الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

جامعة غرداية
University of Ghardaia
كلية العلوم والتكنولوجيا
Faculty of Science and Technology
قسم الرياضيات والإعلام الآلي
Department of Mathematics and Computer Science
مخبر الرياضيات والعلوم التطبيقية
Mathematics and Applied Sciences Laboratory

Registration n°:
...../...../...../...../.....

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master

Domain: Mathematics and Computer Science
Field: Computer Science
Specialty: Intelligent Systems for Knowledge Extraction

Classification of X Users based on their Interests Using Deep Learning and Machine Learning Algorithms. A comparative study

---

Presented by:

BAALIOUSAID Anis & HADJ SAID Aoumer

Publicly defended on 09 22, 2024

Jury members:

Mr. Houssem Eddine DEGHA	MCB	Univ. Ghardaia	President
Ms. Nassira BRAHIM	MCB	Univ. Ghardaia	Examiner
Mr. Abdelkader BOUHANI	MCB	Univ. Ghardaia	Supervisor

Academic Year: 2023/2024

Acknowledgment

First, we thank Allah for giving us the knowledge, patience and health to be able to carry out this graduation project.

We also thank all the members of our Families for their moral, physical efforts and advices that helped in the realization of this project. We extend our deepest gratitude to all our teachers of the Computer Science department in the university of Ghardaia.

Finally, we would like to thank all those who have contributed in one way or another to the realisation of this work.

Dedication

This modest work is dedicated To,
My parents Whom affection, love, encouragement, and prayer of day and night make me able to get such success and honor.

Dedication

In the name of God, the most gracious, the most merciful I dedicate this work: To my dear mother, who endured the difficult days and hardships for my comfort. To my dear father, who taught me the meaning of diligence and persistence. To my honorable family, who did not hesitate to encourage and support me at every stage of my academic life. To my dear friends, who shared with me moments of joy and challenges. To my distinguished teachers, who did not skimp on me with their knowledge and experience, and provided me with guidance and direction throughout my studies. I dedicate this humble effort as an expression of our gratitude and appreciation.

مـــلـــخـــص

في العصر الرقمي الحالي، يعد فهم اهتمامات المستخدمين على منصات التواصل الاجتماعي مثل إكس أمرًا بالغ الأهمية لتعزيز تفاعل المستخدمين وتخصيص المحتوى. تتناول هذه الدراسة التحدي المتمثل في تصنيف مستخدمي إكس بناءً على اهتماماتهم باستخدام خوارزميات التعلم الآلي والتعلم العميق المختلفة. تم تجميع مجموعة بيانات شاملة تغطي فئات اهتمامات متنوعة كالسياسة والرياضة والصحة...، وتمت معالجتها مسبقًا لضمان جودة البيانات واتساقها. نفذت الدراسة عدة نماذج، بما في ذلك خوارزميات التعلم الآلي التقليدية (الغابة العشوائية (RF)، الانحدار اللوجستي (RL)، نايف بايز(Naive Bayes)) وهياكل التعلم العميق (الشبكة العصبية التلافيفية (CNN) ، الشبكة العصبية المتكررة المدمجة مع الشبكة العصبية التلافيفية (CNN + RNN)، وLSTM ثنائي الاتجاه (Bi-LSTM).

تظهر النتائج أنه بالرغم من أن النماذج التقليدية مثل الغابة العشوائية (RF) حققت دقة عالية وكفاءة حسابية 94.13%، إلا أن نماذج التعلم العميق، وخصوصًا LSTM ثنائي الاتجاه، تميزت بقدرتها على التقاط الأنماط المعقدة والمعلومات السياقية داخل البيانات. حقق LSTM ثنائي الاتجاه أعلى دقة بنسبة 92.54%، وإن كان بتكاليف حسابية أعلى وأوقات تدريب أطول. أظهرت مقاييس الدقة، والاسترجاع، وF1-score قوة كل نموذج باستمرار، حيث أظهرت الغابة العشوائية ونماذج التعلم العميق أداءً قويًا وفقًا لمعايير التقييم المختلفة.

كما تناولت الدراسة تحديات كبيرة مثل عدم توازن البيانات والإفراط في التكيف من خلال تقنيات مثل زيادة البيانات، والتنظيم، وضبط المعاملات الفائقة. كشفت تحليلات وقت التنفيذ أن النماذج التقليدية مناسبة للتطبيقات في الوقت الحقيقي نظرًا لسرعتها كنموذج (Naive Bayes) الذي استغرق وقتا قياسيا، بينما تستفيد نماذج التعلم العميق من تسريع المعالج (GPU Acceleration) للتعامل بكفاءة مع مجموعات البيانات الأكبر.

بشكل عام، تؤكد هذه الدراسة المقارنة على أهمية اختيار النماذج المناسبة بناءً على متطلبات المهمة المحددة. تشير النتائج إلى أن النهج الهجين، الذي يستفيد من سرعة نماذج التعلم الآلي التقليدية وقدرات التعرف على الأنماط المتقدمة لنماذج التعلم العميق، يوفر حلاً فعالاً لتصنيف اهتمامات المستخدمين على إكس. تضع هذه الدراسة أساسًا لتطوير أدوات تحليل وسائل التواصل الاجتماعي المتقدمة، مما يسهم في فهم أعمق لسلوك المستخدمين في العصر الرقمي

كلمات مفتاحية: التعليم اللآلي, التعليم العميق, تصنيف المستخدمين, الغابة العشوائية, الإنحدار اللوجيستي, الشبكات العصبية, CNN, RNN, Bi-LSTM, الدقة.

Abstract

In today's digital age, understanding user interests on social media platforms like X is crucial for enhancing user engagement and personalizing content. This study addresses the challenge of classifying X users based on their interests using various machine learning and deep learning algorithms. A comprehensive dataset encompassing diverse interest categories like Politics, Sport, and Health was compiled and preprocessed to ensure data quality and consistency. The study implemented multiple models, including traditional machine learning algorithms (Random Forest, Logistic Regression, Naive Bayes) and deep learning architectures (Convolutional Neural Network, RNN combined with CNN, and Bidirectional LSTM).

The results demonstrate that while traditional models like Random Forest achieved high accuracy and computational efficiency 94.13%, deep learning models, particularly the Bidirectional LSTM, excelled in capturing complex patterns and contextual information within the data. The Bidirectional LSTM achieved the highest accuracy of 92.54%, albeit with higher computational costs and longer training times. Precision, recall, and F1-score metrics consistently highlighted the strengths of each model, with Random Forest and deep learning models showing robust performance across various evaluation criteria.

The study also addressed significant challenges such as data imbalance and overfitting through techniques like data augmentation, regularization, and hyper-parameter tuning.

Execution time analysis revealed that traditional models are suitable for real-time applications due to their speed especially Naive Bayes, while deep learning models benefit from GPU acceleration to handle larger datasets efficiently.

Overall, this comparative analysis underscores the importance of selecting appropriate models based on specific task requirements. The findings suggest that a hybrid approach, leveraging the speed of traditional machine learning models and the advanced pattern recognition capabilities of deep learning models, offers an effective solution for user interest classification on X. This research lays a foundation for developing sophisticated social media analytics tools, contributing to a deeper understanding of user behavior in the digital age.

Keywords: User Interest Classification, X Data, Random Forest, Logistic Regression, Naive Bayes, CNN, RNN, Bi-LSTM, Accuracy, ML, DL.

Résumé

Dans l'ère numérique actuelle, comprendre les intérêts des utilisateurs sur les plateformes de médias sociaux comme X est crucial pour améliorer l'engagement des utilisateurs et personnaliser le contenu. Cette étude aborde le défi de la classification des utilisateurs de X en fonction de leurs intérêts en utilisant divers algorithmes de l'apprentissage automatique et de l'apprentissage profond. Un ensemble de données complet couvrant diverses catégories d'intérêts comme Politique, sport et santé a été compilé et prétraité pour garantir la qualité et la cohérence des données. L'étude a mis en œuvre plusieurs modèles, y compris des algorithmes de machine learning traditionnels (Arbre aléatoire, Régression Logistique, Naive Bayes) et des architectures de deep learning (réseau de neurones convolutifs, RNN combiné avec CNN, et LSTM bidirectionnel).

Les résultats montrent que, bien que les modèles traditionnels comme Les arbres aléatoire aient atteint une haute précision et une efficacité computationnelle 94.13%, les modèles de deep learning, en particulier le LSTM bidirectionnel, se sont distingués par leur capacité à capturer des schémas complexes et des informations contextuelles dans les données. Le LSTM bidirectionnel a atteint la précision la plus élevée de 92,54 %, bien qu'avec des coûts computationnels plus élevés et des temps d'entraînement plus longs. Les métriques de précision, rappel et F1-score ont constamment mis en évidence les points forts de chaque modèle, avec Random Forest et les modèles de deep learning montrant des performances robustes selon divers critères d'évaluation.

L'étude a également abordé des défis significatifs tels que le déséquilibre des données et le surapprentissage grâce à des techniques comme l'augmentation des données, la régularisation et l'optimisation des hyperparamètres.

L'analyse du temps d'exécution a révélé que les modèles traditionnels conviennent aux applications en temps réel en raison de leur rapidité particulièrement bayésien naïf, tandis que les modèles de deep learning bénéficient de l'accélération GPU pour gérer efficacement des ensembles de données plus grands.

En somme, cette analyse comparative souligne l'importance de choisir des modèles appropriés en fonction des exigences spécifiques de la tâche. Les résultats suggèrent qu'une approche hybride, tirant parti de la rapidité des modèles de machine learning traditionnels et des capacités avancées de reconnaissance de schémas des modèles de deep learning, offre une solution efficace pour la classification des intérêts des utilisateurs sur X. Cette recherche pose les bases du développement d'outils sophistiqués d'analyse des médias sociaux, contribuant à une compréhension plus approfondie du comportement des utilisateurs à l'ère numérique.

Mots clés: Classification des intérêts des utilisateurs, Données X, Arbres aléatoires, Régression logistique, Naïve Bayes, Réseaux de neurones, (CNN), (RNN), (Bi-LSTM), Précision, Apprentissage automatique, apprentissage profond.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

Bi-LSTM Bidirectional Long Short-Term Memory

CNN Convolutional Neural Network

DL Deep Learning

GPU Graphics Processing Unit

LDA Latent Dirichlet Allocation

LR Logistic Regression

LSTM Long Short-Term Memory

ML Machine Learning

NLP Natural Language Processing

PCA Principal Component Analysis

RF Random Forest

RNN Recurrent Neural Network

SVM Support Vector Machine

Chapter 1

Introduction

In the fast-paced digital world, social media platforms have become essential for communication, interaction, and information sharing. Among these platforms, X stands out with its vast user base and real-time data flow. With over 330 million monthly active users worldwide, X is a rich resource for researchers and businesses to analyze user behaviors, interests, and trending topics. Understanding user interests on X offers significant potential, providing valuable insights for targeted marketing strategies, personalized content recommendations, and enhanced user engagement.

As the volume of data generated on X continues to grow exponentially, the need for sophisticated analytical techniques to process, analyze, and extract meaningful patterns from this data has become increasingly important. Traditional machine learning algorithms have long been the mainstay of text classification tasks, but the advent of deep learning has brought about a new era of more complex models capable of capturing intricate relationships within textual data. This study aims to explore and compare the effectiveness of machine learning and deep learning models in classifying user interests on X.

Significance of User Interest Classification

Classifying user interests on social media platforms like X is important for several reasons:

- **Personalized User Experience**: Tailoring content to individual user preferences enhances satisfaction and deepens engagement with the platform.
- **Targeted Advertising**: Precision-targeted advertisements based on user interests improve ad relevance and optimize advertising spending.
- **Content Recommendation**: Recommending content aligned with user interests enriches user experience and fosters meaningful interactions.
- **Market Research**: Insights into consumer preferences from user interest classification inform product development and strategic decision-making for businesses.
- **Social Network Analysis**: Understanding user interests facilitates in-depth analysis of social dynamics, influence patterns, and information diffusion within networks.

Research Objectives

This study aims to achieve several key objectives:

- **Dataset Compilation**: Collecting diverse X datasets covering various interest categories, including business, health, entertainment, politics, sports, technology, and travel.
- **Data Preprocessing**: Ensuring data quality and consistency through thorough preprocessing steps, such as data cleaning, tokenization, normalization, and feature extraction.
- **Model Development**: Developing and training both machine learning and deep learning models for user interest classification.
- **Model Comparison**: Evaluating and comparing the performance of different models based on metrics like accuracy, precision, recall, and F1-score.
- **Insights Extraction**: Extracting insights into the relative strengths and weaknesses of machine learning versus deep learning approaches in the context of user interest classification.

Structure of the Thesis

This thesis is structured as follows:

- **Chapter 1: Basic Concepts**: Introducing fundamental concepts related to social media analytics, user profiling, and text classification.
- **Chapter 2: State of the Art**: Reviewing existing literature on user interest classification on X and comparative studies of machine learning and deep learning algorithms.
- **Chapter 3: Methodology**: Describing the datasets used, preprocessing steps, feature extraction techniques, model selection criteria, and training methodology.
- **Chapter 4: Experiment and Implementation**: Presenting the experimental setup and implementation details, or both machine and deep learning models.
- **Chapter 5: Results**: Detailing the results obtained from the experiments.
- **Chapter 6: Discussion**: Interpreting the results, discussing implications for user interest classification, practical applications, limitations of the study, and suggestions for future research.

Significance of the Study

This study contributes to social media analytics by providing a thorough comparison of machine learning and deep learning techniques for user interest classification. It underscores the practical implications of these techniques in real-world applications, such as enhancing personalized recommendations and optimizing targeted advertising. Furthermore, the insights gained from this research pave the way for developing more sophisticated models and methodologies for analyzing user behavior on social media platforms.

By addressing the challenges associated with processing large volumes of social media data and evaluating the performance of various models, this study aims to advance our understanding of how computational techniques can be leveraged to derive actionable insights from user-generated content on X. Ultimately, this research seeks to elevate the capabilities of social media analytics in delivering personalized and relevant experiences to users while providing invaluable insights for businesses and researchers alike.

In conclusion, the exponential growth of social media platforms like X presents both opportunities and challenges in understanding user interests. This thesis endeavors to navigate these challenges and opportunities through a comprehensive analysis of machine learning and deep learning approaches, contributing to the ongoing evolution of effective and efficient user interest classification techniques.

Chapter 2

Basic Concepts

2.1 Introduction

In this current era (digital age), social media has become a part of our daily lives, serving as a platform for communication, entertainment and commerce. So understanding the massive amount of data being generated on these platforms is crucial for businesses, researchers, and marketers. In this chapter we will see the basic concepts of social media analytics, user profiling, and text classification and a comprehensive overview of their definitions, importance, methodologies, and applications.

2.2 Social Media Analytics

Definition: Social media analytics involves collecting, analyzing and interpreting data in a systematic and structured way from social media platforms such as Facebook, X, Instagram and LinkedIn and is intended to extract useful perspectives. This is for the purpose of understanding patterns and trends within the big data that results from publishing, commenting, interacting, and participating on these platforms on a daily basis (Sebei, Taieb, & Aouicha,).

Importance: Social media analytics are of great importance in our time more than in the past because we are in a time of competition, and every company or marketer wants to be the first. It helps companies and marketers to accurately understand customers in terms of their desires, preferences, and behaviors, which helps them produce products suitable for them, services they are comfortable with, marketing strategies suitable for their customers in the best way, develop their advertising campaigns, and improve their content with the ability to monitor and analyze the performance and strategies of competitors on social media, and this is important. Very focused on finding and exploiting opportunities and confronting threats before they occur (Rao et al.,).

Key Areas: Social media analytics has main areas: sentiment analysis, trend analysis, and user behavior analysis. Sentiment analysis is concerned with identifying, studying and evaluating sentiments and emotions in posts and interactions on social media to understand the opinion of users to determine a good way to

respond to them. While trend analysis is concerned with identifying and analyzing trending topics, issues and hashtags, this helps companies keep pace. User behavior analysis studies the way users interact with social media and this provides insights into engagement patterns, content preferences and social networks, which helps in creating accurate strategies (Yadav et al.,).

