste, faith, religion, or ethnicity. It is possible that some remarks might be considered hate speech simply because of the vulgar language used. To eradicate hate speech, we have dove headfirst into natural language processing, and accuracy-based comparison of several artificial intelligence models for selection [12].

The growth of Web services is significantly hampered by malicious websites, which also considerably help the propagation of online crime. There must be an all-encompassing plan to prevent people from visiting these sites. We suggest an approach based on machine learning that classifies websites as either safe, spammy, or harmful. Our approach analyses just the Uniform Resource Locator (URL) of a page to determine its relevance. Therefore, it prevents users from being vulnerable to browser-based exploits in production. Since our method employs learning algorithms, it is more generalizable and has a wider scope than blacklisting services [13].

CHAPTER 6

Classification of Medical Text using ML and DL Techniques

Sulabh Mahajan[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The use of sarcasm in everyday conversation has recently risen to prominence. All the kids of our age utilise sarcasm to convey a negative message in a more nuanced manner. With the advancement of AI and machine learning techniques in the area of natural language processing (NLP), it has become more difficult to reliably and effectively identify sarcasm. This research provides a new method for sarcasm detection using machine studying and deep learning, hoping to make a meaningful contribution to this expanding area of study. In order to prepare the phrase for a hybrid deep learning model for training and classification, this method employs bidirectional encoder representations from transformers (BERT). The combination of CNN and LSTM creates the hybrid model used here. The suggested model has been tested on two datasets, with the goal of identifying sarcasm from nonsarcastic words. The trained model obtained a 99.63% accuracy rate, a 99.33% precision rate, a 99.83% recall rate, and an F1-score of 99.56%. These findings are based on 10 rounds of cross-validation performed on the model that was suggested using the medical datasets.

Keywords: DL techniques, Medical text classification, ML techniques, Real-time applications.

INTRODUCTION

Herein, we built a classification system to categorise the reasons for ER visits based on free-text clinical notes in order to construct a nationwide injury monitoring system [1, 2]. However, a significant number of annotated datasets is necessary for the success of supervised learning methods in this field. Recent advances in neural language modelling (NLM) have been made possible by Transformer-based models that pre-train their neural networks in an unsupervised generative manner. We hypothesise that the number of annotated samples needed

* **Corresponding author Sulabh Mahajan:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sulabh.mahajan.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

for supervised fine-tuning may be greatly reduced by using a generative self-supervised pre-training phase [3- 5]. There are two primary schools of thought tried when attempting the difficult Natural Language Processing job of text categorization of unknown classes. Instances are categorised using similarities between document representations and class description representations in similarity-based techniques. By correctly labelling articles as belonging to unknown classes, zero-shot text classification methods hope to generalise information acquired during a training job [6-8]. Although prior research has explored some of these categories, the trials reported in the literature do not allow for a direct comparison. In recent years, there has been a lot of research into the possibility of automating the process of illness diagnosis in GI-related movies and photos. However, text and location overlays often degrade the quality of photographic data [9, 10]. In this study, we show several approaches to preprocessing such pictures and detail our methodology for classifying GI diseases using the Kvasir v2 dataset. Electronic Health Records (EHR) clinical notes provide extensive documentation of patients, allowing for phenotypic inference for illness diagnosis and patient characteristics analysis through cohort selection. Patients may be encoded into fixed-length vectors *via* unsupervised user embedding, which does not rely on human supervision. The retrieved medical ideas from the clinical notes include several links between patients and their clinical classifications. On the other hand, current unsupervised methods of user embeddings from clinical notes do not explicitly integrate medical ideas. Since its inception, coronavirus has caused widespread destruction and altered the human way of life in many ways. Researchers can monitor the public health reaction to COVID-19 with the use of the great instrument which is Twitter. In light of the massive data flood emanating from social media platforms, automated text mining methods may be of great assistance in locating, analyzing, and synthesizing relevant material.

RELATED WORK

Free-text clinical notes were studied in this case study to see whether they may provide light on whether or not a patient's visit was connected to trauma. Using newly retrained GPT-2 algorithms (without OpenAI's pre-trained weightings), we analysed two scenarios. Scenario A included teaching the GPT-2 on a dataset of up to 161,930 trauma- and non-trauma-labeled medical records (26 studies of varied sizes). Scenario B (19 research cases) includes both an unsupervised (self-)pre-training phase containing up to 151 930 unnamed notes and a supervised (expert-)fine-tuning session with up to 10,000 tagged notes. Scenario B needed 10 times fewer notes to achieve the same level of performance (AUC>0.95) as Scenario A did (6,000 *vs.* 600). Despite having sixteen times data (161,930 *vs*. 10,000), Scenario A only improves on Scenario B in the worst-case AUC and F1

score. In conclusion, a general-purpose NLM model like GPT-2 may be modified to become a very effective tool for free-text note categorization with a relatively modest number of labelled examples [11].

In order to fill this void, this work systematically compares and contrasts many zero-shot and similarity-based methods for text categorization of unseen classes. On four text classification datasets—including a novel dataset from the medical domain—various state-of-the-art algorithms are compared and contrasted. Other baselines utilised in prior work provide poor classification results and are readily exceeded, hence new baselines based on SimCSE and SBERT are suggested. When it comes to unsupervised text categorization, the innovative similarity-based Lbl2TransformerVec methodology is offered as a state-of-the-art method. Our results provide a clear advantage for similarity-based methods over zero-shot ones. Similarity-based classification results are further improved by employing SimCSE or SBERT embeddings as opposed to simpler text representations [12].

Several methods are proposed to inpaint issue regions in the photos in order to enhance the anomaly classification, and the impact of this preliminary processing on the raw data is discussed. Experiments conducted by our team reveal that the suggested strategies lead to an improvement in the Matthews correlation coefficient for GI anomaly categorization of about 7% [13]. We present a concept-aware unsupervised user embedding that uses MIMIC-III and Diabetes clinical corpora's text documents and medical concepts together. Extrinsic tasks such as phenotypic categorization, in-hospital mortality prediction, patient retrieval, and patient relatedness are used to assess the quality of the user embeddings. Our method improves upon unsupervised baselines, as shown by experiments on the two clinical corpora, and the addition of medical ideas may substantially boost baseline performance [14]. The current medical dataset on COVID-19, called CORD-19, was preprocessed and annotated for supervised classification tasks in this study. We provide a preprocessed dataset for the scientific community to use during the current COVID-19 epidemic. This may help provide light on how to improve upon certain COVID-19 social interventions. The cleaned-up version of the dataset in question may be found on GitHub [15].

PROPOSED WORK

Problem Formulation

Assume, (x_1, y_1) ,..., (x_N, y_N) indicates the class-labeled dataset used for training, x_i signifies the specific occurrence of data, and $y_i$indicates the name of the class that isx_i. The training data serves as the foundation for the learning system, which uses it to acquire a classifier $P(Y|X)$ or $Y=f(X)$. The categorization algorithm assigns a

CHAPTER 7

Evaluation of ML and Advanced Deep Learning Text Classification Systems

Tarun Kapoor[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Classifying texts into groups determined by their content is called text classification. In this process, automatic labelling of documents written in natural languages is carried out according to predetermined labels. Both text comprehension systems, which perform transformations on texts such as creating summaries, answering queries, and extracting data, and text-retrieving systems, which obtain texts in fulfillment of a user query, rely heavily on text categorization. In order to learn effectively, current algorithms for supervised learning for text classification need a large enough training set. This research introduces a novel text categorization algorithm that uses artificial intelligence techniques (machine studying and deep learning techniques) and needs fewer documents for training than previous methods. To generate a feature set from already-classified text documents, we resort to "word relation," or association rules based on these words. To classify the data, we use the idea of a Convolutional Neural Network with Deep Convolution to the extracted features and then employ a single genetic algorithm approach. The suggested method has been built and thoroughly tested in a working system. The results of the experiments show that the suggested system is an effective text classifier.

Keywords: Advanced deep learning techniques, Machine learning, Text classification systems.

INTRODUCTION

The fundamental issue and severe difficulty of text classification jobs have always been how to apply cutting-edge deep learning technology to construct an efficient and potent text classification model, retrieve text semantic features, and obtain excellent classification outcomes on large-scale test datasets [1, 2]. Due to factors such as the randomness of incident occurrence in time and space, the lack of information at the outset of a reported traffic interruption, and the absence of

[*] **Corresponding author Tarun Kapoor:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: tarun.kapoor.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

advanced technologies in the transportation field to draw insights from past accidents, estimating how long a traffic incident will last is a challenging problem to solve [3, 4]. Abstract Artificial intelligence (AI), machine learning (ML), as well as deep learning (DL), are becoming increasingly popular in many computer vision applications such as object classification, recognition of objects, human detection, *etc.*, but they place significant computational demands on system resources [5, 6]. The widespread and efficient solutions to the high computational requirement are multicore computers and graphics cards having a large number of specialised processing cores. Very few intrusion detection systems (IDS) use ML as their foundation [7]. These kinds of solutions provide a promising new avenue for strengthening network safety in the cyber realm. These tools may identify previously undetectable threats [8]. Researchers and consumers face these assaults in anonymity, making them unconventional. With the tremendous growth of the Internet comes an increase in sophisticated assaults. Machine learning (ML) and deep learning (DL) were developed to help find these assaults [9]. Therefore, several strategies for identifying sophisticated assaults have been presented by researchers. Many alternative systems have been created, based on diverse approaches to extracting features and classifying speakers, making speaker identification techniques one of the most cutting-edge current technologies. Speech recognition is a complex field that calls for state-of-the-art equipment and a wealth of audio data samples before it can be put to practical use [10].

RELATED WORK

The focus of this work is on using deep learning technology to optimize natural language processing (NLP) applications in the areas of text semantics and text categorization. To solve this problem, we present the WC-GCN classification method, which integrates GCN and an attention mechanism. Three common text classification tasks are described as sentiment evaluation, topic analysis, and news categorization. Experimental evidence supports the algorithm model provided in this work. The suggested algorithm model outperforms competing deep learning-based text categorization approaches on three publicly available datasets. The accuracy rate for the news text was likewise rather good [11].

In this paper, we propose a new fusion framework for incident duration prediction that integrates machine learning with traffic flow/speed as well as incident description as features, encoded *via* multiple Deep Learning techniques (ANN autoencoder as well as character-level LSTM-ANN sentiment classifier). The study develops a transport and information science interdisciplinary modelling strategy. When applied to benchmark incident data, the method outperforms the best ML models in terms of accurately predicting how long an event will last. The results demonstrate that compared to traditional linear or vector regression

support vector models, our suggested strategy may increase accuracy by 60%, with an additional 7% improvement relative to a hybrid deep learning auto-encoded GBDT model, which seems to beat all other models. San Francisco is the setting for the application because of the abundance of data it provides on traffic accidents (the Countrywide Traffic Accident Data Collection) and congestion (5-minute accuracy readings from the Caltrans Performance Measurement System) [12].

This book examines the performance of DL algorithms on various multicore hardware architectures. There are three classification issues for which a model using Convolutional Neural Networks has been developed. All of these trials have been conducted on three distinct types of computing hardware—a Raspberry Pi, an Nvidia Jetson Nano Board, and a desktop PC. Each hardware configuration's performance is examined based on its categorization accuracy and hardware response results [13].

In this research, we offer a comprehensive categorization for Machine Learning (ML) & Deep Learning (DL) algorithms. Moreover, the comprehensive review will shed light on the many techniques used in detecting assaults. The foundations of DL and ML lie at the heart of each of these techniques. In addition, a wide variety of systems and devices will be shown for carrying out DL and ML procedures and recommending security solutions that may be applied to the IoT [14].

In this study, a text-based (pre-defined words or phrases) speaker identification system was developed to efficiently and accurately identify the speaker with little effort, training data, and processing power. The first of the system's four key components is the development of audio databases. The research used two different audio datasets, the first being QSDataset and the second being audioMNIST_meta. The database configuration and processing methods are detailed in the main study text. The second phase of the study involves extracting the features using an algorithm based on pitch coefficients, and the third phase involves using a network of neurons as a classification. The study is complete after the system's performance and output have been confirmed. The test results proved that the MNN network could handle even a small amount of data since it obtained a perfect score. For massive datasets, the range was 80%-81%. In contrast to the CNN network, findings for small samples varied from 60% to 76% negative, whereas for big samples, the positive range was from 91 to 96%, and the outcomes were better than the existing approaches [15].

CHAPTER 8

Machine Learning Method Employed for the Objective of Identifying Text on Tweet Dataset

Sakshi Pandey[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: When it comes to training ML systems, internet-based data is invaluable. Despite the difficulty in collecting this information, teams of experts from academic institutions and research labs have created publicly accessible databases. Twitter and other social media platforms provided large quantities of useful information throughout the pandemic, which was used to evaluate healthcare decisions. In order to forecast illness prevalence and offer early warnings, we suggest analysing user attitudes by using efficient supervised machine learning algorithms. The gathered tweets were sorted into positive, negative, and neutral categories for preprocessing. Hybrid feature extraction is the innovative aspect of our work; we used it to correctly describe posts by combining syntactic features (TF-IDF) and semantic elements (FastText and Glove), which in turn improved classification. The experimental findings suggest that when using Naive Bayes, the combination of FastText and TF-IDF achieves the best results.

Keywords: Fast text model, Machine learning, Text identification, Tweet dataset.

INTRODUCTION

Recent advances in machine learning and deep learning enable novel methods of problem resolution and have far-reaching, game-changing effects across many sectors. The present paradigm shift has affected architectural practices similar to other fields. Hybrid methodology was used to investigate the impact on construction practises [1]. First, a text-mining technique is used for content analysis to undertake a comprehensive literature review of the field [2]. A computerised method of locating and labelling opinions expressed in the literature may help readers gain insight into the author's perspective and comprehension of a topic [3]. The most used language for sentiment analysis is English. However, it has been claimed that several approaches to sentiment analysis exist, preserving

[*] **Corresponding author Sakshi Pandey:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sakshi.pandey.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

the study as a fascinating research into Indonesian writings [4]. Several scholars have resorted to an algorithm for machine learning in an attempt to handle the sentiment analysis difficulty [5], thanks to the broad accessible Twitter data from previous years and the establishment of machine learning models. For a long time, telecom fraud has proven to be a serious financial issue for Indian internet and phone users. Attempts to uncover telecom fraud have traditionally revolved around warning users not to dial unknown numbers [6]. However, hackers may avoid detection by using VoIP (Voice over IP) phone lines to masquerade as legitimate callers [7]. Using the substance of a caller's voice rather than just their sign, this method may be able to detect telecommunications fraud [8]. Spam detection is crucial now more than ever because of the decentralised nature of social media and the ease by which incorrect information may be spread by people. Spam filters provide protection against this kind of fraud [9]. The amount of content released on Twitter and other social media sites about cybersecurity has increased significantly. When correctly analysed, this data might be utilised to create a system for identifying and counteracting cyberattacks. We propose employing deep learning to examine tweets in this research. The numerical content of tweets may be represented in a variety of textual ways [10]. By feeding these features into a machine learning architecture, better feature extraction and classification is possible.

These feelings are highly helpful in developing more rapid disease monitoring systems. However, the link between syntactic as well as semantic data and ML approaches based on feature types was not considered by many research works that helped evaluate tweets written in English. To anticipate and track the spread of COVID-19, we used supervised machine learning techniques to analyse Twitter sentiment about a dataset related to the virus. For the syntactic analysis, we utilised TF-IDF N-gram and word level, and to perform the semantic evaluation, we used Word2vec, FastText, and Glove. This study primarily focused on the following areas:

- Five cutting-edge feature extraction methods, comprising TF-IDF N-gram, TF-IDF community college-gram, Word2vec, Glove, and FastText, are demonstrated in this article. We also provide two state-of-the-art approaches using a TF-IDF strategy: the TF-IDF and Glove methods, and the TF-IDF and FastText approach.
- Using a variety of feature extraction techniques, we evaluate the efficacy of several machine-learning approaches for classifying tweets written in English.
- Many different machine learning classifiers have been studied, and the best combination and fusion of methods to boost their performance have been chosen.

RELATED WORK

This allows for the methodical analysis of current developments and a debate on potential future directions in this area of research. Second, a Scientology study using bibliometric reviews may provide quantitative estimates of global research in the aforementioned area. We were able to get insight into research patterns and pinpoint the most important networks in this dataset by analysing the co-occurrence of terms, scientific collaborations, geographic distribution, and co-citation analysis. Finally, we discuss the pros, drawbacks, and future research directions for using data mining in the building sector based on our findings throughout the study [11].

Using Sentiwordnet and machine learning, this study aims to develop a sentiment analysis model for the next general election in Indonesia. The text and information of the tweet were collected. Joko Widodo and Prabowo Subianto, two contenders for becoming the president in 2019, were the main topic of the tweet. From November 13th, 2018, to January 11th, 2019, data was gathered for the campaign. Indonesian was used for the tweet. The Nave Bayes classification approach was used to analyse public opinion, and the results showed an accuracy of 74.94% with regard to the Joko Widodo issue and 71.37% for the Prabowo topic [12].

Thai civil case outcomes are often recorded in a way that lacks any real structure. A Thai civil case rationale document consists of four primary sections: the conflict, the truth, the decision, and the judgment. The first stage in creating a text synopsis or information extraction from a Thai civil case decision document is identifying the most essential elements of the document. Given this deficiency in the scientific literature, they set out to remedy it by devising a text categorization-based method for extracting four essential elements. We collected the dataset from the Court of Appeals of Thailand's website (http://www.supremecourt.or.th), whereupon we used two weighting methods and three supervised AI algorithms. The key parts of Thai civil court judgement documents were properly recognised in memory, reliability, and F1 tests [13].

This article gathers examples of misinformation spread *via* the media and online platforms. To evaluate current data and choose superior dataset descriptions, our suggested approach employs machine learning techniques. The next step uses Natural language processing to extract characteristics from the incoming text. For further telecommunications fraud detection, the criteria to distinguish similar content in the same call are then developed. Customers may protect themselves against telecoms fraud online by downloading an Android app that works in tandem with the system. The programme does real-time analysis of each call for

CHAPTER 9

Textual Classification Utilizing the Integration of Semantics and Statistical Methodology

Ayush Gandhi[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Effectively classifying texts is possible using several classification techniques. Machine learning constructs a classifier by studying and memorising the characteristics of several classes. For text categorization, deep learning provides similar advantages since it can function with great precision using simpler architecture and processing. In order to categorise textual information, this research makes use of machine learning and deep learning methods. There is a great deal of extraneous details in textual data that must be removed during pre-processing. To prepare it for analysis, we remove duplicate columns and impute missing data. In the next step, we use deep learning techniques for classification, including long short-term memory (LSTM), artificial neural network (ANN), as well as gated recurrent unit (GRU). According to the findings, GRU obtains 92% accuracy, which is higher than that of any other model or baseline investigation.

Keywords: Deep learning technology, Machine learning, Semantic features, Statistical features, Text classification.

INTRODUCTION

Using data mining techniques to analyse photos is just the beginning of image mining. The term "web image mining" refers to a set of tools used to extract data from online images. The demand to mine image data online is expanding as conventional online mining increasingly concentrates on quantitative and textual data. There is an abundance of information available on the web, both written and visual. Web image mining is not as popular as text data analysis [1] due to how difficult it is to tackle the semantics of photos. Real-world data mining operations need efficient methods for building models that predict from extremely structured relational data. In this work, we take on the problem of learning classifiers utilising structured data that is relational and annotated with relevant metadata [2].

[*] **Corresponding author Ayush Gandhi:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: ayush.gandhi.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

We show how to train classes at different levels of abstraction in a relational structure where relational information is organised in a hierarchy of abstractions that characterises the semantic content of the data [3]. When working with structured data, we show how to handle incomplete specification, which arises as a natural result of choosing a certain level of abstraction [4]. There is a nonlinear relationship between a woman's mental state—however subjective—and her labour, which could result in eugenic features, according to studies in obstetrics [5]. Because of these geographical differences, there is a growing interest in examining the connection among survey data from different areas. Research in conventional obstetrics has focused on trying to deduce the health implications of this state of mind [6]. However, much of the research done till now has focused on numerical relationships using statistical methods. This method sometimes runs into trouble while analysing data because of its inability to understand the underlying semantics of the data, leading to problems like data divergence and missing information (such as location data). The idea of textual content as well as semantics may be an obstacle for those seeking to make sense of the chaotic environment of information retrieval, categorization, and filtering. Interestingly, most of us, including myself, continue to consider that developing a linguistically principled approach to text categorization is an interesting research problem [8]. This is relevant to the main point of the book review, as you shall see. Thorsten Joachims proposes a framework for the automatic learning of categorization of text models and training to categorise texts utilisingSupport Vector Machines [9]. The analysis of texts for extracting data, fact, and semantics provides details in asimple machine processing form (such as ontology). It also provides text classification and clustering, including thematic modelling; and retrieval of data (including thematic search), which are all areas that can benefit from knowledge extraction techniques applied to large volumes of conversational texts [10].

Naive Bayes's flaw is that it treats categorical variables with zero probability and treats all characteristics as independent. Below are a few of this paper's key contributions.

- Utilising the Long-Short Term Storage algorithm, suggest a method for effectively classifying textual material.
- To choose the optimal classifier, we use a wide range of machine learning as well as deep learning techniques.
- The findings demonstrate that GRU works well since it has numerous hidden layers, retains relevant information while discarding irrelevant data, and has a firm grip on the data we provided with a considerable accuracy of 95%.

The remainder of the paper is organized into the following subsections. In the second section, we obtained a synopsis of the relevant research. Then, in Chapter

3, we explain in detail the procedures and techniques utilised throughout our investigation. The paper's findings and discussion can be found in Section 4, while its conclusion and suggestions for further research can be found in Section 5.

RELATED WORK

This study offers a novel method for picture identification and classification by automatically collecting a large number of photos from the Internet to use as training data. The system uses content-based image retrieval (CBIR), that does not restrict target images like traditional image recognition methods, and support vector machine (SVM), among the most effective as well as widely used statisticians for general image categorization that fit to the learning tasks. The proposed method outperforms numerous existing search methods, according to experimental results such as support vector machines, image categorization, image collecting, and Internet image mining [11].

Our statistical strategy for partial specification is based on shrinkage. Results from experiments haveshown that (i) how the level of conception chosen can affect the performance of ensuing connect-based classifiers and (ii) how to investigate the impact of partially specific data on learning relate-based naive Bayes estimators for an article classification task [12].

To this end, we look into the extraction of the semantic relationship between data of dubious quality in the wellness information domain by using a bipartite approach called Geo-SPS and generic graph illustrations on geospatially-enhanced obstetrics and obstetric surveys. Our method is novel because we map semantic objects into a graph with two parts using principles from graph theory that were initially created for use in computerised physics. A new approach to determining semantic similarity is provided by Geo-SPS. Using this method, a graph of a convolutional neural can rapidly and accurately evaluate ambiguous textual input across several areas. We further evaluate and establish the practicality of Geo-SPS utilising a case investigation involving obstetric surveys along with medical information data from over 3,000 women across three different areas in China. The findings of this research show that Geo-SPS can reflect the mother's mental health during pregnancy and reliably detect birth defects from this diverse data collection. [13].

Topic text categorization (TC) is proposed as a solution because it is based on a wide language generality of (what appeared to be) a linguistically dependent activity. When applied to linguistic objects (such as the documents), the outcome is a basic linguistic model called the bag-of-words representation, which still achieves remarkable accuracy [14]. While most text categorization research relies

CHAPTER 10

The Use of Machine Learning Techniques to Classify Content on the Web

Dikshit Sharma[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: In text categorization, texts are sorted into groups according to their content. It is the process of automatically classifying texts written in natural languages according to a set of guidelines. Both text comprehension systems, which perform transformations on text such as creating summaries, answering queries, and extracting data, and retrieval of text systems, which collect texts in reaction to a user query on the internet content, rely heavily on text categorization. In order to learn effectively, current algorithms for supervised learning for text classification need a large enough training set. This research introduces a novel text categorization system that makes use of an AI approach and needs fewer articles for training over information found on the web. To generate a feature set from already categorised texts, we resort to "word relation," or association rules based on these terms. The obtained characteristics are then processed by a Support Vector Machine, and ultimately, a single genetic algorithm idea is introduced for classification. The suggested approach has been developed and validated in a working system. The experimental results verify the effectiveness of the proposed system as a text classifier.

Keywords: Machine Learning, SVM and TF-IDF, Web Content Classification.

INTRODUCTION

Thanks to technological developments, routine jobs in biomedical research and development are getting easier to do. In order to read the glucose level on the urine strip, a urine analyser detector may be utilised [1]. Your vision may be impaired due to cataracts, trachoma, corneal ulcers, or another eye condition. These four eye diseases may be halted in their tracks only with proper early identification [2]. Various eye diseases manifest themselves visually in a wide variety of ways [3]. The accurate diagnosis of ocular disorders, however, requires the consideration of a wide variety of symptoms. Due to the severity of the

[*] **Corresponding author Dikshit Sharma:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: dikshit.sharma.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

problem, researchers have been working hard to develop better diagnostic and therapeutic strategies for plant diseases [4, 5]. To increase agricultural production and income, farmers would greatly benefit from the ability to make use of existing equipment to remove the barriers to agricultural productivity [6]. Tools come in many kinds and sizes, and humans have been using them ever since the beginning of time [7]. The ingenuity of humans has allowed for the creation of several helpful implementations. This provides the resources for individuals to address a broad variety of needs, including those in the areas of transportation, industry, housing, and computer [9]. In the current day, we serve as both creators and consumers of knowledge. Users may have a substantial impact on the future of digital information by contributing their ideas and opinions in the social media sphere [9]. These points of view include a wide range of feelings. Computer science studies for recognising and obtaining the sentiment underlying textual input constitute sentiment analysis, a technique for classifying public opinion stated in online forums [10].

Website categorization is a significant challenge with several real-world implications. Parental controls provide a child-friendly online environment. Without clear guidelines for website classification, many Internet users of all ages are exposed to inappropriate content and have a more difficult time finding what they need. There are no comprehensive or adaptable internet security options available right now. The courts' banning orders also do not apply to all problematic sites, and the domains of these sites often change hands. This highlights the significance of dynamic website categorization based on text data. This research involves the categorization of webpages using NLP and ML methods. Websites written in a variety of languages have their content gathered and preprocessed in preparation for machine learning. The research included a total of 17 classes, with the SVM technique yielding the greatest success rate in the classification of 0.8756.

RELATED WORK

The goal of this research is to utilise the MATLAB programme to analyse glucose data from digitally photographed patients. Injections of insulin are necessary for the control of blood sugar levels in diabetics. Injections may cause minor physical harm, which may weaken the body's defences against infection. A lot of research has gone into developing noninvasive methods of detecting glucose, particularly those that use urine. In this study, image processing was used to examine the process of glucose testing that does not include a needle. Noise is reduced using a Gaussian filter and histogram-based feature extraction to facilitate information extraction from picture databases. The support vector machine classification allocates 70% of its time to training and 30% to testing. The processing time for

the SVM classification results was under 0.5 seconds, and they were 85% correct. Medical decisions may be made in light of a person's diabetes, prediabetes, or lack thereof [11].

Utilising digital image processing, machine learning techniques, deep convolution neural network (DCNN) and encouraging vector machine (SVM), we describe a novel approach to automatically detect eye illnesses based on outward symptoms. We employ t-distributed stochastic neighbour embedding (t-SNE) and principal component analysis (PCA) with variables to improve feature selection. The proposed method dissects a frontal face image into its constituent components and automatically removes the eye area. The proposed method diagnoses and classifies a wide range of eye conditions, including but not limited to glaucoma a condition called conjunctiva cornea ulcer, ectropion, periorbital cellulitis, bitot spot of vitamin A deficiency. The experimental results favour the DCNN strategy over SVM models. We also compare our method to popular alternatives. Our method outperforms the closest rival by a factor of 14. An overall precision of 98.79%, a sensitivity rating of 97.1%5, and a specificity of 99.0% are all achieved by our DCNN model [12].

In this work, we used ML techniques to create a mobile app for diagnosing plant diseases and suggesting therapies. A filtering based on recommendation algorithm was then used to suggest treatments for the observed plant ailments, after which ANN and KNN were utilised to classify the illnesses. The results of the implementation confirmed that plant diseases were correctly detected and treatment recommendations were given [13].

Current customised internet search strategies do not take into account unvisited sites that may supply simple responses to the client's information requirements. A page in the results package may not solve the user's problem entirely, but it may point them in the right direction. Only by doing a semantic analysis can the concealed connections be uncovered. This study aims to classify particular connected sites by doing a semantic analysis of the search route and providing an efficient individualised site search. By providing context and a direct link between a user's search query and related websites, this facilitates online discovery. [14].

To identify the tone of Bangla-language Facebook posts about Bangladesh Cricket, this paper presents a sentiment polarity detection method that employs three well-known supervised machine learning algorithms: Naive Bayes (NB), support vector machines (SVM), and logistic regression (LR). Classifiers' performance is also examined and contrasted; employing n-gram as an attribute, LR achieved a higher accuracy (83%) than SVM and NB. [15].

CHAPTER 11

Lexical Methods for Identifying Emotions in Text Based on Machine Learning

Mridula Gupta[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The study of emotions has emerged as an important area of research because of the wealth of information it can provide. Emotions can be expressed in a variety of ways including words, facial expressions, written material, and movements. Natural language processing (NLP) & deep learning concepts are essential to solving the content-based classification problem that is emotion detection in a text document. Therefore, in this research, we suggest using deep learning to aid semantic text analysis in the task of identifying human emotions from transcripts of spoken language. Visual forms of expression, such as makeover jargon, may be used to convey the feeling. Datasets of recorded voices from people with Autism Spectrum Disorder (ASD) are transcribed for analysis. However, in this paper, we specialize in detecting emotions from all of the textual dataset and using the semantic data enhancement process to fill a few of the phrases, or half-broken speech, as patients with Autism Spectrum Disorder (ASD) lack social contact skills due to the patient not very well articulating their communication.

Keywords: Lexical methods, Machine learning, Speech emotions, Text classification.

INTRODUCTION

Whether expressing joy or sorrow, the ability to communicate effectively *via* speech emotion is essential in human interaction. The fraction of languages that express sentimentality varies significantly among areas. In the realm of machine learning and artificial intelligence, emotion recognition is a relatively recent technique. In this publication, we discuss the research behind and performance metrics for an emotion classification system. Enhancing interaction between humans and computers *via* speech-emotion recognition requires carefully selecting relevant signal features and constructing relevant classification models

[*] **Corresponding author Mridula Gupta:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: mridula.gupta.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

[1, 2]. The total recognition rate will decrease as the number of individuals who use several speech emotions increases [2]. To solve this problem, we provide a Fisher feature selection-based choice tree support vector machine (SVM) method for speech emotion recognition [3]. During feature selection, the parameters with higher discriminatory power are omitted using the Fisher criterion [4]. During the stage of emotion classification, an approach is proposed to set up the structure of the choice tree [5]. Decision tree SVMs allow for simultaneous coarse and fine classification [6]. When extraneous information is taken out of the equation, emotion detection becomes more accurate. The capacity to recognise emotions *via* a speaker's voice is becoming an increasingly important area of study [7]. The breakthrough introduces a semantic cells-based method for identifying emotions conveyed in one's speech [8]. The speech emotion recognition process involves constructing a voice library, preprocessing each word indicated in the audio libraries, obtaining the emotional characteristics of every conversation signal, and establishing a vector of traits of every language signal using the extraction outcome, using instruction of the carriers of capabilities to obtain a mixture of models based upon semantic cells, and the data gathered through combined models as a recognition model of a class [9, 10].

