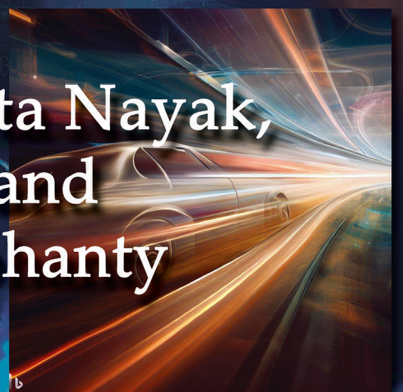


# HOW MACHINE LEARNING IS INNOVATING TODAY'S WORLD

*A Concise Technical Guide*

EDITED BY

Arindam Dey, Sukanta Nayak,  
Ranjan Kumar, and  
Sachi Nandan Mohanty





# How Machine Learning is Innovating Today's World

**Scrivener Publishing**

100 Cummings Center, Suite 541J

Beverly, MA 01915-6106

*Publishers at Scrivener*

Martin Scrivener ([martin@scrivenerpublishing.com](mailto:martin@scrivenerpublishing.com))

Phillip Carmical ([pcarmical@scrivenerpublishing.com](mailto:pcarmical@scrivenerpublishing.com))

# How Machine Learning is Innovating Today's World

**A Concise Technical Guide**

Edited by

**Arindam Dey**

*School of Computer Science, VIT-AP University, Andhra Pradesh, India*

**Sukanta Nayak**

*Dept. of Mathematics, VIT-AP University, Andhra Pradesh, India*

**Ranjan Kumar**

*Dept. of Mathematics, VIT-AP University, India*

and

**Sachi Nandan Mohanty**

*Dept. of Computer Science, VIT-AP University, Andhra Pradesh, India*



Scrivener  
Publishing

**WILEY**

This edition first published 2024 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

© 2024 Scrivener Publishing LLC

For more information about Scrivener publications please visit [www.scrivenerpublishing.com](http://www.scrivenerpublishing.com).

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

#### **Wiley Global Headquarters**

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at [www.wiley.com](http://www.wiley.com).

#### **Limit of Liability/Disclaimer of Warranty**

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchant-ability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials, or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read.

#### ***Library of Congress Cataloging-in-Publication Data***

ISBN 978-1-394-21411-2

Cover image: Pixabay.Com

Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

# Contents

---

<b>Preface</b>	<b>xvii</b>
<b>Part 1: Natural Language Processing (NLP) Applications</b>	<b>1</b>
<b>1 A Comprehensive Analysis of Various Tokenization Techniques and Sequence-to-Sequence Model in Natural Language Processing</b>	<b>3</b>
<i>Kuldeep Vayadande, Ashutosh M. Kulkarni, Gitanjali Bhimrao Yadav, R. Kumar and Aparna R. Sawant</i>	
1.1 Introduction	3
1.2 Literature Survey	4
1.3 Sequence-to-Sequence Models	5
1.3.1 Convolutional Seq2Seq Models	5
1.3.2 Pointer Generator Model	6
1.3.3 Attention-Based Model	6
1.4 Comparison Table	6
1.5 Comparison Graphs	9
1.6 Research Gap Identified	9
1.7 Conclusion	10
References	10
<b>2 A Review on Text Analysis Using NLP</b>	<b>13</b>
<i>Kuldeep Vayadande, Preeti A. Bailke, Lokesh Sheshrao Khedekar, R. Kumar and Varsha R. Dange</i>	
2.1 Introduction	14
2.2 Literature Review	15
2.3 Comparison Table of Previous Techniques	18
2.4 Comparison Graphs	20
2.5 Research Gap	20
2.6 Conclusion	21
References	22
<b>3 Text Generation &amp; Classification in NLP: A Review</b>	<b>25</b>
<i>Kuldeep Vayadande, Dattatray Raghunath Kale, Jagannath Nalavade, R. Kumar and Hanmant D. Magar</i>	
3.1 Introduction	25
3.2 Literature Survey	26
3.3 Comparison Table of Previous Techniques	29

3.3.1	Sentiment Analysis	32
3.3.2	Translation	33
3.3.3	Tokenization Based on Noisy Texts	33
3.3.4	Question Answer Model	34
3.4	Research Gap	34
3.5	Conclusion	35
	References	35
<b>4</b>	<b>Book Genre Prediction Using NLP: A Review</b>	<b>37</b>
	<i>Kuldeep Vayadande, Preeti Bailke, Ashutosh M. Kulkarni, R. Kumar and Ajit B. Patil</i>	
4.1	Introduction	37
4.2	Literature Survey	38
4.3	Comparison Table	41
4.4	Research Gap Identified	45
4.5	Future Scope	45
4.6	Conclusion	45
	References	46
<b>5</b>	<b>Mood Detection Using Tokenization: A Review</b>	<b>47</b>
	<i>Kuldeep Vayadande, Preeti A. Bailke, Lokesh Sheshrao Khedekar, R. Kumar and Varsha R. Dange</i>	
5.1	Introduction	47
5.2	Literature Survey	49
5.3	Comparison Table of Previous Techniques	51
5.4	Graphs	54
5.5	Research Gap	55
5.6	Conclusion	55
	References	56
<b>6</b>	<b>Converting Pseudo Code to Code: A Review</b>	<b>57</b>
	<i>Kuldeep Vayadande, Preeti A. Bailke, Anita Babu Dombale, Varsha R. Dange and Ashutosh M. Kulkarni</i>	
6.1	Introduction	58
6.2	Literature Review	60
6.3	Comparison Table	63
6.4	Graphs of Comparison Done	66
6.5	Research Gap Identified	67
6.6	Conclusion	67
	References	67
	<b>Part 2: Machine Learning Applications in Specific Domains</b>	<b>69</b>
<b>7</b>	<b>Evaluating the Readability of English Language Using Machine Learning Models</b>	<b>71</b>
	<i>Shiplu Das, Abhishikta Bhattacharjee, Gargi Chakraborty and Debarun Joardar</i>	
7.1	Introduction	72