2.3 User Profiling on Social Media

Definition: User profiling on social media includes collecting data about users and analyzing it to create detailed profiles about them that contain their characteristics, preferences, and behaviors (U., Sunithamma, Shenoy, & Venugopal,).

Importance: This process is important because it helps to provide accurate services and to make the user experience acceptable and satisfactory. By understanding individual users well, marketers and companies can create content, ads, and recommendations appropriate to specific needs and distinct interests, all in order to increase the activity and growth of the company or marketer (U. et al.,).

Methods: Popular methods for user profiling are demographic analysis, behavioral analysis, and interest analysis. Demographic analysis collects basic information such as age, gender, and location, to facilitate classifying users into groups to customize each group with its services and content or target it with advertisements. Behavioral analysis examines users' actions on social media, such as posting methods and interaction methods, to facilitate predictions of future behavior. While interest analysis determines the topics that users interact with on a regular basis, this helps platforms organize content and ads for users (Cufoglu,).

Applications: User profiling is used in many practical applications, such as personalized advertising, recommendation systems, and market segmentation. Personalized ads target users with ads that match their preferences and behaviors. Recommendation systems on platforms like Netflix and Spotify also make it easier to recommend specific content to a user based on their previous behaviors and interests. Market segmentation allows companies to divide their audiences into distinct groups based on their shared characteristics to facilitate the implementation of more precise marketing strategies (Kanoje, Girase, & Mukhopadhyay,).

2.4 Text Classification

Introduction: Text classification is a machine learning technique used to automatically classify text into pre-defined categories, structuring and organizing different types of text such as articles, medical research, and customer tickets. This processing used to be done manually and this is time consuming and expensive. That's why automated text classification tools combine natural language processing (NLP) and machine learning to efficiently analyze large amounts of text, saving time and resources (Dien, Loc, & Thai-Nghe,).

Relevance: Text classification is very important in many applications and programs, as it is used to detect spam, analyze sentiment, and classify user interests. In spam detection, it helps a lot in filtering unwanted messages by identifying

common spam characteristics. Sentiment analysis uses text classification to identify the emotion used in text to help companies measure customer satisfaction with their products. As for classifying user interests, the text created by the user is analyzed to determine his interests and preferences so that companies can customize content, recommendations, and advertisements according to his interests (E., P, S, K, & V,).

2.4.1 Applications of Text Classification

Spam Detection: Text classification is used to detect unwanted messages and emails by analyzing their content and identifying patterns that indicate spam. This is done by training a machine learning model on a data set of emails previously classified into spam or legitimate. When a new email arrives, the text classifier evaluates its content based on these patterns it has learned. It is classified as regular or spam mail (Kumar,).

Sentiment Analysis: In sentiment analysis, we can use text classification to determine the emotional tone of text data such as posts or product reviews. By training machine learning models on datasets previously classified into emotions (positive, negative, or neutral), the model learns to recognize patterns and linguistic features associated with each emotion. When applied to new data, the classifier predicts the emotion category based on the learned patterns (Saglani,).

Topic Categorization: Text classification is useful for classifying documents into specific topics. By training machine learning models on datasets labeled with pre-defined topics, the classifier learns to recognize terms and patterns associated with each topic. When new documents are submitted, the classifier assigns appropriate topic labels based on these patterns (Elkan,).

User Interest Classification: Text classification can also be used to infer a user's interests based on their social media activity by analyzing the content of their posts, likes, comments, and shares. Machine learning models are trained on labeled datasets where user interactions are labeled with specific interest categories such as sports, technology, fashion, or music. The classifier learns to recognize patterns, keywords, and context within the text that indicate these interests. When applied to new data, the classifier can accurately predict the topics the user is interested in. This allows platforms to personalize content, recommend products, and tailor ads to match individual user preferences, to ensure user satisfaction (Pérez-Vera, ro Sandy González Alfaro, & Allende-Cid,).

2.5 Text Representation Technique

Word Embedding: Word embedding are a way to represent words as dense vectors in a multidimensional space, where the distance and direction between the vectors reflect the similarity and relationships between words. The included words capture rich meanings and semantic relationships. One well-known method for training word embedding is Word2Vec (Lauren, Qu, Huang, Watta, & Lendasse,).

Word2Vec: Word embedding with Word2Vec takes place in natural language

processing (NLP) and creates dense vector representations of words that capture semantic relationships. Word2Vec creates distributed numerical representations where words with similar meanings are placed next to each other in vector space. This technology enables the model to learn the context and meaning of words based on the words next to them in a given text. Through the use of Continuous Bag of Words (CBOW) or Skip-gram structure, Word2Vec well identifies semantic relationships between words, helping NLP systems better understand and process language data (Nugaliyadde, Wong, Sohel, & Xie,).

2.5.1 Techniques in Text Classification

Artificial Intelligence (AI):

It is the technology that makes computers and machines intelligent devices that attempt to access human intelligence and problem-solving abilities. Where artificial intelligence alone or in combination with other technologies can perform tasks that may require human intelligence or intervention. Artificial intelligence includes machine learning and deep learning (Panesar,).

Machine Learning (ML):

Definition: Machine Learning is a subset of artificial intelligence (AI) that is concerned with developing computer algorithms that learn automatically through experience and use of data. That is, machine learning helps computers learn from data and make decisions or predictions without being explicitly programmed to do so. Machine learning focuses on creating and implementing algorithms that facilitate these decisions and predictions. These algorithms are designed to improve their performance over time to become more accurate and efficient as they process more data, ensuring the quality of the decision or prediction (Vermeulen,).

Types: Algorithms can be classified into four methods depending on the expected output and the type of input: supervised machine learning, unsupervised machine learning, partially supervised learning, and reinforced machine learning (Ayodele,). Among the algorithms used are:

- **Naive Bayes:** Naive Bayes is a probabilistic classifier and adopts Bayes' theorem with the assumption of strong independence between features. Naive Bayes classifiers describe the relationship between conditional probabilities. Although the assumption of independence is rarely true in reality, Naive Bayes performs especially well in cases where features are approximately independent. They are used in text classification, such as spam filtering, document classification, sentiment analysis, and recommendation systems to predict user preferences. However, it faces some challenges, including the assumption of independent predictors, which is rarely true in real life, and the problem of the presence of a class in the test data and its absence in the training data, thus causing unpredictability (Bao,).
- **Random forest:** Random forest is a powerful machine learning algorithm that can be used for a variety of tasks such as regression and classification. The random forest model consists of a large number of small decision trees, called estimators, each of which produces its own predictions. The random

forest model then combines the estimators' predictions to produce a more accurate prediction. A decision tree is a simple way to classify a dataset. First, we choose the attribute that enables us to divide the dataset into different categories most effectively. We then partition the dataset on this attribute and create a new node in the decision tree. For each data division, we repeat the same process, dividing the data set by the best feature. We stop creating new nodes when we find that the samples in the current node belong to the same class or if no attribute provides value or if the tree reaches the maximum allowed depth. Random forest is called the ensemble method because it uses many estimators. Each individual estimator is a weak learner, but when a group of weak estimators are combined together, they can produce a much stronger learner (Kulkarni & Sinha,).

- **Logistic regression:** Logistic regression is a data analysis method that relies on mathematics to find relationships between two data factors. It then uses this relationship to predict the value of one of these two factors using the other factor. The prediction values are usually known in advance, such as yes or no (0,1). The advantages of logistic regression include simplicity, speed, flexibility, and vision, and it is used in manufacturing, health care, finance, and marketing. To understand logistic regression, we must first understand basic regression analysis. The first thing is to define the question within a specific framework to obtain specific results. After that, we collect the data, of course, specifying the relevant data factors. Then we train the regression analysis model, and this processes the data using a regression program where various data will be linked and processed mathematically. Finally, we can now predict unknown values, as the program uses an equation to predict them. There are several types of logistic regression: binary, multinomial, and ordinal (Wang et al.,).

Deep Learning: It is a tool used in artificial intelligence that teaches computers to process data, think, and learn like humans, and it is part of machine learning. Deep learning models learn about complex patterns in images, text, sounds, and other data to provide us with accurate insights and predictions. Deep learning is used in tasks that typically require human intelligence, such as describing images, converting voice to text, fraud detection, automatic facial recognition, and many more uses. Deep learning algorithms are neural networks similar to the human brain, and artificial neurons are software units called nodes that are used in data processing, mathematical operations, and in solving complex problems. It consists of three layers: the input layer, the hidden layers, and the output layer. One of the challenges of deep learning is large amounts of high-quality data and large processing power. To meet these challenges, deep learning can be used in the cloud, as it is characterized by speed and scalability (Mo,). Among the algorithms used in deep learning are:

- **CNNs:** Convolutional neural networks are a type of deep learning model designed to process visual data that mimics the human hierarchical organization of the visual system. It is good at image classification, object detection, and segmentation tasks. The CNN architecture consists of several layers namely convolutional layers that apply filters to detect features, activation functions such as ReLU that introduce non-linearity, pooling layers that downsample

the data, and fully connected layers that make the final predictions. Dropout is used to prevent overfitting, and the softmax layer finally assigns probabilities to each class. CNN training contains forward propagation to calculate features and back propagation to adjust weights based on errors (Crowley,).

- **RNN:** A recurrent neural network is a type of artificial neural network in which the connections between nodes form a directed graph along a time sequence, and this makes its behavior temporal and dynamic. Unlike feedforward neural networks, RNNs use their internal state (memory) to process sequences of inputs, making them particularly useful for tasks where the sequence of data points is important. RNNs achieve this by replicating information within the network, knowing that decisions are influenced not only by current inputs, but also by previous inputs. Although RNNs have strengths, they face challenges such as vanishing gradient and explosion problems and computational intensity. To address these challenges, various forms have been developed such as long-term memory (LSTM) networks to better capture long-term dependencies (Yu, Si, Hu, & xun Zhang,).

- **Bi-LSTM:** Bidirectional long-term memory is a type of recurrent neural network (RNN) that consists of two LSTM layers: the first processes the input sequence from past to future (forward) and the second from future to past (backward). This structure helps the network capture and use information from both directions, enhancing its ability to understand context and dependencies in the sequence. By combining hidden states from both forward and backward passes, Bi-LSTM can preserve information from the past and future at any point in time to be useful for tasks where understanding the full context of the sequence is crucial (Su, Huang, & Kuo,).

2.6 Evaluation Metrics

Evaluation metrics are quantitative measures used to evaluate the performance of machine learning models and their effectiveness in performing specific tasks. Different metrics are suitable for different tasks. It is necessary to understand and choose the appropriate metric to interpret the model results (Novakovic, Veljovic, Ilić, Željko M. Papic, & Milica,). Here are some evaluation metrics:

- **Accuracy:** is the ratio of correctly predicted observations to the total observations. It is one of the simplest and easiest to compute evaluation metrics for classification tasks, but accuracy can be misleading in cases of imbalanced datasets (Brodersen, Ong, Stephan, & Buhmann,).

- **Precision:** also known as positive predictive value, is the ratio of correctly predicted positive observations to the total predicted positive observations. It is particular to the accuracy of the positive predictions made by the model. Precision is important in cases where the cost of false positive results is high (Gray, Bowes, Davey, Sun, & Christianson,).

- **Recall:** measures the ratio of correctly predicted positive observations to all actual positives. It reflects the ability of the model to identify all relevant

cases in the data set. Recall is useful in cases where losing positive cases would be costly (Torgo & Ribeiro,).

- **F1-Score:** is the harmonic mean of precision and recall and provides a balance between the two measures, useful when false positives and false negatives are taken into account. F1-Score is useful in situations with imbalanced data sets where neither precision nor recall alone provide a measure of performance (Makhoul, Kubala, Schwartz, & Weischedel,).
- **Confusion matrix:** is a table to describe the performance of a classification model on a set of test data whose true values are known. It shows the number of true positive, true negative, false positive and false negative predictions, to facilitate detailed analysis of model performance. The confusion matrix helps understand the types of errors the model makes and provides the basis for calculating other evaluation metrics (Ting,).

2.7 Conclusion

In this chapter, we looked at the basic concepts of social media analytics through its definition, importance, and main areas. We also looked at user profiles on social media and learned about their importance, methods, and applications. We discussed the classification of text by clarifying its importance and various applications. We discussed also Word2Vec; a text representation techniques, in addition to various machine learning techniques (Naive Bayes, Random Forest, Logistic Regression) and deep learning (CNNs, RNNs, Bi-LSTM) used in text classification. Finally, we reviewed important evaluation metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix, which are used to evaluate the performance of machine learning models.

In later chapters, we will apply and expand on these basic concepts. We will delve deeper into advanced techniques in social media analytics, explore case studies that illustrate the real-world application of these methods, and present empirical results from experiments. We will discuss integrating user profiling and text classification techniques to develop comprehensive models that can predict user interests and behaviors on social media platforms. Through the concepts introduced in this chapter, the following chapters aim to provide a comprehensive understanding and practical framework for conducting effective social media analytics research.

Chapter 3

State Of The Art

3.1 Introduction

After the emergence of social media platforms there has been a change in the way individuals and organizations interact, communicate and exchange information. This shift has given rise to a huge amount of user-generated data, which provides a great opportunity for analysis and insights. In the field of data science, two important areas have emerged: social media analytics and user profiling.

3.2 Overview of Social Media Analytics and User Profiling

Social media analytics involves collecting, processing, and analyzing data from platforms like X, Facebook, Instagram, and LinkedIn to extract valuable insights and patterns. This process begins with collecting data, including posts, comments, likes, shares, hashtags, and user profiles through APIs. The collected data is then cleaned and organized through processing, by removing duplicates, filtering out irrelevant information, and dealing with missing values. Natural language processing (NLP) is often used to extract keywords, sentiments, and topics from textual data. Then analyze the data using statistical techniques and machine learning techniques. Analytical methods include sentiment analysis, trend analysis, user behavior analysis, and network analysis.

Social media user profiling involves collecting and analyzing data to create detailed profiles based on users' demographics, behaviors and interests inferred from their interactions on platforms to personalize user experiences by tailoring content, recommendations and advertisements to individual preferences and needs, to enhance user engagement and satisfaction. Techniques used to identify user data include demographic analysis, behavioral analysis, and interest analysis.

Integrating social media analytics and user profiling involves combining social media analysis and creating user profiles. By analyzing data such as posts, comments, likes, shares, and browsing history, companies and researchers can develop accurate profiles that reflect users' demographics, behaviors, and interests.

This integration helps deliver highly personalized content, recommendations and ads, to ensure user satisfaction.

3.3 Previous Studies on User Interest Classification on X

X is a valuable platform for user interest classification studies due to its popularity, the public nature of posts, and the rich metadata of each post. With users creating content daily, X offers a dynamic and diverse source of data. And for public access to most posts without privacy restrictions. It provides metadata associated with posts such as timestamps, geolocation, reposts, and hashtags. Therefore, X is an ideal platform for studying user behavior and preferences in various fields. In recent years, the topic of Interest Classification on X has witnessed increasing interest from researchers. Many studies have explored this topic and delved into it with different approaches. We limit ourselves to only 5 studies:

3.3.1 Study 1

The study **Interest classification of Twitter users using Wikipedia** wrote by **Kwan Hui Lim and Amit Datta**(Lim & Datta,) creates a framework for classifying the interests of X users by utilizing information from Wikipedia. The framework begins by compiling a list of popular X celebrities and classifying them into different interest categories using data from their Wikipedia pages, with a particular focus on their professions and textual descriptions. This classification is done automatically with the help of a pre-defined library of categories of interest and their associated keywords. Then determine the relative interests of X users by assigning weights to each interest category based on the number of celebrities users follow within each category. The accuracy of the classification was verified using a dataset of 1,000 X celebrities, and is further evaluated using a dataset of 172,400 X users by analyzing their posts and follower/follower links from November to December 2012.

Results: Evaluation of the celebrity classification component showed a high success rate 83.9% in automatically classifying celebrities into their interest groups. Subsequent evaluation of the user interest classification revealed that the framework effectively identified users' interests. Users with a strong interest in a particular category showed higher engagement with related topics compared to users with diverse interests.