RELATED WORK

We provide a method for deducing the emotional state of a speaker from their words alone. This approach, which employs a feature-selection technique called XGBoost, makes use of a convolutional neural network (CNN) trained on identifiers and a BLSTM (bi-directional long short-term memory) model with a focus on details. It is important to compute the importance score of every characteristic to alleviate the over-fitting problem, reduce training expenses, speed up the procedure of modelling, and improve identification accuracy. Results from experiments conducted on the EmoDB, CASIA, and EMA corpora show that the proposed framework outperforms a baseline model in terms of accuracy in forecasting (86.87-74.17-98.04%) [11].

After establishing the decision tree SVM framework using the emotional perplexity measure, we choose features with higher differentiation ability for each SVM in the tree using the Fisher criterion. This method at last makes it possible to automatically identify emotions conveyed *via* spoken language. To test the reliability of our method, we construct a decision tree SVM with Fisher feature selection and apply it to the CASIA Chinese expressive speech corpora as well as the Berlin speech corpus. The experimental results show that compared to the standard SVM method of categorization on CASIA and the Berlin speech corpus, the proposed approach increases the average emotion recognition rate by 9%. The

proposed method [12] improves emotional recognition accuracy while reducing emotional confusion.

Using machine learning-based strategies for identifying emotions, this paper presumes that determining deltas as well as delta-deltas includes not only useful data but also reduces the impact of emotionally insignificant factors, thereby improving accuracy. Silent frames or unnecessary emotional frames are another common problem with Speech Emotion Recognition (SER). Meanwhile, the focus is very helpful for learning the specific feature representation that is necessary for success in a certain task. In order to generate discriminative features for SER, we are motivated to develop an Attention-based Convolutional Recurrent Neural Network (ACRNN) that takes as input the Mel-spectrogram with deltas and delta-deltas. The experimental findings are detailed in the conclusion, and they indicate that the suggested strategy is effective and achieves the highest accuracy in terms of release-weight average recall [13].

Users are able to identify both the speaker and their mood thanks to a voice emotion detection model built from a mixture of models according to two distinct sets of semantic cells. When applied to speech emotion identification, the recognition model built using the semantic cells-centered approach achieves high accuracy, low memory requirements for storing the r, and fast recognition times [14].

The method presented here uses Berouti spectral removal to restore clarity to muddled audio recordings. Utilising Fast Fourier Transform (FFT) and spectral flux, it extracts the clean speech spectrum from the noisy speech spectrum. The signal-to-noise ratio (SNR) was used to calculate the overall quality of this research. In this investigation, a novel parameter is presented to enhance voice quality. It is suggested to combine spectral reductions with spectral flux for a more precise noise estimation. The new parameter improves the power ratios in spectral subtraction, and unvoiced word frames benefit from this method as well [15].

PROPOSED WORK

Research Gaps

It appears difficult to deduce a person's emotional condition by reading their written language. Human-computer interaction (HCI) relies heavily on determining the reader's emotional state from the text. There are three types of emotional communication used by humans: facial, text-based, and speech-based. Research on text-based emotional recognition systems is necessary since sufficient work has been performed on the face and spoken emotion detection.

CHAPTER 12

Identification of Websites Using an Efficient Method Employing Text Mining Methods

Madhur Taneja[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Herein, we introduce a method for website classification using deep neural networks and mixed data extractors. We use iterative training as well as supervised learning approaches to use a gradient descent methodology to simulate the website categorization. This modern model is comprised of a webpage encoder, a convolutional neural network (CNN) feature extraction, a bidirectional long short-term memory (LSTM) feature extractor, as well as a fully connected classifier. It may retrieve various website features at various granularities. Our model may quickly select a suitable website class by concatenating mixed features obtained from mixed feature extractors. On the realistic website dataset that has been obtained, we conduct in-depth tests. The dataset is compiled using domains that were taken from the telecom operator's DNS records. The proposed categorization schema outperforms state-of-the-art models in comparison to our fresh model as well as a slew of popular machine learning algorithms in terms of accuracy, recall, F1, and precision. Other web apps may benefit from all of this as well, such as detecting fake websites as well as ads.

Keywords: CNN, LSTM, Text Mining, Website Classification, Website Classes.

INTRODUCTION

Website organization is an increasingly common and significant task for a variety of applications, including web security, medical, finance, and other fields. However, there are still significant hurdles to improving the accuracy of classification, which makes it challenging to put into reality. Keyword matching is a method that has been used to overcome this issue. To indicate various categories, they employ specified terms. However, it is impossible to predefine every single keyword required for the classification process. In this instance, some earlier studies used some fresh strategies to complete this work. Initially, this

* **Corresponding author Madhur Taneja:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: madhur.taneja.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

assignment is given to certain traditional machine learning techniques like support vector machines as well as naive Bayes. The bulk of the prior literature on harmful website detection just classifies websites as either harmful or harmless. Because researchers may create and instruct distinct harmful detectors for different types of websites, knowing in advance which category a given website belongs to allows the harmful detectors to offer more precise detection findings. A good website in the "Business & Economics" area, for instance, can appear to be malevolent in the "Health Care" category. Consequently, finding malicious websites relies heavily on the work that we do to categorize websites. Utilizing a combination of deep neural network–based feature extractors, it achieves greater precision in classification. Deploying class-specific malicious detectors to websites might lead to more accurate detection results. Most of the prior research on website classification relied on traditional machine-learning techniques trained on publicly available datasets. Examples of multiple classification tasks that are too complex for some machine learning algorithms include naive Bayes as well as Support Vector Machines. We address this problem by creating a hybrid feature extraction network based on cutting-edge methods like text convolution-gated recurrent units. This network can learn from large datasets in addition to having a voracious capacity for complex input data. We employ website datasets obtained using domains derived from DNS Records as opposed to open datasets, which are a good picture of typical Internet user activity. Latent semantic analysis (LSA)-based text mining requires a lot of storage and processing time [1]. In this article, we go over SyntacticDiff, an innovative, comprehensive, and successful edit-based method for changing word sequences that uses a specific text corpus as a reference [2]. These transformations can be used directly or as features to represent text data in a variety of text-mining applications. The UK Educational Evidence Portal (eep) offers consumers a centralized location from which they may peruse the websites of 33 different organizations, with the goal of transforming the way the education community conducts business. Using the gateway is faster than looking for each piece of information separately. Still, the content of the pertinent websites is made up entirely of almost 500,000 publications [5]. This suggests that the search feature of the site may provide a huge amount of results. Exposure to technical knowledge, news, ads, commercials, electronic commerce, and other information services may be found worldwide through the World Wide Web (WWW) [6]. This makes information [3, 4] retrieval exceedingly difficult [7]. The majority of users usually do not have a thorough grasp of network architecture, so if they have to make too many access hops, they can get impatient while waiting for their information to arrive. Data mining techniques such as web mining have offered successful answers to these issues [8]. Web mining is a data mining technique used to analyze material from the Internet. Standard data mining techniques are used to extract information from

the World Wide Web, which is subsequently incorporated into the functioning of the website. Finding useful data patterns on the internet and extracting them for more in-depth research is the aim of internet mining. Web structure mining, which examines the web of connections connecting sites, web content mining, which examines the content of individual pages, and web use mining, which examines traffic information from server logs, are the three categories that make up web mining.

RELATED WORK

This study suggests a unique text extraction method that offers a framework for the effective use of statistical semantics research in the text extraction process. This method uses centrality to ignore passages of text that are strikingly similar to others in terms of phrasing, data, or meaning. This novel multi-layer similarity technique to similarity detection uses the vector space model and the Jaccard similarity in the first and second layers, and the LSA in the third. Only segments that the first two layers were unable to identify can be compared in the third layer. We created a new assessment method that considers the extract size because the Rouge tool does not consider this. It considers the fraction of sentences that cross in both the automated and reference extracts, as well as the compression rate. In comparison to traditional LSA and conventional statistical extractions, we achieved good accuracy results, reduced the initial matrix dimensions by 65%, and required 52% less time for LSA calculation. The centrality feature, as proposed by the multi-layer design, is proven to significantly increase the effectiveness and precision of text extraction [11, 12].

By using SyntacticDiff for three different use cases—correcting grammatical mistakes, grouping and analyzing student essays, and identifying native languages—we show the tool's value. Since SyntacticDiff is language-neutral, it may be used for any corpus of textual information in any natural language. It is quick, adaptable, and capable of identifying phrase, paperwork, and sub collection-level syntactic differences between a target text collection and a reference one. Therefore, a more complex translation technique and feature representation may be advantageous for many text mining tasks outside of bag-of-words that deal with word usage and grammar [13].

Since users value their time, it would be beneficial if there were better ways to carry out searches and show results so they could rapidly focus on the exact information they were seeking instead of trawling through lengthy lists of papers that were not pertinent to their requirements. The Joint Information Systems Committee (JISC), which is funding the ASSIST project, has created a prototype web interface to demonstrate how particular text-mining algorithms as well as

CHAPTER 13

Machine Learning-based High-Dimensional Text Document Classification and Clustering

Ansh Kataria[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Text classification is a difficult technique. Many techniques have been developed to decrease the dimension of feature vectors for use in text classification due to their enormous size. This work provides a detailed discussion of unique parameters utilising an optic clustering strategy, as well as a review of some of the most essential text categorization algorithms. In this case, the words are clustered according to their level of similarity. Each cluster's membership function is based on the mean along with the standard deviation of its data. Finally, characteristics are chosen from each grouping. Each cluster's extracted feature is the weighted sum of its words. There's also no need to guess or use trial-and-error approaches to determine the optimal number of clusters.

Keywords: Clustering, High dimensionality text, Machine learning, Text document classification.

INTRODUCTION

Classifying and clustering text documents is a significant learning challenge that has applications in data mining and machine learning [1]. When dealing with high-dimensional text materials, the learning job becomes complex with difficulties. The order in which words appear in written materials is an important factor in the learning process [2]. Document sizes are a serious problem and a troubling indicator [3], despite the fact that High Dimension documents are employed for categorization. Both beneficial and detrimental results may result from dimension reduction [4]. Document categorization using reduced dimensions is ineffective if the reduced dimensions are not in the right format [5]. While the basic similarity characteristic has been the focus of our study, we have yet to address text categorization [6]. The use of text mining applications increased in

[*] **Corresponding author Ansh Kataria:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: ansh.kataria.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

significance across several intellectual domains [7]. Many vital processes in many international organisations and different ethnicities are recorded mostly in text documents. Text poetry is a significant kind of media in the fields of culture and education. There are just a few areas where the classification of Arabic text poetry has been in trend, but it has had a significant impact [8]. The rhythmic harmonic measure that distinguishes the various styles of Arabic poetry [9] is the same in both modern and classical Arabic poetry. When it comes to text categorization jobs, deep learning is one of the most effective machine learning techniques. There are various information retrieval activities that benefit from document clustering, including document browsing, organisation, and display of retrieval results. They are being studied extensively on a worldwide scale at the present time. There has been an extensive usage of generative models for text categorization based on multivariate Bernoulli as well as multinomial distributions.

RELATED WORK

The vast majority of the current work on the topic of high-dimensional text document categorization and group relies on the use of classical distance functions without taking into account the word distribution in texts. For the purpose of feature pattern grouping and high-dimensional text categorization, we suggest a new similarity function in this study. The suggested similarity function is utilised to implement dimensionality reduction through supervised learning. What makes this study special is that the term distribution is unchanged prior to and following dimensionality reduction. The experimental findings show that the suggested technique succeeds in reducing the dimensionality of the data, maintaining the word distribution, and obtaining higher classification accuracies than existing techniques.

Several techniques for performing the tasks of classification and clustering are discussed. In this research, we used classification and clustering techniques across a number of datasets. It also included recommendations on how to make Naive Bayes classification and K-means clustering more effective. The effectiveness of the suggested approach is measured using standard indicators including accuracy, recall, and F-score. The experimental results would show that the suggested model is superior to existing techniques.

This study provides a taxonomy of Arabic poetry texts. In order to improve the effectiveness of models, we offer a specialised feature selection that is merged with a clustering technique. Two well-known machine learning techniques, the support vector machine as well as the decision tree, have been used in tandem with deep learning experiments. The suggested feature technique of extraction has

shown excellent precision across all three approaches. The outcomes are superior to many other comparable studies.

In this paper, we substitute the traditional k-mean process's Euclidean distance-based metric with a machine learning approach to create a novel hybrid algorithm dubbed MLK-Means for grouping TMG format document data. Standard k-means with L-2 normalised data and the von MisesFisher prototype-based cluster (vMF-based k-means) were used to evaluate the performance of the proposed approach. The MLK-Means Technique is implemented in this suggested study, and its performance is contrasted to that of the aforementioned techniques. The suggested algorithm has larger and more similar improvements.

In this research, the LDA model with author n-gram texts and cosine similarity is used to offer a unique technique for author authentication in English and Urdu literature. Similarity measures are used to determine which of the many learned representations of stylometric traits best fit a given author's writing style. Classifications of an author's text are prioritised using instances and profiles in the proposed LDA-based Technique. In this case, LDA's ability to represent text more expressively makes it a good fit for dealing with high-dimensional and sparse data. The technique provided here is an unsupervised computing strategy that can deal with the varied dataset, the wide variety of writing styles, as well as the inherent uncertainty of the Urdu language. The provided technique has been put through its paces using a large corpus. Experiments validate the efficacy of the proposed technique compared to previous authorship identification algorithms and state-of-the-art representations. Cosine similarity using n-gram-based LDA themes is used to evaluate the similarity of document vectors, which is one of the offered work's contributions. The authorship identification job was completed with an overall accuracy of 84.52% on the PAN12 datasets and with an accuracy of 93.17% on Urdu news items without the use of any labels.

PROPOSED WORK

Background

Classifying and labelling unprocessed text according to its subject matter is what we mean when we talk about “Text Classification.” Text classification has a wide range of applications, from news subject labelling to user review sentiment analysis. For Instance, “Phone was awful. Super sluggish. Bloatware and intrusive advertisements were at an all-time high. This is not a phone I would suggest to anybody. Based on the aforementioned language, our classification machine may assign the correct category or tag, such as bad reviews, in this example.

CHAPTER 14

The Application of an N-Gram Machine Learning Method to the Text Classification of Healthcare Transcriptions

Pratibha Sharma[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: An integral aspect of natural language processing is text categorization, the goal of which is to assign a predetermined category to a given text. Feature selection and categorization models come in a wide variety of forms. Most researchers, however, would rather utilise the prepackaged functions of existing libraries. In the field of natural language processing (NLP), automated medical text categorization is very helpful for decoding the information hidden in clinical descriptions. Machine learning approaches seem to be fairly successful for medical text categorization problems; nevertheless, substantial human work is required in order to provide labelled training data. Clinical and translational research has benefited greatly from the computerised collection of vast amounts of precise patient information, including illness status, blood tests, medications taken, and side effects, along with therapy results. As a result, the medical literature contains a massive amount of information on individual patients, making it very difficult to digest. In this research, we suggest using N-grams and a Support Vector Machine (SVM) to classify healthcare-related texts. We conduct experiments to determine the viability of our code and analyse it across a variety of categorization methods.

Keywords: Medical text classification, Machine learning, Natural language processing, N-gram methods.

INTRODUCTION

It is crucial to ensure that students have access to adequate laboratory facilities. Using star and solar photometers installed at Lindenberg Meteorology Observatory since 1995 [1, 2], we were able to calculate the atmospheric water vapour concentration (or Integrated Water Vapour, IWV). Several strategies for calculating the required empirical parameters for the recovery are explored [3].

[*] **Corresponding author Pratibha Sharma:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: pratibha.sharma.orp@chitkara.edu.in

The VKM-100 multi-pass pressure cell at Pulkovo Observatory was used to individually calibrate each instrument [4, 5]. For several model atmospheres, the MODRAN-4 software tool was used to simulate absorption by the atmosphere by water vapour, allowing the empirical values to be computed. Surgery affecting the eyelid, cornea, conjunctiva, camera lens, ocular muscle, vitreous, and iris are all included in the broad category known as "ocular surgery" [6, 7]. Other surgeries include those for the elimination of tumours, the treatment of ocular injuries, and the replacement of damaged corneas [8]. Avoiding surgical site infections (SSIs) during ocular surgery is complicated by the lack of standard regimens for antibiotic prophylaxis. Due to the lack of agreement on the appropriateness of ocular antibiotic prophylaxis for children, this consensus paper aims to provide doctors with a set of guidelines on the topic [9]. The following scenarios are taken into account: Common eye surgeries include corneal grafts, ocular surface transplants, ocular surface grafts, intraocular surgeries, and extraocular surgeries for things like trauma and tumours. Models of stars indicate a strong reliance on convection, with the bolometric flux variation characterised by the complex variable f [10]. Global longitudinal strain (GLS) of the left ventricle (LV) has established as a more objective and reliable method for measuring LV systolic function during the last decade, and it may be able to detect slight abnormalities in LV contraction even in the presence of an intact ejection fraction (EF). Recent research, however, has shown that GLS, like LV EF, is highly dependent on load. In recent years, non-invasive methods of quantifying myocardial work (MW) have developed as a viable option for gauging heart health and performance.

Most previous studies relied on either explicit or implicit representation of text to resolve such issues, but both methods are best for sentences and do not readily apply to short text due to their brevity and sparseness. In light of the fact that traditional multi-size filter Support Vector Machine (SVM) during text classification task obtains the simple word vector feature and ignores the important feature, we present a type of short text classification approach by SVM, which can acquire numerous texts including adopting none linear sliding technique and N-gram language model, as well as selects key features by employing the concentration mechanism, in addition to employing the entropy loss function. Experiments comparing the suggested technique to the standard machine learning algorithm demonstrate a significant improvement in categorization results for short texts. Here is a rundown of everything we have accomplished with this paper as a whole.

In this study, the authors introduced a novel classification strategy that combined N-gram with support vector machine (SVM) technology.

To get text characteristics, we used a concentration process based on an N-gram language model, tweaked many superparameters, and drew benefits from the model as a whole.

By enhancing the pooling procedure, our technique can maintain useful characteristics as soon as feasible, which in turn promotes an increase in the accuracy of the text classification job.

Our suggested methodology is detailed and analysed in Section 3, whilst Section 4 discusses experimental findings. Section 2 provides background information on the SVM model that will be used in Section 3. This paper is brought to a close in Section 5.

RELATED WORK

It was suggested that students studying various control engineering disciplines should benefit from an already existing and usable kit. To update the original TCLab presented by Hedengren (Hedengren *et al*., 2019), a temperature control lab kit is built from scratch using everyday electronics components. Matlab/Simulink simulations are used to check the correctness of mathematical frameworks of the system, which are based on theoretical and experimental approaches. The Kit and its corresponding mathematical models are then subjected to a variety of control algorithms, including On/Off, PID, and Fuzzy, to demonstrate the control feasibility of each. Matlab GUI is also used to develop the human-machine interface (HMI), which allows the user to choose a control method, adjust control settings, and monitor process parameters. The resulting models, according to the results of the experiments, are able to replicate the dynamics of the material kit with a temperature discrepancy of no more than 3 degrees Celsius. This proves the suitability of the kit for instruction in a variety of control-related subjects, including but not limited to system modelling, system proof of identity, classical control, and advanced control algorithms [11].

The findings are contrasted to those reported in the literature, which were compiled using a variety of tools and retrieval strategies. We examine the validity of the experimental variables that are integral to the power estimate that connects the humidity level to the measured absorption. The current standard uncertainty in the column with the total water vapour is about 10% due to inaccuracies in observations, calibration, and computations. In order for data acquired by optical photometry to be used as an independent benchmark for other techniques (GPS, MW-radiometers, lidar, *etc*.), we address the prospects for raising the accuracy of calibration to 1% as a crucial requirement [12].

CHAPTER 15

Method for Adaptive Combination of Multiple Features for Text Classification in Agriculture

Jaskirat Singh[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: When applying conventional text classification techniques, the values in agricultural text are converted into characters, which destroys the original semantic representation of numerical aspects. A unique text classification approach is suggested, based on the dynamic fusion of several characteristics, to completely mine the deep latent semantic characteristics in agricultural literature. The global key semantic characteristics of the text were extracted using the Bi-directional Gated Recurrent Neural Networks (GRU) model with attention mechanism, while the local semantic data about the text at various levels was extracted using the multiple windows Convolution Neural Network. Finally, the number that features essential semantic expressions was obtained using a machine learning approach for creating the quantitative value feature vector. To further enhance the deep semantic expression found in agricultural text and successfully improve the impact of farm text categorization with phenotypic numerical type, we use a focus technique to dynamically fuse the derived numerous semantic characteristics.

Keywords: Agriculture, Bi-directional GRU, Multiple features dynamic fusion, Text classification.

INTRODUCTION

Android malware is proliferating so quickly that accurately categorizing it is becoming increasingly difficult [1]. The dynamic approach to classification, on the other hand, is computationally intensive [2], while the classic static approach is vulnerable to confusion and reinforcement. This study provides a method for classifying Android malware families based on the characteristics of the Dalvik Executable (DEX) file format [3]. Color and textural information are first transformed from the DEX file toward an RGB (Red/Green/Blue) picture and then into plain text [4]. Finally, classification is accomplished by the application of a

[*] **Corresponding author Jaskirat Singh:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: jaskirat.singh.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

feature fusion approach based on multiple kernel learning [5]. Speech emotion detection has many practical uses, but it is still a technically difficult subject [6]. Accurate categorization relies on knowing how to make use of the speech data's native multimodality (*i.e.*, audio and text). Existing research often opts to fuse multimodal information at the utterance level, ignoring the dynamic interplay of features across modalities at the fine-grain, time-series level [7, 8]. If the values in agricultural text are classified using conventional methods, the underlying numerical aspects will lose their original semantic representation [9]. The cultivation of crops processes have not been centralized to their full potential, and insufficient use has been made of available online resources. The process of text analysis and categorization relies heavily on the calculation of text similarity. Examples of such metrics are the Jaccard index and the cosine similarity coefficient [10]. Some syntactic and semantic aspects of the sentences are lost since these methods treat the text as a bag-of-words. Existing literature on agricultural text categorization demonstrates a primary emphasis on sectorial analysis and question-answering AI. Due to the paucity of data from standard agricultural trials, it is necessary to construct matching corpora. The conventional approach to artificial feature engineering is not only difficult to execute, but it also lacks flexibility. Deep learning technology has enhanced the effectiveness of agricultural text classification in comparison to conventional machine learning. Yet, values in the phenotypic aspects of agricultural text will be handled as characters in the procedure of automatic extraction of text semantic features, and the numerical semantic data with practical importance cannot be completely acknowledged or ignored. Despite their obvious significance, researchers seldom investigate how ignoring them affects text categorization accuracy. This article utilizes Python to analyse the national audit of wheat feature text, build the wheat text corpus, and classify the text based on whether or not the wheat has cold resistance, all in an effort to make up for the dearth of the agricultural corpus. The article proposes a multi-feature dynamic fusion method for agricultural text classification, combining machine learning and deep learning based on other text classification studies with a wheat farming text corpus constructed in this paper to address the issue that some phenotypic values with substantial semantic information will be ignored. Text categorization is a typical NLP job, and both Bi-GRU and CNN have shown promising results. While CNN focuses on the local feature information of text within the convolution window, Bi-GRU captures the global feature representation of text within the sequence. By combining them, we can ensure that the classification impact is maximized by obtaining global and local feature representation information of the text. The gist of this paper consists of the following points.

- To acquire the universal feature representation across all settings, we utilize a bidirectional gated recurrent unit model with an attention mechanism (Bi-GRU).

- In order to acquire the most important global semantic feature representation, more focus is being placed on keyword semantic information.
- For this purpose, we employ a Mul-CNN, which is a multi-scale convolution kernel neural network, to produce a local feature representation of agricultural text sequences.
- We extract and code the values in the text independently, taking into account the specifics of agricultural language containing numerical data. Phenotypic values with rich semantic expression are extracted in this research, and numerical feature vectors are built.
- To enhance the precision of agricultural text classification with physical values, we adopt the notification mechanism to continually calculate the importance of the three features: local characteristics of varying levels obtained by neural network, the key features in the global scope, and the constructed feature vector.

RELATED WORK

Classifying agricultural texts is a subfield of text classification. Information that may be beneficial for agricultural production and research may be extracted from large amounts of agricultural text data despite their complexity and noise. If the values in agricultural text are converted into characters using conventional text categorization techniques, the underlying semantic representation of numerical aspects would be lost. A unique text classification approach is suggested, based on the dynamic fusion of several characteristics, to completely mine the deep latent semantic features in agricultural literature. Calculations of value characteristics featuring vital semantic expression have been carried out by artificial method to construct a numerical value feature vector, and the Bi-directional Long Short Term Memory network (Bi-GRU) framework with the attention system was used to extract the global key semantic features of the text. In order to enrich the deep semantic expression of agricultural text and efficiently enhance the impact of agricultural text classification with phenotypic numerical type [11, 12], we introduce the attention technique to dynamically fuse the derived numerous semantic characteristics.

The Android Malware Dataset (AMD) was used as the test population for this investigation. This paper's approach was compared to the standard visualization approach and the feature fusion approach in two sets of experimental comparisons. The outcomes demonstrate that our approach produces superior classification results, with accuracy, recall, and F1 scores all above 0.96. In addition, compared to the frequent subsequence technique, the feature extraction time in this work is lowered by 2.998 seconds. In conclusion, the categorization of the Android malware family described in this study is both efficient and accurate.

CHAPTER 16

Deep Learning-based Text-Retrieval System with Relevance Feedback

Simran Kalra[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: We presented an Information Retrieval (IR) system that learns from existing information and creates a single pertinent document that, we presume, has all its indexed pertinent details for a query. Deep learning makes such a system viable. We then asked people to score the query plus word-cloud representation of three randomly selected relevant texts and our new synthetic document. The synthetic document topped all inquiries and users. We then trained a CNN using query-relevant data. We performed ”deep learn” function on a synthetic, relevant material using the CNN. We used crowdsourcing to compare the “deep-learned” material to related documents. Users can see a query and four-word cloud (three relevant documents and our deep learning synthetic document). The synthetic document provides the the most relevant feedback.

Keywords: Deep learning, Relevance feedback, Text-retrieval.

INTRODUCTION

Children's mental health is a public health concern. It determines children's growth, academic success, and productivity [1]. Business Process Improvement (BPI) initiatives help companies succeed in this competitive economy. Most of these programmes fail for reasons recognised in earlier studies, making it an active study topic [2]. This research uses Taguchi to identify telecom BPI project success criteria [3]. Taguchi Method was used to analyse business process management and team data to determine BPI project success criteria. The experimental findings help telecom managers run effective BPI initiatives by identifying crucial success elements [4]. Vision, skills, resources, incentives, and action plan were crucial to the BPI project's success [5]. This study uses the Taguchi Method to estimate key factor levels in the services sector, which is less applicable compared to manufacturing [6]. This study may help telecom BPI

* **Corresponding author Simran Kalra:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: simran.kalra.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

projects succeed by giving clear instructions before implementation [7]. The research's applicability across sectors is noteworthy [8]. Genetic Algorithm and Population-based Incremental Learning are compared on a basic mathematical function [9]. GA and PBIL use evolutionary search and randomised methods to find near-optimal solutions quicker. Stochastic search methods using unpredictability avoid local optima. The accuracy, number of generations, and duration to attain optimum value are compared between the two methods [10].

RELATED WORK

This research optimises power system stabiliser (PSS) parameters for a multi-machine system using Population-Based Incremental Learning (PBIL), a simplified GA. GAs and competing learning-based artificial neural networks are used. The PBIL algorithm optimises PSS parameter tweaking. PBIL-based PSS simulations demonstrate its efficacy [11].

This research examined mental health issues in Malaysian children aged 5–15 and analysed NHMS 2015 data. The SDQ was verified. Malaysian youngsters had 11.1% of mental health issues. Multiple logistic regression showed that mental health issues dropped 5% each year with age. Further investigation indicated that children with dads with non-formal education and private sector jobs, widowed or divorced parents, and both parents with mental health issues were inclined to have mental health issues. Mental health concerns were more common among Malaysian children from low-income households with parents with mental illness [12].

PBIL is a common Evolutionary Algorithm for optimising problems. A recent study suggests that PBIL with a fixed learning rate may lose diversity and prematurely converge. APBIL solves premature convergence in PBIL. Time-frequency domain simulations demonstrate APBIL algorithm performance [13].

Frequency assignment problem (FAP) plagues GSM-Global System for Mobile-Networks. PBIL solves MS-FAP. MS-FAP reduces the frequency range required for communications in an area. This paper describes the issue and its solution. We tested seven well-known PBIL variants and seven MS-FAP difficulties in a full series of studies. Finally, we compare the findings to determine which PBIL version best solves the MS-FAP issue [14].

BGA and PBIL are used to build Power System Stabilisers (PSSs) in this article. A new evolutionary algorithm BGA is introduced, which employs artificial breeding to pass on the best traits from parents, unlike Genetic Algorithms, which work on the concept of survival of the fittest. PBIL is a type of genetic algorithm that explicitly preserves the population's essential components but abstracts the

crossover operator and redefines the population. The research compares PSSs' electromechanical mode damping capabilities. An eigenvalue-based objective function was used to create PSSs to optimise the lowest absorption value under defined operating conditions while considering different methods. Eigenvalue and time-domain simulations show that BGA-PSS and PBIL-PSS work similarly. BGA and PBIL-based PSSs outperform the CPSS in all operating conditions excepting the nominal one when the CPSS was modified [15].

PROPOSED WORK

Research Gaps

Our aim is to show how that may be done when a brief query (a tweet) is extended using a semantic method and an IRS improves the quality of returning relevant content in tweet contextualization. Our tweet contextualization improvement method combines semantic query expansion and relevance feedback. Exploring similar topics from this source expands the original question. WordNet is regarded as one of the greatest knowledge repositories. This big source gives structured information and allows massive expression of knowledge related to a tweet. It allows to select words that broaden the inquiry. Text Razor improves query expansion by extracting themes.

System Model

Given a query and a collection of relevant papers, we ask whether a synthetic document can be automatically constructed that collects all the relevant information. If so, a synthetic document may be more important to a query than any other related document. CNNs generate synthetic documents. The CNN model must capture the query's semantics and relevant texts for learning a new artificial document. We approach this by giving the CNN vector illustrations of characters1, termed embeddings, of the query and its related texts. Words depend on character order. The CNN takes a character's embedding and updates a recurrent state, which is an order-sensitive synthesis of all the data observed up to that syllable (the initial recurrent state being 0). The CNN limits the conditional probability distributed on the space of potential letters given the input encoding, then each recurrent state estimates the likelihood of the next letter in the sequence. The procedure continues until a finish-of-document sign appears (Fig. **1**).

CHAPTER 17

Domain Knowledge-based BERT Model with Deep Learning for Text Classification

Akhilesh Kalia[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Lexical model BERT already trained on BookCorpus and Wikipedia works well on two NLP tasks after downstream fine-tuning. The requirements of BERT model include strategy analyses and task-specific and domain-related data. The problems of task awareness as well as instruction data in BERT-DL, a BERT-based text-classification model, are addressed through auxiliary sentences. The pre-training, training, and post-training steps for BERT4TC's domain challenges are all provided. Learning speed, sequencing duration, and secret state vectors that select fine-tuning are all investigated in extended trials over 7 public datasets. The BERT4TC model is then contrasted using a variety of auxiliary terms and post-training goals. On multiple-class datasets, BERT4TC with the ideal auxiliary phrase outperforms previous state-of-the-art feature-based algorithms and fine-tuning approaches. Our domain-related corpus-trained BERT4TC beats BERT on binary sentiment categorization datasets.