7.2	Contribution in this Chapter	76
7.3	Research Gap	78
7.4	Literature Review	78
7.5	Proposed Model	81
7.6	Model Analysis with Result and Discussion	81
7.7	Conclusion	86
	References	87
<b>8</b>	<b>Machine Learning in Maximizing Cotton Yield with Special Reference to Fertilizer Selection</b>	<b>89</b>
	<i>G. Hannah Grace and Nivetha Martin</i>	
8.1	Introduction	89
8.2	Literature Review	90
8.3	Materials and Methods	92
	8.3.1 Problem Definition	92
	8.3.2 Objectives	93
	8.3.3 Data Collection	93
	8.3.4 Data Preprocessing	93
	8.3.5 Steps Involved in Combined Decision-Making Approach Using Machine Learning Algorithms	93
8.4	Application to the Fertilizer Selection Problem	94
8.5	Conclusion and Future Suggestions	95
	References	96
<b>9</b>	<b>Machine Learning Approaches to Catalysis</b>	<b>101</b>
	<i>Sachidananda Nayak and Selvakumar Karuthapandi</i>	
9.1	Introduction	101
9.2	Chem-Workflow	102
9.3	ML Basic Concepts	109
9.4	ML Models in Catalysis	114
9.5	ML in Structure–Activity Prediction	119
9.6	Conclusion and Future Works	123
	References	124
<b>10</b>	<b>Classification of Livestock Diseases Using Machine Learning Algorithms</b>	<b>127</b>
	<i>G. Hannah Grace, Nivetha Martin, I. Pradeepa and N. Angel</i>	
10.1	Introduction	127
10.2	Literature Review	129
10.3	Materials and Methods	130
	10.3.1 Definition of the Problem	130
	10.3.2 Objectives	131
	10.3.3 Data Collection	131
	10.3.4 Data Preprocessing	131
	10.3.5 Steps Involved in Supervised Learning Classifiers	134
10.4	Application of the Supervised Classifiers in Disease Classification	135
10.5	Results and Conclusion	136
	References	136

<b>11 Image Enhancement Techniques to Modify an Image with Machine Learning Application</b>	<b>139</b>
<i>Shiplu Das, Sohini Sen, Debarun Joardar and Gargi Chakraborty</i>	
11.1 Introduction	140
11.2 Literature Review	141
11.3 Image Enhancement Techniques for Betterment of the Images	142
11.4 Proposed Image Enhancement Techniques	148
11.5 Conclusion	155
References	156
<b>12 Software Engineering in Machine Learning Applications: A Comprehensive Study</b>	<b>159</b>
<i>Kuldeep Vayadande, Komal Sunil Munde, Amol A. Bhosle, Aparna R. Sawant and Ashutosh M. Kulkarni</i>	
12.1 Introduction	159
12.2 Related Works	160
12.3 Comparison Table	162
12.4 Graph of Comparison	167
12.5 Machine Learning in Software Engineering	170
12.6 Conclusion	171
References	171
<b>13 Machine Learning Applications in Battery Management System</b>	<b>173</b>
<i>Ponnaganti Chandana and Ameet Chavan</i>	
13.1 Introduction	174
13.2 Battery Management System (BMS)	174
13.2.1 Key Parameters of Battery Management System	175
13.2.1.1 Voltage	176
13.2.1.2 Temperature	176
13.2.1.3 State of Charge	176
13.2.1.4 State of Health	177
13.2.1.5 State of Function	177
13.3 Estimation of Battery SOC and SOH	177
13.3.1 Methods of Estimating SOC	178
13.3.1.1 Coulomb Counting Method	178
13.3.1.2 Open Circuit Voltage (OCV) Method	178
13.3.1.3 Kalman Filtering Method	178
13.3.1.4 Artificial Neural Network (ANN) Method	178
13.3.1.5 Fuzzy #	179
13.3.1.6 Extended Kalman Filtering Method	179
13.3.1.7 Gray Box Modeling Method	179
13.3.1.8 Support Vector Machine (SVM) Method	179
13.3.1.9 Model Predictive Control Method	180
13.3.1.10 Adaptive Observer Method	180
13.3.1.11 Impedance-Based Method	180
13.3.1.12 Gray Prediction Method	180
13.3.2 Methods of Estimating SOH	181