Challenges: The study also noted challenges, such as ambiguous celebrity names and incomplete Wikipedia articles, which presented difficulties for the automated classification process.

3.3.2 Study 2

In the study **Classification of Arabic Twitter Users: A Study Based on User Behaviour and Interests** written by **Abdullatif M. AlAbdullatif and**

Basit Shahzad and Esam Alwagait(AlAbdullatif, Shahzad, & Alwagait,), the researchers used a systematic approach to classify Arab X users based on their posts, profile attributes, and behavior. The researchers begin collecting posts from Arab users using the X API. They then collect text data from different news websites related to different interest categories and this data helps test the classification algorithm. With pre-processing of posts. The processed posts are grouped together for each user, forming a document representing their post history. The classification algorithm used is a naive Bayesian classifier, trained on a dataset of cleaned Arabic posts. It calculates the probability that a post belongs to a particular category based on word frequency, and also takes into account profile attributes and user behavior, such as repost rate and celebrity status, to enhance accuracy.

Results: The study showed promising results in classifying Arab X users according to their interests, popularity, and posting behavior. The algorithm achieves high accuracy rates across different interest categories. The accuracy ranges from 91.0% to 98.7%, which indicates the effectiveness of the classifier in correctly identifying users' interests. These impressive results underscore the potential of using machine learning techniques to accurately analyze and classify social media behavior.

Challenges: One important challenge is the abbreviations, slang, and misspellings found in some of the posts. Not ensuring the accuracy and reliability of the data collected, especially when dealing with influential people in the real world. Among the challenges are the limitations of current classification algorithms.

3.3.3 Study 3

In the paper **Efficient User Profiling in Twitter Social Network Using Traditional Classifiers** by **Raghuram, M. A. and Akshay, K. and Chandrasekaran, K.** (Raghuram, Akshay, & Chandrasekaran,)., the authors present a method for classifying Twitter users based on their interests by leveraging a combination of user-based, tweet-based, and time-series features. Their approach focuses on developing a more efficient user profiling mechanism by utilizing traditional classification techniques such as Support Vector Machines (SVM), Naive Bayes, Decision Trees, K-Nearest Neighbors, and Logistic Regression.

Methodology: The methodology adopted by the authors involves extracting three main types of features:

1. **User-based features:** These features represent static user profile information, such as gender, location, follower/friend ratio, and reputation score, and are useful for detecting spammy accounts.
2. **Tweet-based features:** These features include metrics derived from tweet content, such as Term Frequency-Inverse Document Frequency (TF-IDF), the number of hashtags, mentions, sensitive tweets, and hyperlinks.
3. **Time-series features:** These features capture the temporal behavior of users' tweets, analyzing patterns like average tweet frequency, and statistical variations over time. The time at which the tweets were made helps identify

distinct user categories, such as Journalism, which tends to exhibit periodic tweeting behavior.

To classify users into one of six categories—Politics, Journalism, Entertainment, Entrepreneurship, Science & Technology, and Healthcare—the authors applied traditional classifiers like SVM, Naive Bayes, Decision Trees, and K-Nearest Neighbors. They tested their model using 10-fold cross-validation and used Principal Component Analysis (PCA) to reduce the dimensionality of the feature space, which helped in improving the classification accuracy.

Results: The experiments yielded varying results for different classifiers and feature sets. The SVM classifier combined with user-based, tweet-based, and time-series features produced the highest classification accuracy of 89.04%, without the need for dimensionality reduction via PCA. After applying PCA, SVM still performed the best with 89.71% accuracy, using just 50 features. Among the six user categories, the Entertainment class exhibited a perfect True Positive (TP) rate, while Journalism posed the most significant challenge, with frequent misclassification into the Entrepreneurship category. This misclassification is attributed to similarities between business-related news tweets and entrepreneurial activity tweets.

Challenges: One of the significant challenges noted was distinguishing between users in the Journalism and Entrepreneurship categories due to overlapping tweet content. This resulted in a lower TP rate for Journalism, where only 68.1% of actual Journalism users were correctly identified. The authors also highlighted difficulties in scaling the Decision Tree and K-Nearest Neighbors classifiers, as they did not perform well with large numbers of features or data instances. Another challenge was the need for periodic model updates to accommodate changes in user behavior over time, which led the authors to propose a real-time classification system that can adapt dynamically.

3.3.4 Study 4

The study **Twitter Users' Classification Based on Interest: Case Study on Arabic Tweets** Presented by: **Noura A. AlSomaikhi and Zakarya A. Alzamil** (?, ?). The initial step involved collecting approximately 150,000 posts from at least ten accounts for each interest category such as sports, religion, technology, health, economics, and literature. This was done using the X API. Preprocessing of the collected posts included removing URLs, usernames, punctuation, emojis, and stop words to enhance the accuracy of the classifier. The collected posts were divided into training and testing sets, with 60% of the data used for training and 40% for testing. The words were then encoded, extracted, and represented using a bag-of-words model with Biggram features to capture word frequencies. The multinomial Naïve Bayes classifier, known for its simplicity and effectiveness in text classification tasks, was selected and trained using pre-processed posts to assign each post to one of the pre-defined interest categories.

Results: Classification results were compared with classes manually assigned by human raters, using metrics such as precision, recall, and F-measure to evaluate classification accuracy. The study achieved significantly high accuracy with a precision of 91% and 80% for F-measure, indicating the effectiveness of the classification

approach. These results demonstrate that the multinomial Naïve Bayes classifier, when trained with properly preprocessed data, can accurately classify posts into specific interest categories.

Challenges: One big challenge is the dynamic nature of X data, where new terms are constantly being coined, requiring frequent updates to the training data to reflect changes in user interests and language use. Relying on X's API to collect data was also a drawback by imposing limits on the number of posts that could be accessed. In addition to the challenge of ensuring the quality and reliability of manually labeled data to train the classifier.

3.3.5 Study 5

In this study **Using Reddit Data for Multi-Label Text Classification of Twitter Users Interests**, the researchers: **Angel Fiallos and Karina Jimenes** (Fiallos & Jimenes,), focused on automatically classifying users' interests on X using a multi-classification text classification model. A dataset of 42,100 posts was collected from several popular Reddit forums covering a wide range of topics such as sports, entertainment, politics, business, technology, gamers, science, and marketing. This dataset was collected using extraction techniques and Python libraries such as Selenium and BeautifulSoup. The collected data served as training data for the text classification model, which used Word2Vec embeddings and latent Dirichlet allocation (LDA). Word2Vec is trained to learn word embeddings from raw text, assigning vectors to unique words. LDA was applied to identify relevant topics in X users' timelines. The latest posts were collected from 1,573 X users using X API algorithms, and each user's timeline was processed to identify the most relevant topics using a pre-trained LDA model.

Results: The effectiveness of the classification model was evaluated by comparing its predictions with manually classified user profiles. Social media analysts categorized profiles into eight pre-defined categories: Sports, Entertainment, Politics, Business, Technology, Gamers, Science, and Marketing. This manual classification served as the ground truth for comparison. The classification process achieved an average accuracy of 75.62% and an average recall of 66.31%, indicating that the model was relatively accurate in predicting users' interests on X. These metrics demonstrate that the combination of Word2Vec and LDA embeddings was effective in identifying and classifying topics of interest.

Challenges: One of the main challenges was collecting and pre-processing data from Reddit and X, ensuring the representativeness of the training data and the generalizability of the model to diverse user groups and topics was also challenging. Manual categorization of user profiles by social media analysts resulted in potential inconsistencies, which could affect the accuracy of the evaluation process.

3.4 Conclusion

In this chapter, we explored advanced methodologies and findings in social media analytics and user profiling, with a focus on X user interest profiling. The studies reviewed used different techniques, from making use of Wikipedia data and

Study	Methods Used	Accuracy of Results	Shortcomings
Study 1	- Framework using Wikipedia for classifying X users' interests. - Pre-defined library of categories and keywords. - Weights assigned based on followed celebrities	- High success rate of 83.9% in classifying celebrities. - Effective user interest identification	- Ambiguous celebrity names. - Incomplete Wikipedia articles
Study 2	- Systematic classification using posts, profile attributes, and behavior. - Naive Bayesian classifier	- High accuracy rates (91.0% to 98.7%)	- Abbreviations, slang, and misspellings in posts. - Ensuring data accuracy and reliability. - Limitations of current classification algorithms
Study 3	- Support Vector Machines, Naive Bayes, Decision Trees, K-Nearest Neighbors, and Logistic Regression classifiers. - Principal Component Analysis (PCA) to reduce feature dimensions.	- Highest classification accuracy of 89.04% with SVM and full feature set. - PCA reduced features increased accuracy to 89.71%.	-Challenges in distinguishing between Journalism and Entrepreneurship. - Misclassification due to similarity in user tweet content .
Study 4	- Preprocessing steps. - Bag-of-words model with Bigram features. - Multinomial Naïve Bayes classifier	- High accuracy and Precision of 91% and 80% for F-measure	- Dynamic nature of X data. - API rate limits. - Ensuring quality and reliability of manually labeled data
Study 5	- Collection of Reddit posts using extraction techniques. - Word2Vec embeddings and LDA for topic identification	- Average accuracy of 75.62%. - Average recall of 66.31%	- Challenges in data collection and pre-processing. - Ensuring representativeness of training data. - Inconsistencies in manual categorization

Table 3.1: Comparison between studies

naive Bayesian classifiers to using social graphs and advanced text classification models such as Word2Vec and LDA. The results focus on the high accuracy of

user classification and the effectiveness of combining multiple data sources and techniques. With all this there are challenges such as data quality, dynamic content, algorithm limitations, and privacy concerns. Although significant progress has been made, continuous innovation is necessary to enhance methodologies, address current challenges, and leverage social media analytics for more personalized and engaging user experiences.

Chapter 4

Methodology

4.1 Introduction

This chapter outlines the methodology used to classify user interests on X. It includes a description of the datasets, preprocessing steps, feature extraction techniques, model selection criteria, training methodology, and the evaluation metrics used in this study.

4.2 Research Design

This study uses a comparative approach to evaluate the effectiveness of different machine learning and deep learning models in classifying posts. The primary research questions are:

- Which machine learning or deep learning model performs best for classifying posts based on user interests?
- What are the strengths and weaknesses of each model in terms of accuracy, interpretability, and computational efficiency?

4.3 Description of the Datasets

We collected a total of 17 datasets from various sources, mainly Kaggle, covering a wide range of topics and interests. Each dataset falls into one of seven categories: Business and Finance, Health and Fitness, Movies and TV Shows, Politics, Sports, Technology, and Travel.

The datasets vary in size and content, including examples like Financial posts, Covid-19 X Dataset, Avengers, BreakingNews, FIFA World Cup, and TravelTuesday. We combined these individual datasets into one comprehensive dataset, keeping four key columns: post_id, post_text, time, and is_reposted. These columns provide essential information for analyzing user interests based on their post content and interactions.

Dataset	Subject	Size
Financial posts	Finance	28264
IndiaWantsCrypto	Finance	36928
Covid-19 X Dataset	Health	115000
Mental-Health-X	Health	20000
Avengers	Movies	15000
movie_post	Movies	100000
Squid Game	Movies	35000
ThorRangarok X Data	Movies	2608
BreakingNews	Politics	33158
US_news_posters	Politics	120000
FIFA world cup	Sports	130000
postsChampions	Sports	25028
ChatGpt_dataset	Science	50000
Data science dataset	Science	50000
Phone dataset	Science	50000
TravelTuesday	Travel	4054
US Airline dataset	Travel	14640

Table 4.1: Dataset Overview

Though there should be more than just seven interest categories for more effective targeted analysis, we opted for these broader categories for several reasons:

1. **Data Limitations**: The dataset does not have enough examples for specific sub-interests.
2. **Relevance**: The selected categories represent significant user engagement on Twitter.
3. **Comparative Focus**: Keeping categories consistent helps isolate model performance differences.

4.4 Data Preprocessing Steps

The dataset went through several preprocessing steps to ensure consistency and quality. We unified columns from the individual datasets and reduced the size of larger datasets to match the dimensions of smaller ones. Additionally, we performed the following text data preprocessing tasks:

- **Dataset Loading and Labeling**: Datasets were loaded and labeled according to their respective interest categories. This step is crucial for supervised learning, where the model learns from labeled data.
- **Dataset Merging**: Merging multiple datasets into a single Data Frame consolidates the data for easier management and analysis. Shuffling the merged dataset ensures randomness, preventing biases during model training and evaluation.

- **Handling Missing Values**: Removing rows with missing values maintains data integrity and prevents issues such as biased analysis or erroneous model training.
- **Duplicate Removal**: Eliminating duplicate rows reduces redundancy, enhances data quality, and prevents skewed analysis results or overfitting.
- **Text Data Preprocessing**: Transforming raw text into a cleaner format suitable for analysis. This includes lowercasing text, removing punctuation and special characters, removing stop words, and standardizing text format.
- **Tokenization**: Breaking down text data into individual tokens, typically words or sub words, converting unstructured text into a structured format for analysis.
- **Tokenization and Padding**: For deep learning models, text data was tokenized into numerical sequences and padded to ensure uniform length.
  - **Tokenization**: Converts text into numerical sequences.
  - **Padding**: Pads sequence to a fixed length, ensuring uniform input size for neural networks.
- **Label Encoding**: Converting categorical labels into numeric format, allowing machine learning algorithms to process and learn from the data. Each category was assigned a unique integer:
  - 0: Business and Finance
  - 1: Health and Fitness
  - 2: Movies and TV Shows
  - 3: Politics
  - 4: Sports
  - 5: Technology
  - 6: Travel

Train/Test Split: Splitting the dataset into training and testing sets allows for independent model training and evaluation. The training set had 582,092 rows, and the testing set had 145,524 rows, totaling 727,616 rows after preprocessing.

4.5 Feature Extraction

Feature extraction is crucial in text classification, transforming raw text data into a format suitable for machine learning algorithms. We a popular methods for feature extraction: Word2Vec. The dataset size in this step was (727,616, 100) using Word2Vec.

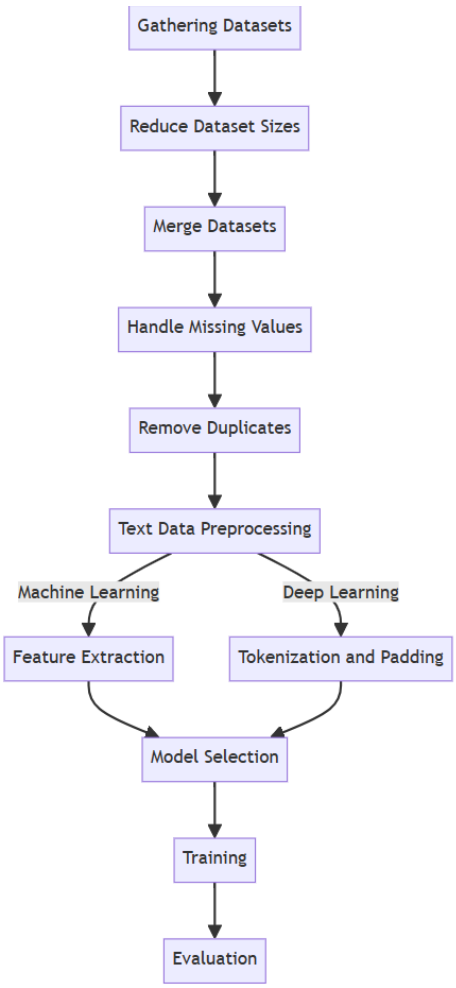

Figure 4.1: Dataset preprocessing Flowchart

4.6 Model Selection Criteria and Justification

The selection of models for user interest classification was based on their performance in text classification tasks, interpretability, and ability to handle various aspects of post data:

4.6.1 Machine Learning Algorithms

Random Forest (RF)

- **Criteria**: Robustness and ability to handle many input features.
- **Justification**: Excellent performance in classification tasks and interpretability. Provides insights into feature importance.