Keywords: BERT Model, Deep learning, Natural Language Processing, Text Classification.

INTRODUCTION

Sentiment analysis for customer feedback and requests is intriguing. Sentiment analysis may indicate public opinions on presidential elections and pandemics. COVID-Twitter BERT (CT-BERT), a domain-targeted BERT language model, has been used in sentiment evaluation on COVID-19. BERT-based language models increase text classification slightly. Thus, BERT was suggested as a supplement. Continuous learning helps NLP models learn and grow. Past continuous learning approaches focused on preserving information from past tasks rather than generalising models to new problems. Pre-trained models (PTMs) have enhanced text categorization by simply using its characteristics. However, task complexity may limit PTM knowledge exploration. On the basic tasks,

[*] **Corresponding author Akhilesh Kalia:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: akhilesh.kalia.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

learning algorithms saturate early. BERT's native sentence representations are compressed, therefore using them for text classification may not completely capture discriminative properties. Deep learning always emphasises generalisation. Such models and domain adaption technologies are often used to solve generalisation. They seek data attributes to increase generalisation and reduce overfitting. Despite their success in other tasks, the models are unsteady when categorising sentences with positive labels but negative words. Few-shot relation categorization is popular now. Identifying relations from a few occurrences solves the long-tail relation issue. Metric learning techniques learn class prototypes and forecast using query-prototype distances. Text variety makes predictions inaccurate. It makes sense that relation and entity text descriptions may enable relation categorization.

RELATED WORK

BERT's text classification performance is improved by converting single-sentence classification to pair-sentence classification. Auxiliary sentences and pre-trained BERT models on COVID-19 tweets increase sentiment analysis classification. Pair-sentence classification improves F1 scores and dataset size. A domain-specific language model should perform better. CT-BERT might not surpass BERT in sentiment understanding [11].

We present an information disentanglement-based regularisation technique for text categorization continuous learning. Our technique first separates text-concealed spaces into generic and task-specific representations and then regularises them differently to better confine the information needed to make generalisation. For improved general and specialised representation spaces, we provide two simple auxiliary duties: prediction of the next phrase and task-id prediction. On large benchmarks, our text classification method beats state-of-the-art baselines. https://github.com/GT-SALT/IDBR has our code [12].

This research proposes a two-stage training text classification approach to overcome these challenges. Auxiliary labels are used in pre-training to raise task difficulty and fully use the pre-trained model. The pre-training textual representation is used in the fine-tuning step to improve classification performance. The methodology outperformed multiple state-of-the-art baselines on six text categorization datasets [13].

This article examined the benchmarks' attention heat maps and discovered that previous models prioritised phrases over sentence semantics. We proposed to emphasise on contradictory emotional concepts to avoid one-sided evaluation. Our two-stream network's auxiliary network included slope reversal and feature projection layers. Slope reversal layers reverse feature gradients while retraining

to improve backpropagation parameters. We blended backward features from an auxiliary network with principal network information. TextCNN, BERT, and RoBERTa baselines were subjectedto sentiment analysis and sarcasm detection. Our method improves both sentiment analysis and sarcasm detection by 0.5% and 2.1%, respectively [14].

TD-Proto adds relation and entity descriptions are used in prototypical networks. An attention module extracts sentence and entity information. To create a knowledge-aware instance, a gate mechanism fuses both information dynamically. Our technique performs well experimentally [15].

PROPOSED WORK

Problem Formulation

The categorization model, f(x)=y, estimates conditional distributions of odds across every tag in the predetermined class set for a source text sequence x$y = \{ y_1, \ldots, y_c$.

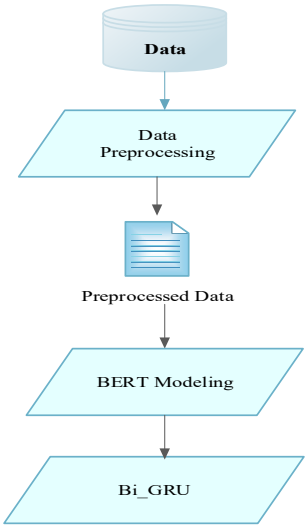


Fig. (1). System model.

Here x is a k-word token input scheme $x_{1:k} = x_1 x_2 \ldots x_i \ldots x_k$, where $x_i (1 \leq x_i \leq k)$is the "th" consecutive word, *etc*. Since the first token is always [CLS] as well as contains special categorization insertion, and since the second token, [SEP], is utilised to divide segments or denote the closeness of the sequence, the BERT model cannot categorically suggest a simple composing request or a pair of portions in one currency collection (*i.e*. [Question, Answer]).

CHAPTER 18

Applying Deep Learning to Classify Massive Amounts of Text Using Convolutional Neural Systems

Shubhansh Bansal[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Supervised learning based on deep learning is often used for mass-scale picture categorization. However, it takes a lot of computing effort and energy to retrain these vast networks to accept new, unknown data. When retraining, it is possible that training samples used before would not be accessible. We present a scalable, gradually expanding CNN that can learn new jobs while reusing some of the base networks and an efficient training mechanism. Our approach takes cues from transfer learning methods, but unlike other approaches, it retains knowledge of previously mastered tasks. Convolutional layers from the early section of the base network are reused in the updated network, and a few more convolutional kernels are added to the later layers to facilitate learning a new set of classes. On the task of categorising texts, we tested the suggested method. Our method achieves comparable classification accuracy to the standard incremental learning method in which networks are updated solely with new training samples, without any network sharing), while also being more resource-friendly and taking less time and space to train.

Keywords: CNNS, Convolutional neural networks, Deep learning, Text classification.

INTRODUCTION

Natural language processing (NLP) relies heavily on text categorization, making it a common and crucial activity. Convolutional Neural Network (CNN) is a member of the greatest effective models in the field of text categorization, and deep learning techniques have recently shown their benefits in this area [1]. The use of deep neural networks for analysing sentiment in the text is a well-studied field. When used for text sentiment categorization, deep learning models may perform at their computational peak [2]. Both well-known open-source data

[*] **Corresponding author Shubhansh Bansal:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: shubhansh.bansal.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

corpora and psyche-extracted short-phrase texts from social media are common resources for these kinds of investigations like Twitter and Reddit, as well as web-scraped text data from other sources [3]. It is unusual to gather and refine that much information about a live event [4]. Scaling the accuracy of the sentiment analysis and developing a predictive analysis of the same would need modelling data from an ongoing event, which is an even more difficult challenge [4]. With Natural Language Processing (NLP), text categorization is a traditional yet crucial work [5]. Convolutional Neural Networks (CNNs) have grown into an important paradigm in deep learning due to their many useful applications. The convolution process is crucial to the performance of a CNN model. Convolution operation-based methods are particularly well-suited to capturing local conjunctions of characteristics in high-dimensional data [6]. The convolution technique is often performed using a sliding window in previous methods [7]. Even though this is problematic given the present problems in the area of Chinese information, such as inconsistent information and a lack of automatic and efficient management, most existing Chinese text classification methods use word characteristics as the basic section of text representation but neglect the beneficial effectiveness of character characteristics. As more and more processes are digitized, massive amounts of data need to be stored [8]. Everyone posts a tonne of personal data about themselves online *via* a smartphone or web app, sometimes without even realising it. There are now privacy concerns due to the ever-increasing need to store individuals' data [9]. The use of another person's private information is not illegal. While the European Union's privacy rule went into effect in 2018, India is currently in the midst of drafting its own [10]. Some businesses are now working on programmes that can determine whether or not a given document is confidential. Such applications might be developed using several deep learning models, such as the Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Term Short Memory (LSTM), *etc.*

RELATED WORK

We present a novel deep learning approach that employs convolutional neural networks with scope information for automatically classifying text content. In contrast to window-based CNN, the scope does not call for the words used to build a local feature to be adjacent. A finer-grained representation of textual data is possible. We introduce a large-scale range-based convolutional neural network (LSS-CNN) by combining scope convolution, aggregation optimising, and the max pooling operation. These techniques might help us reconstruct key regional information from the manuscript. This study also explores methods for effectively computing data based on scope and for parallel training on large datasets. We have conducted comprehensive experiments on real-world datasets to compare our model to other state-of-the-art approaches. The results show that LSS-CNN

can efficiently analyse massive text datasets while maintaining excellent scalability [11].

Utilising real-time tweets analysed by a neural network with deep learning, we provide a novel approach to forecasting the future occurrence of Coronavirus infections. We built a massive Twitter database focused on just the Coronavirus tweets. We perform polarity classification and trend analysis, as well as divide the data into sets for testing and training. With the enhanced outcome from the trend analysis, the data is trained, giving our neural network an incremental learning curve, and this leads to a 90.67% success rate. Finally, we conclude with a statistical forecast for the future expansion of Coronavirus cases. When evaluated with a variety of prominent open-source text corpora, our model not only surpasses multiple past state-of-the-art studies in general accuracy comparisons for comparable tasks but also maintains a consistent level of performance throughout all of the test cases [12].

In this research, we provide a novel approach to convolutional neural networks called a scope-orientated convolutional neural network (SCNN). To reflect the local data, we offer a notion of scope as an alternative to sliding windows. A scope, as contrast to a window, limits simply the spacing between words. More complex local features may be handled with ease. To preserve the most relevant data, we use max pooling. This allows it to learn more nuanced local features that might otherwise be missed by a window-based CNN. We conduct a large-scale experimental study to demonstrate that our model is competitive with the state-of-the-art approach on real-world datasets. This proves that our strategy is superior to others that have been offered [13].

Here we propose a method for automatically classifying Chinese news items utilising character-level convolutional neural networks (char-CNN). We constructed a massive Chinese news corpus and contrasted it to both traditional models and deep learning models to see that personality-level convolutional neural networks (CNN) would achieve the most advanced or competitive results [14].

Methods for representing text in text classification issues are the topic of this study. These challenges include but are not limited to, private data categorization, sentiment analysis, language identification, online abuse detection, and recommendation systems. Having text displayed in a variety of ways improves the efficacy of categorization systems.

CHAPTER 19

An Algorithm for Categorizing Opinions in Text from Various Social Media Platforms

Pavas Saini[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: More individuals are sharing their thoughts and feelings through internet videos as social media platforms proliferate. While successful emotional fusion in multimodal data is a key component of multimodal sentiment analysis, most existing research falls short in this area. Predicting users' emotional inclinations through their expressions of language is made easier by multi-modal sentiment detection. As a result, the field of multi-modal sentiment detection has grown rapidly in recent years. As a result, multimodal sentiment analysis is quickly rising to the forefront of academic interest. However, in actual social media, visuals and sentences do not always work together to represent emotional polarity. Additionally, there are several information modalities, each contributing in its unique way to the overall emotional polarity. A multimodal approach to sentiment analysis that takes into account contextual knowledge is presented as a solution to these issues. The approach begins by mining social media texts for subject information that comprehensively describes the comment material. Additionally, we use cutting-edge pre-training models to identify emotional qualities that span several domains. Then, we provide methods for merging features at different levels, such as cross-modal global fusion as well as cross-modal high-level semantics fusing. At long last, we run our tests on a multimodal dataset that really exists in the real world. Results show that the proposed approach can correctly classify the tone of heterogeneous online reviews, and it also outperforms the standard approach in many other ways as well.

Keywords: Opinion mining, Social media, Sentiment classification, Topical information.

INTRODUCTION

Understanding how to include rich contextual information and multi-modal data into a single model framework is the first step towards user sentiment analysis on social media [1]. The research presents a support vector machine (SVM) based

[*] **Corresponding author Pavas Saini:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: pavas.saini.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

topic model for consumer sentiment evaluation and employs SVMs as well as k-nearest-neighbors algorithms (KNNs) to build models for classifying social sentiment in Japan. Additionally, the study builds an SVM classifier utilizing TE processing information as a case study and picks a radial basis function for the kernel as well as a grid search technique [2]. Media data, comprising text, image/video, as well as social interaction information, has been generated in large quantities by social networking networks in the era of Web 2.0, such as YouTube, Facebook, Google, and Flickr [3]. Multimedia applications (such as annotating and retrieving images and videos and classifying events) may make good use of these data sets [4]. Because these data span several domains and media types (such as text, images, videos, and audio), designing an appropriate feature representation to characterize them is challenging [5]. Many proposed systems focus on determining whether a given social media post is good or negative in tone [6]. People may share their thoughts and feelings through a variety of mediums on social media [7]. User emotional preferences can be better predicted using multi-modal sentiment detection [8]. As a result, the field of multi-modal sentiment detection has grown rapidly in recent years. The effective interplay among the several modalities present in a video is often overlooked in current publications that treat video utterances as isolated modals [9]. Sentiment analysis is the practice of gathering information about how people feel about an issue, product, or person and then categorizing that data as positive, negative, or neutral. The data gleaned from social media platforms like YouTube and online marketplaces like Amazon may be put to good use in recommending items to customers and generating sales, respectively [10]. Most individuals nowadays use a combination of their native language and terms from other languages when they talk about their ideas and opinions. Hence the interpretation of the emotional tone of code-mixed speech is crucial. The field of Tamil mixed sentiment analysis is nascent, with only a few of the published works [12].

To address the problem of insufficient data, one might use the solutions mentioned above. The classification effect may be enhanced using multimodal information by improving input data and refining the structure of the model. Nevertheless, training these models requires a substantial amount of effort. Reduced training time is possible with the discovery of new types of multi-modal data that have characteristics like broad availability as well as fewer variables for feature extraction models. Training a model faster without compromising accuracy is possible using up-to-date multifaceted information. The characteristics of social media users may be used as multi-modal data after thorough analysis. Consequently, we provide a user-oriented sentiment classification method for SM text data. The paper's primary contributions are as follows:

- Since social media user traits are integrated into the sentiment categorization algorithm, gathering the necessary input data is a breeze.
- The model training time is reduced due to a small number of parameters in the CNN network utilized to extract the features of user characteristics.

The bidirectional internet text-based sentiment classification method has recently been upgraded to state-of-the-art, making it ideal for small-scale and time-sensitive data.

The sections of this article are as follows: In Section 2, we examine how the characteristics of the user affect their feelings. In Section 3, we describe a model for short-term and localized data sentiment categorization. The user characteristics of the Convolutional and Recurrent Neural Network model's architecture and formula are detailed below. Various models and their assessments through experiments are addressed in Section 4. Section 5 presents the final findings.

RELATED WORK

Additional evidence linking the remark to the first tweet may be found in a correlation variable inside the feelings of both the comment and the tweet. This study describes a user sentiment assessment model and uses the Twitter real data set as an experimental information set to evaluate its performance. The idea of SVM-KNN is the foundation of the parameter estimation method. The results of the research show the efficiency of the method described here [11].

We present a unique stacked denoising auto-encoder-based cross-domain feature learning (CDFL) technique to address these concerns. In contrast to traditional auto-encoders, our CDFL incorporates a modal correlation constraint and a cross-domain constraint to concurrently optimize correlations between modalities and extract domain-invariant semantic information. We test our CDFL algorithm in three major domains: sentiment analysis, spam detection, and event categorization. Extensive testing shows that the suggested method performs admirably [12].

The purpose of Sentiment Analysis is to mine this information for insights into consumer preferences, which may then be applied to improve marketing strategies, public policy initiatives, and product development. In order to determine if a piece of text is subjective or objective and if it is subjective, whether it is negative or positive, sentiment analysis is performed. Opinion, sentiment, and the subjective nature of texts are the focus of sentiment analysis, which is a branch of text analytics. In addition, we often work together to create cost-effective illation techniques for estimating the parameters of supported folded Gibbs sampling. We usually base our evaluations of SJASM on

CHAPTER 20

Text Classification Method for Tracking Rare Events on Twitter

Prabhjot Kaur[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: A natural catastrophe is an example of a rare occurrence that does not happen often but may have devastating effects on people and their environment when it occurs. People now have a quick and easy outlet for voicing their ideas thanks to social media. Thus, it may be used by researchers to learn about how individuals react to and think about a wide variety of extremely unusual occurrences. Many research works use social media data to investigate how people's reactions to unusual occurrences in the real world translate to their online personas, thoughts, feelings, and actions. In this piece of work, we offer a method for extracting features and classifying tweets on unusual events like Hurricane Sandy. To begin, a new approach to feature extraction is presented, one that may be used to extract relevant features from each communication. The next step is to offer a Score-based categorization system for differentiating between communications about events and those that are unrelated. Finally, the development of a rare event is analyzed using our suggested approach and the popular keyword search method. The findings show that the suggested method is effective in distinguishing between text messages connected to unusual events and those that are unrelated.

Keywords: Rare events tracking, Social media, Text classification, Twitter sentiment analysis.

INTRODUCTION

Rare occurrences, such as natural disasters, may have devastating effects on people and their surroundings when they do occur [1]. Opinions may be voiced quickly and easily through social media [2]. Thus, it may be used by researchers to study how individuals react to and think about a wide variety of extremely unusual situations [3]. Many research works use social media data to investigate how people's reactions to unusual occurrences in the real world translate to their

[*] **Corresponding author Prabhjot Kaur:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: prabhjot.kaur.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

online personas [4]. When it comes to researching the development of unusual occurrences like natural catastrophes, social media provides a wealth of data [5]. Examining the link between social media and occurrences with which humans are involved can be instructive. Many deep learning algorithms rely heavily on semantic word representation [6]. Word angle/distance, word analogies, and statistical data are the mainstays of most word representation methods [7]. However, common models treat each word independently by using a vector representation [8]. This restricts their use for representing uncommon words in languages with extensive lexicons. Text documents frequently provide brief "snippets" of information pertinent to a given area [9]. For example, human analysts frequently use ontology to categorize snippets of text in the social science subject of peace and conflict studies, where it is crucial to identify, classify, and monitor causes of conflict using text sources. The lack of class-conditional evidence in "rare" phrases in snippets is a challenge to automating this approach [10].

There are two key benefits. We begin by proposing a new approach to feature extraction. Second, we propose and validate a fuzzy logic-based classifier with actual social media data. Additionally, there are measures developed specifically for measuring the efficacy of such brief text categorization strategies. The remaining content of this article is as follows. The data and an innovative feature extraction approach are presented in Section 2. In Section 3, we outline the suggested method of text categorization using fuzzy logic. Section 4 presents and discusses the experimental findings. Section 5 provides the summary and findings.

RELATED WORK

In this piece, we offer a method for extracting features and classifying tweets on unusual events like Hurricane Sandy. To begin, a new approach to feature extraction is presented, one that may be used to extract relevant features from each communication. The next step is to provide a classification system based on fuzzy logic that can identify event-related irrelevant communications. Finally, the development of a rare event is analyzed using our suggested approach and the popular keyword search method. The findings show that the suggested method is effective at differentiating between text messages connected to unusual events and those that are unrelated to them [11].

This research does information extraction and text categorization using Twitter text data from a 2012 Hurricane Sandy. Since the original data covers a wide variety of issues, we will need to extract the information pertinent to Hurricane Sandy. To address the issue of text categorization, we provide a method based on

fuzzy logic. The suggested fuzzy logic-based model takes as inputs several characteristics that may be gleaned from each tweet. The final result for each message sent to Sandy is a relevancy score. To achieve the required categorization outcomes, many fuzzy rules are written and various defuzzification approaches are combined. We evaluate the suggested strategy against the standard practice of using keywords to get relevant results. The outcome demonstrates that the suggested fuzzy logic-based strategy is superior to the keyword-word method [12] for classifying Twitter tweets.

Words are proposed to be represented by a vector of domain and semantic properties in this research, with the help of a dynamic model called SemVec. An enhanced word representation containing domain knowledge can be created by adding or removing semantic characteristics based on the issue domain. The suggested methodology is tested by classifying tweets and texts on adverse drug reactions (ADRs). SemVec outperforms other state-of-the-art deep learning approaches in terms of recall score [13], and its results demonstrate that it increases precision for ADR detection by 15.28%.

In this study, we create a strategy to improve a bag-of-words model by including a Word Vector model's related terms to supplement the text's unusual keywords. This technique is then utilized in tandem with common linear text categorization methods. These enriched models outperform the standard classifiers by decreasing bag-of-words sparsity. Second, we show that Paragraph Vectors perform better than the enriched models [14] when applied to the problem of improving performance on “small” classes with few instances.

In this study, we provide a novel approach to event classification using noisy hydrophone data. The 1D hydrophone data is first transformed into a log-frequency spectrogram picture (cepstrum) using an image processing methodology. To further refine this image, we rebuild it using the dominant orientation map's mutual information (MI) criterion. After the cepstrum has been rebuilt, it is improved by applying edge-tracking and noise smoothing to its features. Least-squares support vector machines (LS-SVMs) are used to conduct feature categorization on the refined cepstrum. When applied to noisy hydrophone recordings from the NEPTUNE Canada project, the approach demonstrated a sensitivity for event identification in excess of 99%, with a specificity in excess of 97% and an overall accuracy in excess of 98%. The suggested approach is suitable for automated long-term monitoring of a range of marine animals and human-related activities using hydrophone data [15] because of its cheap processing cost and excellent accuracy.

CHAPTER 21

Text Document Preprocessing and Classification Using SVM and Improved CNN

Jaspreet Sidhu[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Text categorization is a crucial technology in data mining as well as data retrieval that has been extensively investigated and is developing at a rapid pace. Convolutional neural networks (CNNs) are a kind of deep learning modeling that may reduce the complexity of the model while accurately extracting characteristics from input text. Support vector machine (SVM) results have always been more trustworthy and superior to those of other traditional artificial intelligence approaches. Using enhanced convolutional neural network (CNNs) as well as support vector machines (SVMs), we offer a novel approach to online text categorization in this study. Our approach begins with text attribute identification and prediction using a model based on CNN with a five-layer network structure. Databases including both text and images will find it to be a major factor in the long run.

Keywords: Classification, SVM and Improved CNN, Text Document Preprocessing.

INTRODUCTION

Using the naive Bayes technique and the support vector machine (SVM), this study provides an improved hybrid classification strategy [1]. In this study, we employed the Bayes formula to create a probability distribution that vectorizes (as opposed to classes) a document into the most likely categories to which it belongs [2]. If you have a set of subjects (categories), like the ones in the "20 Newsgroups" dataset [3], you can use the Bayes formula to estimate the likelihood that a text falls into each category. The SVM can perform multi-dimensional document classification by employing these probability distributions as the document's vector representation [4]. To prevent the decrease in dimensionality that happens when a naive Bayes classifier uses just the greatest likelihood for classification, the support vector machine (SVM) takes into account

* **Corresponding author Jaspreet Sidhu:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: jaspreet.sidhu.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

all the likelihood values linked with each group for every document [5, 6]. The Formula E racing series has quickly risen to the top of the international motorsports scene. The world-famous Formula E event was held in Indonesia in 2022 [7]. The estimated attendance of 35,000 people has the potential to generate 78 million euros in economic benefits for Indonesia. Many in Indonesia believe that hosting Formula E races will help their country attract tourists and raise its profile abroad [8]. However, there are many who disagree with this move. They believe that the government's resources would be better used to aid those impacted by the COVID-19 pandemic than to fund a Formula E race at this time [9]. To categorize new examples, ensemble techniques generate a series of classifiers and then obtain a weighted vote of their individual predictions. An advanced challenge in ensemble classification [10] is to reduce error while improving accuracy [15].

By combining the improved CNN as well as SVM algorithms, this work aims to overcome the aforementioned problems and finish the Web text categorization assignments in the language analysis sector. In comparison to the current state of affairs, experimental results demonstrate that the suggested Web text categorization approach achieves higher levels of precision in classification.

RELATED WORK

In comparison to the Lsquare technique, this approach requires much less time to train, and it outperforms pure naive Bayes algorithms as well as TF-IDF/SVM hybrids in terms of precision for classification [11].

The Bayes algorithm is used here to assign a document's vector representation to a set of categories based on the probabilities associated with a set of keywords. The Bayes formula provides a probability distribution over the collection of subjects (categories) to which the document may be allocated. This probability distribution can be used as the carriers that represent the documents by text categorization methods utilizing the vector space model, like the Support Vector Machine (SVM) as well as Autonomous Map (SOM). This improves upon the results obtained when the naive Bayes classifier is used alone, which only uses the highest likelihood to group the piece of paper. These techniques allow us to recover from the consequences of a mistaken dimensionality reduction. For high-dimensional data, we evaluate the effectiveness of various classifiers [12].

This research evaluates the effectiveness of the Support Vector Machine and the Naive Bayes algorithms in identifying fans' sentiments toward the Formula E championship. The data in this study comes from public discussions on social media sites like Twitter. Cleaning, folding cases, tokenizing, filtering, and stemming are all part of the text preparation phase. The TF-IDF method of

weighting should be used moving forward. Data testing evaluates the categorization findings using accuracy, precision, and recall tests using a confusion matrix. The SVM algorithm's recall, precision, and accuracy in classifying public opinion are all above 80%. The accuracy is 82%. However, the Naive Bayes method only achieves an accuracy of 87.54 percent. The general public's attitude on Twitter is supportive of introducing Formula E [13].

An ensemble of classifiers for prediction is introduced in this research, along with a revolutionary general object-oriented voting and weighting adaptable stacking architecture. This general-purpose model makes its forecast by tallying the votes of a set of basic learners, with each learner's probability of correctness given equal weight. For demonstration, the proposed stacking framework is used to ensemble classification using three well-known heterogeneous classifiers: the Support Vector Machine, [Formula: see text]-Nearest Neighbor, and Naive Bayes. Additionally, the framework's resulting ensemble classifier is compared to others and assessed across a number of benchmark datasets utilizing a wide variety of cross-validation levels and percentage splits. The result is what sets this framework apart from its rivals. The suggested approach may properly estimate inmates' tendency to commit crimes 99.9901% of the time [14].

As the internet and computing infrastructure continue to advance, text categorization will likely become the most important method for handling massive amounts of text. One major hurdle to perform better text classification results is figuring out how to precisely identify the text as a data collection that can be analyzed. The author discusses a method for characterizing texts using a formula called p-idf, which is derived from the vector space model and tf-idf. The author constructs text classification system with a support vector machine after evaluating Bayes, K neighbors, neural networks, and other common text classification tools. The p-idf formula is fair and valid for a text categorization system [15] after a scientific test compared to its performance to that of the tf-idf and LTC formulas.

PROPOSED WORK

An example of a common type of artificial neural network is the convolutional neural network (CNN). Each layer's output is fed into the following layer's neuron input in this type of network. Nonlinear transformations are applied to the results of each layer using a multi-layer convolution technique. Fig. (**1**) shows the four main parts that make up the convolutional neural network structure utilized for text processing. The four layers that make up the network are the fully linked, collecting, convolution, as well as embedding layers. Convolutional neural net-

CHAPTER 22

Identification of Text Emotions Through the Use of Convolutional Neural Network Models

Vaibhav Kaushik[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Increasing numbers of people are using the Internet to share their feelings and communicate with one another, and the vast majority of these expressions of emotion take the form of text. Using sentiment dictionaries, machine learning, and deep learning are the three most common approaches to text sentiment categorization studies. Due to the exponential growth of textual data, it is crucial to create models that can automatically analyse this material. Labels like gender, age, nationality, emotion, *etc.*, may be included in the texts. Numerous investigations of text categorization have emerged because the use of such labels may be useful in various commercial sectors. The Convolutional Neural Network, also referred to as CNN, was recently utilised to the problem of text categorization, with promising results. In this study, we advocate for the use of convolutional neural network networks for the job of classifying emotions. Using three popular datasets, we demonstrate that our networks outperform existing cutting-edge deep learning models by using successive convolutional layers to process substantially longer sentences.

Keywords: CNN, Deep learning, Text sentiment classification.

INTRODUCTION

Opinion mining, another name for sentiment analysis, is another hot topic in the study of the text. It uses textual subjective information as its study object and makes an effort to classify the many kinds of emotional overtones that may be found in it. It might be a self-assessment or a judgment of oneself, or it could be a feeling or a state of mind. That is why it is important for researchers studying text sentiment categorization to separate the text into its subjective and objective components. There are presently various obstacles in the field of text sentiment classification research, including emotional opaqueness, text formatting, text language, *etc.* Emotional opacity often relates to textual ambiguity and dependent

* **Corresponding author Vaibhav Kaushik:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: vaibhav.kaushik.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

interpretation. “This is a very big noise,” “This playground is very big,” *etc.* Both words use the same wording, “very big,” yet their meanings could not be more different. The vast majority of the writings we examine come from the internet and range in length and presentation. Chinese texts have unique characteristics that are not shared by English writings, such as the requirement for word segmentations that are unnecessary in English texts due to the presence of gaps between individual words. Text classification of sentiment research is often done nowadays using the sentiment dictionary, although this approach requires establishing a high-quality emotion dictionary and can only be categorised manually. This kind of lexicon requires a lot of work and has drawbacks such as not having enough terms to be comprehensive. Methods based on artificial intelligence may improve classification accuracy over those using a sentiment dictionary. This involves the establishment of a high-efficiency computer learning model. However, the features can only be designed by someone with a certain level of expertise. Very little detail can be seen, and it is hard to get in-depth text features. Natural language processing is an area where the method of deep learning has seen significant progress in recent years. With no need for manual design, it is able to automatically generate features in a hierarchical structure and classify data from beginning to finish.

Natural language processing (NLP) applications have seen great success with the usage of convolutional neural networks (CNNs). Since CNN's convolutional layer may be initialised by encoding vital semantic characteristics, this layer can also be used to extract vital semantic features. CNN's Combining layer filters features using the max-combining feature, however, this ignores the sentence's most important details and the text's contextual semantic information. As our capacity to store information digitally grows, these tasks are performed in the form of text messages. Many text classification jobs may be accomplished using a standard artificial intelligence classifier, such as a support vector machine or a naive Bayes classifier. However, because of the scarcity of text in brief text and the constraints of combining and convolutional layers, these classifiers suffer from sparsity difficulties and lack long-term dependencies. In natural language processing, text categorization is one of the most essential tasks. The scope and usefulness of this job are both high. However, most of the older approaches rely on static word embedding, which does not adequately address polysemy. The field of the processing of natural languages known as “sentiment analysis” classifies written material as either positive or negative. In the last several decades, it has found use in a variety of fields. Opinion analysis is challenging since each person has their unique perspective. Information on individuals, businesses, and consumer goods may all be sorted using sentiment analysis. Massive amounts of information is being added to the Internet's websites, blogs, and social media every minute. Virtually every e-commerce site lets customers provide feedback on products they

have purchased. The feedback reflects some users' experiences with the items. When making a quick, informal choice between two items, customers often look to the reviews for guidance. Additionally, businesses are interested in hearing customer opinions on items *via* reviews. Data analysis of online comments is getting difficult due to the usual characteristics of huge information and noisy brief written data.

This paper's remaining sections are structured as follows. Studies on sentiment classification using ML/DL models are reviewed in Section 2. Our planned network is described in depth in Section 3. The experimental conditions and data sets used to test the proposed model are discussed in Section 4. Section 5 then elaborates on the outcomes of the experiments. Section 6 is where this paper comes to a close.

RELATED WORK

In this paper, we propose a novel neural network method that combines the attention mechanism as well as the Combining layer to make the model focus on search terms in the sentence as well as automatically keep the most significant text message, thereby enhancing the model's ability to classify text. In order for the model to pick up on the crucial semantic characteristics, we initialise the convolution filter with key information features. By integrating the attention mechanism into the Combining layer, the model is better able to preserve the vital informational characteristics of the text. Experiments demonstrate that the proposed model performs very well on a variety of text classification tasks, such as sentiment classification as well as topic classification [11].