13.3.2.1	Capacity Fade Model	181
13.3.2.2	Electrochemical Impedance Spectroscopy (EIS) Method	181
13.3.2.3	Voltage Relaxation Method	181
13.3.2.4	Fuzzy Logic Method	181
13.3.2.5	Particle Filter Method	182
13.3.2.6	Artificial Neural Network (ANN) Method	182
13.3.2.7	Support Vector Machine (SVM) Method	182
13.3.2.8	Gray Box Modeling Method	182
13.3.2.9	Kalman Filtering Method	183
13.3.2.10	Multi-Model Approach	183
13.4	Cell Balancing Mechanism for BMS	183
13.5	Industrial Applications	184
13.5.1	Industrial Applications of Machine Learning in Battery Management System	184
13.5.2	Machine Learning Algorithms That Are Used for Industrial Applications in Battery Management System	185
13.5.3	Steps Involved in Machine Learning Approach in BMS Applications	186
13.5.4	Applications of Different ML Algorithms in BMS	187
13.5.4.1	Artificial Neural Networks (ANNs)	187
13.5.4.2	Decision Trees	188
13.5.4.3	Support Vector Machines (SVMs)	188
13.5.4.4	Random Forest	188
13.5.4.5	Gaussian Process	188
13.6	Case Studies of ML-Based BMS Applications in Industry	189
13.6.1	Machine Learning Approach to Predict SOH of Li-Ion Batteries	189
13.6.2	Anomaly Detection in Battery Management System Using Machine Learning	189
13.6.3	Optimization of Battery Life Cycle Using Machine Learning	189
13.6.4	Prediction of Remaining Useful Life Using Machine Learning	190
13.6.5	Fault Diagnosis of Battery Management System Using Machine Learning	190
13.6.6	Battery Parameter Estimation Using Machine Learning	190
13.6.7	Optimization of Battery Charging Using Machine Learning	190
13.6.8	ML Approach to Estimate State of Charge	191
13.6.9	Battery Capacity Estimation Using ML Approach	191
13.6.10	Anomaly Detection in Batteries Using Machine Learning	191
13.6.11	ML-Based BMS for Li-Ion Batteries	191
13.6.12	Battery Management System Based on Deep Learning for Electric Vehicles	192
13.6.13	A Review of ML Approaches for BMS	192
13.6.14	Battery Management Systems Using Machine Learning Techniques	193
13.6.15	Machine Learning for Lithium-Ion Battery Management: Challenges and Opportunities	193
13.6.16	An ML-Based BMS for Hybrid EVs	194
13.6.17	Battery Management System for EVs Using ML Techniques	194

13.6.18	A Hybrid BMS Using Machine Learning Techniques	195
13.7	Challenges	195
13.8	Conclusion	196
	References	196
<b>14</b>	<b>ML Applications in Healthcare</b>	<b>201</b>
	<i>Farooq Shaik, Rajesh Yelchurri, Noman Aasif Gudur and Jatindra Kumar Dash</i>	
14.1	Introduction	202
14.1.1	Supervised Learning	202
14.1.2	Unsupervised Learning	203
14.1.3	Semi-Supervised Learning	203
14.1.4	Reinforcement Learning	204
14.2	Applications of Machine Learning in Health Sciences	204
14.2.1	Diagnosis and Prediction of Disease	204
14.2.1.1	Predicting Thyroid Disease	205
14.2.1.2	Predicting Cardiovascular Disease	205
14.2.1.3	Predicting Cancer	206
14.2.1.4	Predicting Diabetes	206
14.2.1.5	Predicting Alzheimer's	208
14.2.2	Drug Development and Discovery	209
14.2.3	Clinical Decision Support (CDS)	211
14.2.4	Medical Image Examination	212
14.2.5	Monitoring of Health and Wearable Technology	214
14.2.6	Telemedicine and Remote Patient Monitoring	215
14.2.7	Chatbots and Virtual Medical Assistants	215
14.3	Why Machine Learning is Crucial in Healthcare	216
14.4	Challenges and Opportunities	216
14.5	Conclusion	217
	References	217
<b>15</b>	<b>Enhancing Resource Management in Precision Farming through AI-Based Irrigation Optimization</b>	<b>221</b>
	<i>Salina Adinarayana, Matha Govinda Raju, Durga Prasad Srirangam, Devee Siva Prasad, Munaganuri Ravi Kumar and Sai babu veesam</i>	
15.1	Introduction to Precision Farming	222
15.1.1	Definition of Precision Farming	222
15.1.2	Importance of Precision Farming in Agriculture	225
15.2	Role of Artificial Intelligence (AI) in Precision Farming	226
15.2.1	Influence of AI in Precision Farming	226
15.2.2	Challenges and Limitations of AI in Precision Farming	227
15.3	Data Collection and Sensing for Precision Farming	228
15.3.1	Remote Sensing Techniques	228
15.3.2	Satellite Imagery Analysis	230
15.3.3	Unmanned Aerial Vehicles (UAVs) for Data Collection	232
15.3.4	Internet of Things (IoT) Sensors	233
15.3.5	Data Preprocessing and Integration	235

15.4	Crop Monitoring and Management	236
15.4.1	Crop Yield Prediction	236
15.4.2	Disease Detection and Diagnosis	236
15.4.3	Nutrient Management and Fertilizer Optimization	237
15.5	Precision Planting and Seeding	237
15.5.1	Variable Rate Planting	237
15.5.2	GPS and Auto-Steering Systems	237
15.5.3	Seed Singulation and Metering	237
15.5.4	Plant Health Monitoring and Care	238
15.6	Harvesting and Yield Estimation	238
15.6.1	Yield Estimation Models	238
15.6.2	Real-Time Crop Monitoring During Harvest	239
15.7	Data Analytics and Machine Learning	239
15.7.1	Predictive Analytics for Crop Yield	240
15.7.2	Machine Learning Algorithms for Precision Farming	241
15.7.3	Big Data Analytics in Precision Farming	242
15.8	Integration of AI with Other Technologies	242
15.8.1	AI and Blockchain in Supply Chain Management	242
15.8.2	AI and Drones in Precision Farming	243
15.8.3	AI and Robotics Collaboration	244
15.9	Case Studies and Success Stories	245
15.10	Challenges and Future Trends	247
15.11	Conclusion	248
	References	248
<b>16</b>	<b>An In-Depth Review on Machine Learning Infusion in an Agricultural Production System</b>	<b>253</b>
	<i>Sarthak Dash, Sugyanta Priyadarshini and Sukanya Priyadarshini</i>	
16.1	Background Study	253
16.2	Research Methodology	255
16.2.1	Planning the Review	256
16.2.2	Search String	256
16.2.3	Selection Criteria	257
16.2.4	Conduction of Review	257
16.3	Results and Discussion	257
16.3.1	Crop Yield Prediction	257
16.3.2	Crop Disease Management	259
16.3.3	Water Management	261
16.3.4	Soil Management	263
16.3.5	Weather Forecasting	264
16.4	Conclusion	264
	References	266
<b>Part 3:</b>	<b>Artificial Intelligence and Optimization Techniques</b>	<b>271</b>
<b>17</b>	<b>Reinforcement Learning Approach in Supply Chain Management: A Review</b>	<b>273</b>
	<i>Rajkanwar Singh, Pratik Mandal and Sukanta Nayak</i>	
17.1	Introduction	274