Logistic Regression (LR)

- **Criteria**: Simplicity, interpretability, and effectiveness on high-dimensional data.
- **Justification**: Serves as a strong baseline despite being a linear model.

Naive Bayes (NB)

- **Criteria**: Computational efficiency due to feature independence assumption.
- **Justification**: Suited for text classification with high-dimensional data and robustness to noisy data.

4.6.2 Deep Learning Algorithms:

Convolutional Neural Networks (CNN)

- **Criteria**: Extracts local features and patterns in text data.
- **Justification**: Captures spatial hierarchies in text, suitable for extracting relevant post features.

Recurrent Neural Networks (RNN) Combined with CNN (RNN-CNN)

- **Criteria**: Combines CNNs for feature extraction and RNNs for sequence modeling.
- **Justification**: Captures local features and long-term dependencies in text data.

Bidirectional Long Short-Term Memory (Bi-LSTM)

- **Criteria**: Captures context from both past and future tokens.
- **Justification**: Superior ability to model sequential data and retain long sequences' context.

4.7 Evaluation Metrics

Evaluation metrics provide insights into model performance on unseen data and help compare different models. The metrics used include:

4.7.1 Accuracy

Accuracy measures the proportion of correct predictions out of the total predictions made. It is a straightforward metric that is easy to understand but can be misleading if the data is imbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

4.7.2 Precision

Precision is the ratio of true positive predictions to the total positive predictions (both true positives and false positives). It indicates how many of the predicted positive instances are positive. High precision means that the model has a low false positive rate.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

4.7.3 Recall

Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total actual positives (both true positives and false negatives). It measures the model's ability to identify all relevant instances.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

4.7.4 F1-Score

The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, especially useful when the data is imbalanced. A higher F1-score indicates a better balance between precision and recall.

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4.4}$$

In these equations: - TP stands for True Positives. - TN stands for True Negatives. - FP stands for False Positives. - FN stands for False Negatives.

4.7.5 Confusion Matrix

The confusion matrix is a table used to describe the performance of a classification model. It shows the true positives, true negatives, false positives, and false negatives. This matrix helps in understanding the types of errors the model is making.

4.7.6 Classification Report

A classification report provides a detailed breakdown of precision, recall, and F1-score for each class. It is useful for multi-class classification problems where performance across different classes needs to be evaluated.

4.7.7 Learning Curves

Learning curves plot the training and validation accuracy and loss over epochs. These curves help diagnose whether the model is overfitting or underfitting:

- Overfitting: High training accuracy but low validation accuracy.
- Underfitting: Low accuracy in both training and validation.

By analyzing these evaluation metrics, we can gain a comprehensive understanding of each model's performance and identify areas for improvement. They are essential tools for making informed decisions when choosing the best model for a given task.

Metric	**Description**
Accuracy	Proportion of correctly classified instances.
Precision	Proportion of true positive instances.
Recall	Proportion of true positive instances out of actual positive instances.
F1 Score	Harmonic mean of precision and recall.

Table 4.2: Description of Performance Metrics

4.8 Conclusion

This chapter provides a comprehensive overview of the methodology, including research design, data collection, preprocessing, feature extraction, model selection criteria, and evaluation metrics. This framework ensures a systematic and rigorous approach to comparing the performance of various machine learning and deep learning models for post classification.

Chapter 5

Experiment and Implementation

5.1 Introduction

This chapter provides a detailed overview of the experimental setup and implementation processes used in this study to classify user interests on X. The sections cover the dataset, preprocessing techniques, model architectures, training methodologies, and evaluation metrics, offering a comprehensive view of the procedures and rationale behind the chosen methods.

5.2 Working Environment

The development and evaluation of the models were carried out in the following environment:

- Operating System: Windows 10 Professional
- Processor: Intel Core i5-9300H CPU @ 2.40GHz
- RAM: 12 GB
- GPU: NVIDIA GeForce GTX 1650
- IDE: Jupyter Notebook

5.3 Programming Languages, Libraries, and Frameworks

The implementation and evaluation of the deep learning models were performed using the following programming languages, libraries, and frameworks:

- Programming Language:
  - Python: Chosen for its simplicity and extensive support for scientific computing.

- Libraries and Frameworks:
  - NumPy: For numerical computations and array handling.
  - Pandas: For data manipulation and analysis.
  - Scikit-Learn: For data preprocessing, model evaluation, and metrics calculation.
  - TensorFlow and Keras: For building, training, and evaluating deep learning models.
  - NLTK (Natural Language Toolkit): For text preprocessing and tokenization.
  - Matplotlib and Seaborn: For data visualization and plotting learning curves and ROC curves.

5.4 Dataset Description and Preprocessing

The dataset for this study comprises posts categorized into seven distinct interest areas:

- Business and Finance
- Health and Fitness
- Movies and TV Shows
- Politics
- Sports
- Technology
- Travel

After preprocessing the collected dataset as mentioned in the previous section (handling missing values, removing duplicates, text cleaning, etc.), a cleaned dataset was obtained. The used feature extraction method is Word2Vec. The data was split into training, validation, and test sets with an approximate 70-20-10 ratio, providing sufficient data for both model training and evaluation.

5.5 Model Implementation

The study employs several models, including traditional machine learning and deep learning approaches.

5.5.1 Random Forest

Description: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

Implementation Details:

- Hyper-parameters: 100 estimators, a maximum depth of 15, minimum samples split of 50, and minimum samples leaf of 40.
- Training: The model was trained using Word2Vec feature vectors extracted from the posts. Each tree in the forest was built using a different subset of the training data, and the final prediction was obtained by aggregating the predictions from all the trees.

5.5.2 Logistic Regression

Description: Logistic Regression is a linear model commonly used for binary classification tasks, extendable to multi-class classification using techniques such as one-vs-rest (OvR).

Implementation Details:

- **Hyper-parameters:** Regularization strength (C) and solver for optimization.
- **Training:** The model was trained on Word2Vec features, using regularization to prevent overfitting. The solver was chosen based on the dataset's size and complexity.

5.5.3 Naive Bayes

Description: Naive Bayes classifiers apply Bayes' theorem with strong independence assumptions between the features, resulting in fast and efficient text classification.

Implementation Details:

- **Model Variant:** Multinomial Naive Bayes, well-suited for text classification where features represent word frequencies.
- **Training:** The model was trained using word2vec features, calculating the posterior probabilities of each class given a post and assigning the class with the highest probability.

5.5.4 Convolutional Neural Network (CNN)

Description: Though CNNs were originally designed for image recognition, they work well for text classification by spotting patterns in word sequences, just

like they detect shapes and features in images.

Architecture:

- **Embedding Layer:** Convert the words in the text into numerical representations (word vectors), so the CNN can work with them using pre-trained embeddings models like Word2Vec.
- **Convolutional Layers:** Apply multiple filters to capture different patterns and n-grams in the text.
- **Pooling Layers:** Perform max pooling to reduce spatial dimensions and retain significant features.
- **Fully Connected Layers:** Output final classification probabilities.

Training: The CNN model was trained using word embeddings, optimized with the Adam optimizer. Early stopping and dropout were used to mitigate overfitting, ensuring robust model performance.

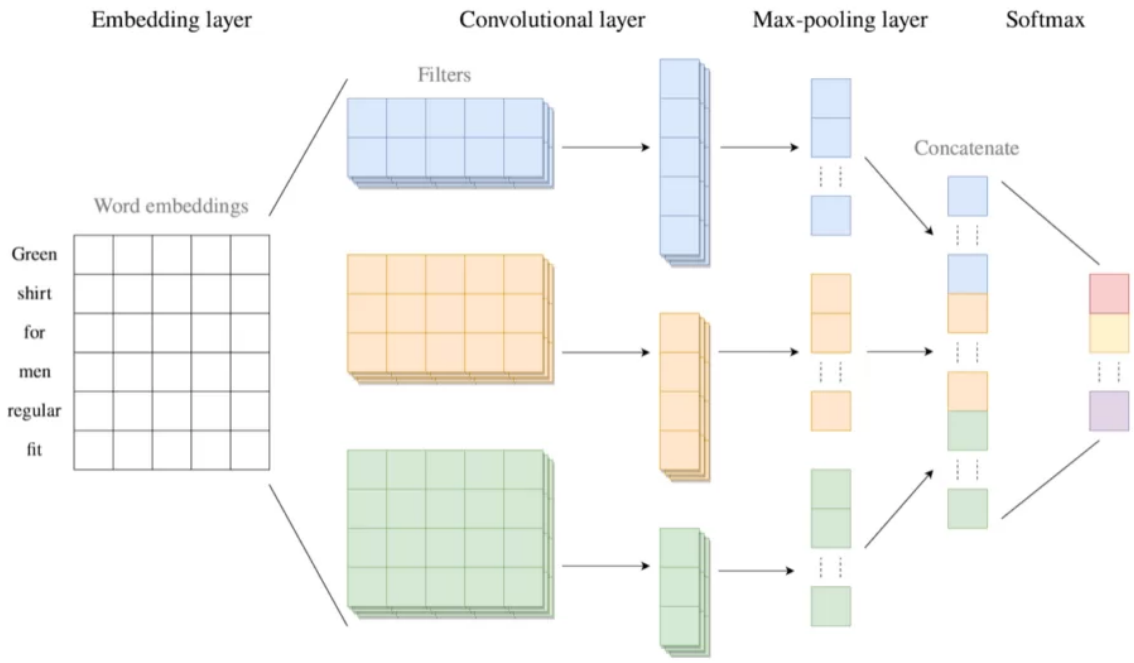

Figure 5.1: CNN Architecture for text classification

5.5.5 RNN Combined with CNN (RNN + CNN)

Description: This hybrid model combines the strengths of CNNs and RNNs to capture both local patterns and sequential dependencies in the text.

Architecture:

- **Convolutional Layers:** Extract local patterns from the input text.
- **Pooling Layer:** Perform max pooling to reduce spatial dimensions and retain significant features.
- **RNN Layer:** The pooled features are fed into an RNN to capture the sequential dependencies in the data.
- **Fully Connected Layers:** Output final classification probabilities.

Training: The model was trained using word embeddings, optimized with the Adam optimizer. Early stopping and dropout were employed to prevent overfitting.

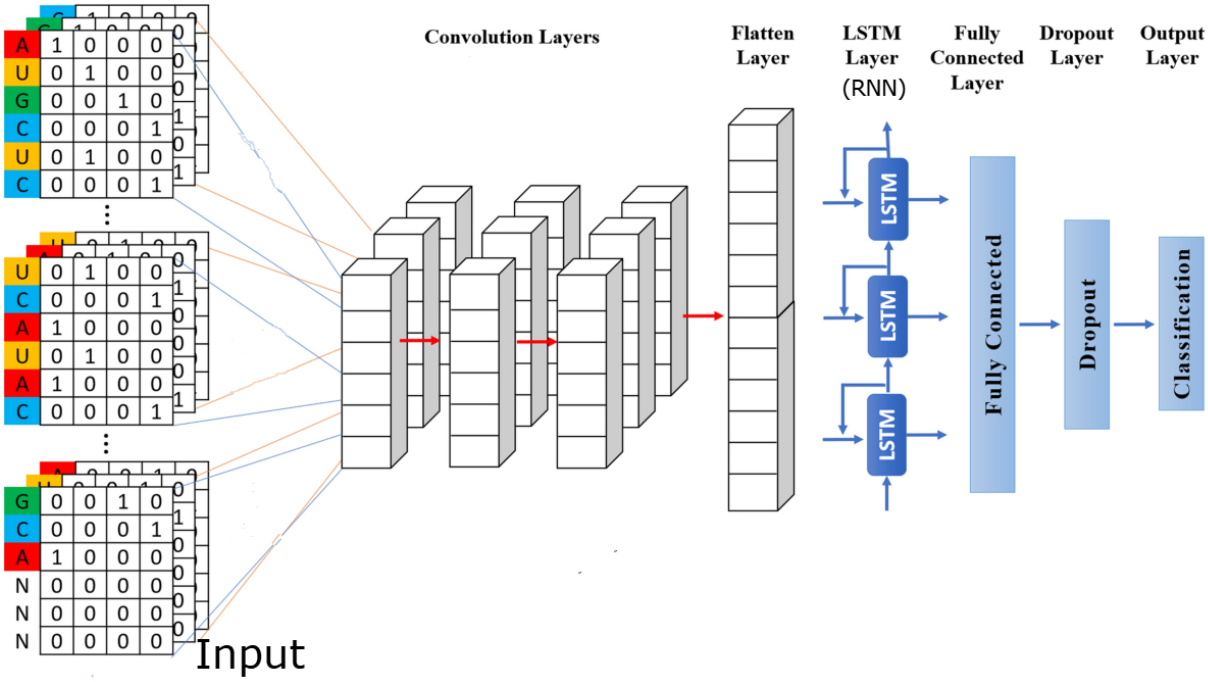

Figure 5.2: CNN + RNN Architecture for text classification

5.5.6 Bidirectional LSTM (Bi-LSTM)

Description: Bi-LSTMs capture contextual information from both forward and backward directions, making them highly effective for understanding context in sequences.

Architecture:

- **Embedding Layer:** Converts input text into dense word vectors using pre-trained embeddings.
- **Bidirectional LSTM Layers:** Capture dependencies in both directions.
- **Fully Connected Layers:** Output final classification probabilities.

Training: The model was trained using word embeddings, optimized with the Adam optimizer. Early stopping and dropout were used to mitigate overfitting.

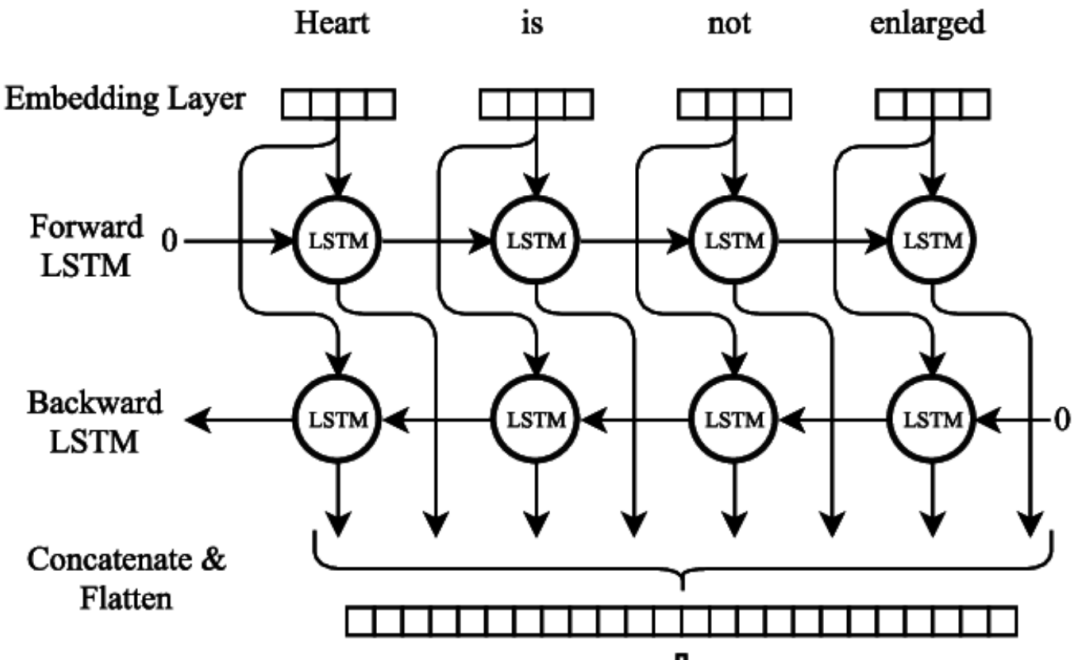

Figure 5.3: Bi-LSTM Architecture for text classification

5.6 Model Training and Evaluation

5.6.1 Training Process

Each model underwent a comprehensive training process:

- **Hyperparameter Tuning:** Techniques such as grid search or random search were employed to find the optimal hyperparameters for each model.
- **Optimization:** For deep learning models, the Adam optimizer was used to minimize the loss function, ensuring efficient convergence.