In this study, we provide a convolutional recurrent neural network architecture that is built based on an enhanced skip-gram algorithm. For the adversarial training of the skip-gram algorithm, we use the L2 regularisation technique. Not only may it boost this model's performance in text emotion classification tasks, but it can also make it more resilient and generalizable. To extract meaningful data from texts while decreasing the influence of irrelevant words, we used a Bidirectional short-term, long-network with attention mechanisms. At long last, CNNS-based text sentiment categorization has been accomplished. When compared to other classifiers used on the IMDB dataset [12], this model and method were shown to be the most efficient and accurate.

To this end, we propose employing contextualised BERT word embedding to efficiently encode the input sequence, followed by the temporal convolutional module, which merely determines 1-D convolutions to extract high-level features, and finally the max-Combining layer, which preserves the most crucial features for the classification of text. We test on six large-scale text classification datasets

CHAPTER 23

Classification & Clustering of Text Based on Doc2Vec & K-means Clustering based Similarity Measurements

Prakriti Kapoor[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: One crucial task in text processing is determining how similar two papers are to one another. A novel similarity metric is suggested in this study. Finding a suitable similarity metric for written materials that permits the development of coherent groupings is a significant difficulty for document clustering. After that, we use TFIDF to build a vector space, and then we use the ward's approach and the K-means algorithm to accomplish clustering. WordNet is additionally employed in the process of semantic document clustering. Visualisations and an interactive website illustrating the connections between all clusters illustrate the findings. The existence (and quantity) of words in texts are all that are taken into account while utilising the traditional bag-of-words paradigm. This process might lead to texts with identical meanings but distinct vocabulary being placed in various groups. The findings acquired using the suggested approach are analysed for their correctness using the F-measure. Comparisons using the sentence vectors model (Doc2vec) and the bag-of-words model are made to confirm the edge of the suggested strategy. The suggested methodology may be used to decipher web chat logs and client feedback posted online. We evaluate our method on a variety of real-world data sets including examples of text classification and clustering problems. The findings prove that the proposed measure outperforms competing strategies.

Keywords: Accuracy, Classifiers, Clustering algorithms, Document classification, Document clustering, Entropy.

INTRODUCTION

The challenge of determining how similar two texts are is an important one in the field of text processing [1]. Similarity measurement for text processing (SMTP) is a technique proposed by Lin *et al.* [1] to make it easier to search for specific

* **Corresponding author Prakriti Kapoor:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: prakriti.kapoor.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

information across text collections. The suggested metric took into account all three possible comparisons of document pairings in terms of similarity [2, 3]. Each of these scenarios is based on whether or not a certain characteristic appears in either of the two texts [4]. In the first scenario, elements from both sources are considered; in the second scenario, elements from just one source are considered; and in the third scenario, elements from neither source are considered [5]. As knowledge and information technology advanced in many different areas, so did the number of scholarly articles published in those areas [6]. Therefore, researchers spend a considerable amount of effort trying to track relevant study publications [7]. As a result, this work proposes a documents classification strategy capable of categorising research paper texts into meaningful groups based on the subjects they cover [8]. Arabic Organising vast quantities of data into just a handful of specified clusters, as carried out in text document clustering [9], enables conjectural navigation and browsing tools to be made available. However, the vector representation of words often utilised in clustering algorithms is not ideal since it fails to take into account the interdependencies between different concepts. For the sake of better comprehension or summarization, data may be clustered using cluster analysis. Statistics, data mining, recognising patterns, and other disciplines have all contributed to the evolution of clustering methods [10]. Clustering was traditionally performed using a syntactic rather than a semantic idea, and relied on statistical characteristics. Traditional document clustering methods employ single, unique, or complex words from the collection of texts as characteristics. Traditional methods, however, overlook semantic relations. Due to issues with polysemy and synonymy in the conventional approach, a collection of unique words cannot accurately represent the document's intended meaning and cannot produce useful clusters. Grouping data points that are semantically related is what semantic clustering is all about. In this context, “clustering” refers to the process of partitioning a data collection into distinct groups, whereby items belonging to the same cluster have the same meaning. On the other hand, you would not find a match between two products from distinct groups. By highlighting differences in meaning, papers that are not relevant may be eliminated thanks to semantic clustering. Numerous critical procedures (description of each file as a phrase, vectoring of every record to be placed in a mathematical structure, and ultimately the application of clustering algorithms for clustering objectives) are required for semantically analogous clustering of content assets. The study's principal objective, to use Ward's method for semantic document clustering, is met. We are unaware of any previous study that used a combination of the Ward's method and semantic similarity document clustering.

RELATED WORK

A novel similarity metric is suggested in this study. The suggested measure considers the following three scenarios when calculating the degree of similarity between two texts with regard to a feature: a) It is present in both papers, b) it is present in just one document, and c) it is absent from both. In the first scenario, closeness improves as the distance between the two feature values becomes less. In addition, the difference's normal contribution is typically scaled. In the second scenario, a constant is added to the similarity. In the third scenario, the characteristic makes no difference in the degree of similarity. The suggested metric is refined such that it may be used to compare collections of documents. Various real-world data sources for text categorization and grouping issues are used to assess the efficacy of our approach. The results demonstrate that the suggested metric yields superior performance compared to other approaches [11].

Despite the fact that the proposed solution measure of similarity was thought to be ideal for finding similarity between written materials based on the presence or lack of features available in text documents, it was discovered that the SMTP similarity measure did not cover the context of measuring the parallel among the pair of analogous paperwork. In view of this discrepancy, this study proposes a modest adjustment to the SMTP in order to bring it into conformity with other common similarity approaches and make it a comprehensive similarity measurement methodology for knowledge discovery [12].

In this study, we provide a measure of the degree to which two clusterings of the same dataset produced by different algorithms (or maybe the same algorithmic rule) are comparable. Calculating the degree of similarity between two papers is an important task in the study of text processing. The authors of this study hypothesised a successor similarity index. When determining the degree of similarity between two publications that have the same feature importance, the suggested metric takes into account the following three cases. Select “a” if the feature is present in every one of the papers, “b” if it is present in some but not all, & “c” if it is absent in all of the papers. In the first case, the similarity will increase as the difference between the feature values decreases. The significance of the difference is also often assessed. In the second case, the similarity is given a predetermined numerical value. In the third case, there is no correlation between the feature and the level of similarity. Numerous practical expertise sets for text categorization and group problems are used to assess the efficacy [13].

Since each of these groups encompasses a wide range of issues, our proposed method is based on their core focus and scopes. Accordingly, we segregate the extraction of category-specific word symbols from these discussions. Using the

CHAPTER 24

Categorization of COVID-19 Twitter Data Based on an Aspect-Oriented Sentiment Analysis and Fuzzy Logic

Tarang Bhatnagar[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: During times of disaster or epidemic, social media has emerged as a vital means of communication. It is difficult to examine the complete situational awareness *via* many elements and emotions to aid authorities due to the unpredictability of these calamities. Currently, systems for aspect recognition and sentiment analysis rely heavily on labelled data and require human curation of aspect categories. To analyze public opinion from a variety of angles, this study suggested a hybrid text analytical approach. Using the popular Latent Dirichlet Allocation (LDA) topic modeling, we first extracted and clustered the elements from the data. We then used the linguistic inquiry and word count (LIWC) lexicon to extract the sentiments and label the dataset. Finally, in the third layer of our structure, we mapped the elements into emotions, and the sentiments were classified using well-known machine learning classifiers. The comparison of our technique with other aspect-oriented sentiment analysis approaches shows encouraging results in experiments with actual datasets, and our method with several variants of classifiers surpasses current methods with top F1 scores of 91%.

Keywords: Aspect-oriented sentiment analysis, COVID-19 classification, Fuzzy logic, Twitter dataset.

INTRODUCTION

During times of disaster or epidemic, social media has emerged as a vital means of communication. Since the causes of these calamities are often unclear, it is difficult to examine the issue from all angles and gauge public opinion to help authorities [1]. The human labeling and categorization of aspects are major weaknesses in the current aspect detection and sentiment analysis method [2]. The information individuals create on various websites reflects their feelings about a wide range of topics, from the cuisine they eat on a daily basis to the items they

* **Corresponding author Tarang Bhatnagar:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: tarang.bhatnagar.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

try. We may learn a lot about a product or service from the feedback and opinions of consumers [3] amid the widespread use of social networks and micro-blogs, especially during the COVID-19 epidemic. As a result, it can be time-consuming and challenging to go through all the evaluations and analyze them to get a review of the overall quality before making a choice [4]. Sentiment analysis has been utilized as a quick and automated technique to assess customer opinion to get around this problem [5]. Large reviews may include both positive and negative commentary on several aspects of a product or service. Purpose. The purpose of this article is to categorize the opinions expressed in online traveler and guest reviews by factor [6]. Based on a neural network model, a novel approach to aspect-oriented sentiment categorization is given. Large amounts of commercially valuable commentary data have been created by Internet users sharing their thoughts and experiences with widely used products and services [7]. Many remark phrases include many comment components, each with a different tone, rendering the sentence worthless for polarization [8]. The goal of sentiment categorization at the aspect level is to identify context-specific sensory extremity in the target. As the field of Deep Learning develops, it has become increasingly fashionable to use these techniques for emotion detection [9]. The COVID-19 epidemic has resulted in profound changes to every facet of human existence. All citizens were affected by the government's regulations in this area. As a result, anticipating the effects of future pandemics requires researching public opinion [10].

The purpose of this research is to catalog the sentiments and discussions around COVID-19 on Twitter. We investigate 1) how to automatically recognize people's attitudes stated on Twitter because of COVID-19, and 2) what themes are most discussed by Twitter users when expressing sentiments regarding COVID-19? We offer an exploratory analysis of the data set, as well as topic discovery and sentiment identification using natural language processing (NLP), to provide insight into these concerns. From February 2020 to March 2020, we used the Twitter API to gather streaming data, and using sentiment analysis, we classified the tweets we collected into three groups: positive, negative, and neutral. To ensure that there is a roughly equal quantity of tweets in each class, we give a sentiment analysis of COVID-19-related tweets. The data collection is utilized to train and test several algorithmic models, which in turn serve as benchmarks for identifying Twitter sentiment about potential COVID-19 therapies. The best-performing model is chosen for further optimization and dissemination once a final verification study has been conducted. We conclude that future study has to do a better job of taking context and the diversity of opinion on COVID-19 therapies into account. The following are some of the main results of this study.

- COVID Senti, a huge, manually annotated data collection of 90,000 tweets analysed in February and March of 2020, was created. Three smaller data sets of the same size make up the whole. Each tweet is assigned a label indicating whether it is favorable, negative, or neutral. The data sets can be accessed by anybody in the scientific community.
- A display of the methods by which public opinions on Coronavirus were tracked, includes the application of sentiment analysis on Twitter data to construct a classifier and the creation of a visualization of textual data. In addition, a qualitative analysis is provided in the form of a word cloud including the most frequently used terms.

We identify themes that are indicative of popular anxiety regarding COVID-19 and present and analyze this prevailing discourse. The results of this research might be used by governments all around the world to better prepare for public health emergencies.

- Results from a benchmarking study comparing the performance of various state-of-the-art ML text categorization techniques are described.

Here's how the rest of the article is laid out. The relevant research is presented in Section 2. The approach that is proposed in this article is discussed in Section 3. In Section 4, we describe the findings, and in Section 5, we draw the necessary conclusions.

RELATED WORK

In this study, we developed a mixed-method text analytical framework for dissecting public opinion at the level of individual aspects. Using the popular Latent Dirichlet Allocation (LDA) topic modeling, we initially extracted and clustered the aspects from the data. We subsequently utilized the linguistic inquiry and word count (LIWC) lexicon to extract the feelings and label the dataset. Finally, in the third layer of our framework, we mapped the aspects into feelings, and the sentiments were classified using well-known machine learning classifiers. The comparison of our technique with other aspect-oriented sentiment analysis approaches shows encouraging results in experiments with actual datasets, and our method with several variants of classifiers surpasses current methods with top F1 scores of 91%.

In this research, we offer a novel approach to sentiment classification that integrates automated labeling (SentiWordNet), an ensemble technique (Stacking), and the prior knowledge topic model algorithm (SA-LDA). Datasets from various fields are used to assess the framework. According to the findings, the suggested

CHAPTER 25

Feature-Level Sentiment Analysis of Data Collected through Electronic Commerce

Preetjot Singh[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: When dealing with data in the form of text, the most popular method for doing analysis and determining sentiment content is called "Sentiment Analysis." Sentiment analysis is also known as Opinion Mining. Suggestions, feedback, tweets, and comments are all examples of the various types of text data that are being created. Customer feedback on e-commerce sites is a constant source of new information. Online stores may better meet client needs, improve their services, and boost sales by analyzing E-Commerce data. Positive, negative, and neutral feedback from customers may be separated using sentiment analysis. Numerous methods for Sentiment Analysis have been developed by academics. Typically, only a single machine learning algorithm is used for sentiment analysis. The purpose of this study, which makes use of Amazon review data, is to extract positive, negative, and neutral review ratings by locating aspect phrases, identifying the Parts-of-Speech, and applying classification algorithms to the collected data.

Keywords: Customer reviews, E-commerce, Features, Sentiment analysis.

INTRODUCTION

When dealing with data in the form of text, the most popular method for doing analysis and determining sentiment content is called "Sentiment Analysis." A synonym for sentiment analysis is "Opinion Mining" [1]. Suggestions, feedback, tweets, and comments are all examples of the diverse text data that are being created [2]. Customer feedback on e-commerce sites is a constant source of new information. By analyzing E-Commerce data, online stores may learn what customers want, improve their offerings, and ultimately boost sales [3]. Positive, negative, and neutral feedback from customers may be separated using sentiment analysis. Many methods for Sentiment Analysis [4] have been developed by researchers. Typically, only a single machine learning algorithm is used for

[*] **Corresponding author Preetjot Singh:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: preetjot.singh.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

sentiment analysis. Demand for automated sentiment analysis systems is skyrocketing [5] in the age of big data, thanks to the growing number of online shoppers and social media users throughout the world. Aspect-based sentiment analysis (ABSA), which considers the sentiment at the aspect level, is a well-liked area of study [6] because of its relevance to business needs. Two more tasks, aspect-term sentiment analysis (ATSA) and aspect-category sentiment analysis (ACSA), make up ABSA. Sentiment analysis is the study of how people feel about something or someone [7]. The importance of analyzing user evaluations for product makers to get insights into the acceptance of items has grown with the rise of social media websites and e-commerce applications. They may learn more about how customers feel about individual product features by performing an aspect-level sentiment analysis [8]. E-commerce, online communities, and social media have all seen significant growth in popularity. Customers' opinions matter greatly in terms of the end result [9]. Providers, sellers, and producers of goods and services benefit greatly from any and all feedback, even if it is simply a comment or review. Since user feedback is so important, sentiment analysis has attracted a lot of attention. A user's characteristics, attributes, and points of view can be gleaned from their writing through a process called sentiment analysis. The proliferation of e-commerce websites as alternative sales channels stimulated the development of several review websites covering various goods and services [10]. As a result, businesses can better track their reputation and gauge customer demand, while consumers may employ aspect-based sentiment analysis to inform their purchasing decisions. Beyond simple phrase or text-level sentiment categorization, further in-depth study may be encouraged with the use of a method called aspect-based sentiment analysis (ABSA).

RELATED WORK

Based on consumer feedback left on Amazon, this study uses Parts-of-Speech tagging and classification algorithms to determine if a review is positive, negative, or neutral [11].

In order to learn both ATSA and ACSA simultaneously, this research provides a multi-task learning architecture that uses a pre-trained BERT model as a shared representation layer. To make the most of the surrounding context data, we build a multi-head self-attention layer over the common BERT model and connect the heads through a skip connection. Our multi-task learning model outperforms baseline multi-task networks and single-task models using SemEval datasets [12], demonstrating improved performance on the ATSA test.

Over the course of 7.8 million Amazon customer reviews, this study proposes a workflow model for topic extraction (aspect identification) and sentiment

detection utilizing a frequency-based approach and unsupervised machine learning techniques. The paper also makes use of traditional techniques including logistic regression, support vector machine, and a naïve Bayesian approach [13] to determine the precision of the analytical model.

The key portion of this procedure is extracting user aspects, which are then utilized to categorize the user aspects. In the field of natural language processing, CNN models have been increasingly popular in recent years. This study offers a unique hybrid CNN model by combining the bidirectional long short-term memory and CNN models to sequentially analyze the data by learning its high-level properties. The loss of vital data is kept to a minimum using the concatenated approach. Experiments utilise benchmark product review and hotel review datasets, with the proposed hybrid model achieving accuracies of 93.6% for the product review dataset and 92.7% for the hotel review dataset, when compared to state-of-the-art methodologies [14].

The objective is to recognize valuations placed on certain things (like laptops) and their characteristics (like cost, functionality, durability, *etc.*). There are extremely few methods that can provide such outcomes based on customer ratings, and even fewer that can do so from free-text reviews. Aside from insufficient review data, the cold start problem is another obstacle. In this research, we offer a method for automatically computing sentiments of dynamic elements based on user-generated evaluations gleaned *via* web scraping from numerous sources, thereby allowing us to avoid the cold start problem. So, our approach is improving upon previous ways of gauging customer opinion in online stores.

PROPOSED WORK

Overview

Since it examines the feelings and opinions expressed in a text, sentiment analysis falls within the broader category of NLP. It pulls the good, negative, and neutral feelings from the text. Due to the nature of working with text data, a great degree of preparation work must precede the actual categorization. The sentences are first preprocessed by assigning a Parts-of-Speech tag to each word, then by identifying and deleting stop words and unnecessary adjectives. The proposed study uses Amazon customer reviews to evaluate the efficacy of the Nave Bayes and Support Vector Machine (SVM), and Machine Learning algorithms for sentiment analysis. The primary focus of Aspect-level sentiment analysis is on product features and aspects. The measures to take are laid out in Fig. (**2**), which depicts the suggested system architecture.

CHAPTER 26

Classification Algorithms for Evaluating Customer Opinions using AI

Saniya Khurana[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: There are now a great deal of consumer reviews of items that are written entirely in text. The reviews express their opinions. Opinion mining is another name for sentiment analysis. A common way for businesses to keep tabs on how customers feel about their brands and goods is through the use of sentiment analysis on textual data. Naive Bayes, Random Forest, Decision Tree, and Support Vector Machine classifiers are all used and compared in this study. In this study, we evaluate the efficacy of several classifiers by measuring their ability to correctly categorize mobile product data sets of varying sizes. Data were collected from popular online retailers like Amazon, Flipkart, and Snapdeal and analyzed to determine categorization accuracy. Naive Bayes, Random Forest, Decision Tree, and Support Vector Machines are some of the categorization algorithms compared here.

Keywords: Artificial intelligence, Customer opinions, Shopping reviews, Text classification.

INTRODUCTION

There are now a great deal of consumer reviews of items that are written entirely in text. Reviews are express customers' opinions about a product or a service [1]. Opinion mining is another name for sentiment analysis. In order to monitor how customers feel about their products or services and gain a better understanding of the industry, several companies use sentiment analysis on textual data [2]. Opinion mining, also known as sentiment analysis, is the practice of sifting through large amounts of user-generated content in search of expressions of positive or negative emotion [3]. Sentiment analysis is crucial because it allows businesses and organizations to quickly gauge the extent to which their customers are satisfied with a certain product on the basis of reviews [4]. The ever-increasing volume of product reviews makes it harder to do this analysis manually

[*] **Corresponding author Saniya Khurana:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: saniya.khurana.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

[5]. Sentiment analysis, commonly known as "opinion mining," is the process of analyzing and categorizing people's feelings about a text. It has to do with text and language analysis [6] as well as natural language processing. With the proliferation of web-based technology comes an explosion in data volume [7]. Expressing one's emotions with terse remarks is becoming commonplace on social networking platforms [8]. Emotions like joy, sorrow, worry, fear, *etc.*, fall under this category. The Bidirectional Encoder Representations from Transformers (BERT) is a cutting-edge language model employed in a wide variety of NLP and SM applications [9]. High amounts of unstructured data make it difficult and time-consuming to analyze consumer review data from different social media sites and make accurate predictions based on context-based sentiment. Recurrent neural network techniques, such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), as well as hybrid, neutral, and classic text categorization algorithms, have seen increased study in recent years. Successful businesses aim to please their customers and earn their favorable endorsement in a number of ways [10]. However, due to the vast amounts of data acquired from a variety of sources, assessing customer evaluations to accurately anticipate sentiment has proven to be a difficult and time-consuming task. Various scholars have developed algorithms, methodologies, and models to tackle this problem. Among them are the Artificial Neural Network (ANN) and the bag-of-word (BOW) regression model, as well as unigram and skip-gram-based algorithms. Numerous investigations and studies have uncovered polarity incoherence, model overfitting and performance difficulties, and high data processing costs.

Machine learning techniques are used to classify reviews. In this study, we identify emotions using six different machine-learning techniques. Research on these techniques' precision and effectiveness has been conducted. Customer feedback has also been subjected to feature sentiment analysis. Sentiment analysis and forecasting are common applications of machine learning. Sentiment analysis is performed on three different levels: document, phrase, and feature. Document-level sentiment analysis determines if a piece of writing is favorable, negative, or neutral. Both document-level and feature-level sentiment analysis are taken into account here.

The sections of this article are as follows: The research that informed this project is discussed in the next section. The analytical model is described in Section 3. The findings are presented in Section 4, and the conclusion is given in Section 5.

RELATED WORK

Naive Bayes, Random Forest, Decision Tree, and Support Vector Machine classifiers are all used and compared in this study. In this study, we evaluate the efficacy of several classifiers by measuring their ability to correctly categorize mobile product data sets of varying sizes. Data was collected from popular online retailers like Amazon, Flipkart, and Snapdeal and analyzed to determine categorization accuracy. Naive Bayes, Random Forest, Decision Tree, and Support Vector Machines are four categorization methods compared to choose the best [11].

The purpose of this research is to present a sentiment analysis model for positively, neutrally, and negatively categorizing product reviews. With the goal of developing the most effective classifier, it uses five well-known machine learning classifiers: Naive Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbor, and Maximum Entropy. The dataset utilized includes 82,815 ratings and comments posted on the Kaggle website pertaining to various mobile phone devices. We employed measures of recall, precision, F1-measure, and accuracy to compare the five different classifiers. The experiments all suggest that Maximum Entropy and Naive Bayes are the most accurate classifiers. Across all experiments, the Decision Tree algorithm performed worst [12].

The analysis of brief texts has a tendency to pick up on the collective mood of the audience. The “Sentiment Analysis” feature on IMDb lets users see how critics generally feel about a film. Costs are increasing as a result of the fact that human opinions matter for product success and that audience reception may make or break a film's box office performance. Sentiment analysis is performed by ML systems as a regular classification problem based on syntactic and linguistic features. In this study, we explore the application of machine learning techniques for sentiment analysis. SVM, RF, ANN, and NB, as well as the DT, BN, and KNN Algorithms [13], are commonly used for sentiment analysis.

In this paper, we report the results of our experimental study into how to improve upon the performance, accuracy, and context-based predictions of sentiment analysis models. Utilizing consumer evaluations from Twitter, IMDB Movie evaluations, Yelp, and Amazon, we present a fine-tuned BERT model to predict user attitudes. In addition, we gave a dashboard report that compared the suggested model's performance to that of our own bespoke Linear Support Vector Machine (LSVM), fastText, BiLSTM, and hybrid fastText-BiLSTM models. This experimental result demonstrates that the suggested model outperforms the competition across a range of performance metrics; transforms model; BERT;

CHAPTER 27

Analysis of Sentiment Employing the Word2vec with CNN-LSTM Classification System

Rajat Saini[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The identification ofproblems has become easier in sentiment categorization using conventional neural network–based short text classification methods . Word2vec, a convolutional neural network (CNN), and Bidirectional Long-term and Short-term Memory networks (LSTM) are used incombination to overcome this issue. Using Word2vec word embeddings, the CNN-LSTM model was able to attain an accuracy of 91.48%, as demonstrated experimentally. This demonstrates that the hybrid network model outperforms the single-structure neural network when dealing with relatively brief texts.

Keywords: CNN, LSTM, Sentiment Analysis, Word2Vec.

INTRODUCTION

In order to decipher public opinion on a wide range of issues, natural language processing is put to use in sentiment analysis [1]. This practice, also known as opinion mining, involves the system collecting, analyzing, and evaluating the opinions expressed in tweets. Today's social media comments contain several ideas and phrases, making sentiment analysis a difficult process [2]. To get public approval, it is helpful to monitor popular sentiment and respond appropriately. Thanks to advancements in Artificial Intelligence, determining the commenter's intent was a breeze. Numerous Neural Networks exist for handling this kind of problem [3]. In contrast, we employed a hybrid Deep CNN-LSTM Neural Network model to improve the comment vectors we obtained in this experiment [4]. Sentiment analysis makes extensive use of distributed word representations [5]. Word embeddings only take into account the meaning of words. As the Internet has grown, so has the amount of information shared by its users across many mediums [6]. Sentiment analysis is given more force when more perspec-

* **Corresponding author Rajat Saini:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: rajat.saini.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

tives and feelings from different people are made available. However, the lack of consistent labeled data in the Bangla NLP area makes sentiment analysis much more difficult [7]. Word2Vec, GloVe, and fastText are just a few examples of deep learning models that rely heavily on context-independent word embeddings, where each word has a fixed representation regardless of the surrounding text [8]. Meanwhile, pre-trained language models that take into account context, like BERT, have lately made tremendous strides in NLP. These days, social media platforms provide a flood of user-generated social data [9]. Conducting text sentiment analysis on the views stated by users is vital to comprehend people's thoughts and emotional inclinations on a commodity or event in a timely manner. The comment data from microblogging sites is especially complicated since it is usually a hodgepodge of different lengths of messages. Particularly abundant in long text data, the association between words is more intricate than in short text [10].

RELATED WORK

When using traditional neural network–based short text classification algorithms for emotion analysis, it is simple to make mistakes. Word2vec, a convolutional neural network (CNN), and Bidirectional Long-term and Short-term Memory networks (LSTM) are coupled to overcome this issue. Using Word2vec word embeddings, the CNN-LSTM model was able to attain an accuracy of 91.48%, as demonstrated experimentally. This demonstrates that the hybrid network model outperforms a neural network with a single structure on short texts [11].

Using CNN-LSTM, a deep learning approach that employs Word2Vec as a word embedding layer, our proposed model is able to extract the sentiment of tweets and classify them. 1.6 million Tweets that make up the Sentiment140 dataset were mined from the Twitter API. Tweets are classified using the memory-based state of LSTM cells. Texts conveying emotions like love, hate, friendship, and pride are shared on social networking platforms. We will determine if a tweet is favorable or bad by employing a deep learning technique. When compared to other methods such as SVM and Naive Bayes Classifier, CNN-LSTM performs exceptionally well [12].

In order to extract the characteristics of the word and build a weighted word vector, we have presented the conventional TFIDF technique in addition to Word2Vec embedding. These word vectors are weighted, and the max pooling layer and the CNN receive them as input. In order to get reliable results for sentiment classification, the produced output is forwarded to the Bi-LSTM neural network. The hybrid CNN-LSTM model, which combines Word2Vec with TF-IDF, was tested on the Stanford IMDB movie review dataset, where it achieved

impressive results. The model suggested here is based on a CNNLSTM variant of the Deep Neural Network.

Here, we applied BERT's capacity for transfer learning to a deeply integrated model CNN-LSTM in order to improve the accuracy and speed with which sentiment analysis decisions are made. In order to compare CNN-LSTM's performance with that of traditional machine learning algorithms, we also created the concept of transfer learning. Furthermore, we investigate several word embedding methods, including Word2Vec, GloVe, and fastText, and evaluate their efficacy in relation to the BERT transfer learning approach. Therefore, we have demonstrated state-of-the-art binary classification performance for Bangla sentiment analysis, vastly surpassing the performance of any embedding or technique [14].

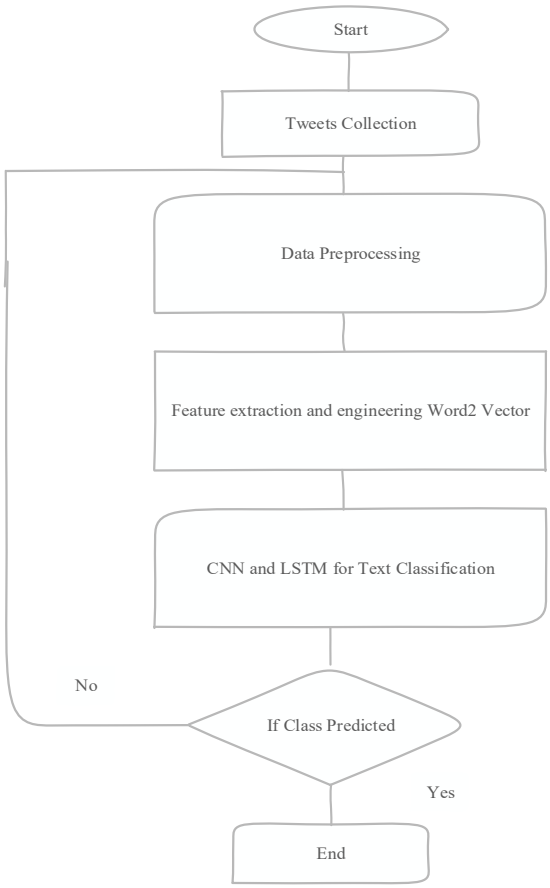


Fig. (1). Proposed work flow diagram.

CHAPTER 28

Hadoop-based Twitter Sentiment Analysis Using Deep Learning

Manpreet Singh[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Sentiment analysis, the practice of classifying and identifying opinions displayed in audio, words, database reports, and tweets to ascertain if the opinion is positive, neutral, or negative, is of great interest to many individuals in the microblogging service arena. It could be challenging to extract sentiment from Twitter data due to its quirks. The research suggests a way to analyze sentiment using a Hadoop infrastructure with a deep learning classifier. In order to gather characteristics, data is distributed across Hadoop nodes. After that, the data from Twitter is parsed for the most crucial parts. The input data from Twitter is sorted into two categories, positive review along with negative review, *via* a deep learning algorithm, including a deep recurrent neural networks classifier. Some of the metrics used to measure performance include classification accuracy, specificity, and sensitivity. With a sensitivity of 0.9404 and a generality of 0.9157, the proposed technique surpassed traditional methods in classification with a precision of 0.9302.

Keywords: Big data, Deep learning, Hadoop, Twitter sentiment analysis.