17.2	Literature Review	275
17.2.1	Challenges	275
17.2.2	Advantages of Using ML Techniques in SCM	277
17.2.3	Limitations of Using ML Techniques in SCM	278
17.2.4	Effectiveness of ML Techniques in Handling Various SCM Activities	279
17.3	Methodology	280
17.4	Reinforcement Learning in Supply Chain Management	282
17.4.1	RL and Its Application in SCM	282
17.4.2	Benefits of RL in SCM	285
17.5	Adoption of Reinforcement Learning in Supply Chain Management	286
17.5.1	Technical Barriers	287
17.5.2	Organizational Barriers	288
17.5.3	Cultural Barriers	288
17.5.4	Economic Barriers	288
17.6	Alternatives to Reinforcement Learning in Supply Chain Management	289
17.7	Conclusion	297
	References	299
<b>18</b>	<b>Alternate Approach to Solve Differential Equations Using Artificial Neural Network with Optimization Technique</b>	<b>303</b>
	<i>Ramanan R., Sukanta Nayak and Arun Kumar Gupta</i>	
18.1	Introduction	303
18.2	Artificial Neural Network	305
18.2.1	Architecture	305
18.2.2	Neuron Architecture	306
18.2.2.1	Single and Multilayer Neural Networks	306
18.2.2.2	Feedforward Neural Network	306
18.2.2.3	Feedback Neural Network	306
18.2.3	Different Training Process	306
18.2.3.1	Supervised Training	307
18.2.3.2	Unsupervised Training	307
18.2.4	Learning Process	307
18.2.5	Activation Function in ANN	307
18.2.5.1	Unipolar Sigmoid Function	307
18.2.5.2	Bipolar Sigmoid Function	308
18.3	Backpropagation Algorithm	309
18.4	Solving Differential Equation Using ANN	310
18.4.1	Structure of Multi-Layer ANN	311
18.4.2	General Formula for Solving ODE	312
18.4.3	Formulation for $n$ th-Order Initial Value Problem (IVP)	313
18.4.4	Case Study: Solving First-Order Linear Differential Equation	314
18.4.4.1	Algorithm	314
18.4.4.2	Example	315
18.4.4.3	Second Approach	316

18.4.4.4	Algorithm	317
18.4.4.5	Example	318
18.4.4.6	Comparison between First and Second Approaches	319
18.5	System Identification Using ANN	319
18.5.1	Problem Structure for System Identification	320
18.5.2	Analysis and Modelling	320
18.5.3	ANN Training for SI	322
18.5.4	Results and Discussion	323
18.6	Conclusion	326
	References	326
<b>19</b>	<b>GPT-3- and DALL-E-Powered Applications: A Complete Survey</b>	<b>329</b>
	<i>Kuldeep Vayadande, Chaitanya B. Pednekar, Priya Anup Khune, Vinay Sudhir Prabhavalkar and Varsha R. Dange</i>	
19.1	Introduction	329
19.2	Understanding GPT-3	331
19.3	Understanding DALL-E	333
19.4	Applications Powered By GPT-3 and DALL-E	334
19.4.1	Copy.ai	335
19.4.2	OpenAI GPT-3 Playground	336
19.4.3	GPT-3 Sandbox	336
19.4.4	AI Dungeon	337
19.5	Challenges and Open Issues	338
19.6	Future Directions	339
19.7	Conclusion	341
	References	341
<b>20</b>	<b>New Variation of Exam Scheduling Problem Using Graph Coloring</b>	<b>343</b>
	<i>Angshu Kumar Sinha, Soumyadip Laha, Debarghya Adhikari, Anjan Koner and Neha Deora</i>	
20.1	Introduction	343
20.1.1	Review of Previous Work	344
20.1.2	Application	344
20.1.3	Main Result	345
20.1.4	Organization of the Paper	345
20.2	Notations and Preliminaries	345
20.3	Description of Algorithm	345
20.3.1	The Algorithm	345
20.3.2	Illustration of the Algorithm	346
20.3.2.1	Problem of Scheduling the Examination	346
20.3.3	Algorithm in Python	347
20.3.4	Algorithm in C++	348
20.3.5	Algorithm in C	349
20.4	Time Complexity of the Algorithm	350
20.5	Concluding Remarks	350
20.6	Acknowledgments	351
	References	351