- **Early Stopping:** The training process was monitored for validation loss, and training was halted when performance stopped improving, preventing overfitting.

5.6.2 Evaluation Metrics

To evaluate the models, several metrics were used:

- **Accuracy:** Proportion of correct predictions out of total predictions.
- **Precision:** Ratio of true positive predictions to total positive predictions.
- **Recall:** Ratio of true positive predictions to total actual positives.
- **F1 Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Table showing true positives, true negatives, false positives, and false negatives.
- **Classification Report:** Detailed breakdown of precision, recall, and F1-score for each class.
- **Learning Curves:** Plots of training and validation accuracy and loss over epochs, diagnosing overfitting or underfitting.

5.7 Conclusion

This chapter provided a comprehensive overview of the experimental setup and implementation processes used in the study to classify user interests on X. It detailed the working environment, including the hardware and software specifications, and outlined the programming languages, libraries, and frameworks employed for the implementation and evaluation of the models.

The chapter described the dataset, including the categories of posts and the preprocessing techniques applied to clean and prepare the data. It also highlighted the feature extraction methods used(Word2Vex), and explained the data splitting strategy for training, validation, and testing.

Various models were implemented, ranging from traditional machine learning techniques like Random Forest, Logistic Regression, and Naive Bayes, to advanced deep learning approaches such as CNN, a hybrid RNN combined with CNN, and Bi-LSTM. For each model, specific implementation details, including hyperparameters and training procedures, were discussed. The training process involved hyperparameter tuning, optimization with the Adam optimizer, and the use of early stopping to prevent overfitting.

To evaluate the models, several metrics were used, including accuracy, precision, recall, F1 score, confusion matrix, and classification report. Learning curves were also plotted to diagnose potential overfitting or underfitting issues.

Chapter 6

Results

6.1 Introduction

This chapter presents the results of our experiments, providing a detailed comparative analysis of the different models used to classify user interests on X. Key metrics, execution times, and practical applicability of each model are discussed, highlighting their strengths and weaknesses. This analysis aims to guide the selection of the most suitable model for this task.

6.2 Dataset

The dataset now (after preprocessing) contains a total of 727616 rows. The distribution of each interest in the dataset along with the representation of world cloud for each interest is showed below:

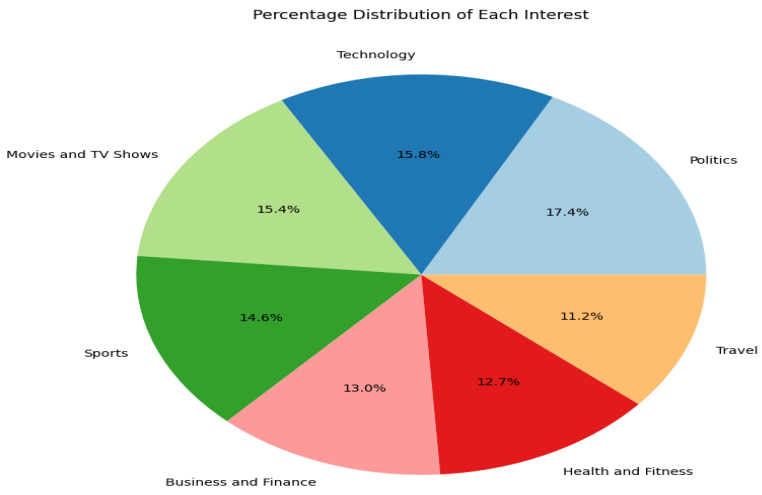

Figure 6.1: Dataset distribution

To visualize common words for each interest we represent Word Clouds:

Figure 6.2: Word cloud for Business interest

Figure 6.3: Word cloud for Health and fitness interest

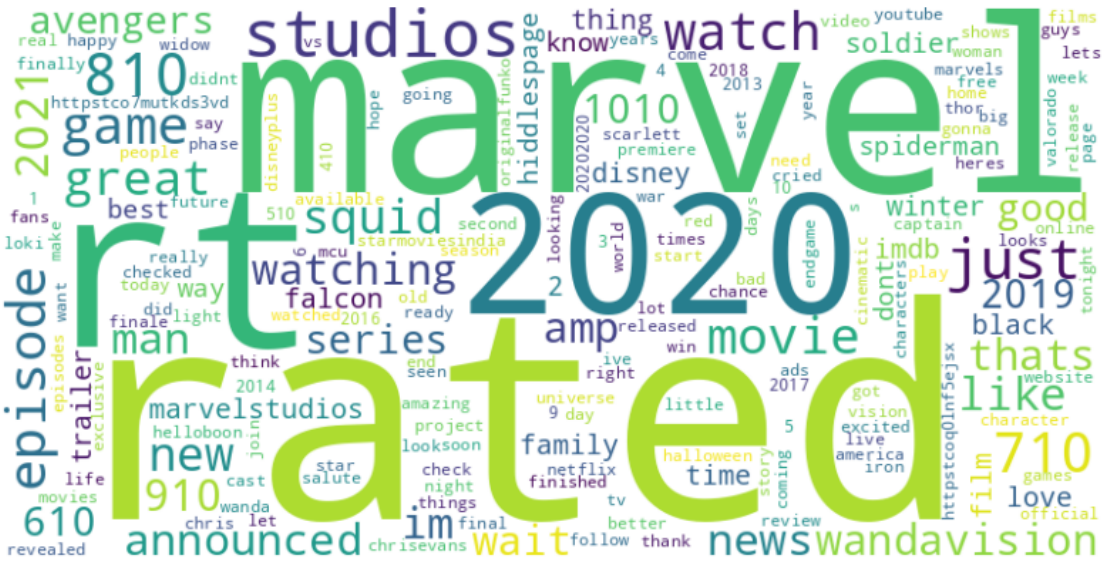

Figure 6.4: Word cloud for Movies and Tv shows interest

Figure 6.5: Word cloud for News interest

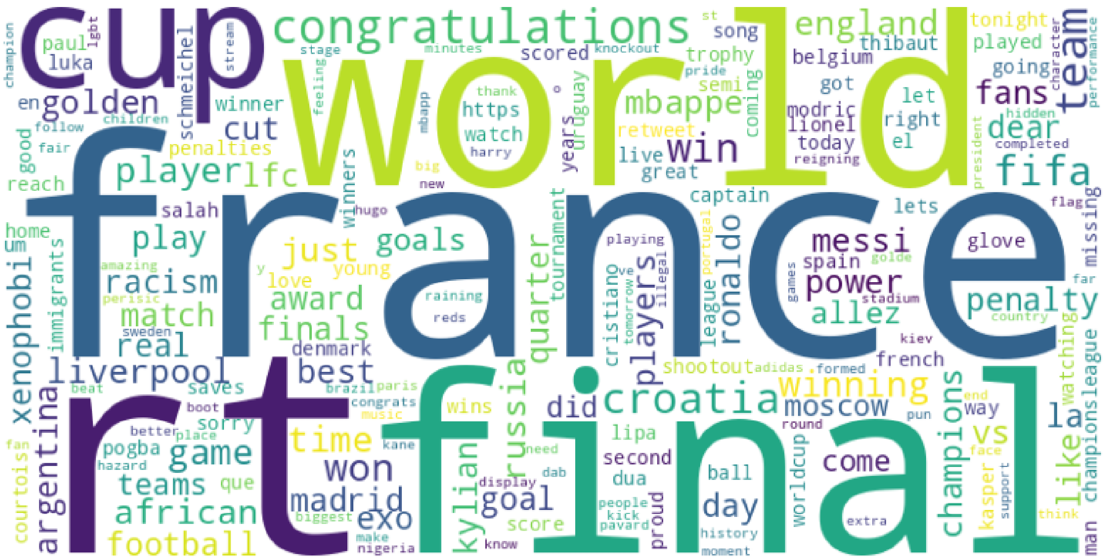

Figure 6.6: Word cloud for Sports interest

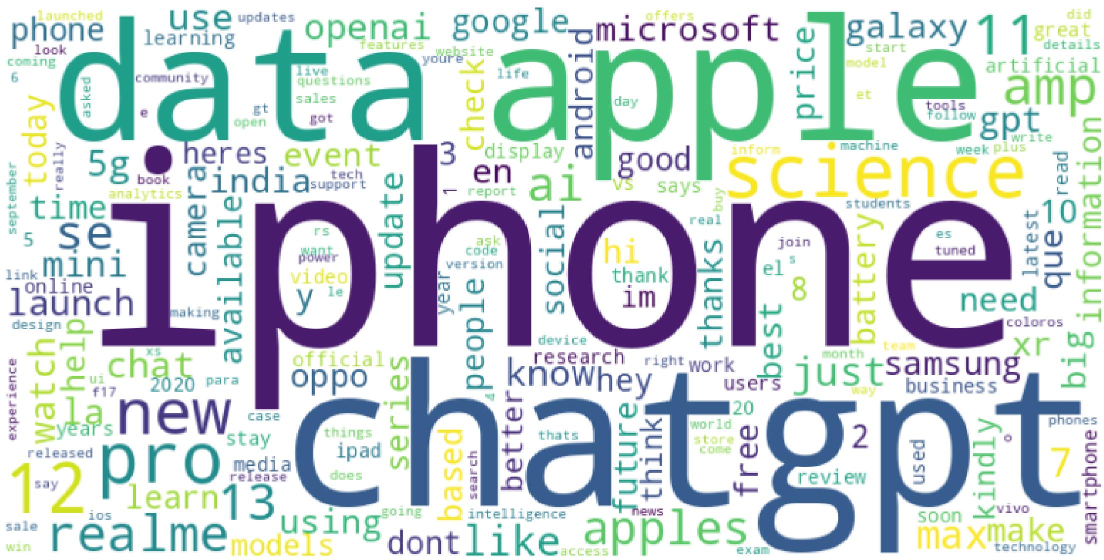

Figure 6.7: Word cloud for Technology interest

Figure 6.8: Word cloud for Travel interest

6.3 Models result

After running the algorithms:

- Random Forest
- Logisric Regression
- Naive Bayes
- CNN
- RNN + CNN
- Bi-LSTM

the evaluation metrics: Accuracy, Precision, recall, and F1-Score were used to evaluate each model. Here are the values of each metrices along with the confusion matrix and the learning curve for each model:

6.3.1 Random Forest

| Accuracy: 0.9413 | Precision: 0.9424 | Recall: 0.9413 | F1 Score: 0.9416 |

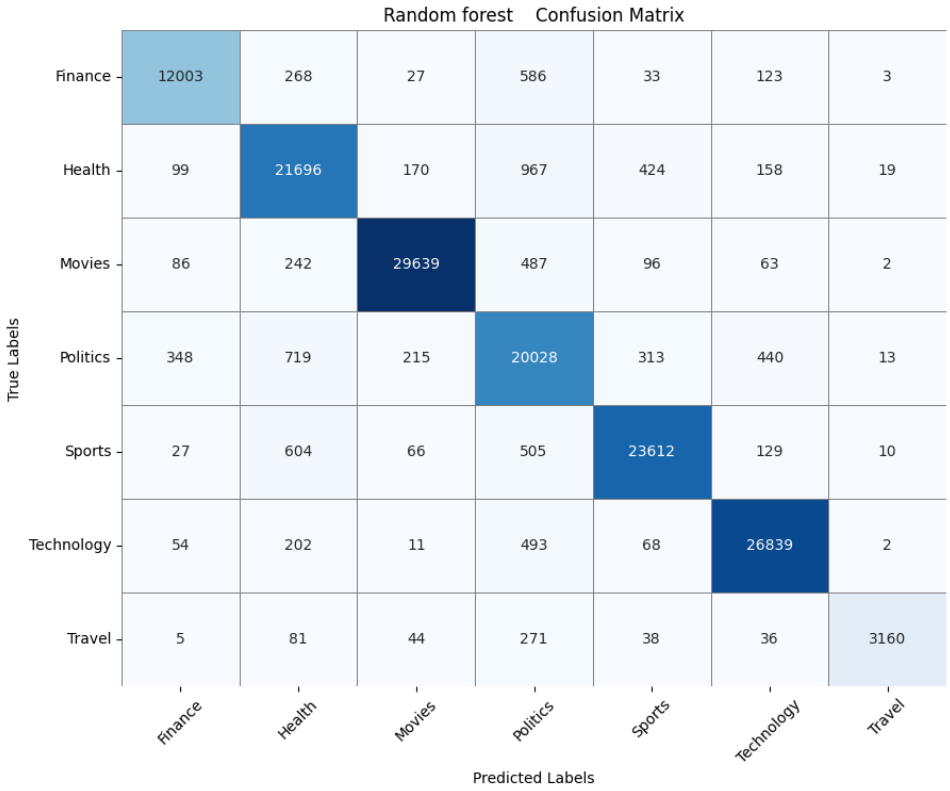

Figure 6.9: Random forest Confusion matrix

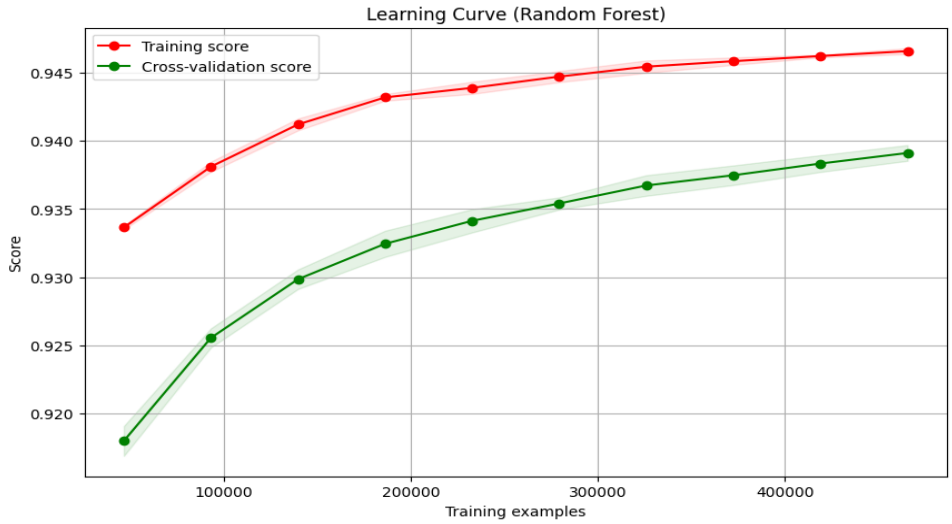

Figure 6.10: Random forest Learning curve

6.3.2 Logistic Regression

| Accuracy: 0.9319 | Precision: 0.9325 | Recall: 0.9319 | F1 Score: 0.9321 |

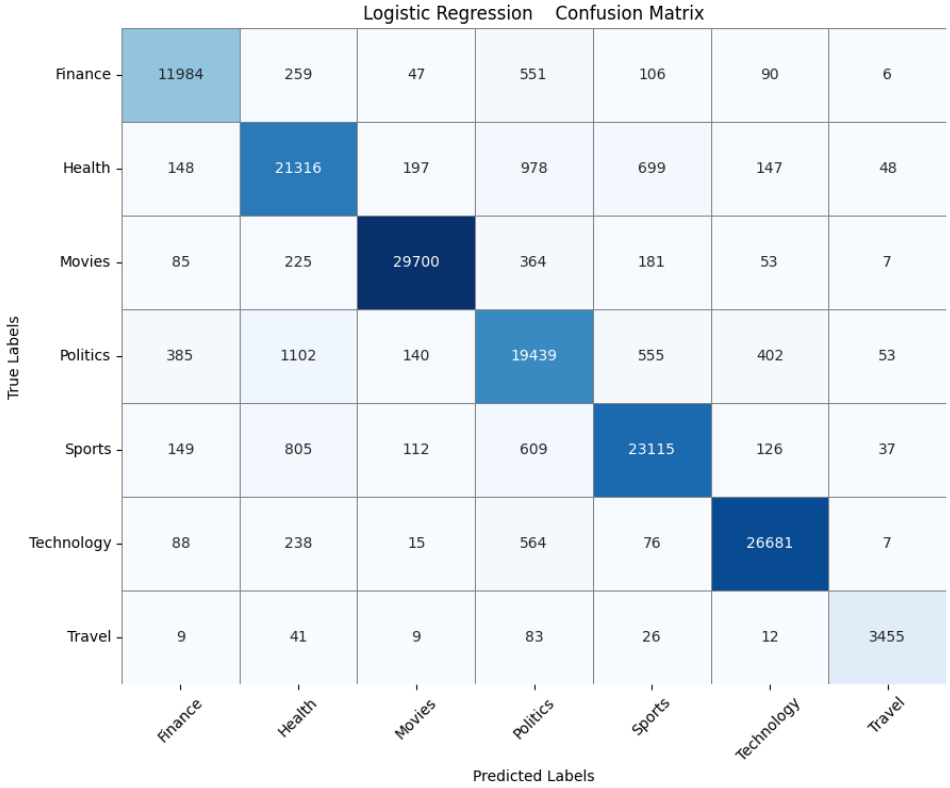

Figure 6.11: Logistic regression Confusion matrix

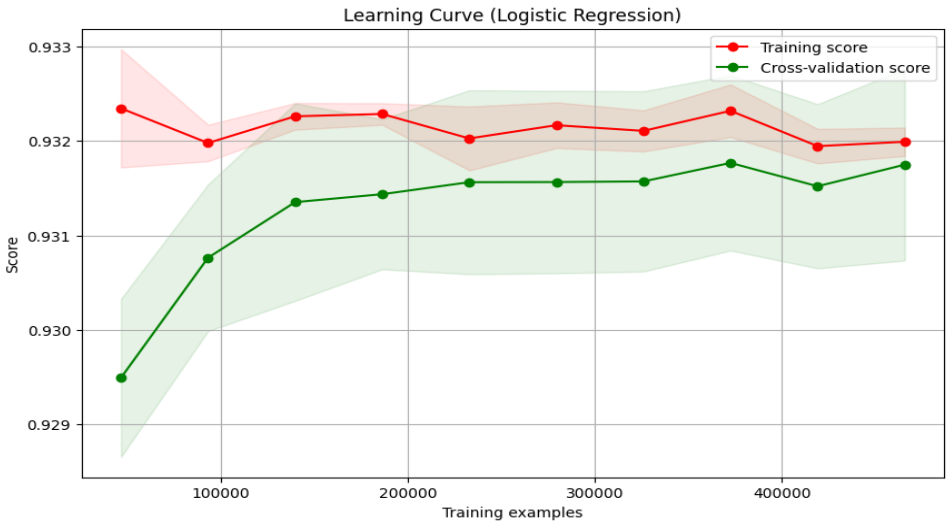

Figure 6.12: Logistic regression Learning curve

6.3.3 Naive Bayes

| Accuracy: 0.9296 | Precision: 0.9314 | Recall: 0.9296 | F1 Score: 0.9291 |

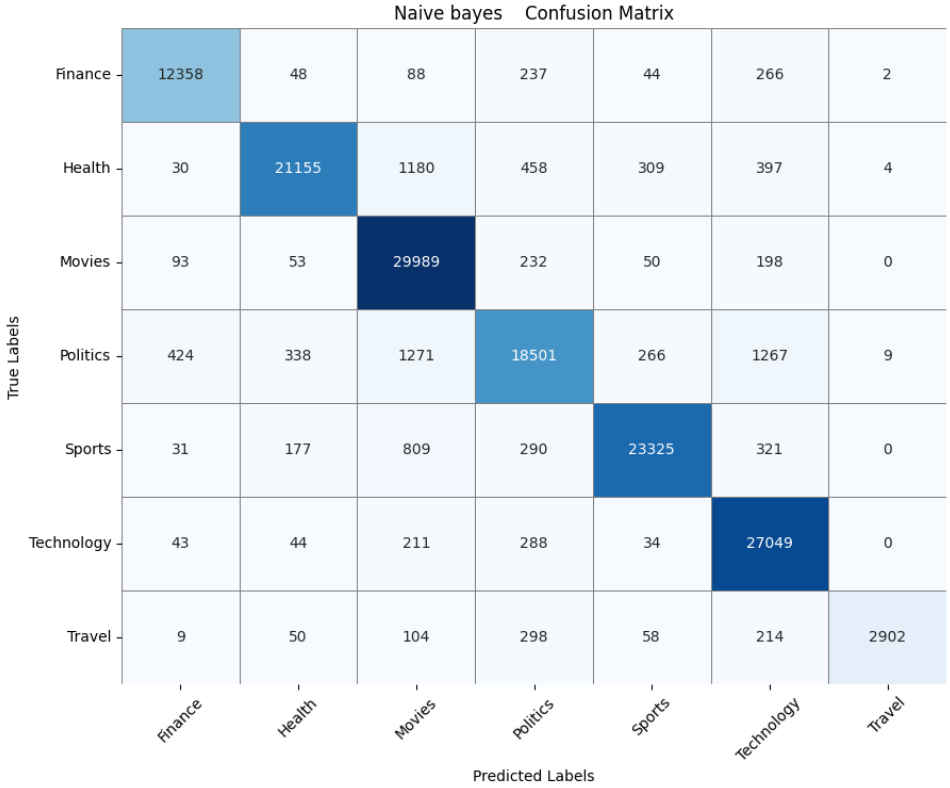

Figure 6.13: Naive Bayes Confusion matrix

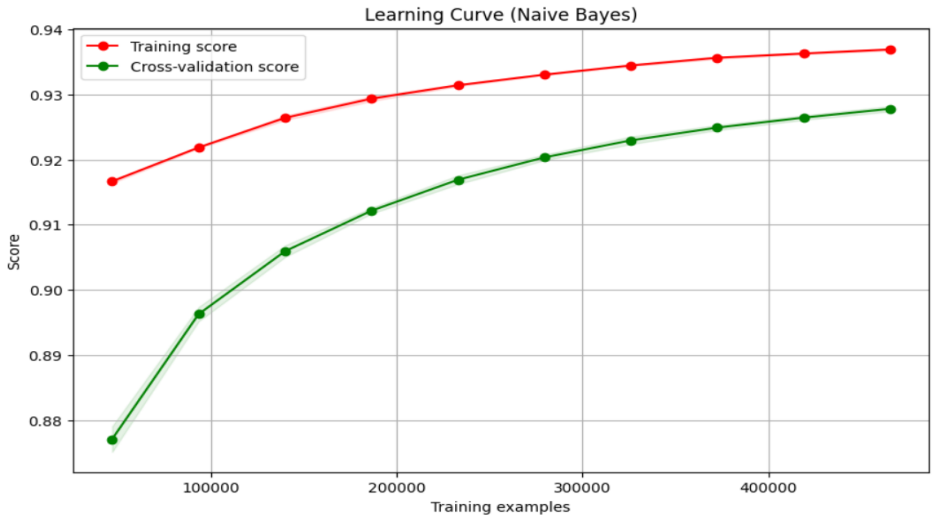

Figure 6.14: Naive Bayes Learning curve

6.3.4 Convolutional Neural Network (CNN)

| Accuracy: 0.9156 | Precision: 0.9193 4 | Recall: 0.9156 | F1 Score: 0.9160 |

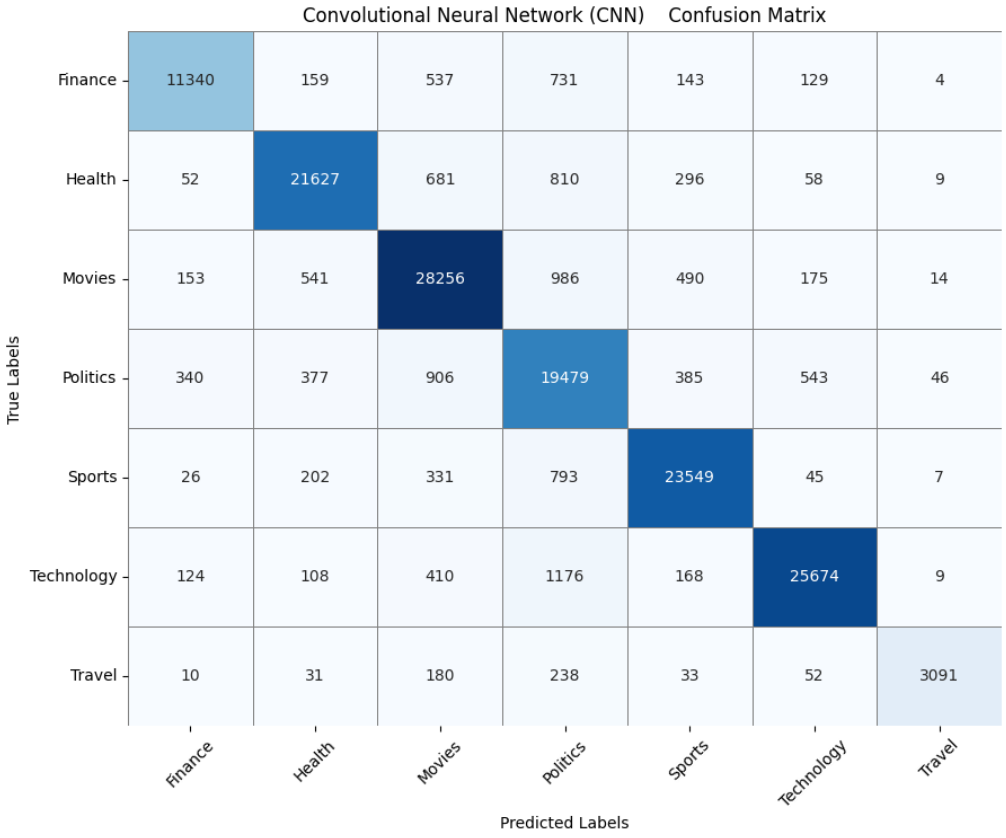

Figure 6.15: CNN Confusion matrix

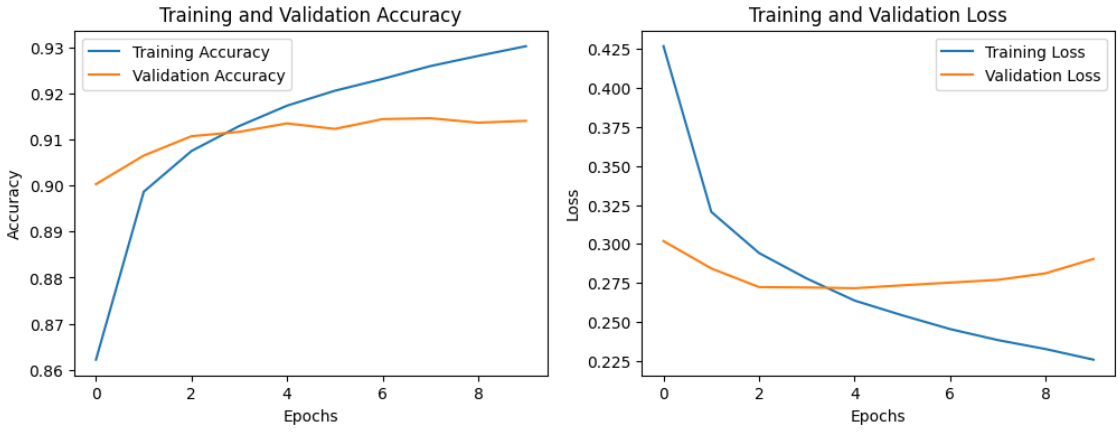

Figure 6.16: CNN Learning curve

6.3.5 RNN Combined with CNN (RNN + CNN)

| Accuracy: 0.9182 | Precision: 0.9217 | Recall: 0.9182 | F1 Score: 0.9185 |

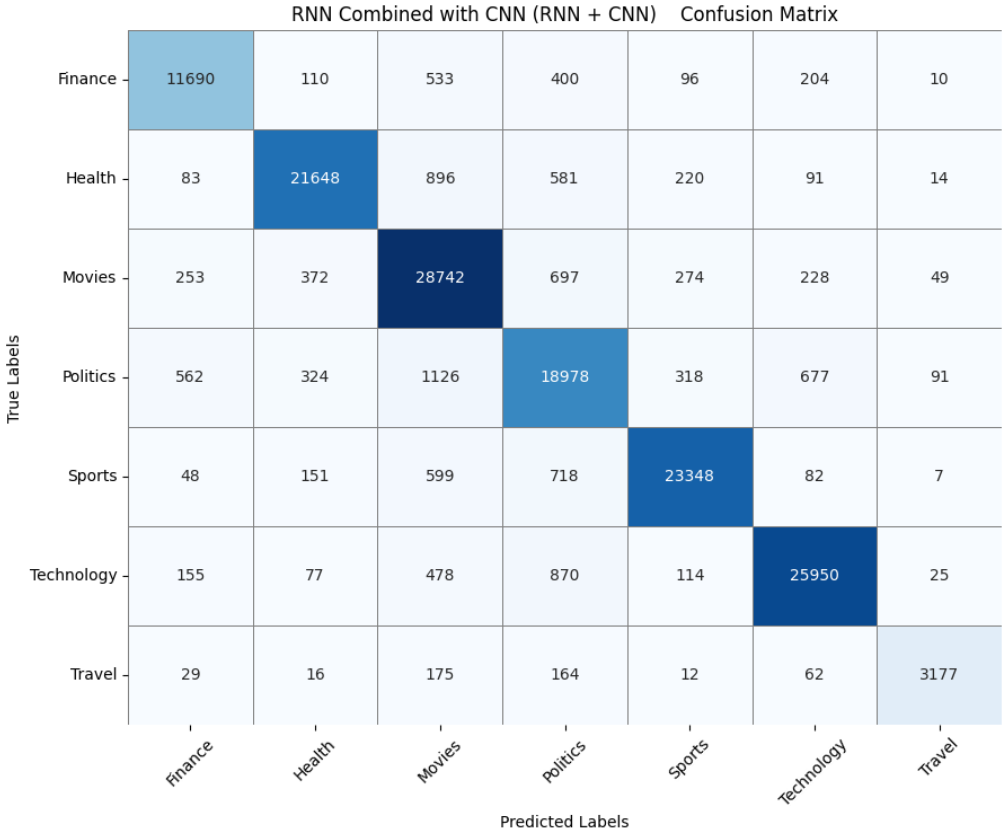

Figure 6.17: RNN + CNN Confusion matrix

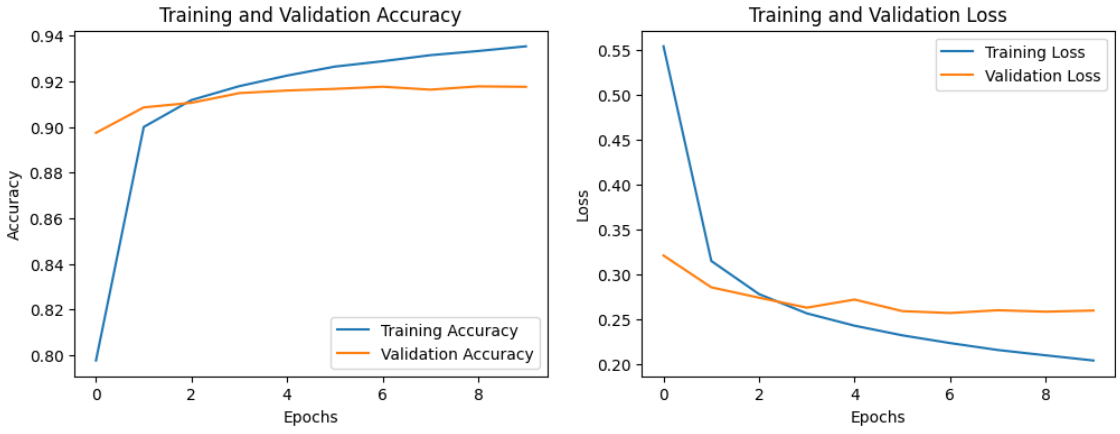

Figure 6.18: RNN + CNN Learning curve

6.3.6 Bidirectional LSTM (Bi-LSTM)

| Accuracy: 0.9254 | Precision: 0.9265 | Recall: 0.9254 | F1 Score: 0.9257 |

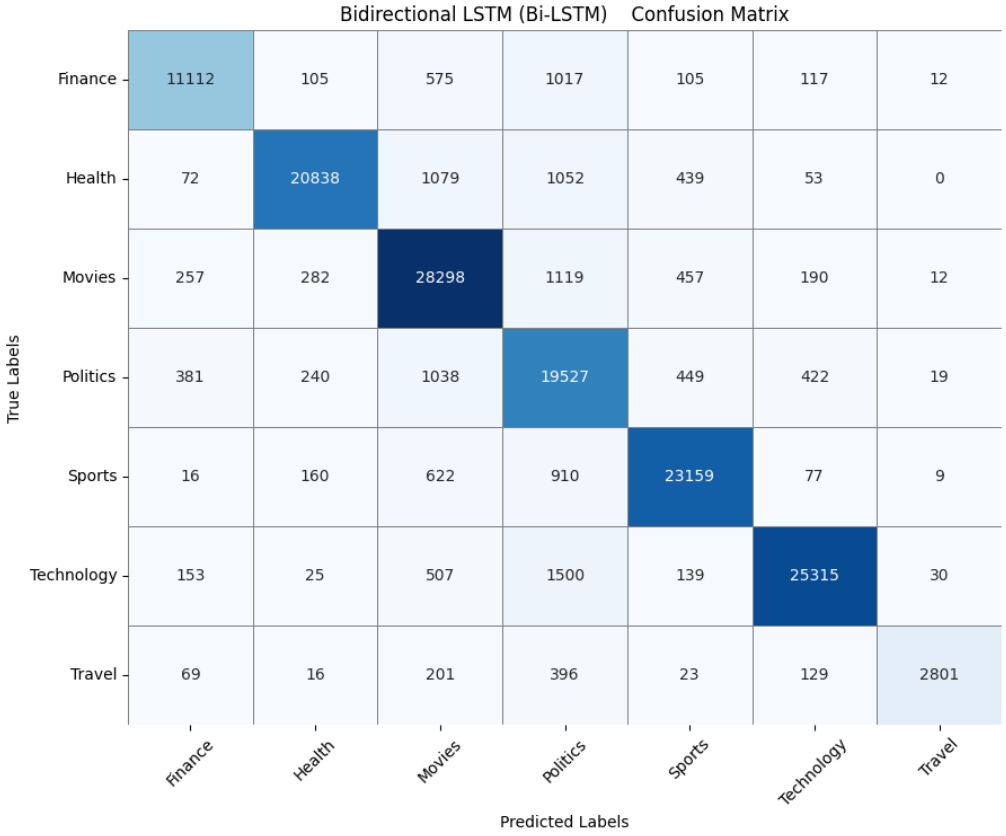

Figure 6.19: Bi-LSTM Confusion matrix

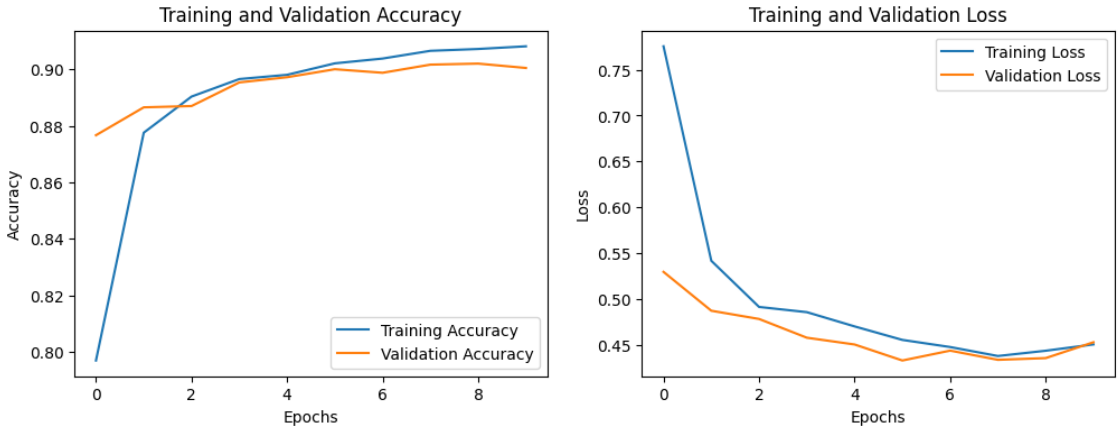

Figure 6.20: Bi-LSTM Learning curve

6.4 Comparative Analysis

This section compares the models based on performance metrics, execution time, and practical applicability. Each model's strengths and weaknesses are evaluated to help determine the most appropriate model for classifying user interests on X.

6.4.1 Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	94.13%	94.24%	94.13%	94.16%
Logistic Reg	93.19%	93.25%	93.19%	93.21%
Naive Bayes	92.96%	93.14%	92.96%	92.91%
CNN	91.56%	91.93%	91.56%	91.60%
RNN & CNN	91.82%	92.17%	91.82%	91.85%
Bi-LSTM	92.54%	92.65%	92.54%	92.57%

Table 6.1: Performance Metrics of Different Models

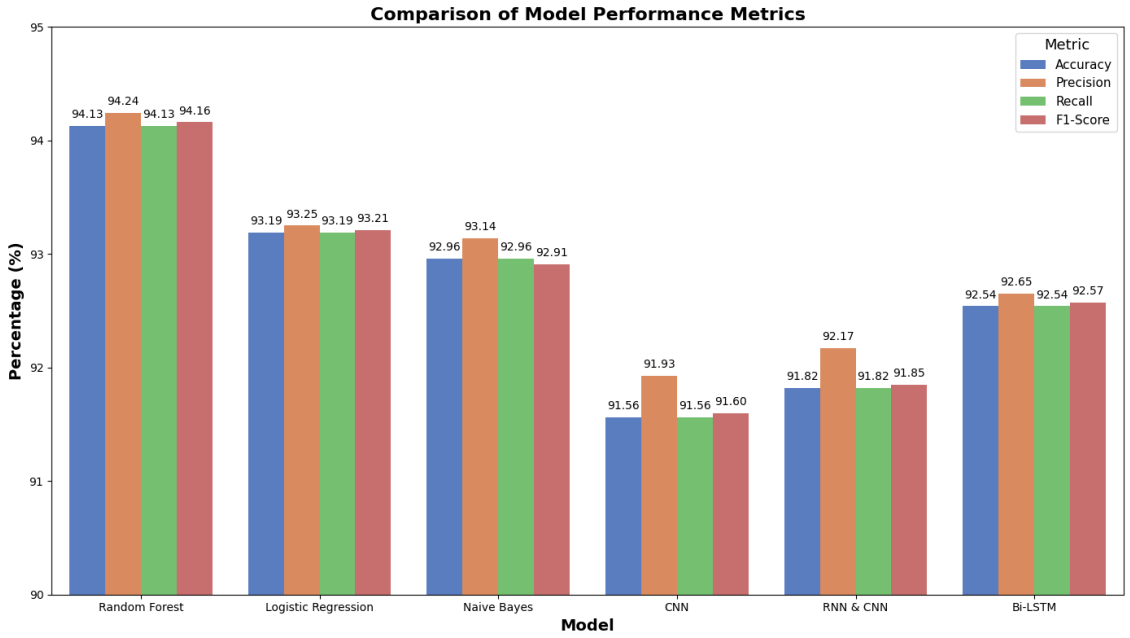

Figure 6.21: Comparative of model performance

Accuracy:

- **Random Forest:** Achieved the highest accuracy at 94.13%, indicating strong performance across various categories.

- **Logistic Regression and Naive Bayes:** Followed closely with accuracies of 93.19% and 92.96%, respectively, showing their effectiveness in linear and probabilistic classification tasks.

- **Deep Learning Models (CNN and RNN + CNN):** Displayed slightly lower accuracies (91.56% and 91.82%), but their ability to capture complex patterns and sequential dependencies makes them valuable for nuanced text classification.

- **Bi-LSTM:** Achieved an accuracy of 92.54%, demonstrating its strength in understanding contextual information, though slightly less accurate than Random Forest.

Precision, Recall, and F1 Score:

- **Random Forest:** Exhibited the highest precision (0.9424) and recall (0.9413), leading to the best F1 score (0.9416) among all models.

- **Logistic Regression and Naive Bayes:** Showed comparable precision and recall, with slight variations in F1 scores (0.9321 and 0.9291, respectively), making them reliable choices for efficient and interpretable models.

- **CNN and RNN + CNN:** Had similar precision (0.9193 and 0.9217) and recall (0.9156 and 0.9182), indicating their capability to handle large and complex datasets.

- **Bi-LSTM:** Displayed good precision (0.9265) and recall (0.9254), with an F1 score (0.9257), effectively capturing bidirectional context.

Confusion Matrix Insights

- **Random Forest and Logistic Regression:** Showed minimal misclassifications across categories, particularly excelling in Business and Finance, and Technology.

- **Naive Bayes:** Performed well but had more misclassifications in categories with overlapping vocabulary, like Health and Fitness, and Politics.

- **CNN and RNN + CNN:** Effective in capturing local and sequential patterns but showed some misclassifications in Technology and Sports.

- **Bi-LSTM:** Had issues with categories having less training data but excelled in categories like Movies and TV Shows and Travel due to its contextual understanding.

Learning Curves The learning curves for each model provide insights into training dynamics and potential issues such as overfitting or underfitting.

- **Random Forest:** Learning curve indicates consistent improvement with minimal overfitting, and steady convergence as training progresses.

- **Logistic Regression:** Learning curve shows clear convergence with stable performance, indicating the model's efficiency and effectiveness in handling the dataset.

- **Naive Bayes:** Learning curve demonstrates rapid learning and quick convergence, reflecting the model's efficiency in learning from the data quickly.

- **CNN:** Learning curve shows steady learning with signs of overfitting towards later epochs, indicating the need for regularization techniques to improve generalization.

- **RNN + CNN:** Learning curve shows balanced training with consistent accuracy improvements, demonstrating effective learning over the epochs with good generalization.

- **Bi-LSTM:** Learning curve shows significant training time with eventual convergence, making it effective over extended epochs, beneficial for tasks requiring sequential data handling.

6.4.2 Execution Time and Efficiency

Model	**Execution Time (s)**	**Execution Time (s) With GPU Acceleration**	**Prediction Time (s)**
Random Forest	1690	932	0.41
Logistic Regression	286	178	0.41
Naive Bayes	0.24	0.24	0.1
CNN	10613	510	9.34
RNN & CNN	19667	1830	15.37
Bi-LSTM	4150	3830	32.69

Table 6.2: Execution and Prediction Times for Models

Random Forest:

- Execution Time: 1690 seconds (about 28 minutes), 932.05 seconds (about 15 and a half minutes) using GPU acceleration.

- Efficient in training with high accuracy but relatively slower due to the ensemble nature of the model.

Logistic Regression:

- Execution Time: 286.07 seconds (about 5 minutes), 178 seconds (about 3 minutes) using GPU acceleration.

Naive Bayes:

- Execution Time: 0.24 seconds, fastest among all models, making it suitable for applications requiring quick training and prediction.
- Extremely efficient, suitable for real-time applications with large datasets.

CNN (With GPU acceleration):

- Training Time: Approximately 51 seconds per epoch (10 epochs total).
- Prediction Time: 9.34 seconds.
- Requires more computational resources but benefits from GPU acceleration.

RNN + CNN (With GPU acceleration):

- Training Time: Approximately 183 seconds (about 3 minutes) per epoch (10 epochs total).
- Prediction Time: 15.37 seconds.
- More complex and slower than CNN alone due to the added recurrent layers but benefits from GPU acceleration.

Bi-LSTM (With GPU acceleration):

- Training Time: Approximately 383 seconds (about 6 and a half minutes) per epoch (10 epochs total).
- Prediction Time: 32.69 seconds.
- Slowest among deep learning models, reflecting its complexity and resource requirements.

6.5 Practical Applicability

- **Random Forest:** Best suited for scenarios where high accuracy and robustness are critical, and computational resources are sufficient to handle longer training times.
- **Logistic Regression:** Ideal for applications needing rapid training and predictions, offering a good balance of accuracy and simplicity.
- **Naive Bayes:** Effective for real-time applications with massive datasets due to its fast computation and simplicity.
- **CNN and RNN + CNN:** Suitable for complex text classification tasks where capturing both local and sequential patterns is essential, though they require more computational power.
- **Bi-LSTM:** Best for applications that need to understand context and long-term dependencies in text, despite its slower training and prediction times.

6.6 Challenges and Solutions

Several challenges were encountered during implementation and experimentation. This section discusses these challenges and the solutions employed to address them.

6.6.1 Data Imbalance

- **Challenge:** Imbalanced data can lead to biased models favoring the majority classes**.**
- **Solution:**
  - **Data Augmentation:** Synthetic data generation for underrepresented classes.
  - **Class Weights:** Assigning higher weights to minority classes during model training to ensure balanced learning.

6.6.2 Overfitting

- **Challenge:** Overfitting occurs when models learn noise in the training data, leading to poor generalization.
- **Solution:**
  - **Regularization:** Techniques like L2 regularization in Logistic Regression and dropout in deep learning models.
  - **Early Stopping:** Monitoring validation performance to halt training when no improvement is observed.
  - **Cross-Validation:** Using k-fold cross-validation to ensure models generalize well to unseen data.

6.6.3 Computational Resources

- **Challenge:** Deep learning models, especially Bi-LSTM and hybrid models, require significant computational power and memory.
- **Solution:**
  - **GPU Utilization**: Leveraging GPU acceleration for faster training.
  - **Efficient Batch Processing:** Using mini-batch gradient descent to optimize memory usage and speed.

6.6.4 Hyperparameter Tuning

- **Challenge:** Finding optimal hyperparameters can be time-consuming and computationally intensive.
- **Solution:**
  - **Automated Hyperparameter Tuning:** Using grid search and random search techniques to systematically explore hyperparameter spaces.
  - **Bayesian Optimization:** Employing advanced optimization techniques to efficiently find the best hyperparameters.

6.7 Conclusion

This chapter detailed the experimental setup, including the dataset, preprocessing steps, and model implementations. It also provided an in-depth comparative analysis of the performance, execution time, and practical applicability of various machine learning and deep learning models. Despite challenges like data imbalance, overfitting, and resource constraints, the models demonstrated robust performance in classifying user interests on X. The insights gained from this chapter form the basis for the subsequent discussion on the implications of the results and potential areas for future research.

The next chapter will delve into the discussion of results, exploring the implications of model performance and potential improvements for enhancing user interest classification on social media platforms.

Chapter 7

Discussion

7.1 Introduction

In this chapter, we delve into a detailed discussion of the results obtained from the results that are in the previous Chapter. We analyze each model's performance, interpret the results, and discuss the implications of these findings. Additionally, we explore potential improvements and future research directions to enhance the classification of user interests on X.

7.2 Analysis of Model Performance

We analyze the models based on key metrics such as accuracy, precision, recall, F1-score, execution time, and their suitability for practical applications.

7.2.1 Random Forest

The Random Forest model achieved the highest accuracy among all models at 94.13%. It also exhibited the highest precision (0.9424), recall (0.9413), and F1-score (0.9416). The confusion matrix revealed minimal miss-classifications, particularly excelling in categories like Business and Finance, and Technology.

- **Interpretation:** The ensemble nature of Random Forest allows it to perform robustly across diverse categories, effectively handling non-linear relationships and interactions between features. Its high precision and recall indicate its ability to minimize both false positives and false negatives.
- **Implications:** Random Forest is highly suitable for applications requiring high accuracy and robustness, such as targeted advertising and personalized content recommendations. However, its longer execution time may be a limitation in real-time applications.

7.2.2 Logistic Regression

Logistic Regression demonstrated an accuracy of 93.19%, with precision (0.9325), recall (0.9319), and F1-score (0.9321). It exhibited minimal miss-classifications and fast execution time (286.07 seconds).