INTRODUCTION

Microblogging platforms are focusing on sentiment analysis, which involves categorizing and recognizing opinions expressed in audio, text, database sources, as well as tweets as positive, neutral, or negative [1]. Twitter data has unique characteristics that make sentiment analysis challenging. There are millions of people using Twitter, which makes it a very popular micro-blogging service [2]. Twitter users update their statuses, or "tweets," and discuss current events and topics by using hashtags. As a result, Twitter has emerged as a powerful and reliable barometer of public sentiment [3]. Twitter produces a massive volume of data, making it tough to manually examine all of it. Twitter sentiment detectors (TSDs) are superior to more conventional methods for gauging customer satisfac-

* **Corresponding author Manpreet Singh:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: manpreet.singh.orp@chitkara.edu.in

tion with the quality of a service or product [4]. Both the accuracy and performance of TSDs in detecting patterns and making accurate classifications are heavily reliant on the efficiency of the classification methods used [5]. Existing machine learning technologies have the potential to change firms into those handled by automated processes, but the time needed is a major concern [6]. Several real-world applications have made use of deep learning methods in areas like sentiment analysis [7]. To extract insights from unstructured data like texts or tweets and represent them in models, deep learning techniques use a variety of algorithms. A person's voice may be heard all around the world *via* their use of online social media [8]. People choose social networking sites because they make it simple to share their thoughts and views and stay abreast of the latest developments in the world of fashion and pop culture [9]. Every day, thousands upon thousands of tweets are sent on Twitter about a wide range of subjects [10]. One objective of the proposed sentiment analysis model is to use deep recurrent neural networks to analyze real-valued input data from Twitter. The assessment emphasizes the potential speed of findings for sentiment analysis. Feature extraction and classification are the two parts of the proposed procedure. Distributing the data *via* the Hadoop cluster is the first stage in obtaining features from input Twitter data. Combining the learned traits with a classifier known as a deep recurrent neural network (DARN) is what happens in the classification module. Using the obtained attributes, the classifier from deep recurrent neural systems produces two outputs: positive review and negative review. Section 2 shows a synopsis of conventional sentiment analysis methods, whereas Part 1 gives an introduction to sentiment analysis in general. Section 3 lays out the approach, Section 4 displays the findings, and Section 5 draws the conclusion.

RELATED WORK

Using a Hadoop architecture and a deep learning classifier, this study proposes a method for sentiment analysis. The distribution of data for feature extraction is handled by the Hadoop cluster. Then, the important elements are culled from the Twitter information. The input data from Twitter is sorted into two categories, positive review, and negative review, *via* a deep learning algorithm, namely a deep recurrent neural networks classifier. Performance evaluations make use of measures like specificity, sensitivity, and accuracy in categorization. The accuracy of classification (0.9302), sensitivity (0.9404), and specificity (0.9157) were all areas in which the proposed method excelled above more traditional methods [11].

Our new approach to sentiment analysis is the Hybrid Lexicon-Naive Bayes Classifier or HL-NBC. We first aggregate related subjects and filter out irrelevant ones before feeding them into the emotion assessment engine. The proposed

method is tested using Lexicon, a Bayesian naive classifier for bi-gram and uni-gram features. The suggested HL-NBC technique performs sentiment categorization more effectively than the other approaches and has an accuracy of 82%. We also saw a 93% reduction in processing time for bigger datasets when compared to conventional techniques of doing sentiment analysis [12].

These models are put to use in order to draw conclusions regarding unmodeled datasets. Our innovative method for sentiment analysis leverages deep learning architectures to improve detection accuracy as well as efficiency. It combines the "universal language modelling fine-tuning" (ULMFiT) with a support vector machine (SVM). This technique reveals a novel deep-learning strategy for assessing the sentiment expressed by users of Twitter on specific subjects. When tested extensively on three different datasets, our model consistently outperforms the state-of-the-art. For example, its accuracy reaches 99.78% when used on the Twitter US Airlines datasets [13].

This massive trove of unstructured data may be organised and processed to serve a variety of social, industrial, economic, and governmental purposes. Hadoop, with its ability to analyse massive datasets in parallel, is the finest technology for analysing Twitter data. Customers' actions may be better understood by analysing the views expressed on Twitter, which cover a wide range of issues. In order to determine the sentiment of each tweet, an analysis is performed utilising Hadoop and its ecosystem. In this study, we examine the efficacy of a Big data technique for doing sentiment analysis on Twitter data. We have utilised Apache Flume to continuously feed HDFS with tweets from Twitter. Tweets are retrieved from raw nested Twitter data using pig scripts. One way to classify tweets according to their attitude is by using a dictionary-based approach [14].

As a result of the incredible growth of social media platforms as a way for people to connect and share ideas, there is a deluge of data available on these sites. To describe this ever-growing database, the phrase "Big Data" has recently surfaced. As a result, sentiment analysis (also known as optimum mining) is one area where the use of new approaches from data mining research has considerable promise to obtain more accurate categorization of hidden information in Big Data. In this article, we build upon existing work by presenting a Nave Bayes/Random Forest hybrid strategy for mining Twitter datasets. In a nutshell, relevant data sets are gathered from Twitter through the Twitter API and the hybrid approach is then shown and compared to a Naive Bayes classifier-only approach. The results of the sentiment categorization suggest that the hybrid strategy is more accurate and time-efficient [15].

CHAPTER 29

A Contrast Between Bert and Word2vec's Approaches to Text Sentiment Analysis

Manish Nagpal[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: A novel approach to creating a dataset to train a neural network that can analyze the tone of social media postings is proposed in this study. The paper goes on to detail how the word2vec and BERT algorithms may be used in a neural network to analyze social media messages and evaluate their emotional tone. A hybrid of cosine similarity as well as ontological mappings based on a tweaked version of the Term Frequency-Inverse Document Frequency (TFIDF) characteristics is used by the algorithm for semantic searching. The execution includes sentiment analysis, phrase extraction process, textual belief indicator, keyword-based search, as well as text summary, among other things. Additionally, trials were carried out proving the efficacy of the methods presented. The efficiency of stemming and lemmatization of the text in creating a training set for sentiment analysis was also tested experimentally.

Keywords: Bert model, Text sentiment analysis, Training tweets, Testing tweets, Word2vector.

INTRODUCTION

The word vector representation of the text is often obtained by the use of tools like Word2Vec, GloVe, *etc* [1]. The word context is lost with these approaches. Sentiment analysis is a highly applicable technique that helps address issues like the disorganisation of online discussion threads and the need for precisely located user data [2, 3]. While there has been some progress in the field of Chinese sentiment analysis, there is still much left to learn. One pressing issue is how different text feature representation methods and feature selection mechanisms affect classification performance [4, 5]. Memes are a recent phenomenon in the realm of online material. Feelings about a topic, item, person, or thing are included in a meme [6]. Memes may take the shape of either text or a picture, or both. Memes may be caustic, critical, humorous, and even political [7]. The majo-

* **Corresponding author Manish Nagpal:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: manish.nagpal.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

rity of existing sentiment analysis techniques concentrate on text. The proliferation of the internet has made it possible to share opinions on a wide range of subjects through various online forums. It is crucial for businesses and data scientists to examine the information included in these remarks [8]. There is a plethora of approaches to data analysis. Recent years have also seen an uptick in the number of research works using language models for the purpose of sentiment analysis and text categorization. This study compares the results of two text categorization methods—Support Vector Machine (SVM) as well as Neural Network (NN) for text classification—with four text embedding methods—word2vec, Glove, BERT, as well as GPT-2—using user reviews from two common e-commerce sites, Zappos as well as Yelp [9, 10].

With Twitter as a case study, this research aims to undertake sentiment analysis as well as opinion identification on social networking information. One way to determine whether a tweet is neutral, good, or unfavorable in tone is to use an analysis of sentiment. Finding out whether a particular tweet contains useful opinions for the creation of products is the goal of attitude detection. Our primary goal is to do this by comparing the sentiment evaluation as well as opinion identification embedded words of Word2Vec as well as BERT. In order to determine which method works better with tiny to medium amounts of data, this research also compares standard machine learning with deep learning. The following are, hence, the aims of the research:

- Word2Vec and BERT both produce word insertions; compare their performance on sentiment evaluation as well as opinion identification.
- Evaluate how well Deep Learning as well as conventional Machine Learning method deals with data sets that are moderate in size.
- Use the results of the application case's sentiment assessment as well as opinion identification tasks to inform product recommendations.

RELATED WORK

A neural network framework for sentiment analysis of text is suggested in this study to solve this problem. It combines a language model that has been pre-trained with a bidirectional short-term long-term memory network as well as an attention mechanism. Bidirectional encoding representation from transformers is the name of this concept. To begin, the word vector including contextual semantic information is acquired using the BERT pre-training algorithm. The two-way long- and short-term memory network is then used in this article to derive context-related characteristics for deep learning. In order to conduct text sentiment categorization, highlight key points, and apply weights to the retrieved information, the attention mechanism is developed. On the SST (Stanford

sentiment treebank) test set, the accuracy rate may reach 89.17%, indicating an increase in accuracy relative to previous approaches [11].

In this research, we have used TF-IDF to rank the words in the text according to their frequency of occurrence, so determining their relative significance. To further investigate the emotional leanings of texts, we used SVM and ELM with kernels. Results from experiments demonstrate that ELM with kernels may provide accurate classifications more quickly than SVM [12].

This study uses text as well as image-based Indonesian memes to examine the effectiveness of four sentiment evaluation methods. Memes in text were first retrieved and then categorized utilizing supervised techniques for machine learning such as Decision Tree, Convolutional Neural Networks (CNN), Naive Bayes, and Support Vector Machines, among others. According to the results, the Naive Bayes method achieved the highest accuracy (65.4% in this case) when it came to analyzing sentiment on memes content [13].

In this research, we use language models to do Turkish sentiment analysis on datasets of hotel and film reviews. The linguistic examples are selected because they are uncommon in contemporary Turkish writing. The pre-trained Turkish language models BERT, ALBERT, ELECTRA, as well as DistilBERT have been trained as well as tested on these datasets. Also included is a text filtering method for sentiment analysis, the objective of which is to exclude terms from positively or negatively tagged content that can elicit the inverse sentiment. This technique also allows for the retraining of datasets using language models and the subsequent evaluation of model accuracy levels. This research draws parallels to others that have used the same datasets. The investigation produces state-of-t-e-art outcomes using language models, as measured by the accuracy values when compared to prior research. Training the ELECTRA language model with the suggested text filtering strategy yielded the best results [14].

The results demonstrate that BERT when combined with SVM and NN, is the top-performing text embedding approach across both datasets. Overall text classification using NN is also shown to be superior to using SVM. Our goal is to present a multidimensional comparison in an exploratory effort to identify the best algorithm for consumer review data, with the conclusion being that BERT and NN are capable of producing adequate results in the vast majority of cases [15].

CHAPTER 30

Text Emotion Categorization Using a Convolutional Recurrent Neural Network Enhanced by an Attention Mechanism-based Skip-Gram Method

Madhur Grover[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Text-based web archives have become increasingly common as technology has advanced. For many text classification applications, classic machine learning classifiers like support vector machines (SVMs) and naïve Bayes (NBayes) perform well. Since short texts have fewer words and convolutional and pooling layers have their limits, these classifiers suffer from sparsity and lack long-term dependencies. In this study, we present a convolutional recurrent neural network architecture that makes use of a modified skip-gram method. For the adversarial training of the skip-gram algorithm, we employ the L2 regularization technique. It can boost the model's performance in text sentiment classification tasks and increase its robustness and generalizability. To extract information from the entire text while dampening the influence of irrelevant words, we deployed a convolutional neural network equipped with attention mechanisms. The CNN-based categorization of text emotion is complete. When compared to other classifiers used on the Twitter dataset, our model and algorithm were shown to be more efficient and accurate.

Keywords: Attention scheme, CNN, Skip-gram model, Text emotion classification.

INTRODUCTION

Text-based web archives have become more common as technology has advanced. For many text classification applications, the tried-and-true SVM and naive Bayes classifiers of traditional machine learning excel [1]. However, owing to the scarcity of text in brief text and the constraints of convolutional and pooling layers, these classifiers suffer from sparsity difficulties and lack long-term depen-

* **Corresponding author Madhur Grover:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: madhur.grover.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

dencies [2]. Convolutional neural networks (CNNs), which are typical examples of deep learning technology, have found widespread use in machine vision tasks including object detection and image processing in recent years [3, 4]. To improve the accuracy and flexibility of traditional CNNs, we introduce an RCNN for object identification in this study [4]. However, RCNN is only capable of handling data at the local level and cannot adapt to a wide variety of scenarios. The rapidly expanding area of neural network models used in processing natural languages is attracting more attention from sentiment analysis [5]. With more people being involved in social media, there is more data to analyse. Sentiment analysis at the aspect level is used to determine the text's polarity in various contexts [6]. As a subfield of NLP, sentiment evaluation is starting to take an interest in models based on deep neural networks [7]. More people using social media means more data, which has made analysis more difficult [8]. Sentiment analysis at the aspect level is useful for determining the overall tone of a text in its various contexts. This study introduces four deep neural network–based algorithms for opinion mining at the aspect level [8], each with a unique input word vector format. We introduce a novel approach to aspect-level opinion extraction that integrates a network of convolutional neural networks, an attention mechanism, as well as a gated recurrent unit, all of which make use of different input vector representations. This study adds to the literature by presenting an innovative method for evaluating service quality based on patron feedback in the hospitality industry [9]. When scientists contrasted their technique to others using gathered feedback from restaurants as a test case, they discovered that it obtained excellent scores of precision, recall, and accuracy, as well as the f-measure [10, 16].

RELATED WORK

To identify emotions in spoken language, we provide a deep convolution recurrent neural network (CNN) that uses the log-Mel filterbank energies. The task of learning distinguishing characteristics falls on the convolutional layers. We also propose a convolutional attentiveness method for learning the task-relevant utterance structure, on the premise that a better understanding of the inner workings of an utterance can help reduce misinterpretation. In order to describe the features of emotional speech, we additionally measure the performance boost produced by every component in our model. Supporting our notion are experimental results on the eNTERFACE'05 emotions dataset that demonstrate a 4.62 percent increase over the existing gold standard technique [11].

In this study, we present the architecture of a convolutional recurrent neural network trained using an enhanced skip-gram technique. For the adversarial training of the skip-gram algorithm, we use the L2 regularisation technique. Not

only may it boost this model's performance in text sentiment classification tasks, but it can also make it more resilient and generalizable. To extract meaningful data from texts while decreasing the influence of irrelevant words, we used a Bidirectional Long Short-Term Memory network with attention mechanisms. CNNS-based text sentiment categorization is at last a reality. When compared to other classifiers used on the IMDB dataset [12], this model and method were shown to be the most efficient and accurate.

We propose a recurrent convolutional neural network (RCNN) that incorporates a self-attention mechanism (A-RCNN) based on findings from cognitive science and neuroscience to solve this problem. In contrast to the majority of methods, which include the mechanism for self-attention in the convolutional parts of a CNN, our technique compartmentalizes it into a distinct layer inside the RCNN. Publicly available datasets like CIFAR-10, CIFAR-100, as well as MNIST, are used to evaluate our A-RCNN. Results from experiments show that when it comes to object identification tasks, our proposed A-RCNN is more accurate than both the initial RCNN as well as the CNN [13].

To tackle the problems of poor accuracy and excessive computation, researchers suggested sEMG gesture recognition approaches using a type of dilated neural network with convolution that combines an attention mechanism with a bidirectional gated recurrent unit. By adjusting the size and expansion rate to a parity hybrid, HDC improves upon normal CNN in several aspects, such as increasing the field of reception, decreasing over-fitting, as well as extracting more features. The BiGRU module can efficiently extract and analyse the data's timing characteristics, while the Attention module prioritises the most crucial ones, resulting in improved accuracy. Using the NinaproDB1 dataset and our dataset, we were able to attain accuracy rates of 92.72 and 97.85, respectively, demonstrating the method's efficacy [14].

This research introduces four deep neural network–based approaches for opinion mining at the aspect level, each with a unique input word vector format. A novel approach is detailed for aspect-level mining of opinions using multiple input vector representations; it integrates a convolutional neural network, a gated recurrent unit, as well as an attention strategy. This paper contributes to the current literature by adding unique methods for evaluating service quality in the hospitality industry based on patron feedback. Research using collected restaurant reviews has shown promising results in terms of accuracy, precision, recall, and f-measure when compared to other methods [15, 16].

CHAPTER 31

Multimodal Sentiment Analysis in Text, Images, and GIFs Using Deep Learning

Deepak Minhas[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: More and more people are inclined to use GIFs, videos, and photographs on social media as a way to convey their feelings and thoughts. We developed a Python-based multimodal sentiment analysis tool for various Twitter formats, taking into account not just the text of a tweet but also its accompanying GIFs and pictures, for more precise sentiment scoring. We employ fine-tuned CNN for image sentiment analysis, VADER for text analysis, and image sentiment and facial expression analysis for GIFs, with each frame individually analyzed. Our research shows that combining textual and picture data yields superior outcomes compared to models that depend only on either images or text. The output scores from our text, picture, and GIF modules will be aggregated to get the final sentiment score for the incoming tweets.

Keywords: DL, GIFs, Images, Multimodal text classification, Text.

INTRODUCTION

GIFs, movies, and photos are becoming more popular ways for social media users to convey their feelings and thoughts. To improve the accuracy of the overall sentiment score for a tweet, we developed a multimodal sentiment analysis programme in Python that takes into account not just the text of the tweet but also GIFs and pictures [1]. We use fine-tuned CNN for image sentiment analysis, VADER for text analysis, and image sentiment and face expression analysis for GIFs across all frames. Instagram is a well-liked social networking platform that is used by many different types of individuals, from hobbyists to professionals [2]. Instagram postings may include text, images, and videos, and are thus quite popular. Many people add captions or other text to the photographs they post online. The emotion of these postings may be analysed by taking into account both the text and the accompanying picture. For this, it is necessary to represent

[*] **Corresponding author Deepak Minhas:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: deepak.minhas.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

the connection between the textual and visual modalities [3]. To solve this issue, we suggest using a multimodal deep learning approach. Sentiment analysis [4] may be used to deduce people's thoughts and views about a topic or product from online social media postings. Data generated every day by users may be in several modalities including text, picture, audio, and video due to the widespread use of the Internet and smart devices. By combining analyses of text, images, and audio, multimodal sentiment analysis may identify the underlying tone of a user's postings [5]. Integrating the sentiment and meaning of a post from several sources is a key problem for multimodal sentiment analysis. The same categorization strategies are used to examine many modalities in numerous research [6]. It turns out that various feature sets may optimally be used by a variety of classifiers. Web data sentiment analysis is a large research topic with several potential uses, including context learning, election outcome prediction, and incident reaction analysis [7]. For the most part, online data sentiment analysis has only considered a particular modality, such as text or images. Multiple modal information, such as an image and various types of text, is easily accessible and may be used together to improve the accuracy of sentiment estimation [8]. It has been shown that integrating visual and textual variables without considering their interrelationships leads to an overly complicated model and lowers sentiment analysis performance [9]. To identify utterances in natural language when the intended meaning varies from the surface meaning, sarcasm detection is used. In the field of natural language processing, sarcasm detection is used for a variety of applications, including sentiment analysis and opinion mining. Many of the seminal studies in the field of sarcasm detection [10] have narrowed their attention to textual input alone. The advent of social media has resulted in a dramatic increase in the volume of multimodal data in the modern world. Research works on visual sentiment analysis especially for GIF videos have only just started to develop. This is mainly because of the fact that the ability to use GIFs on big platforms like Twitter, Instagram, Facebook, and Telegram was introduced recently thus multimodal designs that use GIFs as one of the modals remained unexplored. Our framework includes:

- A multimodal sentiment framework that analyzes the incoming tweet and splits it into different modals like text, image, and GIFs. Based on the type of input, the respective processing module will calculate sentiment for that type.
- An image module consists of a fine-tuned CNN, which is built on top of a pre-trained vgg16 CNN model that is trained with annotated Twitter data set. We use this module to analyze a tweet’s image or a GIF’s separated frames.
- We demonstrate the feasibility of using face detection and emotion detection with the combination of the image module to analyze a GIF for its sentiment.

RELATED WORK

In this research, we show that combining textual and visual characteristics yields superior outcomes compared to models that depend only on visual or textual data. The output scores from our text, picture, and GIF modules will be aggregated to provide the final emotion score for the incoming tweets [11].

The suggested technique employs a 2-dimensional convolutional neural network (2CNN) for image analysis and a bi-directional gated recurrent unit (bi-GRU) for processing text comments. We provide a new dataset of Instagram posts called MPerInst, which contains 512 pairs of photographs and their related Persian language comments, in order to evaluate the effectiveness of the suggested model. Based on experimental data, combining text and picture modalities yields a 23% increase in polarity identification accuracy and a 0.24 increase in F1-score compared to either modality alone. The suggested model also has a higher accuracy and F1-score than 11 other deep fusion models. Our data and model code are accessible for anyone to utilise in the future [12].

To capitalise on the efficient performance of many classifiers across multiple modalities, this research proposes a soft voting-based ensemble model. The proposed model uses deep learning techniques (BiLSTM, CNN) to extract deep features from multimodal datasets. The final feature sets were categorised using a soft voting-based ensemble learning model after feature selection was performed on the features, which are a combination of text and picture data. Two benchmark datasets of text-image pairings have been used to evaluate the proposed approach. The experiments have shown that the suggested model performed better than several adversarial models on the same datasets [13].

This research proposes a paradigm for sentiment analysis that takes advantage of the connections between different types of online data by integrating visual and linguistic signals that are likely to attract the most attention. To discover the connections between the learned salient visual aspects and textual data, a multimodal deep association learner is constructed. In addition, two streams of unimodal deep feature extractors are suggested to automatically learn the visual and linguistic aspects that are most important to the feelings. Finally, a late fusion process is used to estimate the sentiment from the merged characteristics. Compared to both current unimodal algorithms and multimodal approaches that blindly merge the visual and textual elements [14], our proposed framework shows promising results for sentiment analysis utilising online data.

Because of this, users nowadays often use visuals, such as photographs and videos, to supplement their written words while expressing themselves online. In this research, we aim to improve upon existing sarcasm detection systems by

CHAPTER 32

Public Opinion Regarding COVID-19 Analyzed for Emotion Using Deep Learning Techniques

Abhinav Mishra[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: As a result of the COVID-19 epidemic, many individuals are experiencing extreme worry, dread, and other difficult emotions. Since coronavirus immunizations were first introduced, people's reactions have gotten more nuanced and varied. In this study, we will use deep learning methods to decode their emotions. Twitter provides a glimpse into what is popular and what is on people's minds at any given time, and social media is presently the finest means to convey sentiments and emotions. Our goal while conducting this study was to have a better grasp of how different groups of individuals feel about vaccinations. The research period for this research's tweet was from December 21st to July 21st. Of the most talked-about vaccines that have recently been available in various regions of the world were the subject of several tweets. The term Valence Aware Sentiment Dictionary An NLP program called Believed (VADER) was used to examine people's sentiments on certain vaccines. We were better able to see the big picture after categorizing the collected attitudes into positive (33.96 percent), negative (17.55 percent), and neutral (48.49 percent) camps. We also included into our study an examination of the tweets' chronology, given that attitudes changed over time. The performance of the forecasting models was evaluated using an RNN-oriented design that included bidirectional LSTM (Bi-LSTM) as well as long short-term memory (LSTM); LSTM attained an accuracy of 90.59% as well as Bi-LSTM of 90.83%. Additional performance metrics, such as Precision, F1-score, as well as a matrix of confusion, were used to confirm our hypotheses as well as outcomes. The findings of this research provide credence to efforts to eradicate coronavirus across the globe by expanding our knowledge of public opinion on COVID-19 vaccines.

Keywords: COVID-19, Deep learning techniques, Social media networks and Bi-LSTM.

[*] **Corresponding author Abhinav Mishra:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: abhinav.mishra.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

INTRODUCTION

Sentiment Analysis [1] is a classification issue that involves mining social media and microblogging sites for user reviews on popular items, movies, restaurants, political candidates, and other subjects of the day. As the 2019 coronavirus (COVID-19) remains to be a global health problem, people are using social media like Facebook and Twitter to voice their concerns and provide solutions [2]. Photos, texts, videos, and sounds uploaded by users have all seen a rise in popularity on social networking sites in the past few years [3]. To gauge public sentiment towards COVID-19-related activities, Twitter [4] is a popular choice among social media platforms. Because tweets are so short and dynamic, this is the case [5]. To get a sense of how people feel about COVID-19, we provide a deep learning method for analyzing tweets on the film. Worldwide attempts at widespread vaccination against COVID-19 have come across significant skepticism as well as vaccine reluctance or refusal [6], despite the vaccinations' proven effectiveness, safety, as well as accessibility. Twitter analysis of negative COVID-19 vaccine sentiment might lead to novel strategies for boosting vaccination uptake [7]. Social media's popularity is not at all that high in relation to other services on the market today [8]. People all around the globe use Twitter to broadcast their ideas and sentiments to friends, family, and the rest of the world. Extracting public opinions on various topics from social media platforms, such as reviews of films, restaurants, and news stories, is known as sentiment analysis [9]. We are all aware that the epidemic of COVID-19 has affected the whole planet. People's emotions, opinions, and conflicting views about the situation were all on display as the sentiment spread rapidly [10].

RELATED WORK

In this study, we use a neural network made up of convolutions with Bidirectionally Long-Short Term Memory (CNN-Bi-LSTM) to analyze hashtags related to the COVID-19 worldwide epidemic. A deep learning system is employed to ascertain a user's emotional state, which may be positive (80%), negative (16%), or neutral (0%). Using FastText as well as Globe models that were previously trained, the proposed method successfully cleaned the data as well as extracted word embeddings for out-of-the-ordinary phrases in our corpus. Metrics such as precision, recall, accuracy, and f1 were calculated to assess the performance of a CNN-Bi-LSTM hybrids model. CNN-Bi-LSTM trained using the FastText model beat CNN-Bi-LSTM taught with the GloVe approach, 99.33% to 97.55%, in experimental tasks [11].

The suggested method is based on an LSTM-RNN network that uses attention layers to provide more weight to features. This methodology uses the attention

method to enhance the feature transformation framework. Twitter data was collected from the Cagle dataset and categorized as either "sad," "happy," "scared," or "angry." The proposed deep learning model significantly outperformed the baseline methods, with enhanced accuracy of 20%, precision of 10% to 12%, and recall of just 12-13%. Adding attention layers to the preexisting LSTM-RNN method allowed for this enhancement to be realized. The majority (45%) of 179,108 tweets we analysed were positive, while 30% were neutral, and 25% were unfavorable. This proves the viability and efficiency of the proposed deep learning technique for analyzing the opinions expressed on COVID-19 [12].

Therefore, this research analyzed public tweets over a 16-month period to determine the level of opposition to COVID-19 immunization. From April 1, 2021, to August 1, 2022, we culled the original English tweets. The model used here is based on bidirectional encoder reconstructions from transformers (BERT), and it was used to select those tweets that made negative emotional claims. After every round of topic modeling as well as manual thematic analysis, those involved in the study independently assessed the topic labels as well as themes. The total number of tweets examined was 4,448,314 times. Six issues and three themes connected to the widespread disapproval of the COVID-19 vaccine immunisation were identified *via* investigation. Emotional responses to what are seen as unfair regulations or worries about the COVID-19 vaccinations' efficacy and safety are two possible interpretations of the common threads. The findings of the current infodemiology research provide significant public debates to be held and prospective paths for future policy intervention and campaign activities [13], and they are consistent with the rising vaccination model.

The primary goal of this article is to examine the feelings communicated *via* Twitter. Twitter tweets related to Corona are gathered, cleaned using a pre-trained word embedding model, and then fed into a convolutional neural network, a long short-term memory network, and a convolutional neural network trained using a bias network (CNN-BiLSTM). Accuracy, precision, and recall methods are used to assess the model [14].

To estimate the immunological response to repeated exposure to carbon monoxide poisoning, this study proposes a unique probabilistic technique [15].

PROPOSED WORK

System Overview

The dataset used in this study was gathered from Cagle and consists of several tweets on COVID-19 vaccines. Data characters are subsequently detokenized to separate phrases into words as well as label them after validating the unique

CHAPTER 33

CNN-based Deep Learning Techniques for Movie Review Analysis of Sentiments

Prateek Garg[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India.*

Abstract: Twitter, Facebook, Instagram, *etc.* are just a few of the many online discussion platforms that have sprung up as a result of the explosion in internet use and popularity, giving individuals a place to air their views on current events. Films get both acclamation and criticism from the general public. As a major form of entertainment, they inspire user evaluations of film and television on websites like IMDB and Amazon. Scientists and researchers give careful thought to these critiques and comments in order to extract useful information from the data. This data lacks organisation but is of critical importance nevertheless. Opinion mining, also known as sentiment classification, is a growing field that uses machine learning and deep learning to analyse the polarity of the feelings expressed in a review. Since text typically carries rich semantics useful for analysis, sentiment analysis has grown into the most active investigation in NLP (natural language processing). The continuous progress of deep learning has substantially increased the capacity to analyse this content. Convolutional Neural Networks (CNN) are commonly utilised for natural language processing since they are one of the most successful deep learning methodologies. This paper elaborates on the methods, datasets, outcomes, and limits of CNN-based sentiment analysis of film critics' reviews.

Keywords: CNN, Deep learning techniques, Movie reviews, Sentiment analysis.

INTRODUCTION

The proliferation of electronic items has coincided with the rise in popularity of online social media such as Twitter and Microblog [1]. The challenge of efficiently extracting and categorising the sentiment behind such a large volume of views, often conveyed by plain text, has captured the interest of data analysts [2]. Thanks to advancements in deep learning algorithms, NLP models have become more effective in opinion mining and text categorization [3]. Since there is such a wealth of opinionated content on the web, many researchers are focused

[*] **Corresponding author Prateek Garg:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: prateek.garg.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

on sentiment analysis [4]. When considering an e-commerce service or product, the majority of customers read review articles before making a purchase. Unfortunately, the prevalence of false evaluations [5] is the fundamental issue with these reviews, and it is not fully handled. In order to assist companies and organizations in enhancing their marketing strategies and getting a thorough understanding of customers' thoughts on their goods and brands, the suggested approach makes use of opinion mining on user evaluations [6]. There is a vast amount of unstructured information [7] that has accumulated on social media as a consequence of individuals using the platform to engage with one another and express their ideas in the form of comments and opinions on various issues and stories. When it comes to gauging how users will respond, sentiment analysis is the gold standard. Many different techniques based on machine learning and natural language processing have been used to probe these emotions in the past [8]. However, owing to their efficacy, deep learning-based solutions are becoming more popular. Image processing is now a key growing technology in the realm of AI-based systems. Emotion recognition [9] is explored using artificial intelligence machine learning approaches. Emotion recognition is the study of how a person's facial expressions, body language, and vocal tone reveal what they are thinking and feeling at any given moment. One may then ascertain whether or not the person has any desire to continue with the current event. The core purpose of Audience Reaction Analysis is the examination of the emotional state of the audience as a whole that may be obtained by collecting and analysing every attendee's facial expression before, during, and after a seminar, discussion, lecture, or keynote [10].

This study suggests a paradigm for large-scale convolutional neural networks with seven layers. The information is first transformed into word vectors. Word2Vec has been called one of the Keras' most potent embedding layer approaches. The convolutional layer is used twice. In order to create a feature map, the first 1D a convolutional is applied to the input information to extract relevant features. The characteristics chosen by the first layer of convolution are summarised by the second. The output's resolution is lowered and overfitting is avoided thanks to the global max-pooling layer. In order to prevent overfitting and improve generalisation, the layer of dropouts is implemented. In this layer, nodes are removed from the network at random and their connections to the outside world are temporarily severed. In order to determine whether an input feature is positive or negative, the dense layer applies the loss function from the training dataset.

Here is how the rest of the paper is structured: Movie reviews, machine learning, and sentiment analysis literature are introduced in Section 1.1. The suggested solution seven-layer CNN model is outlined in Section 2, and experimental results

and discussions are presented in Section 3, and the paper is concluded in Section 5.

RELATED WORK

In this study, we retrieved 25,000 evaluations and their accompanying sentiment labels in an effort to build an analysis of sentiment for film criticism. To determine whether an audience member liked or disliked the film, we improved upon the traditional Bag of Words (BoW) model by using the TF-IDF method to vectorize the data and trained a Convolutional Neural Network (CNN). We compare the model's performance in identifying positive and negative sentiment separately and report its overall accurateness and stability. The best model has an average 'Accurateness' of 80.62 percent, a standard deviation of 1.33 percent, and a sensitivity and specificity of 76.4 percent for identifying positive sentiment and 84.6 percent for identifying negative sentiment [11].