<b>Part 4: Emerging Topics in Machine Learning</b>	<b>353</b>
<b>21 A Comparative Study of Different Techniques of Text-to-SQL Query Converter</b>	<b>355</b>
<i>Kuldeep Vayadande, Preeti A. Bailke, Vikas Janu Nandeshwar, R. Kumar and Varsha R. Dange</i>	
21.1 Introduction	355
21.2 Literature Survey	356
21.3 Comparison Table of Previous Techniques	358
21.4 Comparison Graphs	363
21.5 Research Gap	364
21.6 Conclusion	364
References	365
<b>22 Trust-Based Leader Election in Flying Ad-Hoc Network</b>	<b>367</b>
<i>Joydeep Kundu, Sahabul Alam and Sukanta Oraw</i>	
22.1 Introduction	367
22.2 Related Work	368
22.3 Discussion of the Proposed Methodology and Results	370
22.4 Conclusion	371
References	372
<b>23 A Survey on Domain of Application of Recommender System</b>	<b>375</b>
<i>Sudipto Dhar</i>	
23.1 Introduction	375
23.2 Background	377
23.3 Study of Recommender Systems	378
23.4 Conclusion	380
References	380
<b>24 New Approach on M/M/c/K Queueing Models via Single Valued Linguistic Neutrosophic Numbers and Perceptionization Using a Non-Linear Programming Technique</b>	<b>383</b>
<i>Antony Crispin Sweety C. and Vennila B.</i>	
24.1 Introduction	383
24.2 Neutrosophic M/M/C/K Queue	386
24.3 Perceptionization of the NM/NM/c/K Queueing Model Using a Non-Linear Programming Technique	407
24.3.1 Classic M/M/c/K Model	407
24.3.2 Neutrosophic M/M/c/K Queue	408
24.3.3 Performance Measures	408
24.3.4 Neutrosophic Extension Principle	410
24.3.4.1 $(\alpha, \beta, \gamma)$ -Cut of Set Neutrosophic Numbers	410
24.3.5 Non-Linear Programming (NLP)	411
24.3.6 Parametric Non-Linear Programming Technique	411
24.3.6.1 Upper and Lower Boundaries of the $\alpha$ -Cuts in $\theta_{(N(x,y))}$	412
24.3.6.2 Upper and Lower Boundaries of the $\beta$ -CUTS in $\theta_{(N(x,y))}$	413
24.3.6.3 Upper and Lower Boundaries of the $\gamma$ -CUTS in $\theta_{(N(x,y))}$	414



Conclusion	421
References	421
<b>25 The Rise of AI-Generated News Videos: A Detailed Review</b>	<b>423</b>
<i>Kuldeep Vayadande, Mustansir Bohri, Mohit Chawala, Ashutosh M. Kulkarni and Asif Mursal</i>	
25.1 Introduction	423
25.2 Web Scraping	425
25.3 Image Searching	427
25.4 News Authentication	430
25.5 Scripting for Video	434
25.6 Audio Generation	436
25.7 Mapping Text and Images	438
25.8 AI-Avatar Generation	440
25.9 Video Generation	443
25.10 Thumbnail Creation	446
25.11 Conclusion	448
References	449
<b>Index</b>	<b>453</b>



## Preface

---

This timely book presents a diverse collection of chapters that delve into the remarkable ways that machine learning (ML) is transforming various fields and industries. It provides a comprehensive understanding of the latest advancements and practical applications of ML techniques.

Machine learning, a branch of artificial intelligence, has gained tremendous momentum in recent years, revolutionizing the way we analyze data, make predictions, and solve complex problems. As researchers and practitioners in the field, the editors of this book recognize the importance of disseminating knowledge and fostering collaboration to further advance this dynamic discipline.

The chapters herein cover a wide range of topics, each contributing a unique perspective to the broader landscape of machine learning. First is a comprehensive analysis of various tokenization techniques and the sequence-to-sequence model in natural language processing. Next, Chapter 2 explores the evaluation of English language readability using ML models, followed by a detailed study of text analysis for information retrieval through natural language processing in the subsequent chapter.

Chapter 4 investigates machine learning's role in maximizing cotton yield with a focus on fertilizer selection, and Chapter 5 delves into the application of reinforcement learning approaches to supply chain management. The following chapter examines the performance analysis of converting algorithms to source code using natural language processing in Java, and Chapter 7 presents an alternate approach to solving differential equations utilizing artificial neural networks with optimization techniques.

The exploration of the subject continues with a comparative study of different techniques of text-to-SQL query conversion in Chapter 8, and the next chapter examines ML approaches to catalysis. After that, Chapter 10 presents the systematic study of text generation and classification using tokenization in natural language processing, followed by the classification of livestock diseases using ML algorithms in Chapter 11.

Chapter 12 provides a closer look at the application of ML in image enhancement techniques, and the following chapter demonstrates the prediction of book genres using natural language processing. Additionally, Chapter 14 delves into efficient leader selection for inter-cluster flying ad-hoc networks, and the subsequent chapter provides a comprehensive survey of applications powered by GPT-3 and DALL-E.

Recommender systems' domain of application is discussed in Chapter 16, and the next chapter reviews mood detection, emoji generation, and classification using tokenization and CNN. Chapter 18 delves into a new variation of the exam scheduling problem using graph coloring, and Chapter 19 examines the intersection of software engineering and machine learning applications.

Moreover, Chapter 20 explores ML strategies for indeterminate information systems in complex bipolar neutrosophic environments, and the rise of AI-generated news videos is scrutinized in Chapter 21. The next section highlights ML applications in battery management systems, while the healthcare industry is covered in Chapter 23. The book's final chapter presents how to enhance resource management in precision farming through AI-based irrigation optimization.

This book will serve as a valuable resource for researchers, scholars, and enthusiasts seeking to understand the cutting-edge advancements in ML. The editors extend our gratitude to all the authors who have contributed their expertise, insights, and knowledge to make this book possible. Their commitment to advancing the frontiers of machine learning has greatly enriched the content and depth of this publication.

We also offer our sincere appreciation to Wiley and Scrivener Publishing for their support and guidance throughout the editorial process. Their commitment to publishing high-quality scientific literature has been instrumental in bringing this book to fruition.