- **Interpretation:** As a linear model, Logistic Regression performs well with linearly separable data. Its simplicity and interpret-ability make it a reliable choice for many applications.
- **Implications:** Logistic Regression is ideal for scenarios requiring quick training and predictions, such as real-time sentiment analysis and trend detection on social media.

7.2.3 Naive Bayes

Naive Bayes achieved an accuracy of 92.96%, with precision (0.9314), recall (0.9296), and F1-score (0.9291). It is the fastest model, showed strong performance but had more miss-classifications in overlapping categories like Health and Fitness, and Politics.

- **Interpretation:** Naive Bayes assumes feature independence, which can be a limitation in cases where features are correlated. However, its simplicity and efficiency make it a strong contender for large-scale applications.
- **Implications:** Naive Bayes is suitable for real-time applications with massive datasets, such as spam detection and email filtering, where speed and efficiency are critical.

7.2.4 Convolutional Neural Network (CNN)

The CNN model showed an accuracy of 91.56%, with precision (0.9193), recall (0.9156), and F1-score (0.9165). The confusion matrix indicated effective handling of complex patterns but some miss-classifications in Technology and Sports.

- **Interpretation**: CNNs are powerful for capturing local patterns in data, making them effective for image and text classification. Their ability to learn hierarchical features is a significant advantage.
- **Implications**: CNNs are suitable for applications involving large and complex datasets, such as image recognition and text classification, though they require significant computational resources.

7.2.5 RNN combined with CNN

The RNN + CNN hybrid model achieved an accuracy of 91.82%, with precision (0.9190), recall (0.9174), and F1-score (0.9179). It showed strong performance

in capturing both local and sequential patterns but required more computational power.

- **Interpretation**: Combining RNNs with CNNs leverages the strengths of both models, capturing sequential dependencies and local patterns effectively.
- **Implications**: RNN + CNN models are ideal for tasks requiring an understanding of context and local patterns, such as natural language processing and video analysis.

7.2.6 Bidirectional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM demonstrated an accuracy of 92.4%, with precision (0.9265), recall (0.9254), and F1-score (0.9257). It showed strong performance in understanding contextual information, the best within deep learning models but slower execution times.

- **Interpretation**: Bi-LSTMs excel in capturing long-term dependencies and contextual information, making them highly effective for sequential data.
- **Implications**: Bi-LSTMs are suitable for applications like language translation and speech recognition, where understanding context is crucial despite longer training and prediction times.

7.3 Implications of Findings

The comparative analysis reveals that while traditional machine learning models like Random Forest and Logistic Regression provide high accuracy and efficiency, deep learning models offer significant advantages in handling complex and sequential data patterns. The choice of model depends on the specific requirements of the application, such as the need for real-time processing, the complexity of the data, and the availability of computational resources.

- **High-Accuracy Models**: Random Forest, Logistic Regression are recommended for applications requiring high accuracy and robustness, such as personalized content recommendations and targeted advertising.
- **Efficiency and Speed**: Naive Bayes and Logistic Regression are ideal for scenarios needing quick training and predictions, such as real-time sentiment analysis and spam detection, especially Naive Bayes.
- **Handling Complex Patterns**: CNNs and RNN + CNN hybrids are suited for tasks involving large and complex datasets, such as image and text classification.
- **Contextual Understanding**: Bi-LSTMs are best for applications requiring an understanding of long-term dependencies and context, such as language translation and speech recognition.

7.4 Potential Improvements

Despite the promising results, there are several areas for potential improvement:

Data Augmentation: Enhancing the dataset with more labeled data, especially for underrepresented categories, can improve model performance. Techniques such as synthetic data generation and data augmentation can help balance the dataset.

Advanced Regularization Techniques: Implementing advanced regularization techniques like dropout, batch normalization, and weight decay can help mitigate overfitting in deep learning models, leading to better generalization.

Hyper-parameter Optimization: Employing automated hyper-parameter optimization techniques, such as Bayesian optimization, can efficiently find the best hyper-parameters, improving model performance and reducing manual tuning efforts.

Ensemble Methods: Combining multiple models using ensemble methods like stacking, bagging, and boosting can leverage the strengths of each model, leading to improved accuracy and robustness.

Transfer Learning: Utilizing pre-trained models and fine-tuning them on the specific dataset can enhance performance, especially when dealing with limited labeled data.

7.5 Future Research Directions

The study opens several avenues for future research:

Exploring Transformer Models: Investigating the use of transformer-based models like BERT and GPT can provide deeper insights into their applicability for user interest classification on X, given their state-of-the-art performance in NLP tasks.

Multi-Modal Analysis: Extending the analysis to include multi-modal data, such as images, videos, and text, can provide a more comprehensive understanding of user interests, leveraging the strengths of various data modalities.

Real-Time Implementation: Implementing and testing the models in a real-time environment can provide valuable insights into their practical applicability and performance under real-world conditions.

Explainable AI: Exploring techniques for making models more interpretable and explainable can enhance their trustworthiness and usability, especially in critical applications like healthcare and finance.

7.6 Conclusion

This chapter has analyzed the performance of various models for classifying user interests on X, comparing traditional machine learning methods and deep learning approaches. Random Forest stood out for its high accuracy, making it suitable for applications requiring precision, such as targeted advertising. Logistic Regression, with its quick execution time, is ideal for real-time tasks like sentiment analysis, while Naive Bayes is efficient for large-scale applications like spam detection. Deep learning models like CNNs and Bi-LSTMs excelled in handling complex and sequential data patterns, suitable for image and text classification, and language translation tasks, respectively.

The choice of model depends on the application's needs, whether it's accuracy, speed, or handling complexity. Recommendations for improvement include data augmentation, advanced regularization techniques, hyper-parameter optimization, ensemble methods, and transfer learning. Future research should explore transformer models, multi-modal analysis, real-time implementation, and developing explainable AI techniques.

In summary, selecting the right model based on specific requirements and exploring advanced techniques can significantly enhance the classification of user interests on X.

Conclusion and Perspectives

This study thoroughly investigates the application of machine learning and deep learning techniques for classifying user interests on X, providing valuable insights into their performance and practical use. By categorizing posts into seven distinct areas of interest and using rigorous preprocessing methods, we created a robust and representative dataset. We evaluated traditional machine learning models such as Random Forest, Logistic Regression, and Naive Bayes, alongside advanced deep learning models like (CNNs), (RNN + CNN), and (Bi-LSTM).

The results reveal that the Random Forest model achieved the highest accuracy at 94.13%, demonstrating its robustness for this task. Logistic Regression and Naive Bayes followed closely, with accuracies of 93.19% and 92.96%, respectively, showcasing their efficiency in text classification. Deep learning models, particularly the Bi-LSTM with an accuracy of 92.54%, proved highly capable of capturing contextual information, although they required more computational resources and longer training times. Evaluation metrics, including precision, recall, and F1-score, consistently highlighted the strengths of each model. The Random Forest model excelled across all metrics, while deep learning models showcased their ability to handle complex patterns and sequential dependencies in the text. Execution time analysis indicated that traditional models, especially Naive Bayes, were extremely fast and suitable for real-time applications. In contrast, deep learning models benefited significantly from GPU acceleration, enabling them to process larger datasets more efficiently.

Addressing challenges such as data imbalance and overfitting was crucial to enhancing model performance. Techniques like data augmentation, regularization, early stopping, and class weighting were effectively employed to mitigate these issues. Furthermore, hyper-parameter tuning using grid search and Bayesian optimization played a pivotal role in optimizing model performance.

In conclusion, this study underscores the importance of selecting appropriate models based on the specific requirements of the task, such as accuracy, computational efficiency, and the ability to handle complex data patterns. The findings suggest that a hybrid approach, combining traditional machine learning models for their speed and deep learning models for their advanced pattern recognition capabilities, may offer the best solution for user interest classification on X. Future research could explore further enhancements in model architectures, the integration of additional data sources, and the application of transfer learning to improve classification performance and scalability. This study lays a solid foundation for the continued development of sophisticated social media analytics tools, ultimately contributing to a deeper understanding of user behavior and preferences in the digital age.

References

AlAbdullatif, A. M., Shahzad, B., Alwagait, E. (2016). Classification of arabic twitter users: A study based on user behaviour and interests. *Mob. Inf. Syst., 2016*, 8315281:1-8315281:11. Retrieved from https://api.semanticscholar.org/CorpusID:22688297

Ayodele, T. O. (2010). Types of machine learning algorithms.. Retrieved from https://api.semanticscholar.org/CorpusID:53061796

Bao, W. (2016). Naïve bayes classification in r. *Annals of translational medicine, 4 12*, 241. Retrieved from https://api.semanticscholar.org/CorpusID:46237516

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *2010 20th International Conference on Pattern Recognition*, 3121-3124. Retrieved from https://api.semanticscholar.org/CorpusID:11557689

Crowley, J. L. (2020). Convolutional neural networks. *Nature Methods, 20*, 1269-1270. Retrieved from https://api.semanticscholar.org/CorpusID:65268752

Cufoglu, A. (2014). User profiling - a short review. *International Journal of Computer Applications, 108*, 1-9. Retrieved from https://api.semanticscholar.org/CorpusID:8122695

Dien, T. T., Loc, B. H., Thai-Nghe, N. (2019). Article classification using natural language processing and machine learning. *2019 International Conference on Advanced Computing and Applications (ACOMP)*, 78-84. Retrieved from https://api.semanticscholar.org/CorpusID:214624101

E., N., P, V. K. B., S, P. V., K, A., V, A. (2019). Survey on classification and summarization of documents. *Other Topics Engineering Research eJournal*. Retrieved from https://api.semanticscholar.org/CorpusID:237526886

Elkan, C. P. (2010). Text mining and topic models.. Retrieved from https://api.semanticscholar.org/CorpusID:7447158

Fiallos, A., Jimenes, K. (2019). Using reddit data for multi-label text classification of twitter users interests. *2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG)*, 324-327. Retrieved from https://api.semanticscholar.org/CorpusID:189824273

Gray, D., Bowes, D., Davey, N., Sun, Y., Christianson, B. (2011). Further thoughts on precision. In *International conference on evaluation & assessment in software engineering*. Retrieved from https://api.semanticscholar.org/CorpusID:18088642

Kanoje, S., Girase, S., Mukhopadhyay, D. (2015). User profiling trends, techniques and applications. *ArXiv, abs/1503.07474*. Retrieved from https://api.semanticscholar.org/CorpusID:5499070

Kulkarni, V. Y., Sinha, P. K. (2014). Effective learning and classification using

random forest algorithm.. Retrieved from https://api.semanticscholar.org/CorpusID:51578605

Kumar, P. (2020). Predictive analytics for spam email classification using machine learning techniques. *Int. J. Comput. Appl. Technol.*, *64*, 282-296. Retrieved from https://api.semanticscholar.org/CorpusID:231589464

Lauren, P., Qu, G., Huang, G., Watta, P., Lendasse, A. (2017). A low-dimensional vector representation for words using an extreme learning machine. *2017 International Joint Conference on Neural Networks (IJCNN)*, 1817-1822. Retrieved from https://api.semanticscholar.org/CorpusID:10066606

Lim, K. H., Datta, A. (2013). Interest classification of twitter users using wikipedia. *Proceedings of the 9th International Symposium on Open Collaboration*. Retrieved from https://api.semanticscholar.org/CorpusID:6005906

Makhoul, J., Kubala, F., Schwartz, R. E., Weischedel, R. M. (2007). Performance measures for information extraction.. Retrieved from https://api.semanticscholar.org/CorpusID:15827348

Mo, D. (2012). A survey on deep learning: one small step toward ai.. Retrieved from https://api.semanticscholar.org/CorpusID:221081343

Novakovic, J., Veljovic, A., Ilić, S., Željko M. Papic, Milica, T. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, *7*, 39-46. Retrieved from https://api.semanticscholar.org/CorpusID:125586327

Nugaliyadde, A., Wong, K. K. W., Sohel, F., Xie, H. (2019). Enhancing semantic word representations by embedding deep word relationships. In *International conference on computer and automation engineering.* Retrieved from https://api.semanticscholar.org/CorpusID:58981651

Panesar, A. (2020). What is artificial intelligence? *Machine Learning and AI for Healthcare*, 1 - 18. Retrieved from https://api.semanticscholar.org/CorpusID:261070249

Pérez-Vera, S., ro Sandy González Alfaro, Allende-Cid, H. (2017). Intent classification of social media texts with machine learning for customer service improvement. In *Interacción.* Retrieved from https://api.semanticscholar.org/CorpusID:27520907

Raghuram, M. A., Akshay, K., Chandrasekaran, K. (2016). Efficient user profiling in twitter social network using traditional classifiers. In S. Berretti, S. M. Thampi, S. Dasgupta (Eds.), *Intelligent systems technologies and applications* (pp. 399–411). Cham: Springer International Publishing.

Rao, S., Jagdale, K., Miceli, L., Pizzetti, M., Dang-Xuan, L., Brown, J. S. (2018). Social media analytics and intelligence. *International Journal of Computer Applications*, *179*, 1-6. Retrieved from https://api.semanticscholar.org/CorpusID:62816793

Saglani, K. (2020). Machine learning based sentiment analysis on twitter data. *International Journal of Emerging Trends in Engineering Research.* Retrieved from https://api.semanticscholar.org/CorpusID:216654917

Sebei, H., Taieb, M. A. H., Aouicha, M. B. (2018). Review of social media analytics process and big data pipeline. *Social Network Analysis and Mining*, *8*. Retrieved from https://api.semanticscholar.org/CorpusID:256098552

Su, Y., Huang, Y., Kuo, C.-C. J. (2018). Dependent bidirectional rnn with extended-long short-term memory.. Retrieved from https://api.semanticscholar.org/CorpusID:67452153

Ting, K. M. (2010). Confusion matrix. In *Encyclopedia of machine learning.*

Retrieved from https://api.semanticscholar.org/CorpusID:16307526

Torgo, L., Ribeiro, R. P. (2009). Precision and recall for regression. In *Ifip working conference on database semantics*. Retrieved from https://api.semanticscholar.org/CorpusID:13560342

U., D. V. G., Sunithamma, K., Shenoy, P. D., Venugopal, K. R. (2017). An overview on user profiling in online social networks. *International Journal of Applied Information Systems*, *11*, 25-42. Retrieved from https://api.semanticscholar.org/CorpusID:55374775

Vermeulen, A. F. (2019). Classic machine learning.. Retrieved from https://api.semanticscholar.org/CorpusID:214138005

Wang, Q., Yu, S., Qi, X., song Hu, Y., Zheng, W. J., Shi, J., Yao, H. (2019). [overview of logistic regression model analysis and application]. *Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]*, *53 9*, 955-960. Retrieved from https://api.semanticscholar.org/CorpusID:201717776

Yadav, A., Alahmar, M., Singh, A., Sharma, K., Agrawal, R., Sharma, C. B. (2023). Analyzing user behavior in social media through big data analytics. *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 1-5. Retrieved from https://api.semanticscholar.org/CorpusID:268544572

Yu, Y., Si, X., Hu, C., xun Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, *31*, 1235-1270. Retrieved from https://api.semanticscholar.org/CorpusID:160013244

Appendices

Appendix A

Deposit Permission

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur Et de La Recherche Scientifique

Faculté des Sciences et de la Technologie

Département des Mathématiques et d'Informatique

جامعة غرداية

Université de Ghardaïa

كلية العلوم و التكنولوجيا

قسم الرياضيات والإعلام الآلي

Ghardaïa le 25/10/2024

Rapport de correction de mémoire de Master SIEC

Je soussigné M/~~Mme/Mlle~~ : **Houssem-Eddine DEGHA**

Président du jury du mémoire de Master intitulé :

Classification of X Users based on their Interests Using Deep Learning and Machine Learning Algorithms. A comparative study

Après les corrections apportées au rapport, je déclare que les étudiants :

Anis BAALIOUSAID & Aoumer HADJ SAID

Sont autorisés à déposer leur manuscrit au niveau du département.

Fait et délivré pour servir et valoir ce que de droit

رئيس قسم الرياضيات و الإعلام الآلي
الحاج موسى ياسين

Signature