To better assess a given system, it is necessary to develop new models that can identify and categorise the emotions communicated *via* an electronic text. Many different approaches have been offered to address this issue, the vast majority of which make use of Machine Learning methods. The capacity to learn and extract meaningful information from data has made Deep Learning a fast-expanding area that has already been shown to be helpful in solving many difficult issues. As a result, a number of works seek to use this strategy in areas of sentiment analysis such as sentiment categorization. In this paper, we propose a novel approach to user sentiment analysis using NLP and a deep convolutional neural network. The 50,000 movie reviews that make up the IMDB dataset have been used to test the suggested approach. With a training phase Accurateness of 99% and a testing phase Accurateness of 89%, the findings obtained were highly persuasive [12].

To analyse the sentiment of online customer reviews, we use a deep learning a convolutional neural network with a long short-term memory (CNN-LSTM) approach to this research. Consumers' reviews of cameras, PCs, cell phones, tablets, TVs, and video security systems on Amazon were included in the real-time data collection utilised to test and evaluate the system. In the course of the data pretreatment process, we used lowercase the process, stopword deletion, punctuation removal, and tokenization. To determine if a consumer is feeling happy or sad, we used LSTM and CNN-LSTM algorithms to clean the data. There was a 94% increase in accuracy using the LSTM method and a 91% increase using the CNN-LSTM method. To sum up, we find that the deep learning methods utilised here offer the most reliable forecasts of how buyers would evaluate a product's quality [13].

CHAPTER 34

Machine Learning and Deep Learning Models for Sentiment Analysis of Product Reviews

Saket Mishra[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: In our research, we use sentiment analysis to determine how well ratings and reviews are compared on Amazon.com. The process of determining whether a text's tone is favorable or negative and labelling it as such is known as sentiment analysis. Consumers may write evaluations on e-commerce sites like Amazon.com and indicate the polarity of their opinion. There is a discrepancy between the review and the rating in certain cases. We used deep learning to analyze the sentiment of Amazon.com product reviews in order to find reviews with inconsistent star ratings. A paragraph vector was utilized to transform textual product evaluations into numeric data that was then fed into a neural network with recurrent equipped with a gated recurrent unit for training. We built a model that takes into account the review text's semantic connections to the product data. Additionally, we built a web service that uses the trained model to predict the rating score of a submitted review and gives feedback to a reviewer if the anticipated and submitted ratings do not line up.

Keywords: Amazon, Deep learning techniques, Machine learning, Product reviews, Sentiment analysis.

INTRODUCTION

Sentiment analysis, also known as opinion mining, has emerged as a hot topic in academic circles in recent years, thanks to the proliferation of social media and e-commerce platforms. Sentiment analysis is a method used to determine how someone feels about a certain topic based on the words they choose to use [1, 2]. On e-commerce sites like Amazon.com, customers may provide ratings alongside their reviews. Natural language processing (NLP) and machine learning techniques are used in sentiment analysis to predict the speaker's intent from a given statement. Sentiment analysis entails dividing text into positive, negative, and neutral categories. Opinion mining [3] is the practice of systematically collec-

* **Corresponding author Saket Mishra:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: saket.mishra.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

ting and analysing user feedback on a certain subject, product, or issue. Sentiment analysis is the process of analysing the feelings reflected in the collected opinions. In today's highly connected world, e-commerce now accounts for the vast majority of all retail activity [4]. Customers may avoid the mall altogether by ordering products from the convenience of their smartphones and having them delivered to their doorsteps [5]. However, this commercial platform has the downside that consumers have little to no idea of the worth of the goods they are purchasing. So, such a platform service often has a place for clients to offer feedback once they've purchased the goods [6]. You may learn a lot about how satisfied your customers are just reading their evaluations. Customer feedback scores may help business managers determine whether a review trend is favourable or bad, but human analysis of this data is time-consuming and error-prone [7]. The trend of the modern world is towards digitalization. The thriving e-commerce sector of the digital economy is helping people everywhere save time and effort by allowing them to shop for goods without leaving the comfort of their own homes [8]. E-commerce systems may utilise sentiment analysis to determine which goods are popular with customers, which ones need to be tweaked or scrapped entirely, and why. The company product owners may use this input to better understand the market and make informed business choices [9]. The proliferation of online product evaluations has contributed to the rise of e-commerce in recent years. Customers' purchasing decisions are heavily influenced by the recommendations and complaints of other buyers. Interpreting and categorising textual data are called sentiment polarity analysis [10].

RELATED WORK

Our goal in writing this study was to use Amazon's dataset to develop sentiment analysis for product ratings and text reviews. Several different machine learning methods, including Linear Support Vector Machine Ma-chine, a Random Forest, Multinomial Neural Networks, Bernays Naive Bayes, and Logistic Regression, were implemented. Using a Random Forest classifier, we were able to achieve a 91.90% success rate. In addition, we employed RNN using LSTM as an approach to deep learning and obtained the highest accuracy (97.52%) possible. RNN-LSTM is the best method for our model [11].

Opinion mining and Sentiment Analysis go towards this end by teaching computers to understand and convey human emotions. Users may easily and quickly express their thoughts and opinions using platforms like Facebook, Twitter, Amazon, Flipkart, and others. Using a machine learning technique called sentiment analysis, consumers' feelings, thoughts, and opinions about a product may be categorised and analysed automatically. Naive Bayes, Support Vector Machine (SVM), and Decision Tree are the most often used categorization models

in product analysis. When compared to existing machine learning methods, the suggested method will provide superior results [12].

In this study, we use machine learning approaches to build computational models that can automatically identify favourable and negative evaluations of a product. We perform our studies using Amazon.com book review data. The bag of words approach was used to adapt the data for a machine learning-oriented strategy, and then models were developed using logistic regression, naive Bayes, support vector machine, as well as neural network techniques. Word embedding was one of the methods we utilised to convert the data for a deep learning approach before moving on to LSM and gated recurrent units. We evaluate the preprocessed and unprocessed datasets to evaluate the efficiency of machine learning and deep learning models, respectively. As a consequence, on both preprocessed and unprocessed datasets, the bag of phrases in neural network achieves better results than any other method [13].

This research conducts a comparative investigation of sentiment analysis of Amazon customer product evaluations using SVM and LSTM and CNN, two machine learning and deep learning classification algorithms [14].

Our primary objective is to classify each customer review according to how favourable or negative we think it is. Each of these phases is essential to the success of our sentiment polarity detecting system: preprocessing, the extraction of features, training, classification, and generalisation. Tf-Idf and Tokenizer were used to first convert the reviews into a vector format. Then, we used LSTM, Linear SVM, RBF, and Sigmoid kernel deep learning models for training. The models were then assessed according to their accuracy, f1-score, precision, and recall. For Amazon customer reviews, our LSTM model forecasts 86% accuracy while for Yelp reviews, it predicts 85% accuracy [15].

PROPOSED WORK

System Model

We developed a model employing recurrent neural networks (RNN) with gated recurrent unit (GRU) which learned low-dimensional vector representation of reviews utilising paragraph vectors and product embeddings in order to interpret the sentiment of Amazon.com reviews. To begin, we used paragraph vectors to turn Amazon.com product reviews into data with a known length. These vectors of features were grouped according to their products and ranked by their timestamps. An RNN with GRU was trained using each cohort. Vectors produced by the RNN's second-to-last layer are known as product embeddings. Product attributes and chronological interactions among reviews are only two examples of

CHAPTER 35

Sentiment Analysis of Hotel Reviews Based on Deep Learning

Jagmeet Sohal[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Punjab, India*

Abstract: Many hotel evaluations have been written and shared online these days. Machine learning sentiment classification requires complicated artificial design features and feature extraction technique, whereas emotion dictionary-based sentiment classification requires a large amount of emotional database resources. In this study, we present the idea of long short-term memory. The text categorization method is used to determine the general tone. First, the brief comment text is processed into the LSTM network using word2vec and word segmentation technology; next, a dropout technique is implemented to avoid overfitting in order to get the final rating model. By using the LSTM network's superior short-term memory, a positive impact has been realized on sentiment categorization of reviews of hotels, with a precision of more than 95%.

Keywords: Deep learning, E-commerce, Hotel reviews, Sentiment analysis.

INTRODUCTION

When comparing word vector sizes, word embedding vectorization outperforms Bag-of-Word. Sentence-level context, word order, and semantic links between words are all lost in conventional NLP approaches, although word embedding may compensate for this [1]. Word2Vec and Fast Text are only two examples of the many Word Embedding methods that may be used for sentiment analysis. While Word2Vec is word-based, Fast Text operates on N-Gram. Customers' reliance on online evaluations posted to a variety of digital platforms has grown substantially in recent years [2]. Products are generally approved or rejected based on reviews and ratings by customers on E-commerce platforms like Flipchart, Amazon, *etc.* [3]. Reviews of services, such as those offered by hotels, airlines, and restaurants, are just as important to consumers as reviews of items. Using sentiment analysis, programmers may quickly sort user feedback into positive and

[*] **Corresponding author Jagmeet Sohal:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Punjab, India;
E-mail: jagmeet.sohal.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

negative categories [4]. Customer opinion is a major factor in how products are reviewed, but there is still room for development in sentiment analysis that accounts for specifics [5]. There are often five things to consider while choosing a hotel: proximity, food, service, comfort, and cleanliness [6]. This study suggests techniques for analyzing hotel reviews in terms of their underlying sentiment. Preprocessing a hotel review corpus into a term set [7]. After a word list's hidden subjects have been identified using Latent Dirichlet Allocation (LDA), the list is then categorized into the five facets of a hotel with semantic similarity [8]. The word Frequency-Inverse Cluster Frequency (TF-ICF) technique is then used to broaden the word list for the purpose of determining the similarity measurement. Finally, word embedding and Long-short Term Memory (LSTM) are combined to classify consumer mood (satisfied or unhappy) [9]. The evaluations were successfully categorized into the five hotel features using the suggested approach. We created a method for analyzing guest feedback on hotels based on their mood. Through the usage of themes, the system is able to provide precise reports on user-selected attributes [10]. Users are able to rate the quality of a hotel's services based on a variety of factors with the use of this system. In addition, a traveler might examine hotels in a variety of ways. Sentiment detection refers to the technique of distinguishing between biased and nonbiased evaluations. To improve the classifier's efficiency, sentiment detection is performed before sentiment classification. Sentiment analysis may be performed on several levels, including the word, phrase, paragraph, and document. Sentiment analysis may be conducted at two different levels: the document level, where the whole content is given a single sentiment, and the paragraph level, where the sentiment for every paragraph is calculated independently. There are further methods that include using sentiment analysis on a word-by-word basis. Since sentences include fewer words than paragraphs and papers, it is more difficult to correctly extract polarity at the sentence level. Deep learning has outperformed traditional methods on many categorization tests.

Contributions

Sentiment analysis is the automated examination of emotional overtones in written or spoken communication by use of natural language processing techniques. Sentiment analysis may help tourist businesses monitor client feedback more effectively. In this research, we provide a deep learning-based method for automating the sentiment analysis of hotel reviews by making use of word embeddings and a gated recurrent unit. Our use of a deep learning model achieved 89% accuracy and 92% F-score in sentiment categorization of hotel reviews, outperforming the performance of classic machine learning approaches.

Section 2 provides a brief overview of current research and developments in the area of sentiment analysis. In Section 3, the overall structure of the system is outlined. The system's implementation is shown in Section 4. Methods and outcomes are evaluated in Section 5. Our findings are summarized in Section 6.

RELATED WORK

Hotel ratings and comments posted online by previous guests are essential when deciding where to stay. As a result, hotel management would benefit greatly from having access to this information. We typically provide a system that gathers user feedback, organizes that feedback into coherent, structured overviews, and makes that information easily accessible [11].

The purpose of this study is to evaluate Word2Vec and Fast Text as two competing sentiment analysis models and compare their accuracies. The TripAdvisor hotel review sentiment analysis dataset is used to compare the two methods. The Skip-gram model is used by Word2Vec and Fast Text. The number of features, the minimum number of words, the number of concurrent threads, and the size of the context window are all the same in both approaches. The ensemble learning techniques of Random Forest, Extra Tree, and AdaBoost are used to combine these vectorizers. Both models' efficacy is compared to that of the Decision Tree as a standard. Accuracy on random forest and Extra Tree was shown to be significantly improved by using either Fast Text or Word2Vec. When compared to Word2Vec, Fast Text's accuracy with extra tree and random forest classifiers was superior. Accuracy of 93% with 100 estimators [12] demonstrates the efficacy of Fast Text's 8% accuracy gain (baseline: Decision Tree 85%).

This article analyses the service of a hotel by looking at the positive and negative reviews to draw conclusions about the business. The primary goals of this effort are aspect identification and sentiment categorization. The topics utilized in aspect identification are constructed using latent Dirichlet allocation (LDA). The naïve Bayes classifier, the support vector machine, the decision tree, and logistic regression are just some of the machine learning classifiers that may be used to categorise reviews. These algorithms are evaluated using metrics including accuracy, recall, precision, as well as the F score [13].

Using the proposed method, the reviews were efficiently sorted into the five distinct hotel amenities. The highest performance in aspect classification (85%) is achieved by LDA + TF-ICF 100% + Semantic Similarity, the best performance in sentiment analysis based on aspects (93%) is achieved by Word Embedding + LSTM, and the comfort aspect gets the most unfavorable sentiments (60%) of any aspect. Moreover, the results show that a feeling is influenced by a certain component [14].

CHAPTER 36

Utilizing Machine Learning for Natural Language Processing to Conduct Sentiment Analysis on Twitter Data in Multiple Languages

Rahul Mishra[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Retailers, market analysts, and other users of the web are greatly influenced by user views. Arranging the unstructured data gathered from various social media networks correctly is necessary for doing relevant analysis. Emotional evaluation as a method for cross-lingual data classification has received considerable attention. Textual organization is a subfield of natural language processing, or NLP, that may be used to classify an individual's emotional or mental condition as positive, negative, beneficial, or detrimental, like a thumbs up or thumbs down, *etc.* A combination of sentiment analysis as well as deep learning techniques might be the key to solving this kind of problem. Deep learning models, which are capable of machine learning, are particularly useful for this. One of the most widely used deep learning architectures for analyzing sentiment in text is called Long Short Term Memory. These frameworks have potential applications in NLP. In this study, we provide algorithms to solve the problem of multilingual sentiment analysis, and we evaluate their precision factors to determine which one is the most effective.

Keywords: LSTM, Multiple languages, Machine learning, Natural language processing, Sentiment analysis, Twitter data.

INTRODUCTION

One way to get to the underlying feelings in a piece of writing is *via* sentiment analysis. It uses machine learning and NLP (natural language processing) to identify, extract, and assess how consumers perceive a service or product. This is why “opinion mining” or “emotional AI” are common terms for this kind of analysis.

[*] **Corresponding author Rahul Mishra:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: rahul.mishra.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

Finding out if a piece of text is favorable, negative, or indifferent is the main objective of mining opinions. An example of this would be a statement like:

- *"We stayed at this hotel for five days"* is neutral,
- *"I liked staying here"* is positive, and
- *"I disliked the hotel" is* negative.

The opinions of WEB users carry a lot of weight with other users, product dealers, and market researchers. In order to conduct useful studies, it is required to properly organise the vast amounts of unstructured data collected from different social media sites [1]. When picking an annotated data set, there are two main considerations:

- **The length of texts.** A dataset that uses mood tagging on long reads (articles, blog posts, *etc.*) would not be applicable to shorter texts (tweets, feedback, *etc.*), and the inverse is also true.
- **The topic or domain.** Even with a sufficiently large library of tagged political messages, training a machine to assess hotel reviews would provide subpar results. The tourism area is best served by using labeled samples of evaluations, such as those of restaurants or airlines, to achieve a reasonable degree of reliability. But if you are looking for hotel reviews, your best bet is...a database that was created from guest evaluations of hotels.

There has been a lot of focus on emotion analysis as a way to categorise data across languages. In Automatic Language Processing (NLP), this is known as the textual organisation and may be used to categorise a user's emotional state or reaction as either positive (thumbs up), negative (thumbs down) or neutral (no reaction) [2, 3]. When it comes to text analysis, sentiment analysis (SA) remains a popular topic of study. Word processing before SA is used to measure its correctness [4]. This study presents a text analysis framework for SA [5] using natural language processing techniques applied to data from Twitter. Collecting data, cleaning the text, pre-processing, extracting text features, and classifying utilizing SA methods are the basic processes in the field of machine learning. The information is retrieved using a method that is unique to the domain, particularly for Twitter data [6]. At this point, the data has been cleaned for things like typos, punctuation, tags, and emoticons. Tokenization and stop word removal (SWR) are pre-processing steps [7]. In natural language processing, sentiment analysis is a crucial job [8]. The field that receives the greatest attention from academics nowadays is Twitter sentiment analysis (SA). Twitter sentiment analysis in languages other than English has received very little attention from academics. There seems to have been a recent spike in the volume of data sets disseminated *via* social media platforms in a wide range of languages [9]. The governments of

the world's nations might benefit much from mining social media data, and machine learning methods like NLP (Natural language processing) play a crucial part in this [10]. Social media may be mined for valuable 'mindset' information on the population, which is essential for every government in the globe.

It seems like every single second, someone new joins Twitter and starts tweeting about some social occasion. Gaining insight into the user's current sentiment on a range of societal issues, from the contentment of an airliner to the image brand, might be greatly facilitated if our suggested model could anticipate sentiment tags for live tweeting. Neural networks with sophisticated structures, such as LSTM, and a simple logistic regression model could be used. Considering the characteristics of tweets, we begin by preprocessing them, and creating a binary dependence tree to feed into LSTM. We used regularization techniques and fine-tuned our hyper-parameters. Documents may be easily categorized as well as predicted to convey positive or negative emotions with the use of machine learning techniques. There are two main categories of machine learning algorithms: supervised and unsupervised. An algorithm that is supervised makes use of a tagged database to train on documents that have been appropriately annotated. In contrast, independent learning makes use of a dataset that has not been tagged with the appropriate emotions. Artificial learning methods applied to tagged datasets are the primary focus of the suggested research. The majority of studies only look at tweets written in one language. In contrast, Twitter is a global social network that facilitates online microblogging. You may find linguistic tweets on a wide range of social topics, and by examining them all, you can deduce the blogger's mood.

RELATED WORK

Two related solutions to this sort of difficulty are techniques based on deep learning as well as sentiment analysis. In this case, deep learning models—which can learn new things—are invaluable. Sentiment analysis in words is a popular task for deep learning architecture like naive Bayes as well as recurrent neural networks (LSTMs). One possible use of such structures is in the processing of natural languages. We provide a suite of methods for multilingual sentiment evaluation and employ accuracy factor comparisons to distill the findings and choose the best method [11].

The suggested article delves into the mechanisms used for pre-processing text and how they impact the dataset. Classification algorithms' accuracy has been improved by using text pre-processing and dimensionality reduction. Market research, consumer behaviour, survey analysis, and brand tracking are just some of the many possible applications for the suggested corpus. Text pre-processing

CHAPTER 37

The Use of Machine Learning to Analyze the Sentiment for Social Media Networks

Darleen Grover[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The amount of textual information on the internet has increased significantly with the debut of social media platforms like Twitter, including news stories and historical records. This is due to the growth of Web 2.0. More individuals are using the internet and different forms of social media to share their thoughts and feelings with the world. As a result, more phrases with emotional nuance were created by the general public. It is natural that researchers will look into new approaches to understanding people's emotions and reactions. In addition to providing a novel hybrid system that combines text mining and neural networks for sentiment categorization, this study evaluates the efficacy of many machine learning and deep learning techniques. More than a million tweets from across five different topics were utilized to create this dataset. Seventy-five percent of the dataset was used for training, while the remaining twenty-five percent was used for testing. When compared to traditional supervised learning methods, the system's hybrid approach displays a maximum accuracy of 83.7%.

Keywords: Machine learning, Social media, Sentiment analysis.

INTRODUCTION

The quantity of written content, such as news articles and historical documents, on the internet, has dramatically increased with the advent of social media platforms like Twitter. The number of people who express their feelings and opinions online continues to rise [1]. The general populace responded by coining a greater number of expressions loaded with sentiment. It is inevitable that scientists will try out new methods to decipher human cognition and behavior [2]. When it comes to sharing information, collaborating on projects, and having meaningful conversations, social media platforms are where it lies [3]. Sentiment analysis is a subfield of sociology that examines how people feel and what they

* **Corresponding author Darleen Grover:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: darleen.grover.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

think about certain brands, products, and services [4]. As the usage of Arabic on social media continues to grow, scholars have been increasingly interested in the topic of Arabic sentiment analysis [5]. Arabic is one of the most widely used languages for internet interaction [6]. Sentiment analysis is challenging in this language because of its complex morphology, broad range of dialects, and lack of available materials [7]. Because of the sheer volume of text available on social media platforms, researchers can now more accurately quantify the emotional value of this data in order to gain a deeper understanding of the mindsets, views, and emotions expressed [8]. In order to provide light on the several approaches that may be taken to social media sentiment analysis [9], this study uses Machine Learning techniques to analyze data from Twitter. It is argued that the Fuzzy Set and Rough Set techniques for classification are especially effective mathematical components of artificial intelligence that provide a new angle to the well-established topic of sentiment analysis. A mountain of work remains in text mining for natural language processing, and there is a shortage of review papers discussing the use of rough-fuzzy classifiers in sentiment analysis [10].

When it comes to readily accessible information, Twitter ranks high. Instantaneous global communication is made possible, which is especially useful in regions with limited opportunities for public expression. There were 135 million people who used the internet in 2014; more than 71 million of them actively used social networks [8]; it was estimated that these people sent out more than 17 million tweets per day. It is natural that researchers will look into alternative approaches to better understand how people think and react.

Therefore, the goal of this study is to categorize which tweets include good feelings and which contain negative sentiments by comparing different approaches applied to a million tweet records. Following is the outline of the paper: In Part II, we examine the relevant literature. In Section III, we present the suggested system. Section IV concludes the paper and discusses what comes next.

RELATED WORK

An innovative hybrid system that utilizes text mining and neural networks to categorize sentiment is described, and many well-known deep learning and machine learning approaches are compared and contrasted. The tweets in this dataset are in the millions and cover five distinct themes. Seventy-five percent of the data set was utilized for training, while the remaining twenty-five percent was used for testing. By reaching a maximum accuracy of 83.7%, the system proved that its hybrid learning technique was better than more conventional supervised approaches [11].

The major objective of this study is to examine the behavior of participants in online debates under varying conditions. Negative influences, such as those from peers, the media, and the surrounding culture, have the power to sway individuals. These days, most people use social media as their primary means of rapid communication. Readers might be inspired or depressed by the words they read. When faced with a problem, it may be best to take proactive measures rather than merely react. People's reactions have been unpleasant, and they've been expressing their discontent in a broad range of ways, both online and off. This will lead the person to think and feel irrationally. Sentiment Analysis can observe and analyze people's actions by utilizing data mining and deep learning methods. In this piece, we explore the potential of computer algorithms for training robots to recognize human behavior. In addition, it will offer suggestions on how those who use social media might modify their perspectives in order to better weather the current crisis [12].

This research explores the potential of contemporary technology to better equip humankind to save lost souls through the use of natural language processing techniques, which may involve the installation of Machine Learning models. Because of this, there has been a rise in the number of young people using the Internet and the many social networking sites it hosts. People sometimes find it easier to express their feelings *via* writing than by talking to specialists about their problems. Social media analysis takes into account the poster's location, the subject matter, the tone of the post, and other characteristics to make inferences about the poster's mental state, predict the poster's future actions, and prevent a possible crime. Kazakhstan has the third-highest suicide rate in the world. As a strategy for decreasing suicide rates, researchers settled on sentiment analysis, a technique from the field of natural language processing used to ascertain whether a given utterance should be seen in a positive or negative light. Every database project necessitates data collection from widely used Kazakh social networks. The fundamental aim of this work is to integrate NLP and ML techniques to collect the necessary data for model development [13].

Using machine learning and other sophisticated learning models, our study seeks to automatically find the emotions and classify them as positive or negative. This study implements and assesses four different machine learning techniques: the long short-term memory (LSTM) models, the support vector machine (SVM), the logistic regression (LR), and the K-nearest neighbors (KNN) approach. The Arabic-Review (ARev) database is used to evaluate these classifiers; it is a large corpus of hand-annotated text from a wide range of Arabic sources. The results [14] indicate that SVM and LR algorithms are the most accurate classifiers, with a 92% and 93% success rate, respectively.

CHAPTER 38

Sentiment Classification of Textual Content using Hybrid DNN and SVM Models

Abhishek Singla[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The proliferation of Web 2.0 has resulted in a deluge of real-time, unstructured data such as user comments, opinions, and likes. The lack of structure in the data makes it difficult to create a reliable prediction model for sentiment analysis. There have been promising applications of several DNN architectures to sentiment analysis, however, these methods tend to treat all features identically and struggle with high-dimensional feature spaces. In addition, existing techniques fail to effectively combine semantic and sentiment knowledge for the purpose of extracting meaningful relevant contextual sentiment characteristics. This paper proposes an integrated convolutional neural network, or CNN, architecture that takes sentiment as well as context into consideration as a means of intelligently developing and highlighting significant components of relevant sentiment contextual in the text. To start, we use transformers' bidirectional encoder representations to create sentiment-enhanced embeddings of words for text semantic extraction using integrated emotion lexicons with broad coverage for feature identification. The proposed approach then adjusts the CNN in a way that it can detect both word order/contextual text semantics data as well as the long-dependency relationship in the phrase sequence. Our approach also employs a system to prioritize the most important portions of the phrase sequence. One last step in sentiment analysis is the use of support vector machines (SVMs) to reduce the complexity of the space of features and identify locally significant characteristics. The accuracy of existing text sentiment categorization is greatly improved by the use of the proposed model, as shown by an evaluation of real-world benchmark datasets.

Keywords: Hybrid DNN, SVM, Textual content sentiment classification, Text documents.

INTRODUCTION

A paradigm shift was initiated by deep neural networks (DNNs), but their high processing requirements have limited their deployment on edge devices like

* **Corresponding author Abhishek Singla:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: abhishek.singla.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

phones and wearables [1]. In light of this, a hybrid edge-cloud computing architecture is proposed [2] to naively divide the DNN processes on the assumption of an ongoing connection state, hence enabling the relocation of some computation to the cloud. Nevertheless, there is a lack of comprehensive approaches to DNN segmentation at the moment [3], and the state of practical networks might change drastically based on their surroundings. One of the most important technologies that will allow 5G wireless networks to function is cloud radio access networks (CRANs) [4]. In order to reduce power consumption while also satisfying the long-term demands of mobile consumers, CRAN allocation of resources has to be further improved [5]. In light of DRL's track record of success in solving challenging control problems, we provide ReCARL, a novel paradigm for DRL-based environmentally friendly utilization of resources in CRANs [6]. At each decision epoch, the DRL agent's action-value functions is approximately determined using a deep neural network, or DNN, once the state distance, action space, as well as reward function have been established [7]. Promoting offline get-togethers as well as attracting new faces has never been easier than on event-based social media platforms (EBSN), which have seen explosive growth in the last few years [8, 9]. An ever-growing event-based social network (EBSN) relies on being able to consistently suggest members' preferred events. There is more competition for limited resources as the intelligence of edge devices like smartphones, wearables, and IoT devices has increased owing to machine learning. Training models on edge devices is a resource-intensive process, making on-device customisation a formidable obstacle [10, 16].

To determine the reviewer's emotional stance, we use Stanford POS Tagger to label the review text's nouns, adjectives, adverbs, and verbs with appropriate POS labels. Sentiment orientation is determined by looking up the aforementioned keywords in large-scale sentiment lexicons and obtaining their associated sentiment orientations. We use linguistic semantic criteria to label sentences/review texts with positive and negative sentiments appropriately. The categorization of ambiguous opinions can benefit from linguistic semantic principles. They contribute to a more precise modeling of the SA issue. Six popular sentiment classification benchmark datasets were used to assess the proposed method's predicted ability. In comparison to existing baseline methods, the suggested method significantly improves the efficiency and effectiveness of sentiment analysis. The rest of the paper is structured as follows. Section 2 displays further work of relevance. In Section 3, we go deep into the approach, architecture, and technological specifics. In Section 4, we discuss the methodology and findings from our experiments. The study finishes with a section on what comes next.

RELATED WORK

We explore how the modular design of DNN may be utilized to modify the edge model for usage in different settings and with different deployment strategies. To achieve this aim of model consistency and computational delay simultaneously, we built a decision engine based on reinforcement learning. The DNN uses a context-aware model tree constructed by the engine to make decisions about which model branch to switch to in real-time. Our approach has been shown to minimize latency by 30% - 50% in emulation and field testing without compromising model accuracy [11].

Short-word detection issues are addressed by creating an HMM/DNN architecture. By including a context-aware keyword theory as well as a 9-state filler framework into the first-stage phrase hypothesizer, we could increase the figure-of-merit (FOM) for short words from 6.08% to 21.88% as well as decrease the percentage of missing issues from 80% to 6%. To further narrow down potential matches, a second-stage keyword verification based on MLP is used following the hypothesizer. Improved short word recognition was achieved by re-engineering the verifier with three cutting-edge techniques: based on knowledge features, deep artificial neural networks, and the use of a hidden Markov model (HMM) for the MLPs' features conversion. Using nine short keywords from the TIMIT collection, we achieved the best FOM of 42.79% with the suggested KWS technique; this compares well with the FOM of 42.6% for lengthy content words and is significantly higher than the FOM of 18.4% for short keywords reported in earlier works [12].

We provide two distinct agents for DRL inside ReCARL: ReCARL-Basic, which uses a conventional deep Q-learning method to train a conventional DNN structure, and ReCARL-Hybrid, which uses a hybrid deep qualitative learning method to train a context-aware DNN architecture. In a series of extensive simulated tests, we evaluated ReCARL and the two DRL agents to two widely used benchmarks. The results of the simulation show that ReCARL is efficient in meeting the demands of its customers while conserving energy [13].

Using semantic content analysis and contextual event impact, our research proposes a hybrid collaborative filtering approach to event suggestion that can aid users in making informed decisions about events in their immediate locations. Specifically, we analyze text from event descriptions using the latent topic model, and then we utilize each user's event registration history to develop a long-term interest model and a short-term interest model. Then, we give each occurrence a value that takes into account both its uniqueness and its social relevance to users. We prioritize a user's long-term interests over the effects of the immediate

CHAPTER 39

Big Data Analysis and Information Quality: Challenges, Solutions, and Open Problems

Sahil Suri[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Big Data is here, thanks to the proliferation of social media and IoT devices. As a result of the immense benefit that Big Data has brought to the public as well as businesses, the question of how to handle and utilise it more effectively has captivated people from every walk of life. Big Data processing has been plagued by difficulties due to the 4V features of Big Data. Solving the data quality issue is crucial to Big Data processing, as is ensuring data quality, which is a precondition for Big Data to play its worth. Two examples of where Big Data has been put to good use are in recommendation and prediction systems. In this research, we investigate Big Data recommendation and prediction systems and endeavour to ascertain the quality of data at each stage of the process: collection, preprocessing, storage, and analysis. The proposed remedy follows from a thorough description and examination of the issue at hand. We have left a few questions unanswered at the conclusion of the report.

Keywords: Big data processing, Big Data, Data quality, Issues, Open problems, Solutions.