We hope that readers find this book insightful, engaging, and thought-provoking. May it inspire you to explore new horizons in machine learning and contribute to the ongoing advancements that are reshaping our world.

**The Editors**  
March 2024

# **Part 1**

# **NATURAL LANGUAGE PROCESSING (NLP) APPLICATIONS**



# A Comprehensive Analysis of Various Tokenization Techniques and Sequence-to-Sequence Model in Natural Language Processing

Kuldeep Vayadande<sup>1\*</sup>, Ashutosh M. Kulkarni<sup>1</sup>, Gitanjali Bhimrao Yadav<sup>1</sup>, R. Kumar<sup>2</sup>  
and Aparna R. Sawant<sup>1</sup>

<sup>1</sup>Vishwakarma Institute of Technology, Pune, India

<sup>2</sup>VIT-AP University, Inavolu, Beside AP Secretariat, Amaravati AP, India

---

## **Abstract**

This research paper provides an in-depth examination of various tokenization techniques and Sequence-to-Sequence (Seq2Seq) models, with an emphasis on the LSTM, Transformer, and Attention-based LSTM models. The process of tokenization, which breaks down text into smaller units, plays a vital role in natural language processing (NLP). This study evaluates different tokenization methods, including word-based, character-based, and sub-word-based methods. It also explores the latest advancements in Seq2Seq models, such as the LSTM, Transformer, and Attention-based LSTM models, which have been successful in tasks like machine translation, text summarization, and dialog systems. The paper compares the performance of different tokenization techniques and Seq2Seq models on benchmark datasets. Additionally, it highlights the strengths and limitations of these models, which helps in understanding their suitability for various NLP applications. The aim of this study is to comprehensively understand the current advancements in tokenization and sequence-to-sequence modeling for NLP, particularly with regard to LSTM, Transformer, and Attention-based LSTM models.

**Keywords:** RNN, CRNN, LSTM, bidirectional-LSTM, text augmentation, tokenization, attention-based LSTM

## **1.1 Introduction**

Tokenization is a fundamental step in natural language processing (NLP) that entails breaking down text into smaller units, such as words or characters. This process is critical for many NLP tasks, including text classification, machine translation, and text summarization. Different levels of granularity, such as word-level, character-level, and sub-word-level, can be used for tokenization.

In recent years, various tokenization techniques have been proposed, each with their unique advantages and disadvantages. The Multi-head Self-attention Mechanism in [1] is

---

\*Corresponding author: kuldeep.vayadande1@vit.edu

a type of attention mechanism that allows the model to concentrate on multiple parts of the input text simultaneously. Tokenization is the most straightforward approach, and it is widely used in many NLP tasks. However, it may not be as effective for languages with complex morphological structures, such as agglutinative languages. On the other hand, character-level tokenization can handle such languages better, but it may also introduce more noise into the data. Sub-word-level tokenization, such as byte-pair encoding (BPE) and unigram language modeling (ULM), has been proposed as a compromise between word-level and character-level tokenization.

The goal of this study is to provide a thorough understanding of the tokenization methods that have been introduced recently. The research will evaluate different tokenization methods, including word-based, character-based, and sub-word-based methods, and compare their performance on a set of benchmark datasets. Furthermore, the research will delve into the details of these techniques, their working principle, and their performance on various NLP tasks. Additionally, the research will also analyze the advantages and limitations of these techniques, which will assist in understanding their suitability for different types of NLP applications. The objective of this research is to gain a complete insight into the latest developments in tokenization for NLP. This research will be a valuable resource for researchers and practitioners in the field of NLP, supplying students with a thorough comprehension of the most advanced tokenization algorithms available at the time. The popularity of Sequence-to-Sequence (Seq2Seq) models in NLP has grown in recent times because of their capability to process input and output sequences of varying lengths. Seq2Seq models, also known as encoder-decoder models, have produced noteworthy outcomes in a number of NLP applications, including dialogue systems, machine translation, and text summarization. Different Seq2Seq models have been put forth through time, and each has merits and faults of its own. One such model is the Long Short-Term Memory (LSTM), a popular Seq2Seq model that has shown promise in a variety of NLP applications. Its limitation is that it is computationally expensive. An alternative to the LSTM model that has proven to be more effective is the Transformer model, which is built on the attention mechanism. Attention-based LSTM models are also proposed, which combine the advantages of LSTM and attention mechanisms. With an emphasis on the LSTM, Transformer, and Attention-based LSTM models, this survey seeks to provide a thorough understanding of the many Seq2Seq models that have been suggested in recent years. We will cover the details of these models, their working principle, and their performance on various NLP tasks. Furthermore, we also cover the advantages, limitations, and performance comparison of these models, which helps in understanding their suitability for different types of NLP applications.

Also, some non-tokenization technique as mentioned in [4] is focused on pre-training an efficient encoder that can operate without tokenization.

## 1.2 Literature Survey

The Multi-head Self-attention Mechanism [1] is a type of attention mechanism that enables the model to concentrate on multiple sections of the input text simultaneously. This is achieved by utilizing multiple “heads” to attend to different segments of the input. This can aid the model’s understanding of the context and relationships between different parts of the input text, resulting in more accurate and coherent summaries.



The pointer network is a type of encoder–decoder [2] model that uses an attention mechanism to point to the part of the input text that should be included in the summary. It helps the model to generate the summary by copying relevant words from the input text, rather than generating them from scratch.

In summary, the research paper [3] focuses on improving the language generation performance by calibrating the likelihood of the generated sequences, while the research paper [4] focuses on Long Document Summarization and uses a combination of top-down and bottom-up inference to extract high-level concepts and specific details from the input text. Both papers provide different approaches to improve the performance of NLP tasks.