INTRODUCTION

The big data age is here, thanks to the proliferation of social media platforms, IoT devices, cloud computing, and other technological advancements [1]. The public and businesses alike have benefited greatly from the explosion in data availability. Meanwhile, the topic of how to effectively handle and utilise big data has spread to all corners of society. Many problems in big data processing may be traced back to the 4V features of large data. To guarantee the proper implementation of the big data approach, it is crucial to address the problem of poor data quality. WBAN, or wireless body area network, is a great tool for long-term healthcare monitoring [2]. Wherein data collection, processing, and transmission are handled by sensor nodes that are wirelessly linked within, on, or around a person. Since

[*] **Corresponding author Sahil Suri:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sahil.suri.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

most sensor nodes run on relatively tiny batteries with finite power, it is crucial to WBAN that we find ways to reduce their energy consumption and increase their longevity [3-6]. Quality of Service (QoS) relies heavily on the work done at the Channel Access layer. We can now gather, store, and analyse mobility data at an unparalleled scale and velocity. The proliferation of GPS-enabled gadgets, such as smartphones and other internet-connected items, is largely responsible for this. Mobility data is becoming more important as pervasive computing becomes the norm in our culture [7-9]. The footprints people leave behind may be useful for a variety of purposes, including contact tracking for communicable diseases (such as COVID), traffic engineering, and risk management, particularly when paired with societal data. Over the course of the previous two decades, researchers have shifted their focus from spatial data to spatiotemporal data and finally to information about mobility. What should we do now? Maybe not, but the combined Big Data analytics that take into account mobile devices certainly are. The foundational technology for Big Data has now come of age. The research community's interest in all matters concerning the Big Data and Value Analytics (BDVA) standard model, including handling data, data processing, analytics, visualisation, and user engagement, has recently reached a height. New ways of thinking about and handling mobility data are opening up as a result of advances in artificial intelligence. Therefore, the time has come for Big Mobility statistical analysis to catch up. The field of large mobility data analytics is ripe with untapped potential, but there is a pressing need to encourage the sharing of novel perspectives on pressing real-world issues, the evaluation of proposed solutions, and the discovery of new avenues for study. By 2020, it was expected that almost 26 billion gadgets, with a large majority being cars, will be linked to the Internet. The web of Vehicles (IoVa) is a concept that describes the interconnection as well as cooperation of smart vehicles as well as devices in a network through the creation, transmission, as well as processing of data with the goals of enhancing traffic flow, shortening travel times, and enhancing passengers' levels of comfort while also decreasing pollution and the likelihood of accidents [10].

In this article, we will examine the problems associated with big data and the methods currently being used to address them. In addition, big data's open questions are listed. The first section provides an overview of large data and the difficulties it presents. The difficulties encountered in large data analysis and processing are discussed in Section 2. The current approaches to our problems are discussed in Section 3. In Section 4, we investigate the outstanding questions that might aid us in processing massive data and extracting actionable insights. Tools and technology for large data are discussed in Section 5. The potential of big data in the future is discussed in Section 6. The last section includes a brief overview and final thoughts [11-14].

LITERATURE REVIEW

In this study, we examine common big data applications like recommendation and prediction systems to attempt to identify data quality concerns at the data collecting, data preprocessing, data storage, and data analysis levels. Solutions are provided that make sense in light of the concerns that have been elaborated upon and analysed. Finally, several future research questions are posed. The purpose of this work is to bring light to the current difficulties encountered by WBAN MAC protocols in the areas of power consumption and latency. Then, they discussed some of the unanswered scientific questions, along with potential answers and developments. Finally, the study wraps up by discussing the knotty problems in WBAN. Major increases in manufacturing, shipping, and farming have increased not just air pollution and climate change but also the likelihood of catastrophic weather events. One of the main causes of these environmental issues is the release of dangerous gases into the atmosphere, in particular the VCD of CO, SO_2, and NOx. By employing remote sensing (RS) methods to keep tabs on air quality, our study hopes to provide decision-makers with useful information. The volume, complexity, diversity, and velocity of RS data make them challenging to handle. Therefore, our publication provides details of how the various satellite data sources were used. Furthermore, the evidence presented in this paper indicates that RS data may be classified as large data. As a result, we have implemented a Hadoop big data framework and detailed the steps necessary to analyse RS environmental data effectively. The editors of the ACM Journal of Information and data Quality are pleased to present a unique volume on quality evaluation and leadership in big data. After a thorough peer review procedure, 11 of the 27 original articles we received might be considered for publication in this issue's two sections. To accompany Part I, whereby we featured articles on machine studying and quality control in big data settings, we provide this editorial. Today, in the “big data” age, businesses must deal with massive volumes of data that change quickly and come from a wide variety of sources, including social media, unstructured information from a wide variety of websites, and raw inputs from sensors. In order to increase productivity, businesses are turning to big data solutions to streamline their procedures and quicken the decision-making process. The amount of data quality issues faced by big data professionals is staggering. These may be difficult to fix, and may even cause erroneous data analysis. It has become difficult to control quality with huge data, and previous studies have only touched on a few areas of the problem. Traditional methods of data quality management are inadequate for the more difficult task of managing the quality of large data. To protect cars and their drivers from attackers who could exploit this information, the transfer of sensitive data (such as position) must take place with specific security features. In dispersed, untruthful contexts like IoV networks, blockchain is a relatively new technology that assures trust between nodes using

CHAPTER 40

Using Deep Learning Techniques to Detect Traffic Information in Social Media Texts

Sourav Rampal[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Given the real-time and pervasive nature of social media, data mining for traffic-related insights is a newly emerging area of study. In this work, we discuss the challenge of mining social media for traffic-generating microblogs on Sina Weibo, the Chinese equivalent of Twitter. It is recast as a classification issue in short texts for machine learning. In the first step, we use a dataset of three billion microblogs and the continuous bag-of-word model to learn word embedding representations. Word embedding, as opposed to the standard one-hot vector representation of words, has been shown to be useful in natural language processing tasks because of its ability to capture semantic similarity between words. Then, we suggest feeding the learned word embeddings into convolutional neural networks (CNNs), long short-term memory (LSTM) models, and their combination LSTM-CNN to extract traffic-related microblogs. We evaluate the suggested techniques against state-of-the-art methods such as the support vector machine (SVM) model using a bag of n-gram features, the SVM model using word vector features, and the multi-layer perceptron model using word vector features. The proposed deep learning methods are shown to be useful in experiments.

Keywords: Deep learning techniques, Social media, Traffic information detection, Text mining.

INTRODUCTION

User-generated information, especially that available on social networking platforms, is more important in today's culture. With the rise in popularity of social media platforms comes an increased need for transportation systems that can extract useful information about traffic conditions in real-time from users' posts. It has been demonstrated that deep learning algorithms excel in processing large volumes of text input and extracting relevant information [1, 2]. Traffic

* **Corresponding author Sourav Rampal:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: sourav.rampal.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

mishaps are frustrating for drivers and must be found quickly so that traffic may resume its usual flow. Social media posts on traffic conditions are a gold mine of information [3]. It is possible that traffic statistics may be extracted from social media posts by using deep learning techniques and sophisticated neural network models that can learn from large amounts of training data [4]. These formulas might help us figure out if a piece of social media material includes any traffic-related information at all. A flood of geo-referenced data on localized occurrences may emerge *via* social media. Social media messages may also be mined for information, such as the location and severity of road accidents, thanks to their usage of these tools [5]. A deep learning-based information extraction approach (termed BERT-BiLSTM-CRF) is described and proven to surpass existing state-of-the-art methods [6] in their ability to create time-related data from social media discussions. Real-time traffic information made available by the efficient use of deep learning algorithms for traffic information identification from social media texts [7, 8] can help transportation systems better manage traffic, reduce congestion, and improve public safety. Therefore, in this era of Big Data, we can satisfy people's mobility needs by delving into and making the most of social media traffic data. Natural Language Processing (NLP), cloud and open platforms (*e.g.*, mobile Internet), and human-computer interfaces (HCI) are only some of the tools outlined in this study to mine social media for traffic data. In this study, we suggest a solution in the form of a model architectural design based on deep learning. A bidirectional LSTM method is used to handle the word level. While character recognition relies on the CNN approach. When combined with word embedding [9, 10], these two deep learning methods provide an F-measure of 0.789.

In this research, we present our work on employing convolutional neural networks (CNN), a long short-term memory (LSTM) model, and an LSTM-CNN with pre-trained word vectors to extract traffic-related tweets from Sina Weibo. Using three billion microblogs gathered from Sina Weibo between 2009 and 2011, the CBOW model is used to learn Chinese word vectors. Microblogs are classified as either traffic-relevant or traffic-irrelevant using a combination of convolutional neural networks (CNNs) and long short-term memory (LSTMs). The suggested approaches may make use of the semantics in microblogs and abstract deep features, making them superior to conventional one-hot word representations and usual feature-based text classification techniques. Experiments demonstrate the effectiveness of the offered strategies in identifying traffic-related microblogs. The following are the most significant results of this study: (i) We implement deep learning strategies for mining social media for traffic data. (ii) We apply the CBOW model to a dataset of 3 billion microblogs to get the word embedding, which captures word semantics, in contrast to previous approaches that extract traffic-relevant microblogs with bag-of-words characteristics. When compared to

other models, we show that deep learning models are superior at recognizing data linked to traffic.

This paper's remaining sections are structured as follows.

In Section II

We examine current developments in research that mine traffic data from social media. The bulk of our methods, such as data collection and preprocessing, word vector models, and classification models, are detailed.

In Section III

In Section IV

We provide experimental verification of the suggested approaches.

Section V

Contains our last thoughts and recommendations for moving forward.

RELATED WORK

The instantaneous and ubiquitous nature of social media has made traffic data mining a hot issue. This paper focuses on a specific subset of social media mining: the difficulty of extracting traffic-relevant microblogs from Sina Weibo (a Chinese microblogging site). Machine learning reframes it as a classification problem for brief texts. We begin by creating word embedding representations using the continuous bag-of-word model and a three billion microblog dataset. In contrast to the typical one-hot vector representation of words, word embedding has been found to be beneficial in natural language processing tasks and can capture semantic similarity between words. To extract traffic-related microblogs, we propose feeding the learned word embeddings into convolutional neural networks (CNNs), long short-term memory (LSTM) models, and their combination LSTM-CNN. We compare the proposed techniques with several others that have already been developed, such as the bag-of-n-grams feature SVM model, the word-vector feature SVM model, and the multi-layer perceptron feature SVM model. It has been demonstrated experimentally that the suggested deep-learning approaches are successful [11].

This study's goal is to evaluate urban development projects in Kuwait using social risk analysis theory. A case study of 17 incidents of social conflict on construction sites in Kuwait evaluated 12 aspects of social risk. Eight stakeholders associated with important societal risk factors were identified through a review of the

CHAPTER 41

Deep Sentiment Classification in COVID-19 Using LSTM Recurrent Neural Network

Jatin Khurana[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: Users (people/patients) concerned about health concerns have an easy outlet in online medical forums along with other public social media on the Internet. The World Health Organization declared a global public health emergency in response to the emergence of a new coronavirus (infection which causes the disease termed COVID-19) in late December 2019. In this research, we employed a natural language processing (NLP) technique based on topic modeling to automatically extract COVID-19-related talks from social media and discover numerous concerns linked to COVID-19 from public viewpoints. As an added bonus, we look into the possibility of employing a long short-term memory (LSTM) recurrent neural network to accomplish the same task with COVID-19 remarks. Our research highlights the value of incorporating public opinion and appropriate computational approaches into the process of learning about and making decisions on COVID-19. The trials also showed that the study model was able to reach an accuracy of 81.15 percent, which is greater than the accuracy attained by many other popular machine-learning methods for COVID-19 sentiment classification.

Keywords: COVID-19, LSTM, NLP, Topic discovery model.

INTRODUCTION

Misinformation spread *via* social media in the wake of the COVID-19 pandemic had a negative effect on people, inducing fear and skepticism of the unknown [1]. Insights in to vaccination recipients' thoughts and sentiments can be gained by sentiment analysis [2], especially when the RNN-LSTM deep learning approach is used. Sentiment analysis has made great strides forward thanks to the advent of mature deep learning neural networks in the field of natural language processing (NLP) [3, 4]. A novel approach has been published [5] that employs natural language processing and deep learning to automatically classify the tone of

[*] **Corresponding author Jatin Khurana:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: jatin.khurana.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

tweets. This research focuses on the many text representation strategies needed to solve text categorization challenges [6]. The purpose of this study is to illuminate the range of emotions experienced by people throughout the world in response to the COVID-19 pandemic [7]. Since so much information and opinions are generated, disseminated, and carried out every day *via* the Internet and other media, this is an active field of study. Latent Dirichlet allocation and two-layer bidirectional long short-term memory (LSTM) are used to extract aspects of a text and classify sentiments, respectively. For tweaking model hyperparameters, the hill-climbing approach is suggested [8]. Using the Twitter API, we can trawl for data, which we then gather and compare using standard Machine Learning methods. In terms of accuracy, precision, and recall, the Support Vector Machine is superior to other techniques [9]. The recurrent neural network achieved an accuracy of 92.70 and 91.24 percent on the COVID-19 dataset. Bi-LSTM and GRU both performed at 92.48% and 93.03% accuracy for the categorization of vaccination tweets, respectively. When faced with a future pandemic, healthcare providers and governments can benefit from the existing models [10].

- RQ1) How can key principles in natural language processing techniques, such as topic modeling, be used in online discussions to expose different difficulties linked to COVID-19?
- RQ2) What is the best way to find out which way people on COVID-19 feel about things?
- RQ3) How effective are several machine learning algorithms for determining the emotional tone of online discussions related to COVID-19, and which method is most effective?

In order to get answers to these concerns, we looked at the public opinions of Redditors on Reddit and analyzed their comments about COVID-19 to discover sentiment and semantic concepts linked to COVID-19. In particular, we employed a topic-modeling approach to natural language processing (NLP) to automatically extract COVID-19-related talks from social media and discover numerous concerns linked to COVID-19 from public opinion. The following are the most significant results of this study: We introduce a structured, NLP-based methodology that can glean important subjects from Reddit threads on COVID-19.

For the purpose of sentiment categorization of COVID-19-related comments, we present a deep learning model based on Long Short-Term Memory (LSTM), which outperforms many other well-known machine learning approaches. As part of our core study into COVID-19, we monitor Reddit for relevant discussions and identify emerging themes.

We analyze the emotion and opinion of COVID-19 comments from 10 different subreddits and determine their polarity. Our research highlights the value of incorporating public opinion and appropriate computational approaches into the process of learning about and making decisions on COVID-19. This paper follows the general outline below.

We begin with a primer on the basics of healthcare discussion boards on the web. In part 2, we discuss COVID-19-related concerns and compare and contrast other relevant publications.

In Section 3, We detail the NLP and deep-learning techniques used on the COVID-19 comments database, as well as the data pre-processing procedures we used. The analysis and commentary follow. Finally, we wrap off by discussing the future of this study and other NLP-based efforts to analyze the online community surrounding COVID-19.

RELATED WORK

This article explores the use of online data for the study of public opinion's emotional leanings *via* the lens of textual sentiment categorization research. Because of its superior semantic representation and its ability to deal with words that might have various meanings, the BERT (Bidirectional Encoder Representation from Transformers) model is used as a tool for text vectorization in this study. In order to simplify and speed up the processing of multi-label data, the BR (Binary Relevance) method is used. This technique takes a multi-label classification problem and breaks it down into many binary classification problems. For further text feature extraction, researchers have developed a BiLSTM-Attention model that combines bidirectional long- and short-term memory networks with the attention mechanism. The BiLSTM-Attention model's efficacy is demonstrated through an analysis of its experimental results. The work addresses the problem of an uneven data set by adjusting the loss function and other parameters to improve classification accuracy. In conclusion, this research offers a strategy for undertaking sentiment analysis with little wasted effort by leveraging state-of-the-art methods like BERT and BiLSTM-attention models [11].

Worldwide lockdowns were implemented due to the global spread of the COVID-19 epidemic. Several pharmaceutical companies, notably Pfizer, Moderna, and AstraZeneca, have produced vaccines against COVID-19. However average people have begun discussing the efficacy and safety of these vaccinations on social media sites like Twitter. The tweets were acquired by researchers using a Twitter API access token and then analyzed using natural language processing methods. The tweets were then classified into favorable, negative, and neutral

CHAPTER 42

Machine Learning-Based Data Preprocessing as well as Visualization Techniques for Predicting Students' Tasks

Pratik Mahajan[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: A student's chances of landing a job after graduation are influenced by their performance in school and their history of academic accomplishment. Students who want to succeed in the working world need to develop skills including technical proficiency, critical thinking, and effective communication. Here we are making an attempt to figure out how students' academic performance affects their future opportunities. Algorithms for data mining play a key role in analysing and forecasting the chance for students' placement according to their previous academic achievement. We polled students at a prominent technical institution for this piece. Multiple factors that influence students' probabilities are included in the dataset, and these factors are examined and shown graphically. We made an effort to assess the data and provide visualisations and insights before running or applying algorithms for machine learning to this dataset. Data analysis, understanding, and preparation are the key focuses of this research.

Keywords: Data preprocessing, Python, Student placement detection, Visualizations.

INTRODUCTION

A student's chances of landing a job after graduation are influenced by their performance in school and their history of academic accomplishment. Students who want to succeed in the working world need to develop skills including technical proficiency, critical thinking, and effective communication. However we are trying to figure out how students' academic performance affects their future opportunities. Machine learning algorithms play a key role in analysing and predicting the possibility of learners in an assignment based on their previous

[*] **Corresponding author Pratik Mahajan:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India; E-mail: pratik.mahajan.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

academic achievement [1, 2]. Colleges and universities cannot function without the placement procedure. Placements have a significant impact on a school's reputation and acceptance rates. As a result, educational institutions are making concerted efforts to expand and enhance their placement capabilities. Model Development with ML for Predicting Pupil Placement examines the efficacy of algorithms for machine learning across a range of dimensions in order to predict a student's placement. The creation of educational curricula and the placement of pupils are both aided by reliable forecasting. This way, the student's development can be monitored, and they may even be ahead of schedule in responding to the business's demands. Various algorithms have been investigated and developed, including Logistic Regression (LR), Support Vector Machine (SVM), and Naive Bayes. Data mining is one of the most useful tools for problems involving prediction and classification [3, 4]. Machine learning algorithms are becoming the backbone of modern economies. This is due to the fact that they provide the groundwork for the age of data analytics. Due to the absence of an appropriate and efficient solution, managing datasets that use machine learning algorithms to perform prediction and visualization is challenging. Teachers may benefit from knowing their students' personalities in order to better anticipate their students' academic development. The unique characteristics, thought patterns, and behavioural quirks of an individual may be better understood if we have a firm grasp of their personality. However, the time, effort, and resources necessary to conduct a traditional personality test must be considered [5, 6]. There are several ways in which an individual's character may be seen in his language usage, including word choice and sentence structure. Utilizing the service's linguistic tools, users of the microblogging website 'Twitter' may create and read short messages. This study overcomes problems associated with utilising conventional personality tests by developing a prediction model using Twitter data and classifying users according to their DISC traits utilizing the naive Bayes classifier strategy [7, 8]. We used 9,044 retweets from 70 distinct accounts as data for training to develop a prediction model. In order to create and assess a model from tweets, extensive preprocessing on several levels is required. The simulation's classification results were compared with those from the expert, and a precision rate of 76.19% was calculated for the data prediction system [9, 10].

Due to their pervasiveness, it is important to have a firm grasp of the principles of any machine learning method or model. The purpose of this article is to demonstrate how a variety of models based on machine learning may be used to the problem of student prediction. This section provides details on the methodology used in the experiment, the results, and an analysis of the several models used in the study. The rest of the paper is organized as follows: In Section 2, a short review of relevant earlier work is presented; in Section 3, the details of the intended work are outlined; in Section 4, the results and discussion on the

models utilized for machine learning are presented; and in Section 5, the results and summary are presented.

LITERATURE REVIEW

We surveyed students at a prominent technical institution. Multiple factors that influence students' probabilities are included in the dataset, and these factors are examined and shown graphically. We made an effort to examine this dataset and provide visualisations and insights before executing or implementing machine learning strategies on it. Data analysis, understanding, and planning are the main concerns of this investigation [11]. The goal of this study is to use historical student data to better anticipate where current students should be placed. This model incorporates a technique for making predictions. The more equipped a university is to find jobs for its students, the more successful it will be. The school and its pupils will benefit in the long run from this. Prediction functionality is included in this model. The entity that would ultimately be in charge of the placement prediction also gathered and preprocessed the study's data. The accuracy of the proposed models was contrasted with the results of certain established categorization strategies. The results showed that the proposed approach outperformed its nearest competitors' algorithms [12]. The proposed research takes the dataset into consideration, uses preprocessing approaches that render the data more accessible for training models of prediction using Decision Tree (DT) and XG Boost, and then evaluates the findings alongside those obtained from previously used methods. There are many different ways to evaluate performance, but some common ones include accuracy, F1-score, precision, and recall. Success rates for other approaches ranged from 88% with the LR & DT algorithms to 86% for the Naive Bayes and 84% for the XG Boost technique, with the SVM approach coming in at 91%. We can choose the most effective algorithm currently in use. Most teachers prefer a more precise approach of forecasting [13]. In this study, we provide a novel approach to interacting with machine learning techniques for predictive data and analytics visualisation. Creating a tool that can simultaneously clean data and present the findings visually is crucial for preparing data for analysis. The developed programme will take as input an organized dataset comprising both textual and numerical data, and then analyse it using machine learning techniques to provide a preprocessed dataset. Depending on the method used for maximum efficiency, several stages of data visualisation and prediction may be involved here [14].

CHAPTER 43

The Prediction of Faults Using Large Amounts of Industrial Data

Jagtej Singh[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: An essential feature of an intelligent workplace is error-free manufacturing. The data-driven approaches rely on fundamental status data with minimal volume, while the conventional model-based methods rely heavily on accurate equipment models. Not only are these aspects problematic, but they also prevent them from meeting the real-time need of evaluating industrial big data in an IoT setting. In this research, we provide a fault prediction approach that uses industrial big data to unearth the connection between data (such as status and sound data) and equipment failures using machine learning techniques. Not only that but the breakdown could be investigated promptly since the equipment's status could be tracked in real-time. Our solution outperforms the current ones in terms of accuracy and real-time capabilities, according to the simulation results.

Keywords: Big data, Fault detection, NN.

INTRODUCTION

Intelligent workshops are distinguished by their faultless production. Conventional model-based fault detection systems rely heavily on comprehensive machine examples, whereas data-driven approaches can only utilize basic status data using a relatively restricted quantity [1]. These characteristics render them unfeasible, in addition to being incapable of meeting the real-time demand of processing commercial big data in the Industrial Internet of Things framework. A Multiscale fuzzy entropy and backward propagation neural networks (or MFE-BPNN) based technique was given to deal with the problems of poor accuracy and lengthy prediction time in failure detection for industrial equipment. Since a great deal of data is now unstructured or disorganised, IoT-based industrial devices are creating and collecting data in real-time. Using IoT-based sensors is the most effective method for keeping tabs on these production processes in the industrial

[*] **Corresponding author Jagtej Singh:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: jagtej.singh.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

sector. However, a major challenge for IoT applications is how to store and make sense of data collected in real-time. Since the quality of the processed goods and production efficiency will be impacted by the degradation of the tools over time, knowing how to utilise data gathered through condition monitoring to properly anticipate the life that remains of the equipment in issue is vital. Predicting the degree of wear on a CNC machine tool's cutting cutter is one use of machine learning in the setting of producing big data as well as industrial equipment fault diagnosis and health management [2, 3]. Features were retrieved from the initial information obtained from the cutting cutters throughout the research and evaluation phase. To eliminate feature sets associated with the level of tool wear loss, a feature set screening approach was used. Ultimately, a machine learning algorithm successfully forecasted how long a milling cutter would last. The huge scope of industrial production and the advancement of technologies like sensing allow for the vast collection of industrial tool data [4, 5]. While it is possible to gather enough information on an induction motor's wellness, finding about its faults is more difficult. As a result, it was decided that the main goal of this article would be to recognize engine operating conditions and forecast potential problems using information about motor health [6, 7]. Adaptive thresholds and a state recognizer are components of the induction condition architecture that were proposed here for use in a motor. To train the condition model for induction engines, we used health data, and to get over the problem of unlabeled motor data, we used an upgraded SOM-FCM Two-Layer clustering method. At last, ordinary motor load variation state identification along with rotor broken motor fault prediction was used to validate the efficacy of the model and approach, with accuracy rates of 97.5 percent and 90.2 percent, respectively [7, 8]. Intelligent industrial facility management relies heavily on mechanization and precise fault identification and evaluation of heating, ventilation, and air conditioning (or the HVAC system) to save money, time, and resources. As a result of their robust portability and superior feature extraction capability, convolutional neural networks are gaining widespread use in a variety of domains to solve classification and prediction challenges. In this research, we create a Federated Learning-based Convolutional Neural Network (Fed_CNN) to handle HVAC defect identification and diagnosis. The system uses multi-party data for multi-scale joint simulation [9, 10]. The impact of fault detection and diagnosis of our presented Fed_CNN model for chillers is greater than 0.9, making it capable of simultaneous multilayer fault prediction and diagnosis. Chilled air handling unit (AHU) and chiller faults can be detected and diagnosed by the Fed_CNN model. The suggested proprietary learning federation architecture outperforms several existing fault identification and diagnosis algorithms for HVAC systems, as shown by a series of experimental comparisons.

A novel CNN-based intelligent defect diagnostic approach is suggested in this study. Before feeding data into a learning system, most methods need a specific, sophisticated preparation step. Changing signals from to realm to the frequencies realm is a novel technique, while being widely used in conventional deep learning. The revolutionary step introduced by the new method is a simple method for preparing information in the chronological domain. To classify images, this method first maps signals in the temporal domain to the visual domain, and then it suggests using a sophisticated convolutional neural network (CNN). The result proves that the method used is effective. Here's how the remainder of the paper is laid out: In Section 2, we present the novel CNN-based intelligent fault diagnostics. The outcomes of an experiment are discussed in detail in Section 3. Division Four is the final analysis.

RELATED WORK

In this research, we describe an approach to fault prediction that makes use of industrial big data, including acoustic and state facts, and uses machine learning techniques to directly uncover the connection between those datasets and equipment failures. Further, failure-causing equipment states can be tracked in real-time for prompt inspection. In comparison to state-of-the-art alternatives, our approach performs better in both judgement and in real time, as demonstrated by the simulation outcomes [11]. After missing, unusual, and highly noisy data from manufacturing machinery has been preprocessed, the primary variables are extracted using recursive elimination of features and cross-validation techniques, and the training and learning rate weights are designed. Preconditioning commercial data and developing a prediction model that utilizes a neural network, this method enhances the judgment of defect prediction in industrial machinery. By isolating the most salient defining features, we can speed up the prediction process. Experiments in the realm of power generation have shown that this strategy is successful at predicting fan blade icing faults [12]. Today's data-driven industrial landscape places a premium on Big Data processing's ability to boost company output. Smart gadgets like the actuators and sensors are becoming increasingly common in manufacturing machinery. These devices are tasked with continuous tracking of machine condition and the implementation of corrective actions prior to the deterioration of component grade and equipment breakage. Unfortunately, not all factories leverage the Big Data that is generated by their manufacturing processes. Because it might take a lot of time and effort, Big Data analytics is often overlooked. In addition, the true value of immediate processing of real time industrial data is frequently underappreciated. In order to increase the value created for businesses, the purpose of the TOREADOR European initiative is to increase the utilisation of Big Data analysis in manufacturing. With an emphasis on the aviation production procedure (as one of the case research topics

CHAPTER 44

Comparison Analysis of Logical Regression and Random Forest with Word Embedding Techniques for Twitter Sentiment Analysis

Dhiraj Singh[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: In order to carry out categorization and generate new categories, the huge volumes of textual materials produced nowadays must be immediately organised. The fundamental method of gaining insights from organising textual data is text classification. Then, we further classify the classes based on the discovered text types. Separated into four stages—pre-treatment, text representation, classifier execution, and classification—we use a wide range of machine learning approaches to classify texts. In this study, we utilise real-world data from Twitter to evaluate and compare several sentiment analysis approaches. We clean the data and divide it into train and text set before developing models using various vectorising approaches and compare the outcomes. Based on a comparison of the models with various vectorizations, it was found that the best performance was provided by the Logical Regression (LR) models using TF-IDF, with an f1 value of 0.81 and good accuracy and recollection values.

Keywords: Logical regression, Random forest, Twitter sentiment analysis, TF-IDF.

INTRODUCTION

Textual sentiment classification experiments employing online data have been conducted extensively in order to analyse the public's prevailing mood [1]. Sentiment classification at the aspect level, a more granular task that looks at both the sentence's overall content and its aspect information to determine its sentiment polarity, has lately received a lot of interest [2]. The main reason for this is that a statement may include elements of opposing polarity. In the sentence “This apple looks nice, yet the flavours are bland,” the dichotomy is positive with respect to the observation that “This apple looks nice,” but it changes to negative with

[*] **Corresponding author Dhiraj Singh:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: dhiraj.singh.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

respect to the observation that "The taste is bland." To address the problem that existing emotion classification algorithms exhibit certain imperfections in obtaining as well as integrating both global and local characteristics of short text, in addition to a lack of attention to the crucial details in sentences, this paper provides a Weibo short-text opinion classification approach [3, 4]. Information may be efficiently extracted from time sequences using RNNs. Designing a perfect deep RNN is tough [5] because of setup and training issues including disappearing and exploding gradients. Comparative investigation reveals that the capsule network lacks the flexibility to quickly adopt the model compression method used by the traditional convolutional neural network. Based on the structural characteristics of the Capsule Networks, a proposed solution is an IPC-CapsNet compression method. The method has the potential to minimise computing complexity and condense the size of model computing [6] while preserving the reliability of model classification. In this paper, gradient boosting, a machine learning method that is currently gaining a lot of popularity, is also used. The harmony search technique has been used to optimize a number of tree parameters for the random forest and gradient boosting machine [7].On two separate datasets, namely the IMDb dataset and the polarity dataset, the performance of the proposed technique has been assessed using a variety of performance evaluation criteria, including Exactness, Recollection, f-measure, and accuracy [8]. The primary goals of this work are to describe the Random Forests construction process and the current state of Random Forests research in terms of capacity expansion and performance metrics [9]. It has been shown experimentally that this model requires less time to train than TextRNN. In the realm of short-term and low-volume data sentiment categorization, our 90.2% accuracy is the most advanced [10].

With an ever-increasing number of online conversations and transactions, it is important to be able to decipher the underlying feelings and perspectives expressed in online writing. Particularly applicable in the realm of social media, it may be used to gauge general sentiment toward certain issues. Data gleaned from social media sentiment analyses has a wide variety of potential uses, ranging from fine-tuning individual products and marketing strategies to informing public policy and forecasting economic performance.

- Information on Twitter and other social media platforms is a treasure mine of expression data, including information on individuals' feelings, thoughts, and perspectives on their everyday lives.
- Sentiment analysis is useful since so much of this stuff informs choices. The automated process of mining attitudes, opinions, thoughts, and emotions inside the text provides insights into the message's positivity, negativity, or apathy.
- The goal of this website is to examine how various sentiment analysis

techniques perform on actual Twitter data. Kaggle datasets (containing information about genuine tweets) are obtained, labelled (where zero represents a negative sentiment as well as one indicates an optimistic one), and then used to train a machine learning model.

LITERATURE REVIEW

It is difficult to analyze sentiment on Malaysia's social media because messages are typically written in a combination of English and Malay, with embedded jargon and different district dialects. Each tweet was classified using the Malaysian halal certification scheme in order to determine the frequency value of the class label based on the polarity results of the sentiment analysis technique. It will show the propensity of social media users to post, and users can use it as a resource when making judgments. 500 tweets with the hashtag #sijilhalal were analyzed to learn more about how individuals felt, thought, and acted toward different halal certification-related topics in Malaysia. A person's feelings about halal subjects are discovered and visualized. Muslims' perspectives are crucial for raising awareness of the local dialects [11].

- In this study, we transform the creation of a language for expressing emotions into a training-optimization procedure. In our system, the precision of sentiment categorization serves as the optimization aim, therefore candidate emotion lexicons are seen as the ones that have to be optimized. You may find two genetic algorithms developed to alter the value assigned to words in the lexicon based on their emotional connotations at /en. The sentiment lexicon is ultimately chosen as the best individual that has evolved in the evolutionary algorithms that have been provided. Our approach simply requires a few marked texts, and it doesn't require any prior knowledge or linguistic expertise. It offers a quick and straightforward method for creating a sentiment lexicon in a certain industry. The results of the experiments demonstrate the proposed method's versatility and its ability to produce high-quality sentiment lexicon in particular domains [12].
- The study's four main findings are as follows: Here, we examine the relative efficacy of four distinct methods for conducting sentiment analysis: (1) a proposed method for preprocessing data for emotions categorization; (2) extra features to enhance the Exactness to feelings' categorization; (3) the application of singular as well as principal component assessment for data dimension reduction; and (4) the development of five modules based on various features, with or without a corresponding originating feature. The experimental results show that the suggested method is more efficient and accurate than its predecessors [13].
- This essay suggests a system for categorizing literature into six different emotional states: joy, sorrow, fear, anger, surprise, and disgust. To successfully

CHAPTER 45

The Classification of News Articles Through the Use of Deep Learning and the Doc2Vec Modeling

Himanshu Makhija[1,*]

[1] *Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*

Abstract: The exponential growth in internet use has also led to the proliferation of textual information in large quantities. Since handling unstructured material manually is difficult, there is a need to explore novel techniques for automated categorization of textual information. The primary goal of text categorization is to teach a model to correctly categorise an unseen text. In this research, the Doc2vec word embedding technique was used to classify stories in Turkish from the TTC-3600 database of Turkish news and BBC news stories in English. In addition to the CNN based on deep learning, traditional machine learning classification methods including Gauss Naive Bayes (GNB), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) are used. The best classification results using CNN were achieved with the proposed model, scoring 94.17% on the Turkish database and 96.41% on the English database.

Keywords: CNN, News articles, TTC-3600 databases.

INTRODUCTION

The newspaper articles provide us with a perspective on many current events [1]. The proliferation of online material has led to a direct impact on people's everyday lives, and digital news is no exception [2]. The availability of such a news story calls for the necessity to categorise it. The number of Indonesian news stories published online has exploded thus far this decade [3]. Data on financial systemic danger is included in the data on economy. Automatically classifying systemic risk documents is a necessary step towards obtaining real-time data on finance systemic risk [4]. In this work, we use deep neural networks and proactive learning to automatically categorise documents relating to systemic risk [5]. We use the Bank of Indonesia's 15 categories of financial systemic risk. There may be more than one piece of systemic risk information inside a text document, hence

* **Corresponding author Himanshu Makhija:** Centre for Interdisciplinary Research in Business and Technology, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
E-mail: himanshu.makhija.orp@chitkara.edu.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

this is a tasking-label classification issue [6]. We have made numerous variations in the deep learning strategies of CNN, Bi-LSTM, and Bi-GRU [7]. In addition, we have drawn parallels to a two-stage categorization system. The internet has evolved into a platform for instantaneous global communication [8]. However, this exponential expansion of the World Wide Web has also facilitated the widespread dissemination of false information. Many classifiers have been proposed in the current literature to differentiate between fake and real news. However, they need a lot of information before they can learn well, and such data is difficult to collect [9]. The growth of bogus and falsified material across several internet web platforms in recent years has led to the establishment of skewed and sometimes influenced public opinion on political, social, and other daily life topics. As a result, there have been several attempts to create fake news detection algorithms to aid in exposing and refuting false information [10].

The purpose of this research is to evaluate Deep Learning and Doc2Vec techniques in comparison to previously explored techniques for classifying articles on Turkish and English data sets. Based on the pre-processing processes used with the TTC-3600 and BBC-News databases, four distinct data sets were generated and documented for this study. After developing the Doc2Vec training model, we used CNN, GNB, RF, NB, and SVM to label each dataset.

For the Turkish database, the created model achieved higher rates of accuracy than the previous research. The average accuracy rate for the English dataset came out rather well. The article's remaining sections are as follows: The methods used are described in Chapter 2, and the data set, prior treatment phases, and constructed models are described in depth in the materials and methods section of Chapter 3. The technique discussed in Chapter 4 is evaluated in light of past research, and the paper is concluded .

RELATED WORK

In this study, we present a new deep learning model called Long Short-Term Memory-Gradient Boosting (LSTM-GB) for news article categorization and evaluate its performance in comparison to that of established techniques. The suggested model also outperforms existing classifiers [11] with an accuracy of 99.8 percent.

Using 1752 articles for training and 228 articles for testing, the maximum F1 value was reached by using a Bi-LSTM topology with a single classification step and a large common corpus for word embeddings. For 15 courses, the greatest F1 value was 45.37 percent, using a probability threshold of 0.15. The first phase of the two-stage classification process achieved an accuracy of 82.46%, classifying items into one of the two categories (containing risk information or not). Because

of the small sample size, we used active learning to choose which data points would come next for labelling. The experiment demonstrated that employing active learning did not increase performance [12] for 420 new data items for each iteration adding 20 new data items.

This research aims to identify flooding occurrences by analyzing both the text and visuals of web news articles from Africa. MediaEval 2019 provides this information as part of the Multimodal Satellite Task. Both the picture and text-based tasks benefit from our efforts. We construct models able to obtain features from photos and texts independently, and then integrate them to generate a complex classifier, which is more convincing proof of floods, and so allows us to perform the necessary classification subtasks. In particular, we use a strong text processing approach determined by long short-term storage cells and the convolutional layers-based MobileNet architecture for image processing. On the company's test sets, our final models performed well, with an average F1-value of 85.26 percent for the first subtask as well as 66.19 percent for the second task. [13].

In this research, we explore how deep learning approaches, including testing using various databases and implementation across multiple categorization models, may be used to detect and categorise bogus news throughout the web. We improved accuracy by 96.25 percent using the deep learning model we presented. The techniques of text categorization have been enhanced by the use of word embedding techniques. By minimising the loss value, the model design for fake information classification has been improved. To properly categorise data, the model needs hyper parameters after training on two distinct real-world databases [14].

The goal of this thesis is to visualise language distinctions between false and genuine articles across three distinct fake news databases. The goal of this research is to determine which of the five machine learning classifiers provides the best universal results across all three data sources by comparing them and developing an ensemble approach comprising multiple combinations of classification models. The outcomes of a deep learning approach were then compared to those of statistical machine learning models, thus a basic long short-term memory neural network was built. As the first step in this project's approach, we use NLP and feature extraction tools to clean and prepare the data that will be "fed" into the various classification models throughout the training and tuning phases. Bar graphs, confusion matrices, and extractness-recollotion curves are then used to display and compare the findings of Data Mining, Data Visualisation, Machine Learning, and Deep Learning [15].

CHAPTER 46

Investigating the Utility of Data Mining for Automated Credit Scoring

Amarpal Yadav[1,*]

Department of AI, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India

Abstract: 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 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 Learners (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 methods 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 used 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. 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.

Keywords: Algorithms, Automated Credit Scoring, Data Mining, Investigation Utility, Support Vector Machines (SVMs).

INTRODUCTION

The banking industry is particularly vulnerable to credit risk, making credit risk management more crucial than ever in light of recent economic developments. Previous research has shown that interest earned on credit equals banking principle. Improvements in banking risk management are the focus of credit risk

* **Corresponding author Amarpal Yadav:** Department of AI, Noida Institute of Engineering & Technology, Greater Noida, Uttar Pradesh, India; E-mail: amarpal@niet.co.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

management. Customers' credit risks may be categorised and predicted using information gleaned from their personal data fed into scoring algorithms. Decisions are made by credit specialists based on prior experience; nevertheless, this process yields different results when applied to the first techniques of credit risk assessment. Statistical programmes have come a long way, allowing for more precise scoring models to be developed as a result of technical progress. Automated processing in the banking industry leads to greater precision and lower operational expenses. Banks have historically used regression and neural network-based credit risk rating algorithms due to their excellent accuracy [1].

Linear Discriminant Analysis [2, 3], Decision Tree [4], Logistic Regression [5], F-Score [6], and Genetic Programming [4] are only some of the classification models that have been produced by constructing credit risk scoring models using data mining technologies. Desai *et al*. [7] studied the decision-making process behind credit scores by comparing the efficacies of Neural Networks, Linear Discriminant Analysis, and Logistic Regression. They discovered that when compared to neural networks and logistic regression, linear discriminant analysis yielded worse results. In conclusion, the two most popular credit risk assessment models are those based on neural networks and logistic regression.

Today's business climate is unlike any other time in history, thanks to the proliferation of modern information technologies. Thanks to developments in information technology, advancements have been made in the capacity of information systems to handle large amounts of data quickly and accurately. Instead of concentrating on data compilation and collection transformation, the company here is searching for the most effective means of intercepting information from databases. Data mining refers to the practise of methodically exploring huge data sets in multiple dimensions utilising computing power as well as processing speed to discover previously undiscovered discoveries and trends that may drive corporate and consumer choices. Data mining has found use in many different areas, including but not restricted to the manufacturing, retail, healthcare, and biotech sectors. Data mining's practical applications in business include the financial sectors (insurance, banking, and credit cards). Data mining has potential uses in the biomedical industry, including illness diagnosis. Scholars like Frawley [8] have described data mining as the act of scouring databases for possibly understandable and relevant information. Frawley reframed data mining as a technique for gaining insight into the world in 1996. The algorithmic application, transformation, analysis, and identification of data features and models constitute data mining. Data mining, as described by Grupe and Owrang [9], is the process of extracting previously undiscovered information from databases. Finding significant correlations and laws by automatically or semi-automatically analysing large amounts of data is what data mining is, according to

Berry and Linoff [10]. Data mining was first described by Kleissner [11] as an innovative analysis procedure that might unearth previously unknown useful information to be used as a resource for businesses. Previous researchers had varied points of view on data mining, yet their defining tendencies are generally consistent. The phrase of data mining used here is comparable to Kleissner's.

There are a variety of uses for data mining. Data mining, for instance, may be broken down into subfields such as classification, prediction, organisation, clustering, and so on. To "classify" anything is to create subsets in feature sets based on the characteristics of the targets. Information may be categorised and used as a foundation for decisions using models that describe characteristics and cross-category correlations. The computation of variable attribute values using categorised data might provide prospective rules and classification outcomes. This study's primary objective is to present four new categories. Training data is used in the Decision Tree classification approach to make predictions about categorical and continuous variables. The idea of decision trees is to categorise things based on what we already know. Each category has a unique framework for making calls. A tree diagram depicting the procedure is shown. ID3 [12], C4.5 [13], Segmentation among Reconstruction Tree; and CART [14], among many more, are all examples of decision trees that use text symbols or independent information types. To paraphrase Berkson [15], the inventor of the logistic regression model describes how it might be used to date with just two possible outcomes from a test: success or failure. One of the goals of model building is to define classification rules and make reliable predictions about the connection between the response variable and independent variables. When a single sample is used to foretell the likelihood of success, an attribute of the sample may be determined. In order to solve classification difficulties, logistic regression is often used to aggregate the class variable while creating predictions between 0 and 1. To circumvent the erroneous assumption that the matrix of covariance has to be identical to the binary class [16], logistic regression is often employed as an alternative to linear discriminant analysis when constructing binary classification. One of the most precise binary output approaches is logistic regression technology [17]. To create an output, a Neural Network simply multiplies the values of its input neurons by its link values. Initial link values are generally generated at a rate between +1 and -1 before being trained and adjusted by the neural network. The worth of a link may be thought of as a multiplicative effect. An overactive connection is more likely to have an effect on the neural network if the link's value is high. If a link is too tiny, it may generally be deleted to save up system resources. The constituent parts of the neural network are as follows. Vapnik [18] provided a helpful dimensional categorization method called the Support Vector Machine, which has seen widespread usage in recent years. It has been used for things like credit score calculations [1, 21, 22] and the identification of diseases

CHAPTER 47

Investigating the Use of Data Mining for Knowledge Discovery

Sover Singh Bisht[1,*]

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

Abstract: The practice of "lifelogging" involves documenting an increasing amount of one's subjective 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 an Improved Life Graph (ILG), a first method for building a Knowledge Graph-based lifelog representation and retrieval solution, which can capture a lifelog in a graph structure and augment it with external data to help with the connection of higher-level semantic information.

Keywords: Data mining, Information retrieval, Knowledge graph, Lifelog data.

INTRODUCTION

The possibilities for mobile data collection, processing, and storage have substantially increased because of the various technological advancements over the previous few decades. One possible effect is the rise of related movements, such as quantified self and lifelogging. The term "lifelogging" is used to describe the act of documenting one's daily activities, whether through the use of cutting-edge technology like a digital camera or sensors that measure one's bio-feedback or the attributes of the immediate environment or through more conventional means, such as writing in a diary. As a consequence, the resulting lifelog might take several shapes and have many connections [1, 2]. Eventually, we settled on

[*] **Corresponding author Sover Singh Bisht:** Department of DS, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India; E-mail: soversingh@niet.co.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

the simple but mostly untested idea of encoding a lifelog as a data graph. Lifelogs serve a variety of purposes, including improving memory recall and inspiring self-care. The need for a quick and simple search option inside lifelogging logs grows with the size and complexity of the data collected. The Lifelog Search Challenge (LSC) [3, 4] was a workshop organised in 2018 to competitively assess interactive lifelog retrieval technologies. Within the framework of the 2020 LSC, this research describes and demonstrates LifeGraph [5], an interactive information graph-based lifelog retrieval system. This section provides a high-level overview of the graph, how the system makes use of the network to answer questions, and the potential ways in which users could engage with the system. Recent advances in retrieval technology have made broad use of this method possible. This higher level of intelligence is very useful in a variety of application fields (data fusion, geographical analysis, early warning, *etc*.) [6, 7]. Since present methods for extracting geographical data fail to fully take into consideration the spatial and semantic dimensions, there is a room for improvement. Due to this diversity, it is presently not viable to directly get geographical data from many sources [8, 9]. However, since it is challenging to account for implicit semantic data and complicated spatial linkages, it might lead to disorientation and information overload [10,11]. Ontology-based data access (OBDA), another popular data access paradigm, often uses a commonsense knowledge base, hence it lacks spatial semantics [12,13]. A knowledge graph (KG) that accounts for the link between geographical data and semantics is required for efficient storage and retrieval of multi-source heterogeneous data. KG's massive data set is used to enhance intelligent retrieval techniques [14] and deal with the semantic gap. The combination of KG and GIS shows promise as a means of systematic geographical data collection. In this study, we present a new retrieval method that makes use of spatial semantics to increase effectiveness while still meeting strict criteria. The purpose of KG is to uncover the hidden meanings in geographical information in order to better structure and combine data from various sources. By elaborating on connected geographical elements, you may get data that conforms to the suggested semantics of search queries. Finally, we explored the concept through empirical evidence and comparative analysis.

RELATED WORK

Here are some potential research directions and some contextual information to get you started.

Analysing Spatial Information 2.1: When referring to information that is both geographically and spatially based, the term “geospatial data” is often employed. Traditional information retrieval techniques include a subset that deals with geospatial data retrieval [15]. Understanding the geographical data model is

crucial in today's age of big data for the purpose of knowledge sharing and retrieval. The two most well-known geospatial data models are the geographical vector model and OpenStreetMap (OSM).

Geospatial Information System Model 2.1.1 Data is shown using the geographic vector model, which consists of geographical layers, attributes, and property fields. In order to handle geographical data (*i.e*. geometry data) and attribute data, two distinct sets of files are often employed [9]. Spatial data is stored in spatial data files and attribute data is stored in attribute files in Fig. (**1**) geographical vector models like the one used to illustrate the road in Fig. (**1**). Using a certain map projection, spatial information files record the locations of features on the ground as a collection of points, lines, polygons, as well as other geometric primitives. Therefore, when using these characteristics, a two-dimensional Cartesian system of coordinates is the most accurate representation of geographical distribution and topological structure.

Version 2.2.1 of the tried-and-true Attribute-Based Data Retrieval System from Data Rescuing. The technique of collecting geographical data based on attributes is comparable with the standard method based on character traits. Early studies largely concentrated on name retrieval techniques, leading to the proposal of many search algorithms dependent on domain dictionary entries, such as Hash indexing. Hash index retrieval uses a structure with three levels (*i.e*., the primary text, a word indexing table, and a Hash table holding the start characters) to filter results in a binary form. However, the fact that it relies on global matching is perhaps its worst flaw. The Trie index technique employs a tree structure (composed of a Hash table for the first character and a tree index) to hasten the matching process while maintaining high accuracy. There are, however, constraints, such as the need for sophisticated indexes and substantial memory. In order to speed up data retrieval, the dbl-word Hash index approach combines the most useful aspects of Hash and Tries. In this approach, the Trie index is used to return phrases with less than or equal to two individuals, while the Hash index is used to find words having at least three characters. This inspired scholars to delve even deeper into the meanings of place names. Characters are viewed as the building blocks for indexes, and therefore Zhang *et al*. devised a character-features-based technique for retrieving geographical names.

Space-Based Information Localization. The principal uses of spatial information retrieval are in the areas of acquisition, display, and analysis of ground objects. The spatial index primarily functions to speed up the retrieval process and to filter out superfluous ground elements. In recent years, several spatial indexes have been devised to provide faster and easier access to two-dimensional geographical data. These spatial indexes include the quadtree, the R-tree, and the R *-tree. Until

CHAPTER 48

Exploring the Role of Big Data in Predictive Analytics

T. R. Mahesh[1,2,*]

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

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

Abstract: Cardiovascular illness is afflicting enormous monetary and psychological costs. The development of an ASHRO-based model for forecasting healthcare resource use and its link with clinical outcomes was driven by a desire to improve the economy and provide a high-quality evaluation of the healthcare system. Data included in this analysis were taken from a big database that included doctor visits, insurance claims across several years, and results of preventive health screenings. Hospitalized patients with heart illness (ICD-10 I00-I99) comprised the study population. Broadly defined composites compliance served as the explanatory variable, while medical as well as long-term care costs served as the objective variable. Using a combination of random forest learning (AI) and multiple regression analysis, predictive models were calibrated. These models were then used to create ASHRO scores. Two measures, the area under the curve as well as the Hosmer-Lemeshow test, were used to assess the prediction model's effectiveness. After controlling for clinical risk variables, we compared the two ASHRO 50% threshold groups' total morbidity at 48 months of follow-up using matching propensity scores. Heart disease affected 61.9% of the 48,456 patients surveyed, with an average age of 68.3 9.9 years at hospital release. For the purpose of adherence score classification, machine learning was employed to combine eight factors into a single index: generic drug rate, interconnecting outpatient visits/clinical laboratory as well as physiological tests, the proportion of days addressed, secondary mitigation, rehabilitation magnitude, direction, and a single index that adjusted for eight factors. In the end, the multiple regression study yielded a 0.313 (p 0.001) coefficient of determination. Medical as well as long-term care expenditures had a statistically significant total coefficient of determination (p 0.001) in a logistic regression study using 50% along with 25%/75% cut-off values. At the 50% level of significance (2% *vs*. 7%; p 0.001), the relationship between ASHRO score and mortality rate was statistically significant.

[*] **Corresponding author Mahesh T. R.:** Department of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, India; E-mail: t.mahesh@jainuniversity.ac.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

Keywords: Artificial Intelligence, Big Data, Multiple regression study, Predictive analytics, Predictive models.

INTRODUCTION

Both Cappelli (2000) and Naraynan *et al*. (2019) found that human resources were the most important factor in a company's success. To make the most of this asset, several firms throughout the world have implemented talent management techniques (Cappelli, 2000). Evidence suggests that efficient and effective human resources administration (HRM) practices boost employment stability and retention rates (Irshad as well as Afridi, 2012). Keeping current staff members on board is a critical HR task. Armstrong (2006) and Paille (2013) both stress the importance of staff retention in giving businesses a competitive edge. For a business to maintain a competitive edge over the long term, investing in a bright and skilled workforce is crucial (Kumar & Kaushik, 2013). In today's competitive job market, keeping top personnel is becoming an increasingly pressing challenge for a company’s executive (Naris alongside Ukpere, 2010; Olckers through Du Plessis, 2012). In this regard, many references are cited: Davenport & Harris (2007), Chen *et al*. (2012), Watson (2014), Frisk & Bannister (2017), Kar & Dwivedi (2020), as well as Bag *et al*. (2021). Big data and data-driven decision-making are gaining traction in the management sphere. Companies are becoming more interested in data-driven decision-making thanks to Big Data Predictive Analytics (BDPA) and its sophisticated Big Data applications (Dawson *et al*., 2007; Secundo *et al*., 2017; McAfee *et al*., 2012; Agrawal *et al*., 2021). According to Wong (2012), FossoWamba *et al*. (2016), and Du bey *et al*. (2019), the area of big data has the potential to completely transform whole business operations, which has piqued the attention of both academics and lawmakers. Notwithstanding researchers' best efforts, studies on information management and human resource management have failed to provide any significant findings (Garcia-Arroyo & Osca, 2019; Rombaut *via* Guerry, 2020; Hamilton with Sodeman, 2020). Notwithstanding several attempts to integrate research and practice (Shah *et al*., 2017; Calvard and Jeske, 2018; Huang *et al*., 2021), the two remain seen as distinct entities. When looking at what influences employee retention, there is a lack of data in the existing literature. On top of that, the literature is lacking in concrete evidence about the BDPA's effect on staff retention.

Various pathogenic states, intricate disease processes, and a propensity toward chronicity alongside the recurrence of acute phase events are hallmarks of circulatory illnesses. The need for long-term care is a substantial financial burden on society, and these variables significantly reduce the patient's quality of life and deteriorate their prognosis [1, 2]. The tremendous advancements in medicines and

medical technology have not prevented the enormous increase in the unit price for medical care. Improvements in lifestyle and other factors are also contributing to the 10,000 annual increases in the number of Japanese inpatients suffering from cardiovascular disease along with other ailments [3]. Cardiovascular medical expenditures were 19.7 percent of overall healthcare expenditures in Japan in 2018 (for individuals aged 65 and over, a 5.7 percent increase from the prior year), ranking them highest within this framework [4]. The efficient use of social capital—which encompasses medical costs associated with long-term care—and the enhancement of clinical outcomes has given rise to a policy disagreement in the cardiovascular profession [5]. The Japanese government keeps track of how much people spend on healthcare and nursing home care. Over the previous decade, medical expenditures have climbed by more than 2% yearly, reaching 8.0% of GDP [6]. This is because of the world's aging population and rapid developments in medical science. Hospitalization rates and lengths of stay among the elderly are major contributors to regional disparities. Long-term care expenses, meanwhile, have been rising at a rate of about 5% annually since 2010, which is faster than the 2% annual growth in GDP in the Fiscal Year 2016 [7]. Improving the administration of clinical quality and healthcare resources is crucial. This would be a great use for the prediction models. Predicting the severity of patient problems and the results of therapeutic treatments is an active element of clinical practice, and so is the creation of models to do so. Comprehensive risk assessments are used as models to forecast the critical prognosis of heart failure; such examples are the MEESSI-AHF, the HFSS, as well as the GWTG-HF [8, 9]. The Spanish assessments of patients having acute heart failure in the emergency room are the basis for these ratings. Some more factors for determining the prognosis of chronic heart failure include the Seattle Cardiac Failure Simulations, which include risk stratification. Public insurers seldom attempt to construct model for risk assessment as well as prediction; private corporations, on the other hand, handle medical along with long-term care insurance money.

RELATED WORKS

In the fields of biomedicine and healthcare, early detection is crucial. Accurate medical outcomes can be predicted with the use of well-analyzed data. Furthermore, several regional illnesses have distinctive local manifestations, which might hamper the ability to foretell disease epidemics globally. In order to successfully anticipate the start of chronic illness outbreaks in disease-frequent societies, this study will provide a rationale for the machine learning methods needed to do so. The rising area of big data hopes to tackle every type of analytical difficulty. Big data is the accumulation of data that will preserve the data qualities to anticipate the disease, and traditional data will have the three

CHAPTER 49

Implementing Automated Reasoning in Natural Language Processing

N. Sengottaiyan[1,*] and **Rohaila Naaz**[2]

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

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

Abstract: One deep learning method is the Convolutional Neural Network (CNN). Natural language processing problems like text classification are simplified using this approach. In this study, we use a deep learning strategy, namely the CNN method to deal with the issue of text classification. CNNs, which require a large deal 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 learned 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.

Keywords: CNN, Deep learning, Logical reasoning, NLP, Sentiment analysis.

INTRODUCTION

To analyse this data, a machine learning technique called Natural Language Processing (NLP) is needed. Since NLP enables companies to utilise huge data in creative ways to get significant insights into present and future market patterns, it is frequently seen as the market's next big thing.

The area of NLP has been studied for decades, but it is only in the past three that major advancements have been made. Partnering with a big data consulting firm, businesses are increasingly using machine learning techniques that make use of natural language processing.

Natural language processing (NLP) analyses the linguistics and semantics of the big data's text entries using statistics and machine learning, then extracts the rele-

[*] **Corresponding author N Sengottaiyan:** School of Computer Science and Engineering, JAIN (Deemed-to-be University), Bangalore, India; E-mail: sengottaiyan.n@jainuniversity.ac.in

Pankaj Kumar Mishra and Satya Prakash Yadav (Eds.)

vant entities and connections in the context of the customers' postings. Rather than analysing individual words or phrases, NLP looks at whole sentences to determine their meaning. Automatic synthesis, disambiguation, component-of- speech tagging, relations extraction, entity extraction, and, most crucially, natural language comprehension and recognition are some of the most prevalent techniques used in NLP.

The use of CNN techniques has led to remarkable developments in several areas, including computer vision and pattern recognition. The convolutional model's capacity to learn an ordered representation of features from pixel to line, contour, form, and object has been cited in many papers as the driving force behind these advancements. Different networks have been suggested in the literature to apply this paradigm to text classification, with positive results. The process of classifying texts into predetermined groups based on their content is known as text categorization or text classification. Parsing, semantic evaluation, information extraction, and Web searching are just a few of the many fields that benefit from its conceptual perspectives of document collections [1].

As a consequence, more researchers are becoming curious about it. Immense-scale classification systems, such as Google's spam filters [13] or Netflix's [14], however, need massive amounts of computer resources due to the immense training data and the tremendous number of variables that need to be fine-tuned. This is the fundamental issue with using CNNs in practical applications of context-dependent words. The words immediately around a search phrase are referred to as "context." The problem of using an Out-Of-Vocabulary (OOV) term, or an unfamiliar word, is prevalent in languages with extensive vocabularies. Since each word is reduced to a collection of letters in character embeddings, this problem is easily solved. Establishing structures at a character level is an obvious decision to avoid word division [2, 3] in works that use applications of deep learning on dialects where text cannot be made up of surrounded words but rather individual characters as well as the symbolic significance of words map to its chemical composition. The initial stage in character embeddings is to define a set of characters to work with. The alphabet plus a few other symbols, for instance. After that, we'll have a series of vectors and the characters would be transported using a one-hot encoding. Each character is represented by a vector of constant length in the final output. Next, the sequence is learned using 1D CNN layers.

One of the first things a model that uses deep learning does is learn how to represent words or characters. As the first machine learning layer in a CNN, words or embeddings of characters are often utilised. Here, we introduce two types of distributed representation—word and character embeddings—and discuss

their advantages and disadvantages. Word incorporation is a technique for representing words in a vector space using a distributional assumption [4 - 7].

It suggests that comparable meanings might be attributed to words when they appear in the same contexts or the same places in two separate sources. Word embeddings are learned by propagating back an error function *via* a neural network, which is an adaptation of this notion from what we term distributed representations of words. In statistical language modelling, this similar concept was subsequently used to discover both word vectors and probabilities for word sequences [8-10]. When it comes to capturing the semantic and syntactic similarities between words, this approach has shown to be quite effective. Distributed representations provided an answer to the sparsity and curse of dimensionality issues by allowing words to be expressed in a stable, real-valued compact region with much lower dimensions than the entirety of the vocabulary.

There have been several efforts to improve word embedding learning in terms of accuracy and computing efficiency of the neural deterministic language model. In particular, Word2vec [11] has excelled in both accuracy and processing efficiency because of its straightforward and decentralised design. The purpose of Word2vec is to develop a model for vector space word embeddings with few dimensions. To do this, we may train a neural network to predict the probability of a certain word sequence within a specified time frame. The words immediately around a search phrase are referred to as "context." The problem of using an Out-Of-Vocabulary (OOV) term, or an unfamiliar word, is prevalent in languages with extensive vocabularies. Since each word is reduced to a collection of letters in character embeddings, this problem is easily solved. Building systems at the protagonist level is an obvious decision to avoid word division in languages (like Chinese) where text cannot be made up of separated words but rather individual characters alongside the linguistic significance of words maps to its compositional characters [12 - 14].

RELATED WORK

In order to extract each entity in biomedical datasets, Chen, A. *et al*. created an open-source platform called DataMed [1]. The primary goal of this study is to identify useful datasets for the data reuse procedure. This methodology needs a new optimum document ranking strategy to identify and extract important documents from biomedical databases as the size of these sets grows. DataMed is a tool for indexing, ranking, and searching things in various biological datasets. There are two crucial parts to this method, and they are: The first step in building a model is the data intake pipeline, which is in charge of collecting and processing the raw metadata needed for building the model. DatA Tag Suite (DATS) is the

SUBJECT INDEX

D

E

F

G

H

I

L

M

N

O

P

R

S

T

V

W

Pankaj Kumar Mishra

Pankaj Kumar Mishra is a dynamic and innovative director with a doctoral degree in mechanical engineering. He has an exceptionally brilliant two decade long experience in teaching and research. He is a multifaceted leader and an expert in curriculum development, program coordination and staff supervision. His ability to facilitate collaborative environments, improve student outcomes and foster a culture of academic excellence is profound. He is a vivid researcher and has more than 40 publications in refereed international/national journals and conferences and edited several conference proceedings. He has authored and reviewed some books and has been an active member of many professional societies. He was awarded with 'Teachers Excellence Award 2017 by Confederation of Education Excellence for outstanding contribution in Education.

Satya Prakash Yadav

Satya Prakash Yadav (SMIEEE) is currently the associate professor of the School of Computer Science Engineering and Technology (SCSET), Bennett University, Greater Noida (India) and has completed postdoctoral research fellow from Federal Institute of Education, Science and Technology of Ceará, Brazil. He was awarded a Ph.D. from Dr. A.P.J. Abdul Kalam Technical University (AKTU) (formerly UPTU). Currently, 6 students are working for Ph.D. under his guidance. He has more than 17 years of experience as academician, and has published four books (Programming in C, Programming in C++ and Blockchain and Cryptocurrency) under I.K. International Publishing House Pvt. Ltd. His area of specialisation is image processing, information retrieval and features extraction. Besides, he is editor in chief of the Journal of Cyber Security in Computer System & Journal of Soft Computing and Computational Intelligence (MAT journals).