CANINE [4] is focused on pre-training an efficient encoder that can operate without tokenization, making the training process faster and more scalable. It uses a simple linear-layer-based architecture and employs a binary masking strategy to hide specific words during training, in order to predict them during inference.

FNet [5], on the other hand, introduces a new method of mixing tokens with Fourier transforms to capture long-range dependencies. This method can produce highly expressive representations and has the advantage of being computationally efficient. The paper shows that the approach outperforms traditional pre-training methods on various NLP tasks.

Charformer [6] focuses on the tokenization stage of pre-processing and proposes a novel method of sub-word tokenization that utilizes gradient information to identify the best sub-word splits. The authors show that their method is fast and results in improved performance on several NLP tasks.

The paper [7] proposes a new pre-training method for language models that leverages retrieval-based techniques. The authors show that this approach can effectively pre-train models on large-scale text corpora, leading to improved performance on a variety of NLP benchmarks.

The paper [8] focuses on enhancing the training efficiency of large-scale transformers, which are frequently employed in NLP applications like language modeling, text classification, and machine translation. The authors propose a new training method, “Random-LTD,” which involves randomly dropping tokens and layers during the training process to speed up convergence and reduce memory requirements. The authors show that this method can effectively train large-scale transformers with improved efficiency.

Detecting Label Errors in Token Classification Data [9] focuses on a different challenge in NLP, which is detecting label errors in token classification data. The authors propose a method for detecting label errors in token classification datasets, which can negatively impact the performance of NLP models. The authors present experiments demonstrating the effectiveness of their method in detecting label errors in real-world datasets.

## 1.3 Sequence-to-Sequence Models

### 1.3.1 Convolutional Seq2Seq Models

Convolutional Seq2Seq (ConvSeq2Seq) models [10] are a variant of Seq2Seq models that incorporate (CNNs) into model architecture. ConvSeq2Seq models are particularly useful for processing sequences of data with a grid-like structure, such as image sequences or spectrograms.

Compared to traditional Seq2Seq models that use recurrent neural networks (RNNs), ConvSeq2Seq models have the advantage of being able to process sequences in parallel, which can lead to faster training and inference times. However, they may be less effective at capturing long-range dependencies in the data compared to RNN-based models.

### 1.3.2 Pointer Generator Model

Pointer Generator [11] models are a type of Seq2Seq model used for text summarization and other tasks where the output sequence is a subset of input sequence. The Seq2Seq model comprises of a decoder that generates the target sequence and an encoder that analyzes the input sequence and generates a fixed-length vector representation.

Pointer Generator models have been utilized for a wide range of tasks such as text summarization, headline generating, and question answering. Compared to traditional Seq2Seq models, attention-based models offer a more adaptable approach to handle the issue of Out-of-Vocabulary (OOV) words and can produce output sequences that are more reflective of the input data.

### 1.3.3 Attention-Based Model

Models that rely on attention [12, 13] are a category of Seq2Seq models that have recently become widespread for various NLP jobs such as machine translation, text summarization, and answering questions. The attention mechanism, which enables it to concentrate on various segments of the input sequence while producing the output sequence, is a key aspect of these models. The attention mechanism computes a score for each element in the input sequence that reflects its importance for the current task, and the decoder uses these scores to weight the contribution of each element when generating the output.

However, attention-based models can be computationally expensive and difficult to train, especially for large sequences, due to the need to compute attention scores for every element in the input sequence. Additionally, the attention mechanism can sometimes be unstable, leading to poor performance in some cases.

## 1.4 Comparison Table

Table 1.1 shows different Approach-based comparison [1–9]. Table 1.2 shows advantages and disadvantages of Tokenization and Seq2Seq Model [1–9]. Table 1.3 shows model performance [1–9].

**Table 1.1** Approach-based comparison.

Paper	Approach used	Performance
[1]	Multi-head self-attention mechanism	Improved understanding of context and relationships in input text
[2]	Pointer network	Use of attention mechanism to point to relevant input text for summary generation

(Continued)

**Table 1.1** Approach-based comparison. (*Continued*)

<b>Paper</b>	<b>Approach used</b>	<b>Performance</b>
[3]	Improved language generation performance by calibrating likelihood of generated sequences	[Performance not specified]
[4]	Long Document Summarization using top-down and bottom-up inference	[Performance not specified]
[5]	Pre-training an efficient encoder with binary masking strategy	Improved performance on NLP tasks
[6]	Novel sub-word tokenization method using gradient information	Improved performance on several NLP tasks
[7]	Retrieval-based pre-training method for language models	Improved performance on a variety of NLP benchmarks
[8]	Improving training efficiency of large-scale transformers	Improved efficiency in training large-scale transformers
[9]	Detecting label errors in token classification data	Effective method for detecting label errors in real-world datasets

**Table 1.2** Advantages and disadvantages based on tokenization and Seq2Seq approach used.

<b>Paper</b>	<b>Approach used</b>	<b>Performance</b>
[1]	Multi-head self-attention mechanism	Improved understanding of context and relationships in input text
[2]	Pointer network	Use of attention mechanism to point to relevant input text for summary generation

*(Continued)*

**Table 1.2** Advantages and disadvantages based on tokenization and Seq2Seq approach used. (Continued)

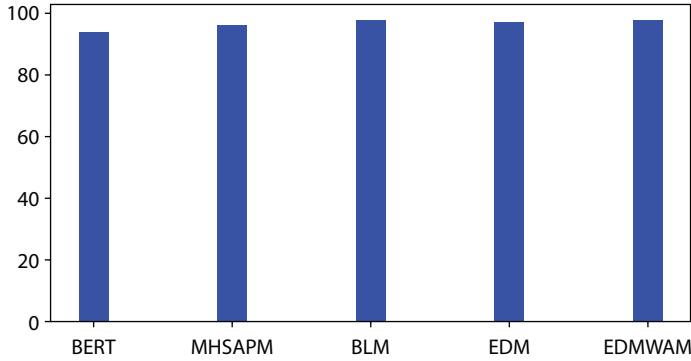
Paper	Approach used	Performance
[3]	Improved language generation performance by calibrating likelihood of generated sequences	[Performance not specified]
[4]	Long Document Summarization using top-down and bottom-up inference	[Performance not specified]
[5]	Pre-training an efficient encoder with binary masking strategy	Improved performance on NLP tasks
[6]	Novel sub-word tokenization method using gradient information	Improved performance on several NLP tasks
[7]	Retrieval-based pre-training method for language models	Improved performance on a variety of NLP benchmarks
[8]	Improving training efficiency of large-scale transformers	Improved efficiency in training large-scale transformers
[9]	Detecting label errors in token classification data	Effective method for detecting label errors in real-world datasets

**Table 1.3** Model performance comparison w.r.t accuracy and RMSE.

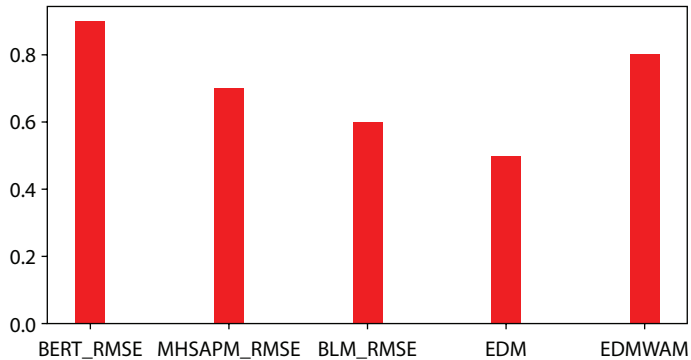
Model	Accuracy
BERT model	94
Multi-head self-attention pointer model	96
Bidirectional LSTM model	97.5
Encoder–decoder model	97
Encoder–decoder model with attention layer	98

## 1.5 Comparison Graphs

This section contains comparison graphs of various techniques. Figure 1.1 shows accuracy comparison graph of five different sequence-to-sequence models [1–9] and Figure 1.2 shows RMSE comparison graph of five different sequence-to-sequence models [1–9].



**Figure 1.1** Accuracy comparison graph of five different sequence-to-sequence models [1–9].



**Figure 1.2** RMSE comparison graph of five different sequence-to-sequence models [1–9].

## 1.6 Research Gap Identified

The current available models lack multi-sentence summarization tasks, and the existing model does not work for large datasets and is limited to character-level data. Handling long sequences: Seq2Seq models tend to struggle with long sequences, especially when the sequence is much longer than the training data. This leads to a drop in performance, and a need to address this challenge. Despite the impressive results achieved by Seq2Seq models, it is often difficult to understand how the model makes its predictions. Improving the interpretability of these models is an important research area. Seq2Seq models are trained on large amounts of data, and any biases in the data are likely to be reflected in the model's predictions. Addressing data bias is an important research area in Seq2Seq models.

Overall, the research gap in Seq2Seq models is to continue to improve the models' accuracy, interpretability, scalability, robustness, and ability to handle diverse data types, while addressing data bias and other challenges.

## 1.7 Conclusion

Tokenization is an important step in pre-processing text data for NLP tasks. It involves dividing a text into smaller units, called tokens, which can be words, characters, sub-words, or even bytes. There are different tokenization techniques that have been proposed in the literature, each with its own advantages and limitations.

Character-level tokenization, for example, can handle rare and out-of-vocabulary words effectively, but is computationally expensive and can result in a large number of tokens. Word-level tokenization is computationally efficient, but may not handle rare or out-of-vocabulary words effectively. Sub-word tokenization combines the advantages of both character-level and word-level tokenization, and has become a popular choice in many NLP tasks.

In addition, recent research has proposed novel tokenization techniques such as gradient-based sub-word tokenization and token dropping to address the challenges of processing large amounts of text efficiently. These techniques show promise in improving the efficiency of NLP models while maintaining their accuracy.

In conclusion, tokenization is a critical step in NLP pre-processing, and different tokenization techniques have been proposed to address the challenges of processing large amounts of text effectively and efficiently. The choice of tokenization technique depends on the specific NLP task, the size of the text corpus, and the computational resources available.

## References

1. Qiu, D. and Yang, B., Text summarization based on multi-head self-attention mechanism and pointer network. *Complex Intell. Syst.*, 8, 555–567, 2022, <https://doi.org/10.1007/s40747-021-00527-2>.
2. Li, Z., Peng, Z., Tang, S., Zhang, C., Ma, H., Text summarization method based on double attention pointer network. *IEEE Access*, 4, 1–1, 2016.
3. Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., Liu, P.J., Calibrating sequence likelihood improves conditional language generation. *ArXiv*, 2022, <https://doi.org/10.48550/arXiv.2210.00045>.
4. Clark, J.H., Garrette, D., Turc, I., Wieting, J., CANINE: Pre-training an efficient tokenization-free encoder for language representation. 5, 11 March, 2021.
6. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S., FNet: Mixing tokens with fourier transforms, NAACL, in: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, 2022.
7. Tay, Y., Tran, V.Q., Ruder, S., Gupta, J., Chung, H.W., Bahri, D., Qin, Z., Baumgartner, S., Yu, C., Metzler, D., Charformer: Fast character transformers via gradient-based subword tokenization, ICLR, 2022.
8. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan,