

**SEMANTIC INFORMATION EXTRACTION USING
DEEP LEARNING FOR AGRICULTURAL DOMAIN**

A thesis submitted to the
University of Petroleum and Energy Studies

For the Award of
Doctor of Philosophy
in
Computer Science and Engineering

By
SUNIL KUMAR

December 2020

Supervisor(s)
Dr. HANUMAT G. SASTRY
Dr. VENKATADRI M.



School of Computer Science
University of Petroleum and Energy Studies
Energy Acres, P.O. Bidholi via Prem Nagar,
Dehradun, 248007:Uttarakhand, India.

**SEMANTIC INFORMATION EXTRACTION USING
DEEP LEARNING FOR AGRICULTURAL DOMAIN**

A thesis submitted to the
University of Petroleum and Energy Studies

For the Award of
Doctor of Philosophy
in
Computer Science and Engineering

By
SUNIL KUMAR

December 2020

Supervisor
Dr. HANUMAT G. SASTRY
Professor, School of Computer Science
University of Petroleum & Energy Studies

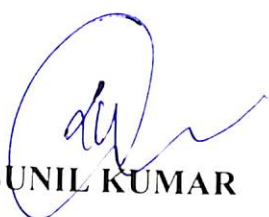
Co-Supervisor
Dr. VENKATADRI M.
Professor, Dept of Computer Science & Engineering,
Amity School of Engineering & Technology
Amity University Madhya Pradesh, Gwalior, India



School of Computer Science
University of Petroleum and Energy Studies
Energy Acres, P.O. Bidholi, via Prem Nagar,
Dehradun, 248007: Uttarakhand, India

DECLARATION

I declare that the thesis entitled “**Semantic Information Extraction Using Deep Learning for Agricultural Domain**” has been prepared by me under the guidance of Dr. Hanumat Sastry G., Professor at School of Computer Science, University of Petroleum & Energy Studies, Dehradun, India and Dr. Venkatadri M., Professor, Department of Computer Science & Engineering, Amity University, Madhya Pradesh, Gwalior, India. No part of this thesis has formed the basis for the award of any degree or fellowship previously.



SUNIL KUMAR

School of Computer Science,

University of Petroleum & Energy Studies,

Bidholi via Prem Nagar, Dehradun, UK, INDIA

DATE: 14 May 2021

CERTIFICATE

I certify that **Sunil Kumar** has prepared his thesis entitled “*Semantic Information Extraction Using Deep Learning for Agricultural Domain*”, for the award of the Ph.D. degree of the University of Petroleum & Energy Studies, under our guidance. He has carried out work at the School of Computer Science, University of Petroleum & Energy Studies.

Supervisor



Dr. Hanumat G. Sastry
Professor

School of Computer Science,
University of Petroleum & Energy Studies,
Bidholi, via Prem Nagar,
Dehradun, Uttarakhand, INDIA



AMITY UNIVERSITY

MADHYA PRADESH

Established vide Government of Madhya Pradesh Act No. 27 of 2010

Ref: ASET/AUMP/2021/009

Date: 09 Aug 2021

CERTIFICATE

I certify that the thesis entitled "*Semantic Information Extraction Using Deep Learning for Agricultural Domain*" by **Sunil Kumar (SAP ID: 500024316)**, a research scholar at the University of Petroleum and Energy Studies, Dehradun, submitted thesis in partial completion of the requirements for the award of the Degree of Doctor of Philosophy in School of Computer Science is an original work carried out by him under my supervision and guidance. It is certified that the work has not been submitted anywhere else for the award of any other diploma or degree of this or any other University.

External Supervisor

DR. VENKATADRI M.

Professor

Head of the Department- Computer Science and Engineering

Amity School of Engineering and Technology

Amity University, Madhya Pradesh



ABSTRACT

Information extraction is a special class of Natural Language Processing techniques. It comprises extraction of entities and events from unstructured data, whereas semantic information extraction automatically mines meanings from extracted entities and their relationships. The explosion of unstructured data made it essential to distill significant information. In most cases, this unstructured data exists in the form of natural language. This has to prompt the development of semantic information extraction techniques to process natural free-text corpus and get the significant information in a structured form.

The traditional information extraction approach involved manually designed large feature sets from various information extraction problems. Whereas deep learning is a branch of machine learning techniques that uses multiple hidden layers in a deep network to find out the essential data automatically. The thesis explores successfully a different approach for semantic information extraction by implementing deep learning to automatically represent the learning process, especially for the agricultural domain.

The thesis presents a detailed introduction to information extraction and its importance in the agricultural area. Further, a detailed literature review has been conducted based on information extraction along with deep learning techniques for the various domains. The survey found some existing techniques proved better in Named Entity Recognition while others in Event Extraction. Furthermore, it has been observed that exponential data growth requires a model that can absorb multiple corpora and gives desired results. And hence, we proposed a novel framework for semantic information extraction using the deep learning technique. The proposed framework accepts different natured corpora and converts them into a single unified corpus using data normalization techniques and a data disambiguation algorithm for Word Sense Disambiguation (WSD). Further, to

extract named entity recognition and event extraction for the agricultural domain, deep learning-based Long Short Term Memory (LSTM) and Rectified Adam Optimizer (RAO) were used. Apache Solr and Lucene tools were used to improve the overall efficiency of the proposed model. The experimental results were compared with the Weighted Self Organizing Map (SOM) and Ensemble Neural Network (ENN) and it was observed that the proposed algorithm outperforming with 1.09% & 1.32% on the accuracy, 1.09% & 1.01% on sensitivity, and 1.11% & 1.00% on specificity as compared to weighted SOM and ENN respectively. The value of the Nash-Sutcliffe Efficiency Coefficient was 0.99.

In a gist, the present research program “Semantic Information Extraction using Deep Learning Technique for Agricultural Domain” provided a framework along with approaches to leverage the benefits of information extraction and semantic in between, from the agricultural domain using deep learning-based LSTM and RAO.

ACKNOWLEDGMENT

Ph.D. journey is truly a life-changing experience and a confluence of multiple learning for me. It is not possible to do without the support and guidance that I received from many people. I am extremely thankful to all of them.

I express my heartfelt gratitude to my research supervisor Dr. Hanumat G. Sastry for his unconditional support and motivation during the entire research program. His guidance always lent a hand to me not only in research but in real life also. Under his guidance, the quality of my research work has continuously improved. He taught me how to understand a problem (in various aspects) and exploring the best possible solutions. His simplicity, positivity, honesty, and discipline inspire me in every second of my life. I wish and try to be as honest and disciplined as he is, for my entire life. Without his continuous support, guidance, and constant feedback, this Ph.D., would not have been achievable.

I express my deep sense of gratitude to my co-supervisor Dr. Venkatadri M. who guided me not only like a mentor but a true friend also. During all the discussions, his suggestions and instructions were undoubtedly helped me a lot. Because of him, I am pursuing my doctoral degree. I am feeling blessed and amazingly fortunate to have a mentor like him, who gives me the freedom to explore various aspects of research in my way. I am indebted to him for always supporting me with this patience and encouragement.

I gratefully acknowledge the Indian Meteorological Department, Dehradun for sharing the weather corpus for my research work. I am thankful to the officials of IMD, Dehradun especially to Mr. Manmohan Salpani, Mr. Gaurav Negi, and Mr. Bikarm Singh for helping me assist in my research work with their valuable suggestions.

This Ph.D. study would not have been possible without the support extended by Dr. Monit Kapoor, who helped me always with his wise advice, useful discussions,

and comments. He always motivated me with his positive inputs and always supported me whenever I got stuck.

My deep appreciation goes to Dr. Divya Srivastava, Dr. Madhushi Verma, Mr. Ankit Khare, Ms. Ambika Aggarwal, and Mr. Prem Kumar Ch. for their much need support during this Ph.D. program. They always so helpful and provided me with their assistance throughout my research journey.

I am blessed with a beautiful and supportive family, I am unable to express my gratitude in words to my father, mother, elder brother (father figure) Mr. Savdesh Sharma, Ms. Savita Sharma, younger brother Mr. Kanwar Pal, Ms. Nidhi Sharma, my nephews and niece. I would also like to say a heartfelt thank you to Gayatri Sharma and two naughty sons Lavya and Chakshu for always believing in me and encouraging me to follow my dreams.

My Wife, who has been by my side throughout this Ph.D., living every single minute of it, and without whom, I would not have had the courage to embark on this journey. She holds my hand even though in her absence. She is the only secret to my happiness and gave meaning to my life. It's not easy to describe my feelings in words for the woman who gives every day the toughest efforts to keep motivated.

Above all, I owe it to all almighty GOD for granting me the wisdom, health, and strength to undertake this research task and to shower his blessings on me to complete my thesis.

SUNIL KUMAR

TABLE OF CONTENTS

DECLARATION	ii
CERTIFICATE	iii
ABSTRACT	v
ACKNOWLEDGMENT	vii
TABLE OF CONTENTS	ix
LIST OF TABLES	xvi
LIST OF ABBREVIATIONS	xviii
LIST OF PUBLICATIONS	xxii
1 INTRODUCTION AND MOTIVATION.....	1
1.1 Natural Language Processing	3
1.2 Information Extraction	5
1.3 Information Extraction and Allied Fields.....	7
1.3.1 Information Retrieval	7
1.3.2 Information Summarization	8
1.3.3 Data Mining.....	9
1.3.4 Text Mining.....	11
1.3.5 Content Extraction.....	11
1.3.6 Terminology Extraction	11
1.3.7 Ontology Learning	12
1.3.8 Text Analytics	12
1.4 Tasks of Information Extraction.....	13
1.4.1 Named Entity Recognition	13

1.4.1.1	Named-Entity typology	14
1.4.1.2	Types of NER Systems	16
1.4.2	Event Extraction	17
1.4.2.1	Event Extraction	19
1.4.2.2	Event Annotation	20
1.4.2.3	Event Extraction System	21
1.4.3	Relation Extraction	22
1.4.3.1	Typology of Relations	22
1.4.3.2	Relation Detection Systems	23
1.4.3.3	Template Filling	23
1.5	Application Areas of Information Extraction	24
1.5.1	Importance of Information Extraction in Agricultural Domain	25
1.5.2	Information Extraction Challenges in Agricultural Domain	25
1.6	Deep Learning	26
1.6.1	Deep Learning Methods	28
1.6.2	Deep Learning Framework	31
1.6.3	Common Deep Learning Algorithms	32
1.6.3.1	CNN- GAN Algorithms	32
1.6.3.2	Recurrent Neural Networks with Long Short-Term Memory	34
1.7	Research Gap and Direction	34
1.7.1	Problem Definition	34
1.7.2	Research Objective	35
1.8	Thesis Navigation	35
2	LITERATURE SURVEY	36
2.1	Data Definition Framework	36

2.2	Significance of Information Extraction	37
2.3	Various Methods of Information Extraction	38
2.4	Information Extraction-based on NER Approaches.....	39
2.4.1	Pattern-based Approach	40
2.4.2	Rule-based Approach	40
2.4.3	Statistical Learning Approach.....	40
2.5	Information Extraction-based on Event Extraction.....	41
2.5.1	Supervised Learning Approach	41
2.5.2	Unsupervised Learning Approach.....	42
2.5.3	Reinforcement Learning Approach	43
2.6	Existing NER Techniques	48
2.7	Existing Event Extraction Techniques	52
2.8	Summary	65
3	PROPOSED FRAMEWORK FOR SEMANTIC INFORMATION EXTRACTION	66
3.1	Introduction	66
3.2	Existing Frameworks of Information Extraction.....	67
3.3	Tools & Techniques used in Proposed Framework.....	67
3.3.1	Preprocessing Methods.....	68
3.3.1.1	Knowledge-based WSD	68
3.3.1.2	Corpus-based WSD	68
3.3.1.3	Tools for WSD.....	69
3.3.2	Long Short Term Memory	70
3.3.3	Adam Optimizer	72

3.3.4	Post-Processing Rules	74
3.4	Proposed Semantic Extraction Framework for Agricultural Corpus	74
3.4.1	Corpora Concatenation.....	76
3.4.1.1	Corpora Collection	76
3.4.1.2	Corpora Amalgamate.....	80
3.4.2	Deep Network-based NER and Event Extraction	80
3.4.2.1	Semantic Extraction.....	80
3.4.2.2	Post-processing	81
3.5	Summary	81
4	DATA PREPROCESSING	82
4.1	Introduction	82
4.2	Corpora Collection	82
4.2.1	Weather Corpus.....	83
4.2.2	Soil Corpus.....	85
4.2.3	Agricultural Corpus.....	87
4.2.4	Pest and Fertilizer Corpus	88
4.3	Data Preprocessing	90
4.4	Corpora Concatenation.....	91
4.4.1	Integration of KB-WSD System into CB-WSD System.....	92
4.4.2	Proposed Disambiguation Algorithm.....	93
4.5	Summary	96
5	NAMED ENTITY RECOGNITION AND EVENT EXTRACTION	97
5.1	Introduction	97
5.2	Existing Rules for AGNER and AGEE.....	98

5.3	Information Extraction using Deep Learning.....	99
5.3.1	Proposed Algorithm for Semantic Information Extraction using LSTM-RAO.....	99
5.3.2	Deep learning Technique for AGNER	101
5.3.3	Deep Learning Technique for AGEE	104
5.3.4	Apache Solr for AGNER and AGEE	105
5.3.5	Semantic Information Extraction using Deep Learning.....	107
5.4	Summary	108
6	RESULTS AND DISCUSSION	109
6.1	Experimental Setup and Design	109
6.2	Results of Proposed Model.....	110
6.3	Comparison with Existing Methods	112
6.3.1	Parameter Metrics.....	112
6.3.2	Confusion Matrix and Other Observations	114
6.3.3	Cross-Validation.....	119
6.3.4	Comparative Analysis	122
6.4	Summary	123
7	CONCLUSION AND FUTURE DIRECTION	124
7.1	Summary and Contributions.....	124
7.2	Research Contributions	125
7.3	Future Research Directions	126
	REFERENCES	128

LIST OF FIGURES

Figure 1.1 Approach of Information Extraction from Unstructured Data	2
Figure 1.2 Idea behind the Turing Test.....	3
Figure 1.3 Evolution of Natural Language Processing.....	4
Figure 1.4 The Idea behind Information Extraction	6
Figure 1.5 Information Retrieval System.....	8
Figure 1.6 Automatic Information Summarization System.....	9
Figure 1.7 Architecture of Data Mining	10
Figure 1.8 Overview of Text Mining.....	11
Figure 1.9 Traditional Information Extraction Framework	13
Figure 1.10 SoNaR Named Entity Typology.....	15
Figure 1.11 Map of Uttarakhand State.....	18
Figure 1.12 Output of Sentence in TimeML.....	21
Figure 1.13 Ballot-Paper Template Filling	24
Figure 1.14 Relation between AI and Deep Learning	27
Figure 1.15 Working Model of Deep Learning Architecture	27
Figure 1.16 Working Model of Convolutional Neural Network	33
Figure 1.17 Working Model of Generative Adversarial Network.....	33
Figure 2.1 Types of Data Formats	36
Figure 2.2 Hierarchical Representation of Information Extraction Techniques...	39
Figure 2.3 Supervised Learning Approach	42
Figure 2.4 Unsupervised Learning Approach.....	43
Figure 2.5 Reinforcement Learning Approach	44
Figure 3.1 Structure of Long Short-Term Memory Deep Network.....	70
Figure 3.2 Proposed Framework for Semantic Information Extraction using Deep Learning Technique for Agricultural Domain.....	75
Figure 4.1 Output of KB-WSD System into CB-WSD System using NLTK Functions	92

Figure 4.2 Relation between Cosine Distance (d) and Cosine Similarity (Θ)	94
Figure 5.1 Extracted AGNER using Proposed Algorithm.....	103
Figure 5.2 Architecture of Apache Solr System	105
Figure 5.3 Use of Apache Solr Framework for Agricultural Information Extraction	106
Figure 5.4 Semantic Information Extraction from Unstructured Unified Input Corpus	108
Figure 6.1 Output-1 of Agricultural-based Named Entity Extraction	111
Figure 6.2 Output-2 of Agricultural-based Named Entity Extraction	111
Figure 6.3 Output of Agricultural-based Event Extraction.....	112
Figure 6.4 Confusion Matrix.....	114
Figure 6.5 Graphical Representation of Nash-Sutcliffe Efficiency Coefficient .	115
Figure 6.6. Accuracy of the Proposed LSTM-RAO Method.....	116
Figure 6.7. Precision Value of the Proposed LSTM-RAO Method.....	116
Figure 6.8. Recall Value of the Proposed LSTM-RAO Method	117
Figure 6.9. F-score Value of the Proposed LSTM-RAO Method.....	118
Figure 6.10. Cross-Validation of Proposed LSTM-RAO Method.....	119
Figure 6.11. Precision, Recall, and F-score Analysis of the Proposed LSTM-RAO in Cross-validation	120
Figure 6.12 Performance of the Proposed LSTM-RAO Method in Crop Yield Prediction.....	122

LIST OF TABLES

Table 1.1 Related Fields of Information Extraction	12
Table 1.2 Various Event Ontologies	19
Table 1.3 Relation Ontologies Adopted from MUC-1 to MUC-7	22
Table 1.4 Advantages and Disadvantages of Deep Learning Methods	30
Table 1.5 Comparison of Deep Learning Framework	32
Table 2.1 Strength and Weakness of Information Extraction Sub-categories	44
Table 2.2 Survey on Information Extraction Tasks in Various Application Areas	55
Table 2.3 Survey on Agricultural Domain in related to Deep learning Techniques	57
Table 2.4 Survey on Single-Solution and Population-based Metaheuristic Approaches	63
Table 3.1 Recommended Values of Default Parameters	74
Table 3.2 Sample Data for Weather Corpus	77
Table 3.3 Zinc (Zn) Recommendation for Corn Crop Production.....	78
Table 3.4 Sample Data for Seasonal-based Crop Productivity.....	78
Table 3.5 Usage of Pesticide Classified by Crop Types	79
Table 4.1 Sample Data of Weather Corpus-IMD, Dehradun.....	83
Table 4.2 Soil Health Card Issued by DACFW	86
Table 4.3 Using DPTA Extractable Zn Extraction Method, Zn Recommendation for Corn Crop Production.....	87
Table 4.4 Sample Data of Seasonal-based Normal Rainfall.....	87
Table 4.5 Sample Data for Seasonal-based Crop Productivity.....	88
Table 4.6 Fertilizer-based Comparison between Requirement, Availability, and Sales of Kharif Crop.....	89
Table 4.7 Usage of Pesticide Classified by Crop Types	89
Table 4.8 Output of Min-Max Algorithm Applied on Weather Corpus	90

Table 4.9 Output of Proposed Disambiguation Algorithm.....	95
Table 5.1 Types of Tags, Tags Descriptions, and Suitable NER Examples.....	102
Table 6.1 Basic Requirements for Experimental Setup and Design.....	109
Table 6.2 Result of Proposed Disambiguation Algorithm.....	110
Table 6.3. Performance of the Proposed LSTM-RAO Method in Crop Yield Prediction.....	121
Table 6.4 Comparison of Existing Techniques with Proposed Algorithm	123

LIST OF ABBREVIATIONS

Acronym	Meaning of Abbreviation
WSD	Word Sense Disambiguation
LSTM	Long Short Term Memory
RAO	Rectified Adam Optimizer
ICT	Information and Communication Technologies
IMD	India Meteorological Department
NER	Named Entity Recognition
IE	Information Extraction
CNN	Convolutional Neural Network
GAN	Generative Adversarial Networks
EE	Event Extraction
CB	Corpus-Based
AGEE	Agricultural Event Extraction
ML	Machine Learning
AI	Artificial Intelligence
NLTK	Natural Language Toolkit
GUI	Graphical User Interface
MUC	Message Understanding Conferences
DACFW	Department of Agriculture, Co-operation and Farmers Welfare
DPTA	Diethylene Triamine Penta-acetic Acid
GDP	Gross Domestic Product
NLP	Natural Language Processing
IR	Information Retrieval
IS	Information Summarization
ACE	Automatic Content Extraction
RE	Relation Extraction
DARPA	Defense Advanced Research Projects Agency
TIMEX	Temporal Expression
NUMEX	Numeric Expressions
ENAMEX	Entity Names Expression
CoNLL	Conferences on Computational Natural Language Learning
ACL	Association for Computational Linguistics
PER	Person Name
ORG	Organization Name

Acronym	Meaning of Abbreviation
LOC	Location
MISC	Miscellaneous Name
EVE	Event
PRO	Product
SVM	Support Vector Machine
MaxEnt	Maximum Entropy
MBL	Memory-Based Learning
HMM	Hidden Markov Models
DL	Deep Learning
RNN	Recurrent Neural Networks
EDI	Electronic Data Interchange
SQL	Structured Query Language
PBNER	Pattern-Based Named Entity Recognition
NE	Named Entity
SLNER	Statistical Learning Approach based Named Entity Recognition
MLE	Maximum Likelihood Estimation
SLA	Supervised Learning Approach
SLEE	Supervised Learning approach for Event Extraction
FBCM	Feature-Based Classification Method
KLM	Kernel Learning Model
ULEE	Unsupervised Learning approach for Event Extraction
CA	Clustering Approach
OIE	Open Information Extraction
WIA	Wrapper Induction Approach
RLEE	Reinforcement Learning for Event Extraction
LA	Learning Agent
BA	Bootstrapping Approach
GBA	Graph-Based Approach
POS	Part-Of-Speech
NERC	Named Entity Recognition and Classification
AGROVOC	A portmanteau of Agriculture and Vocabulary
GCM	Global Climate Model
ANN	Artificial Neural Network
SLR	Stepwise Linear Regression

Acronym	Meaning of Abbreviation
MPE	Mean Percent Error
SMLR	Standard Mixed Lymphocyte Reaction
PCA	Principal Component Analysis
ACC	Australian Community Climate
ACCESS	Australian Community Climate and Earth-System Simulator-Seasonal
PSO	Particle Swarm Optimization
MI	Mutual Information
MRA	Multi-Resolution Analysis
MODWT	Maximal Overlap Discrete Wavelet Transform
CSSI	Compound Specific Stable Isotope
EWE	Extreme Weather Events
GHS	Greenhouse Gas Emission
KNN	k-nearest neighbors algorithm
FCN	Fully Convolutional Network
GBIF	Global Biodiversity Information Facility
BPNN	Back Propagation Neural Network
NN	Neural Network
COCO	Common Objects in Context
RCNN	Region-based Convolutional Neural Networks
BLSTM	Bidirectional Long Short-Term Memory
EA	Evolutionary Algorithms
GA	Genetic Algorithms
SA	Simulated Annealing
TS	Tabu Search
ACO	Ant Colony Optimization
FA	Firefly Algorithms
OWL	Web Ontology Language
RDF	Resource Description Framework
MRD	Machine-Readable Dictionaries
RDFS	Resource Description Framework Schema
DAML	DARPA Agent Markup Language
DAPRA	Defense Advanced Research Projects Agency
OIL	Ontology Inference Layer
AdaGrad	Adaptive Gradient Algorithm

Acronym	Meaning of Abbreviation
SHOE	Simple HTML Ontology Extensions
NLTK	Natural Language Toolkit
SGD	Stochastic Gradient Descent
RMSProp	Root Mean Square Propagation
CEC	Cation Exchange Capacity
OM	Organic Matter
DACFW	Department of Agriculture, Cooperation & Farmers Welfare
NFL	National Fertilizer Limited
MAFW	Ministry of Agriculture and Farmers Welfare
KVK	Krishi Vigyan Kendra
PAF	Pest and Fertilizer
AGNE	Agricultural Named Entity
HDFS	Hadoop Distributed File Storage
TP	True Positive
FN	False Negative
TN	True Negative
MAE	Mean Absolute Error
ENN	Extension Neural Network
DRN	Deep Residual Network

LIST OF PUBLICATIONS

Journal Publications

1. Sunil Kumar and Hanumat Sastry G., “ *A Comprehensive Review of Semantic Information Extraction using Deep Learning Technique for Agricultural Domain*”, International Journal of Grid and Distributed Computing, vol. 13, no. 1, pp. 1219–1231, 2020.
<http://sersc.org/journals/index.php/IJGDC/article/view/21366/10829>
2. Sunil Kumar, Hanumat Sastry G., Venkatadri Marriboyina, Dinesh Goyal, Madhushi Verma., “ *A Novel Deep Learning Approach for Semantic Information Extraction from Medicinal Crops*”, European Journal of Molecular & Clinical Medicine, vol. 7, no. 8, pp. 1363–1378, 2020
https://ejmcm.com/article_4302_406ca7199c4f00c0a3fc9654ca5a05f4.pdf

CHAPTER-1

INTRODUCTION AND MOTIVATION

The agriculture sector contributes a major share in the Indian economy [1]. More than 70% of rural households in India are dependent on agriculture. In India, agriculture is the largest employment-providing sector that provides 60% employment to the total population along with a 17% contribution to the country's Gross Domestic Product (GDP) [2, 3]. Therefore, to improve the quality and quantity of agricultural products, emerging technologies are highly appreciated. By minimizing the manual labor efforts and maximizing overall productivity, modern information management tools and techniques play an incredible role in the agricultural domain. At present, the agriculture sector is transforming towards smart agriculture by utilizing modern Information and Communication Technologies (ICT).

In recent years, there is an exponential growth in data production in the information industry especially in agricultural and its allied sectors. It is everyone's need to extract unswerving information with increasingly non-significant, complicated, large, unclear, and raw data. Government and private organizations also appreciate the importance of extraction of valuable information from free text. Especially, when the technology access to such information more feasible, not only Information Extraction (IE) but also finding the semantics between different corpora is present-day requirements [4, 5]. Figure 1.1 represents the holistic approach of information extraction from an unstructured corpus [6].

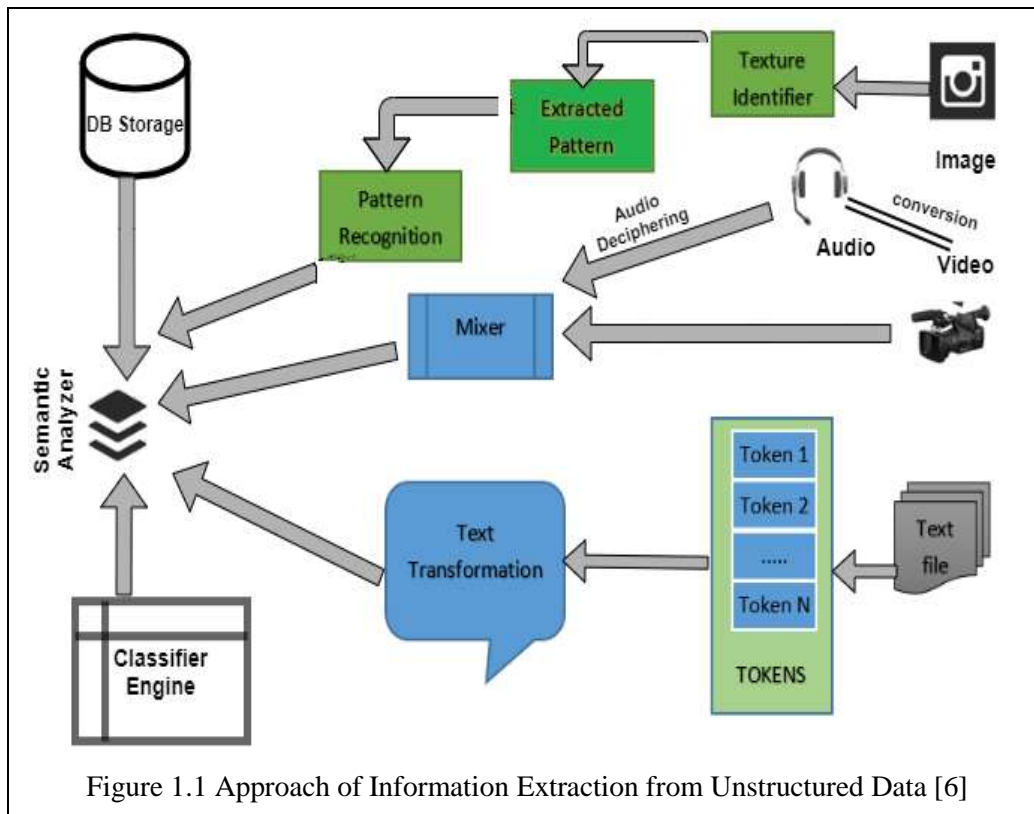


Figure 1.1 Approach of Information Extraction from Unstructured Data [6]

Among various existing machine learning techniques, deep learning emphasized the computational part of the corpora of different fields including the agricultural domain. The agricultural sector directly and indirectly affected by various factors such as climatic factors (light, water and rainfall, temperature, air, relative humidity, and wind), physical factors (topography/relief, soil and climate affect farming), technological factors (tools & techniques applying on agricultural corpus), social factors (land ownership & inheritance and type of farming in practices) and education/farming knowledge affect farming. Above said factors raise the complexity of agricultural data.

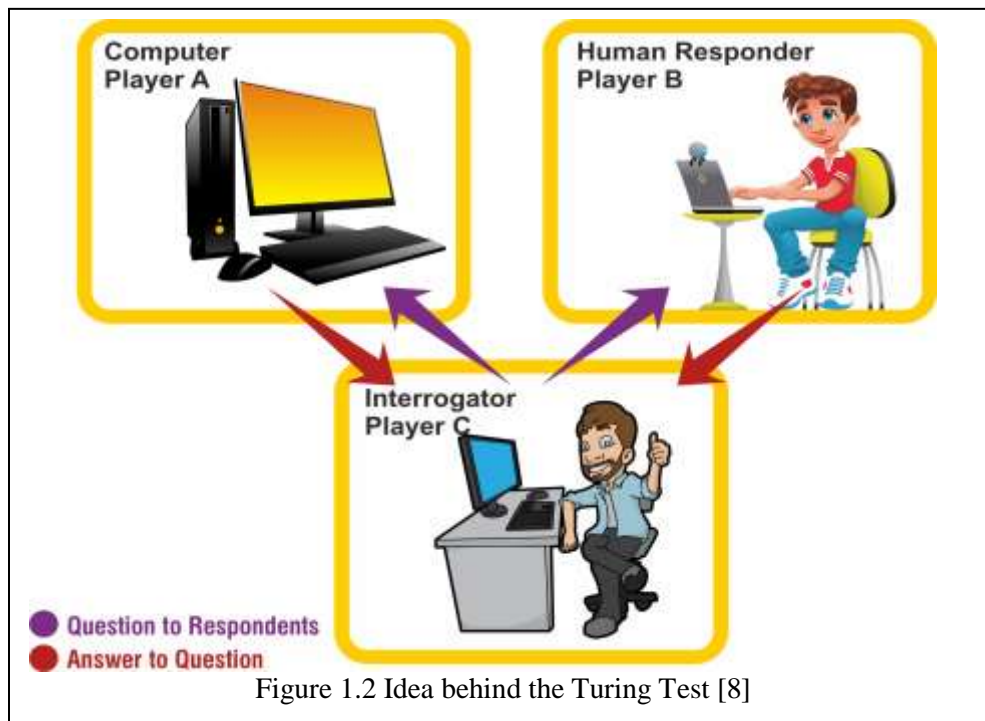
The agricultural dataset comprises semi-structured and unstructured values. To understand the behavior of agricultural corpus where more than 90% of the data is unstructured in nature, various Natural Language Processing (NLP) methods and

fine-tuned machine learning techniques are essential to extract and analyze the agricultural corpus.

The present research program is targeted to establish the relation (semantics) between agricultural, soil, weather, and pest & fertilizer corpora and to extract the information using the deep learning technique. Resulting from this, the farmers may advise with the right crop for better yielding for a particular session with consideration of soil quality. Section 1.1 provides the theoretical fundamentals of NLP and information extraction techniques as defines in section 1.2. In section 1.3, the chapter distinguishes information extraction from other correlated tasks.

1.1 NATURAL LANGUAGE PROCESSING

Turing test (1950) is considered as the first step in NLP and while answering the question, a machine has to make fool the examiner that it is a human being, not a machine [7]. Turing claims that if a machine can pass the test and trick the interrogator then it can consider as “Intelligent” as shown in Figure 1.2 [8].



These intelligent machines were able to understand and produce human-understandable languages that were used in several research areas of computer science. NLP and computational linguistics are significant research areas of computer science [9].

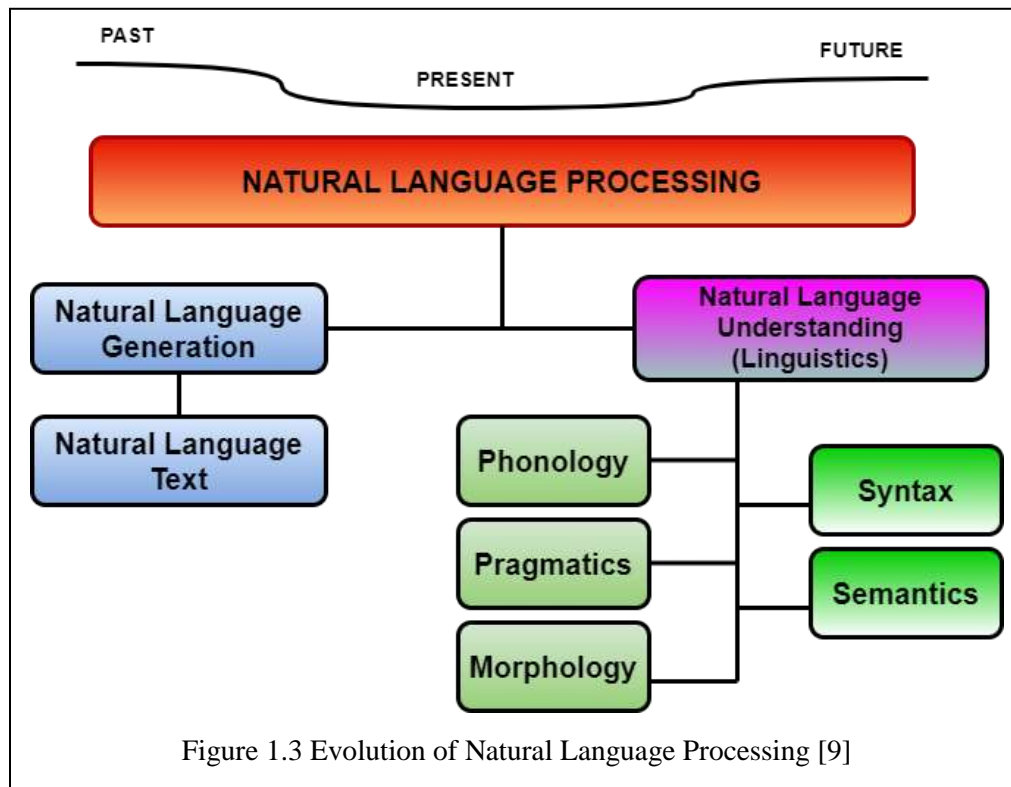


Figure 1.3 presents the evolution of NLP and its subfields [10, 11], the following subsections details it further.

- **Phonology** deals with sound and speech patterns and considers sound as a physical input entity.
- **Pragmatics** deals with the different uses of natural languages.
- **Morphology** studies the word structure of a language and logical relations between them.
- **Syntax** deals with sentence structure.

- **Semantics** is the science that deals with the literal meaning of the words, phrases, and sentences as well.

In a natural language, the communication between two entities can take place through different ways like making gestures, listening, speaking, by specialized hand gestures (while driving or on traffic signals), barrel language for the deaf, and through different forms of text [12]. In day-to-day conversation, textual communication is a very common medium of sending and receiving information from one end to another. This may include document processing, mails transaction, article publication, reports writing, etc. based on a particular language such as English, Hindi, Punjabi, French, Spanish, etc.

Various existing NLP techniques have advantages to analyse the unstructured/semi-structured data and convert it into structured data [13]. The representation of structured data may be as a knowledge base, syntactic base, and semantic text [14].

The use of NLP in various widespread areas such as information retrieval, sentimental analysis, Information extraction, machine translation, and question answering. IE is one of the prominent research areas of NLP. The following section presents the overview of IE

1.2 INFORMATION EXTRACTION

In an IE system, event and entity extraction are major components. In a casual discussion on a politician or a sportsman i.e. entity and any news related to COVID-19 for a particular country is an event. Useful information like an entity – Sachin Tendulkar is one the best batsman and events like Corona positive patients are exponential increases in generally available repositories. The extraction of essential information from unstructured data is always a primary concern in NLP.

Nowadays, large repositories of the semi-structured or unstructured corpus contain a vast amount of substantial information. These repositories can be used

more significantly if they can reliably answer the queries related to the relevant entities such as person name, organization, date, place, etc., and the relationship between these entities such as owner, employee, affiliation, etc.

To explore raw data and extract useful information effectively is very essential task in data mining [15]. The vast amount of data becomes useless without extracting information from it so, it is the primary goal becomes a researcher to extract factual information from unstructured data like online news, medical reports, legal acts, organizational documents, court rulings, social media, sports stats, government documents, etc [16]. The main objective of IE is to extract structured information from a human-readable corpus and to store the extracted information in relational and graph databases so that it can use for future tasks.

The data eruption has made it more crucial while extracting significant entities or relations between entities from a corpus. Most data corpora are unmanaged, unstructured, free-text, scattered documents. Therefore, there is a need to develop technique(s) to extract important information automatically from the unstructured corpus. The distill data from various corpus should be in a structured manner.

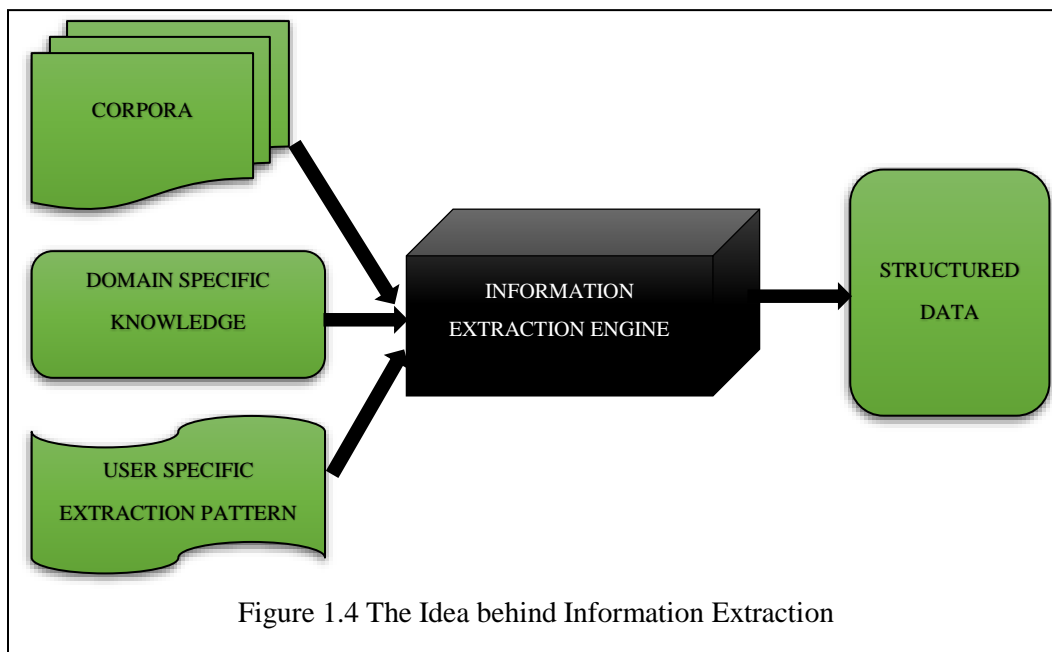


Figure 1.4 The Idea behind Information Extraction

As per Figure 1.4, inputs can be domain-dependent or independent unstructured /semi-structured corpus (or corpora), domain-specific knowledge, and user-specified extraction patterns [17]. IE engine processes the input data to extract knowledge, and save it into a structured database (relational and graph databases).

IE is also considered an endless area of NLP. It is the process of identification and concurrent classification of an unstructured dataset, resulting in semantic classes of specific information [18, 19, 20]. IE classified into two major categories:

- Identification: identify a given text as valuable information
- Classification: putting identified information into a predefined semantic category

The following section highlighted the details of various information processing techniques.

1.3 INFORMATION EXTRACTION AND ALLIED FIELDS

The term IE refers to the process of transformation from unstructured data to organized information. IE should have a separate area of study as compared to its related tasks that overlay in parts with IE but may initiate from other domains, in another period, or from other researcher communities. IE is a most debatable area of NLP because it is closely related to Information Retrieval (IR), Information Summarization (IS), Data Mining, Terminology Extraction, Text Mining, Content Extraction, Ontology Learning, and Text Analytics [21]. These terms are also important to define the introductory point of this thesis and necessary for integration into the NLP tradition, so these terms are briefly described in the following subsections.

1.3.1 INFORMATION RETRIEVAL

Information Retrieval coined parallel with the database system. The searching base of IR is content-based indexing or full text. The searching results may be an

amount of information from a given document, documents themselves, metadata, or databases of images, texts, audio & video. Unlike traditional database systems that are focus on query and transaction processing of organized data values, IR extracts the related documents from a corpus of documents based upon the keywords entered by a user.

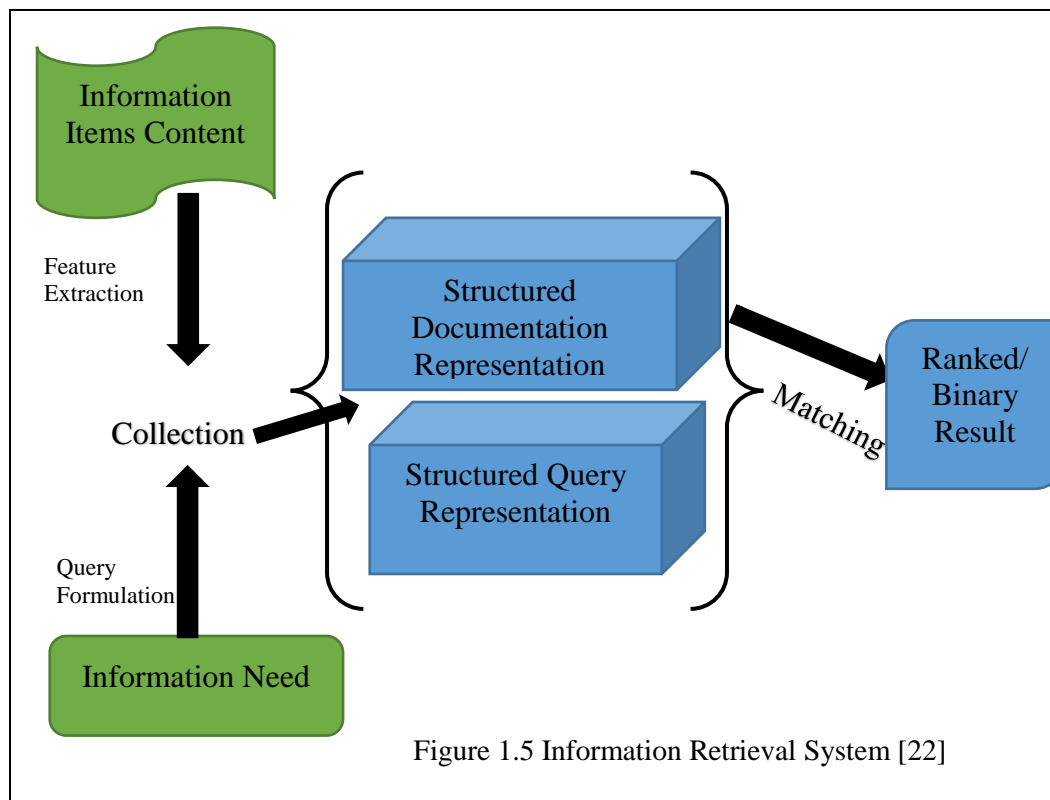


Figure 1.5 Information Retrieval System [22]

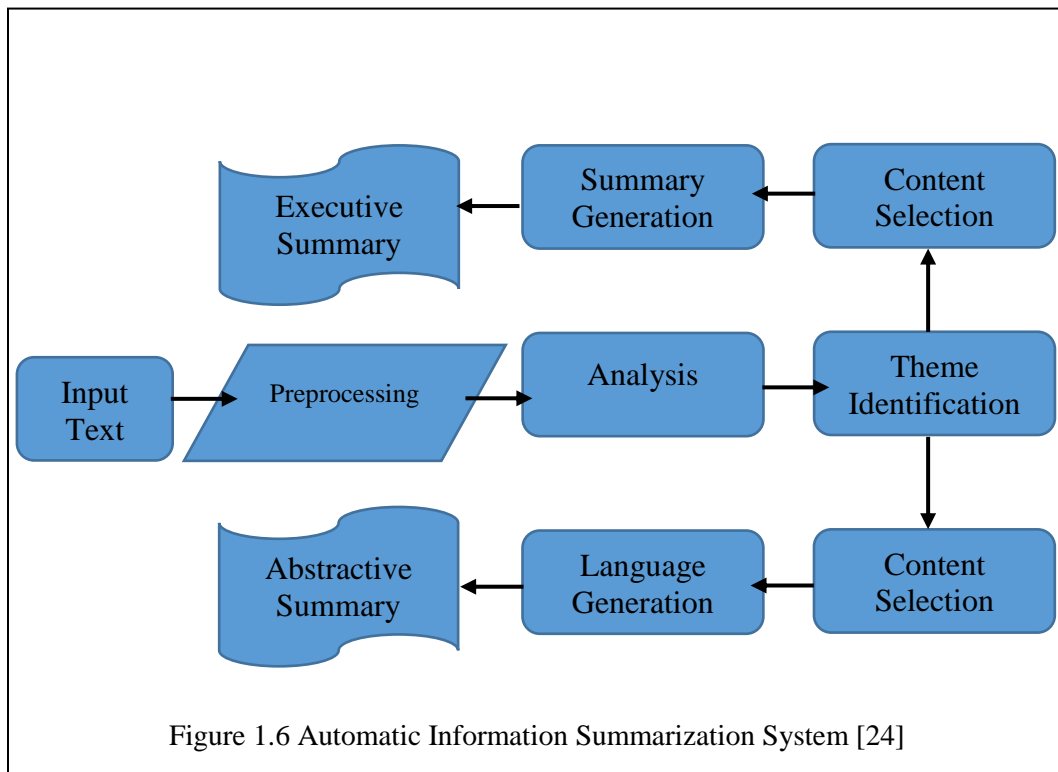
Figure 1.5 shows the typical IR system architecture, where the feature extraction and query formulation are merged and send as input to the IR system [22]. This system assigns the document ranking as per the similarity index of structured documents.

1.3.2 INFORMATION SUMMARIZATION

It is the method to find significant information in the form of a summary from a document or a set of related documents [23]. For a human being, it becomes more

difficult to extract the summary of a large document due to the information overloading problem on the Internet. Therefore, the automation process for information summarization has become important. There are two general approaches for information summarization i.e. Extraction-based Summarization and Abstraction-based Summarization.

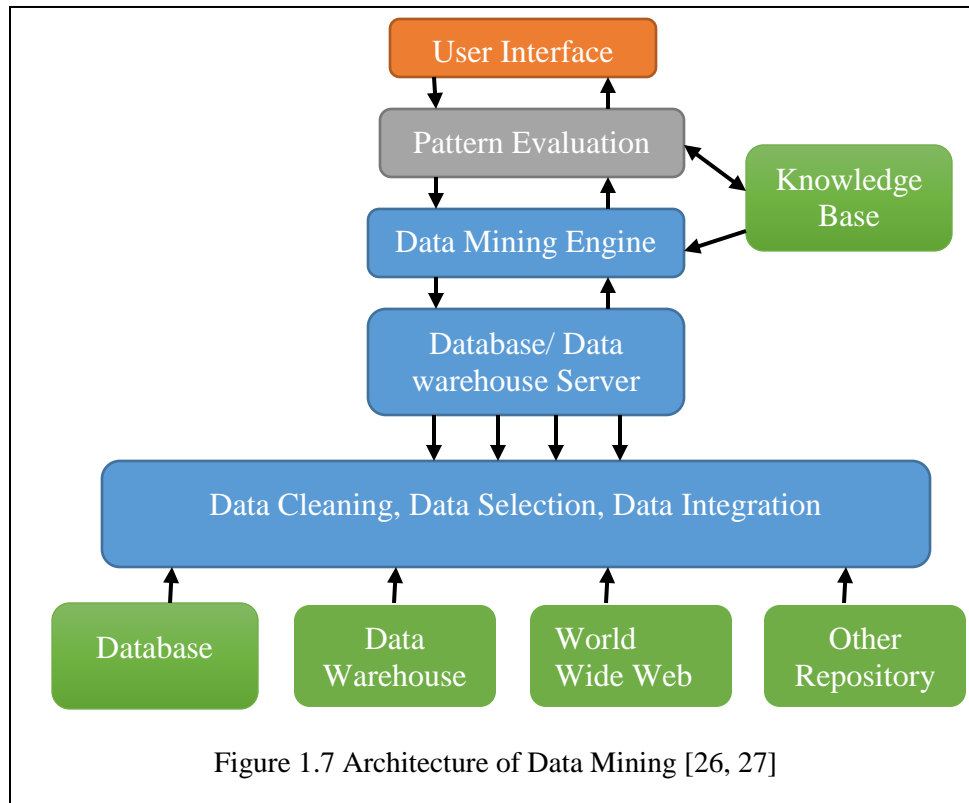
Figure 1.6 presents the process of the Automatic information summarization system. Various steps for information summarization are clearly shown in the following figure [24].



1.3.3 DATA MINING

Data mining extracts some valuable information, valid facts, and useful patterns from a large number of data sets. Data mining involves sophisticated data analysis techniques that can apply to data and discover hidden or earlier unknown meaningful information. Various application areas of data mining are the Insurance

industry, Education sector, Medical industry, Scientific data, Manufacturing, Banking & Finance Sectors, Retail sector, Automobile industry, E-commerce, Super Marketing, Crime investigation, Bioinformatics, Sports, Agricultural sector, and many more [25]. Figure 1.7 Architecture of Data Mining [26, 27] shows the architecture of the data mining system [26, 27].



Data mining tasks are classified into two categories; descriptive data mining tasks define the fundamental properties of existing data and predictive data mining tasks that perform prediction based on inference on available data [28].

Data mining techniques are available in various forms of mining such as web mining, text mining, pictorial data mining, social networks data mining, relational databases, and audio & video data mining [29].

1.3.4 TEXT MINING

Text mining is the mining of meaningful/structured information from semi-structured, unstructured, and uncleaned textual data whereas the manipulation of data is also difficult. This raw data contains a huge amount of facts or information that cannot be directly used for further processing. Therefore, by using the algorithms of computational linguistics, it is easy to extract the expected results from unstructured or semi-structured corpus [30, 31]. A fundamental overview of the text mining technique is explained in the figure below.

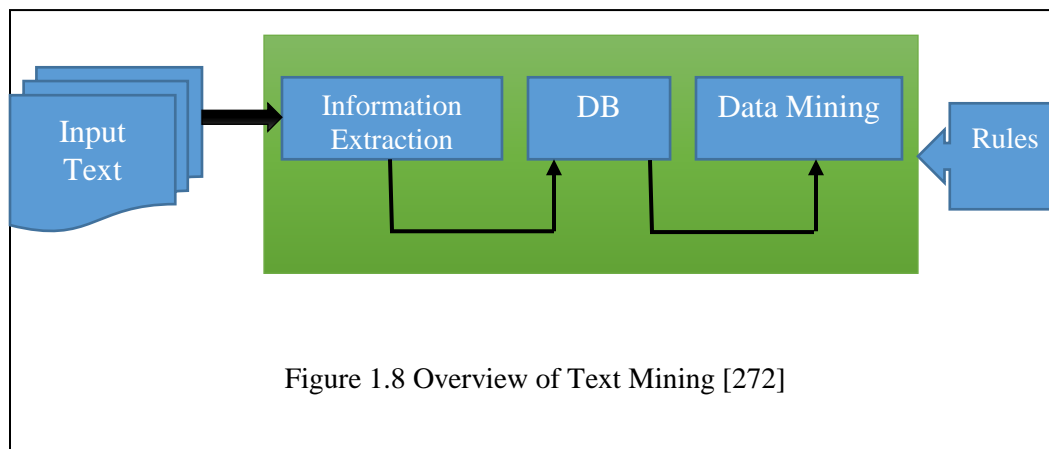


Figure 1.8 Overview of Text Mining [272]

1.3.5 CONTENT EXTRACTION

Word content extraction started using in the campaigns of Automatic Content Extraction (ACE) in the last two decades. Like IE, the ACE also focuses on entities, events, and the relationship between them [32].

1.3.6 TERMINOLOGY EXTRACTION

Terminology extraction is also called terminology mining, term extraction, glossary extraction, and term recognition in NLP [33]. It is interested in common nouns (terms) rather than proper nouns (entities). Language-independent and

multilingual approaches successfully applied to the sub-sentential linguistic alignment and allowed the complex multi-word extraction [34].

1.3.7 ONTOLOGY LEARNING

Ontology learning arises in the early 2000s, also known as ontology generation, ontology extraction, or ontology acquisition. Ontology learning is similar to terminology extraction but for a given domain and it focused on the semi-automatic ontology creation [35].

1.3.8 TEXT ANALYTICS

Text analytics is a famous term in corpus linguistics. Somewhere it is considered as similar to IE whereas few authors consider it is closer to text mining [36, 37]. When merged with tools and techniques of data visualization, text analytics enables organizations to understand the secrets behind the numbers and helps them to make better decisions.

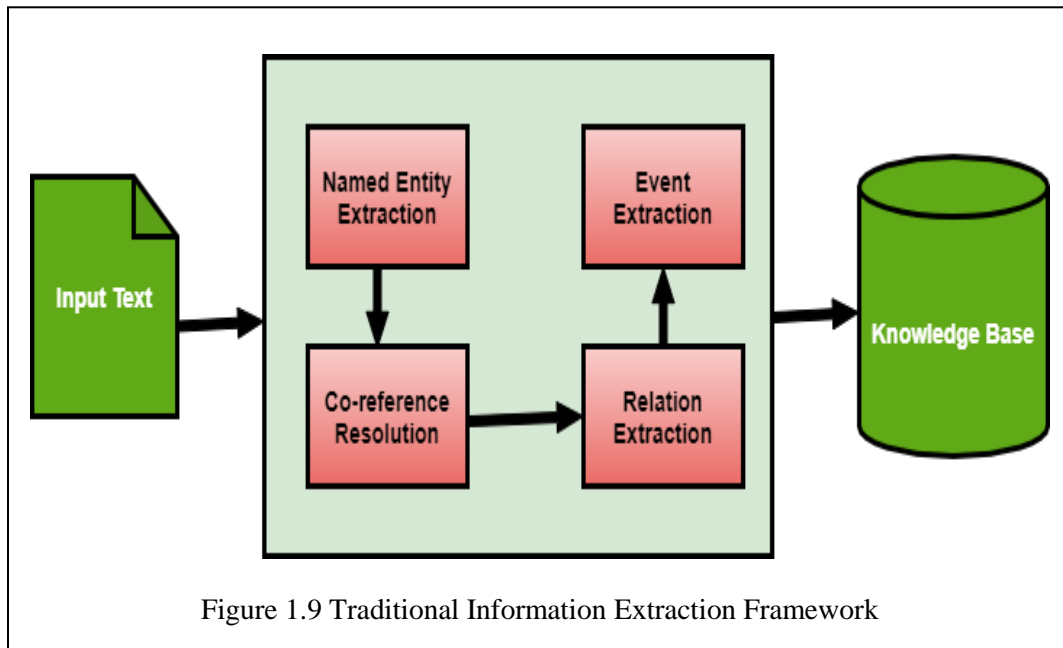
Table 1.1 Related Fields of Information Extraction [38]

Area of Study	Year	Domain of Origin
Information Retrieval	1950	Information Science
Information Summarization	1958	Information Science
Information Extraction	1982	NLP
Data Mining	1989	Computer science
Terminology Extraction	1996	Machine translation
Text Mining	1999	Computational linguistics
Content Extraction	1999	ACE campaign
Ontology Learning	2001	Information science
Text Analytics	2004	Corpus linguistics

Table 1.1 presents a summary of IE-related areas along with an approximate year of arises and the main origin of the domain. The next section discusses various tasks related to the IE.

1.4 TASKS OF INFORMATION EXTRACTION

There are four fundamental tasks in an IE system i.e. used to extract structured information from semi-structured or unstructured data [39]. These tasks are Named Entity Recognition (NER), Co-reference Resolution (CO), Relation Extraction (RE), and Event Extraction (EE). Figure 1.9 depicts the basic IE framework [40]. The following subsections detail it further.



1.4.1 NAMED ENTITY RECOGNITION

It is an essential subtask of information extraction that consists of automatic identification and classification of proper nouns i.e. named entities from an unstructured corpus. NER is to extract the entities like person, organization, location, date, time, numerical, currency, and percentage expressions from the unstructured data. These extracted entities may include additional information e.g., a person may contain name, title, sex, nationality, position, etc. [15] [41]. It is a core component of knowledge discovery and semantic enrichment system and

therefore NER requires more discussion before moving on to other complex tasks of IE.

The typologies used in this section are Named Entities and NER systems. These typologies reduce the flexibility of a NER extraction tool but useful for the evaluation of applications. Therefore, it is highly important to understand the requirements of these classifications by discarding their rigidity.

1.4.1.1 NAMED-ENTITY TYPOLOGY

In addition to the above-said typologies (name, title, sex, nationality, position), there are several other typologies related to NER that have been proposed by researchers from 2013 onwards [42, 43]. There are no fixed and final set of named entities classes and still, it is the point of argument that every NER task has its typologies. Because the researcher's granularity level varies widely e.g., some NER researchers define a crisp difference between mountains, cities, countries, etc. whereas other sets of authors merge all these entities into a single location entity [44]. Below details mentioned vast evaluation programs for developing the widespread typologies.

Message Understanding Conferences (MUC): The MUC-1 to MUC-7(1987-1997) initiated by the Defense Advanced Research Projects Agency (DARPA), aiming to inspire advanced methodologies in the area of IE. A survey paper on MUC-7, reveals an expression division into temporal expressions (TIMEX), numeric expressions (NUMEX), and Entity names (ENAMEX) [45]. Location, organization, and person name are also added in the latter stages.

Conferences on Computational Natural Language Learning (CoNLL): CoNLL was launched in 1997 (ended in 2008), initiated by the Association for Computational Linguistics (ACL) with a focus on natural language learning. The CoNLL 2002-03 tasks were an emphasis on language-independent NER questions. The main objective of the CoNLL conference is the division of named entities into

four categories that were PER (person names), ORG (organization name), LOC (locations), and MISC (miscellaneous names) [46].

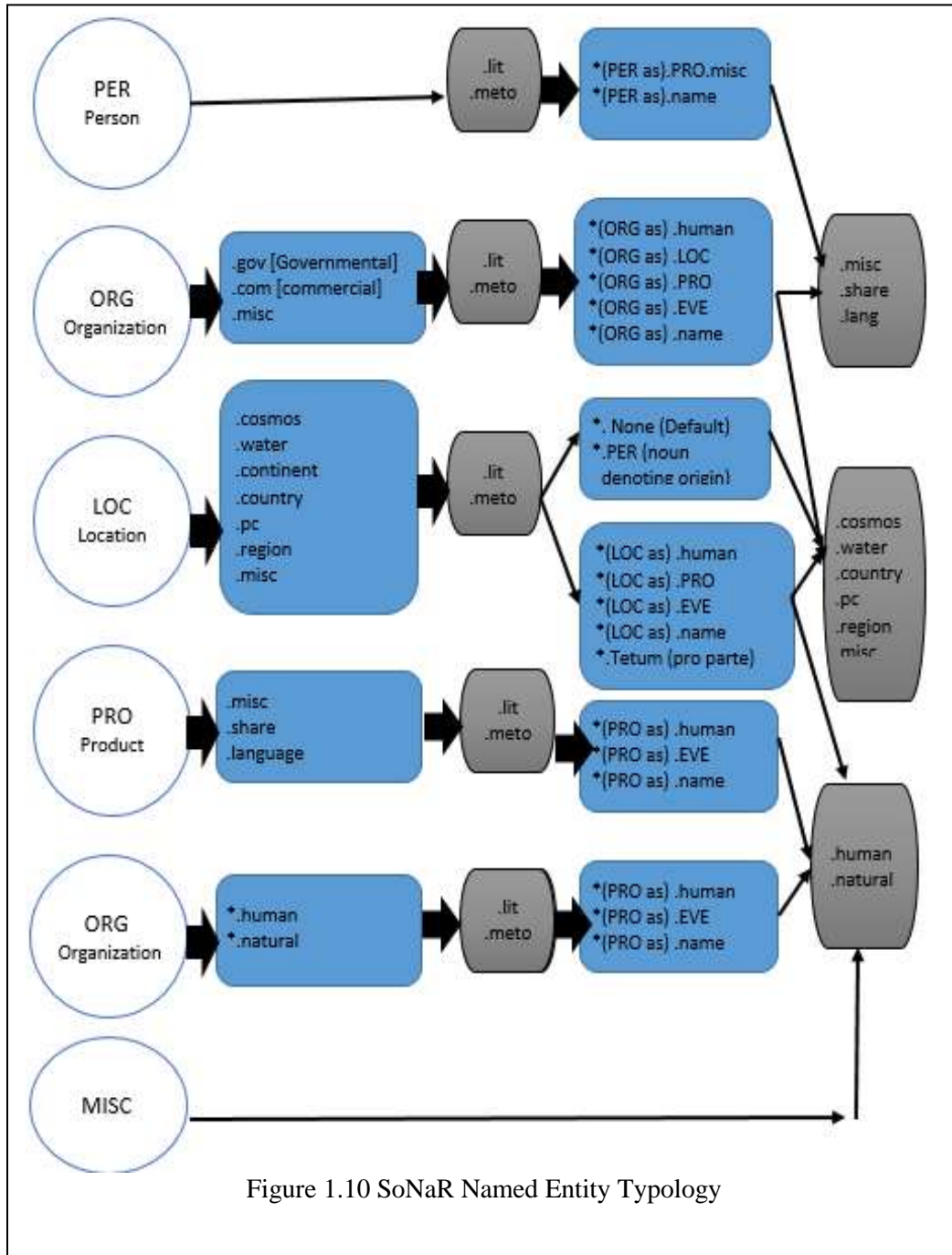


Figure 1.10 SoNaR Named Entity Typology

Automatic Content Extraction (ACE) initiated in 1999 (ended in 2008). The main agenda of ACE was to develop information extraction technologies and focused entities, events, and relations extraction. Like CoNLL, ACE categorized the given entities into seven different types: Person, Organization, Location, Facility, Vehicle, Weapon, and Geo-political entities [47].

Few regional and national programs also started their own specificities e.g the ESTER campaign further added political, military (functions) and award, work of art (Human Production) to PER, LOC, and ORG in France. Similarly, in the Netherlands, the project SoNaR categorizes the named entities into six categories: PER, LOC, MISC, ORG, EVE (events), and PRO (products). Figure 1.10 presents the SoNar NER model [48, 49].

1.4.1.2 TYPES OF NER SYSTEMS

NER classified into three approaches that are linguistic, probabilistic, and hybrid approaches. The following sub-sections show how recent approaches make better robustness and generalizability in the NER system, along with the current challenges faced by NER.

Rule-based or Linguistic Approaches: In the early 90s, based on symbolic, relying on advanced linguistic rules and knowledge base, the first NER system was designed. These rule-based approaches were large time consuming and hard to maintain due to manually encoding by linguists. Gazetteers also used in few linguistic systems help to increase the coverage area of rules that may be in long lists of common locations or first names. Whereas, for some different types of entities (products or organizations), gazetteers are not that much capable of producing a better list with current possibilities. Additionally, to some extent, designed patterns are also incomplete to recognize these entities.

Data-driven or Probabilistic Based Approaches: Due to the availability of large corpora, rules-based approaches shifted towards statistical approaches and had

benefits of machine learning approaches on these huge corpora. Resulting from this, these approaches automatically extracted large listed company names (in thousands) from financial news containing more than one million words, with a 25% more recall rate as compared to human annotators [50]. Due to the lower percentage of precision values in unsupervised approaches, researchers use the semi-structured approaches (manual intervention required in the automated process).

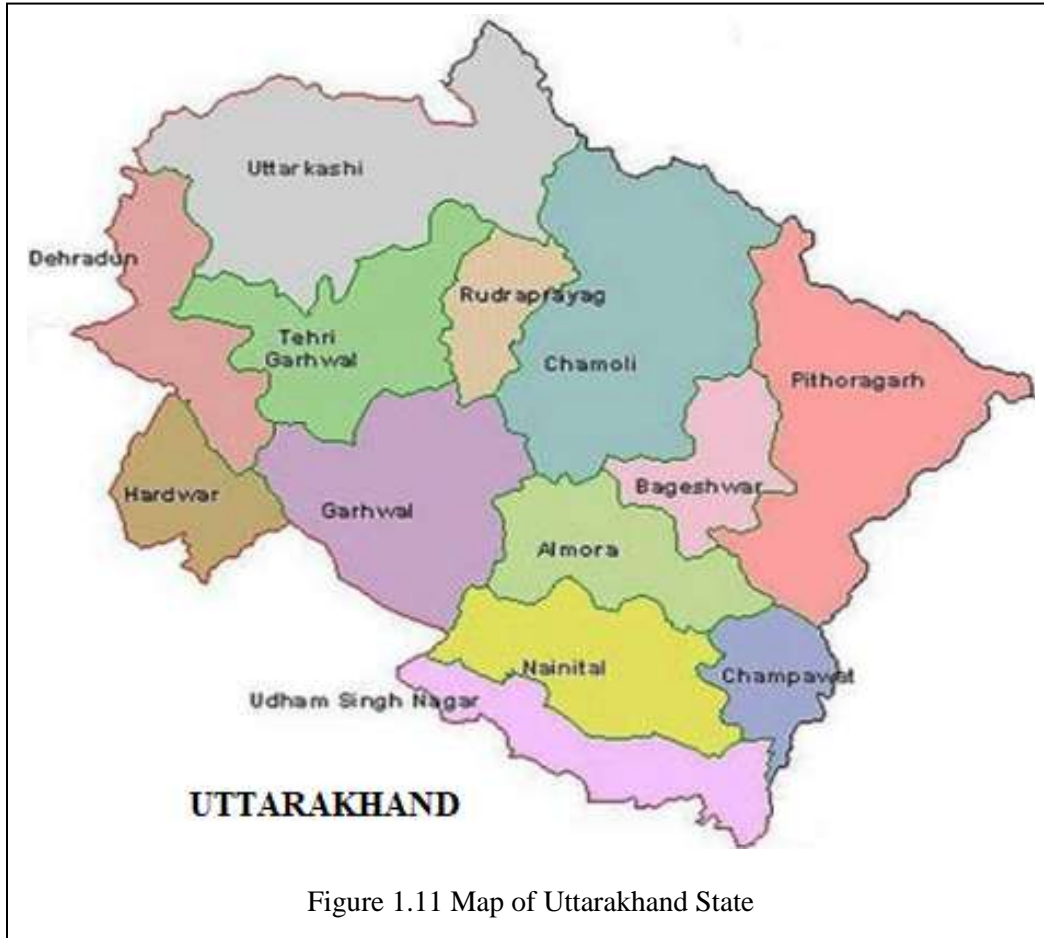
Hybrid Approaches: Due to the benefits of data-driven approaches, most IE systems combine it with linguistic-based approaches for getting better results e.g., maximum entropy [51]. In this IE system, relevant features are extracted by the human being and combining with machine learning approaches where weighting tasks are computed by a machine. Classifiers are also combined for solving complex problems with the above said hybrid approaches [52]. Maximum Entropy (MaxEnt), Memory-based learning (MBL), and Hidden Markov Models (HMM) methods used for a voting strategy on Spanish dataset, resulting in that for named-entity identification accuracy was 98.5% and approx. 85% for classification [53].

As the system trained with a specific algorithm for a specific language dataset, it is doubtful that how these systems would behave with other languages dataset [54]. The following subsections discussed the event extraction, event annotation, and event extraction system.

1.4.2 EVENT EXTRACTION

After discussion on the NER system, event extraction and relation between entities are two additional components of IE application. A semantic enriched information extraction system is the process of extraction of the event along with its relationship between entities. E.g., the following letter to the Press by Trivendra Singh Rawat –Chief Minister of Uttarakhand has signaled the twentieth Sthapana

Diwas of Uttarakhand. The Map of Uttarakhand state as per political classification is shown in Figure 1.11 [273].



Here the fact is that Trivendra Singh Rawat is the chief minister of Uttarakhand is a relationship between two entities. This relation can also represent like:

[PER Trivendra Singh Rawat] IS-CHIEF MINISTER-OF [ORG Uttarakhand]

Remember that at the initial step, relation extraction requires resolving a co-reference between two different entities “Trivendra Singh Rawat” and “Chief Minister of Uttarakhand” refer to a single person. Additionally, two events can also extract from the date of the article (19 Nov 2020).

First Event

EVENT-TYPE : Sathapana Diwas
 SUBJECT : Uttarakhand
 DATE : 9 Nov 2000

Second Event

EVENT-TYPE : Announcement
 ANNOUNCER : Trivendra Singh Rawat
 RECIPIENT : The Press
 DATE : 19 Nov 2020

1.4.2.1 EVENT EXTRACTION

Event Extraction (EE) is a task that is used to find out and analyze the happening of the event from the corpus [55]. It mainly includes who did what to whom, when, where, through which method, and why. An event is a complex mixture of relations that are linked to a set of empirical findings from a database. EE extracts multiple entities from unstructured data and the relationship between the extracted entities. The following table demonstrates a critical overview of some events. Table 1.2 represents various existing event ontology for events like vehicle, personal changes, crime, conflict, transaction, business, financial, political, and family.

Table 1.2 Various Event Ontologies [56]

Event			
Vehicle	Vehicle Transaction	Transaction	Buy artifact
	Vehicle departs		Sell artifact
	Vehicle arrives		Import artifact
	Spacecraft launch		Export artifact
	Vehicle crash		Give money

Personnel Change	Hire	Business	Start business	
	Terminate contract		Close business	
	Promote		Make artifact	
	Succeed		Acquire company	
	Start office		Sell company	
Crime	Sexual assault	Financial	Sue organization	
	Steal money		Merge company	
	Seize drug		Currency moves up	
	Arrest		Stock moves up	
	Try		Stock moves down	
	Convict		The stock market moves up	
	Sentence		The stock market moves down	
	Jail		The stock index moves up	
Conflict	Kill	Political	The stock index moves down	
	Injure		Nominate	
	Hijack vehicle		Appoint	
	Hold hostages		Elect	
	Attack target		Expel person	
	Fire weapon		Reach agreement	
	Weapon hit		Hold meeting	
	Invade land		Impose embargo	
	Move forces		Topple	
	Retreat		Family	Die
	Surrender			Marry

1.4.2.2 EVENT ANNOTATION

TimeML introduced a robust specification language in NLP for event annotation and temporal expression [57] and ISO has recognized this language as a standard for the time, event markup, and event annotation. XML-like syntax and a high number of possible attributes used by TimeML, make this a more complex markup language. E.g., a simple sentence in TimeML [58] is shown in Figure 1.12

```

Sunil
<EVENT eid = "Et1" class = "OCCUR">
Taught
</EVENT>
<MAKEINSTANCE eiid = "ETI1" eventID = "Et1" pos = "VERB" tense = "PRESENT
aspect = "ZERO" polarity = "POS"/>
<TIME3 tid = "T1" type = "DURATION" value = "X30T">
15 minutes
<TIME3 tid = "T2" type = "SET TIME" value = "XX-WX-1" quant = "EVENT">
Every Monday
</TIME3>
<TLINK timeID = "T1" relatedToTime = "T2" relType = "IS_INCLUDED"/>
<TLINK eventInstanceID = "ETI1" relatedToTime = "T1" relType = "DURING"/>

```

Figure 1.12 Output of Sentence in TimeML

1.4.2.3 EVENT EXTRACTION SYSTEM

EE methods can be categorized into three approaches i.e. data-driven, knowledge-driven, and hybrid approach methods. Data-driven methods fail to find valid meanings for various events and these techniques rely on probabilistic and Big Data models. Knowledge-driven approaches need domain expert linguists for designing linguistic patterns. Patterns can be designed either through regular expressions (lexicosyntactic) or by using gazetteers or ontologies (lexico-semantic). Hybrid approaches are a combination of data-driven and knowledge-driven approaches. This approach requires a large corpus with expert knowledge.

In an IE system, after extracting NER and EE, it becomes more crucial to find the relationship in between. The next subsection presents the detail about the relation extraction.

1.4.3 RELATION EXTRACTION

Relation Extraction (RE) or Semantic Information Extraction is a combination of relation detection and relation classification. It is focused on the relationships between different entities extracted by the NER system. These extracted relationships may be of very different kinds and unpredictable though many systems still rely on the findings of the relations based on the pre-established patterns. For a semantic enriched extraction system, the present research program is focused on typologies of relations and relation detection systems.

1.4.3.1 TYPOLOGY OF RELATIONS

After extraction of NER and EE from semi-structured or unstructured text, it becomes more important to find out the semantics between the extracted data. Various models arise to categorize the relationship between extracted entities.

The MUC and ACE-like conference included five general types (located, part, near, social, and role) of 24 subtypes of relations [59]. Whereas MUC proposed an ontology with 37 relation types, mentioned in Table 1.3 [56].

Table 1.3 Relation Ontologies Adopted from MUC-1 to MUC-7

Relation Type	Relation	Relation Type	Relation
Place	Name & Aliases	Person	Name & Aliases
	Type		Type
	Subtype		Subtype
	Descriptor		Descriptor
	Country		Honorific
Artifacts	Name & aliases		Age
	Type		Year
	Subtype		PhoneNumber
	Descriptor		Nationality
	Maker		Affiliation
Artifacts	Owner		Sibling
			Spouse
Organization	Name & aliases		Parent
	Descriptor		

	FoundationDate	Person	Grandparent
	Nationality		OtherRelative
	Location		BirthPlace
	ParentOrg		BirthDate

1.4.3.2 RELATION DETECTION SYSTEMS

In the late 1990s, some domain-specific research was conducted on relation identification and extraction. At IBM Watson Research Center, in a lexical network, researchers find out the key features of the task like the use of patterns, organization, sectional restrictions, and frequency filters of the relations. The following subsection defines the template filling of relation extraction for an IE system.

1.4.3.3 TEMPLATE FILLING

In several IE methods, the information uses well-known patterns for the representation of conventional real-world situations. Template filling also called slot filling, is the simplest subtasks of Information extraction. For a given set of attributes, Template filling is used to find out the values in the text. For example, in a country with a complex political structure like India, the election process at every level happens often.

Indian election system “event’ typically consists of a geographical entity, time-period, a winner, and administrative level. Figure 1.13 shows the following ballot paper template filling of an election event [60]. The representation of template filling can also be defined as follows:

ELECTION:

LEVEL : municipal
ENTITY : Jammu & Kashmir
DATE : 2014
WINNER : BJP

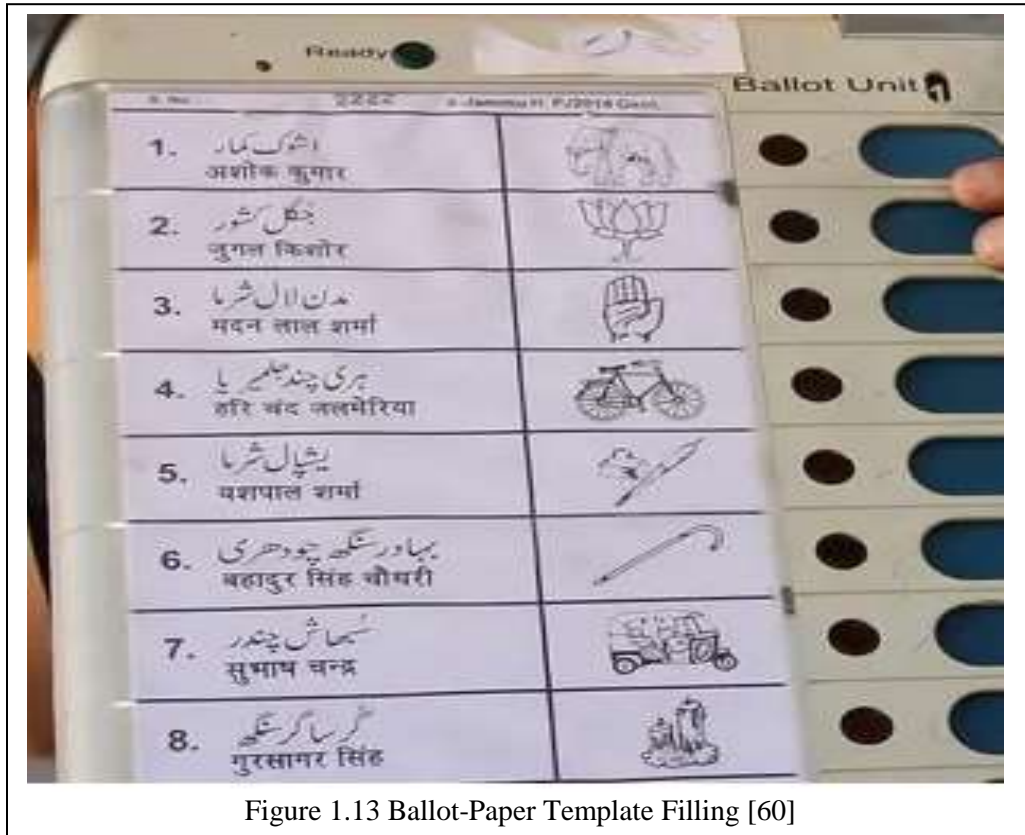


Figure 1.13 Ballot-Paper Template Filling [60]

For similar recurrent events, template filling shows better comparative results. However, it is very hard to predict what is expected to extract from unknown documents. For domain-specific events, open extraction technologies will give better results as compare to simple templates. The following section and subsections discuss various application areas of IE along with its importance in the agricultural domain and finally the existing problems in the agricultural domain in related to IE.

1.5 APPLICATION AREAS OF INFORMATION EXTRACTION

Information Extraction plays a vital role in various fields of human life. The following areas are prominent application areas of IE such as Social Media Mining, Medical Science, Decision Making, Financial Domains, Security System, Sports,

NLP Search Engine, Biodiversity Conservation, Speech Processing, Regional Planning, Agriculture, Cultural Heritage, Geology, Investigation, Warfare, Intelligence & Criminal, Insurance, Business Applications, Resume Processing, Research and Education, Banking Systems, Text Summarization, Text Mining, Sentimental Analysis, Opinion Mining, Online News, Fraud Detection, National Intelligence and Cyber Security [15, 16] [61, 62, 63, 64, 65, 66, 67]. Among all the said application areas, the current research work focused only on information extraction techniques in the agricultural domain.

1.5.1 IMPORTANCE OF INFORMATION EXTRACTION IN AGRICULTURAL DOMAIN

Knowledge management is very crucial in the agricultural domain. Significant agricultural information is unstructured (different website format) corpus and scattered (different geographical locations) in hard documents. It is very hard to find the desired information from the corpus because the traditional search engine shows the list of documents as per their ranked record. Moreover, if web search engines found the semantic information, then often that information is not content digestion [64] [68]. Therefore, it is important to extract structured information from unstructured data that reduce time in browsing and reading, to find out the semantic information and overall structure of the domain knowledge.

1.5.2 INFORMATION EXTRACTION CHALLENGES IN AGRICULTURAL DOMAIN

Various problems have existed in the agricultural sector, but focusing on IE, the following are the key issues in the agricultural domain in related to Information Extraction [15, 16] [61] [64] [69] [70, 71, 72]:-

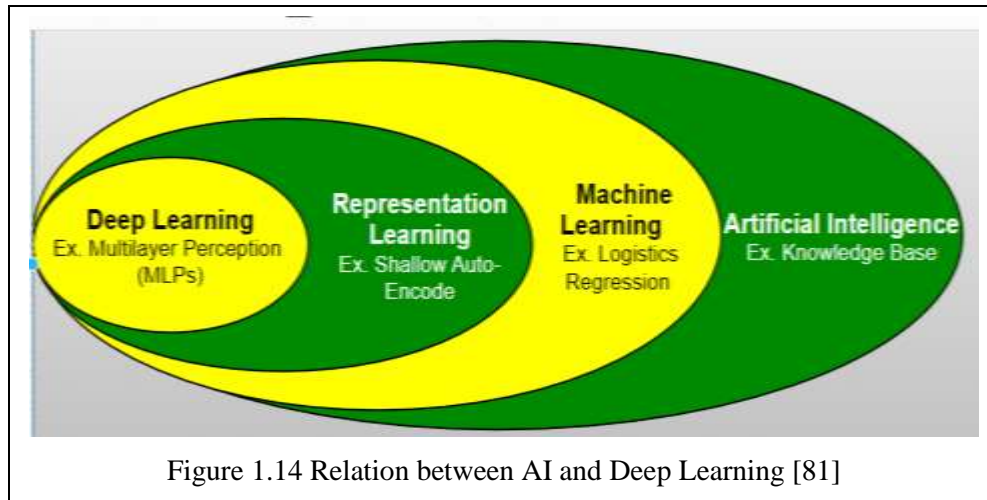
- i. Unavailability of the data for training.
- ii. Ambiguity due to the multilingual users.

- iii. To provide support to naïve users that requires time-critical and domain-specific data.
- iv. Informal Language includes lack of punctuation, non-standard abbreviations, grammatically incorrect, etc.
- v. To extract the semantic information from a multilingual corpus.
- vi. To design a set of rules by domain experts.
- vii. To extract the information about the entities and events.
- viii. To express the unstructured and scattered data into the structure information
- ix. Rich morphology complicates finding the exact relationship between the entities.

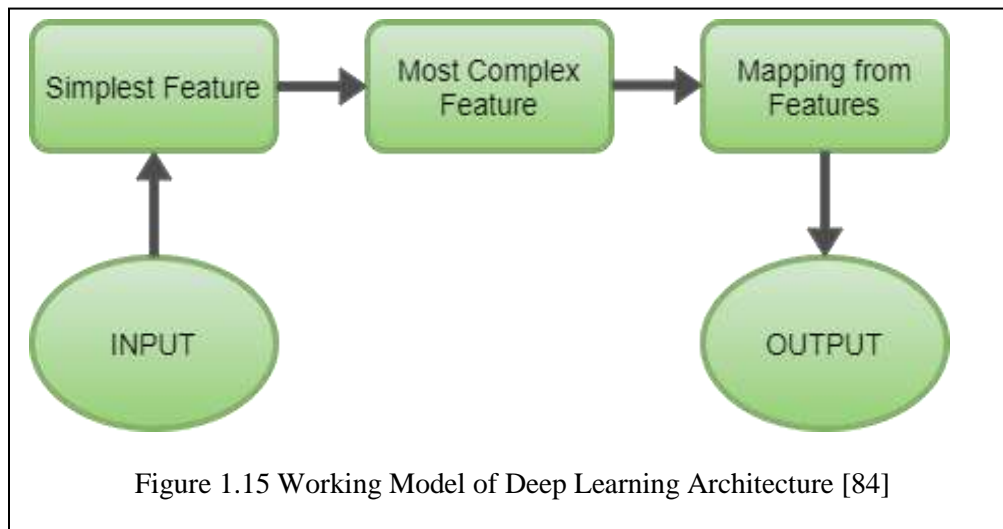
The contemporary development in big data technologies and machine learning algorithms, specifically in deep learning help to address the above mentioned [73, 74]. The following section presents an overview of the deep learning domain.

1.6 DEEP LEARNING

Deep Learning is one of the machine learning techniques i.e. used to make the system automatic by learning and analyze the Big Data (structured, unstructured, or semi-structured) [75, 76, 77, 78]. Application areas of deep learning are computer Vision, Large Scale Deep Learning, Speech & audio processing Recognition, NLP, Neural Language Model, robotics, bioinformatics and chemistry, video games, search engines, online advertising & finance, and many more [79, 80].



Deep learning provides the complex representation of data and makes system automatics (machine independent). Figure 1.14 shows the relation between Deep Learning, representation learning, machine learning, and artificial intelligence. These learning architectures are used for many approaches to Artificial Intelligence [81]. Each circle in the above figure includes an example and related approaches of Artificial Intelligence [82, 83].



An efficient, distributed and hierarchical data representation can create through deep learning architecture.

Figure 1.15 shows Deep Learning, data representation, and feature extraction trained with the help of many non-linear transformation layers [84].

1.6.1 DEEP LEARNING METHODS

To optimize the deep learning algorithms, the following techniques can apply to reduce the training time of the network model.

Backpropagation: To calculate the gradient function for each iteration backpropagation technique can be used. This deep learning technique uses the gradient-based method to solve optimization problems [85].

Stochastic Gradient Descent: Without consideration of local minimum, stochastic gradient descent finds the optimal minimum by using the convex function. Learning rate, step size, and activation function values are used to find the optimal minimum in different paths [86].

Learning Rate Decay: By altering the learning rate, the training time of gradient descent algorithms is reduced and increases the overall performance of a model. This technique is used widely because large changes can be made at the initial level of the training process that reduces the learning rate very gradually. This technique also permits adjustments of weights in the later iterations [87].

Dropout: The dropout technique is specially designed for the outfitting problem of neural networks. By dropping units and their connection randomly during the training process, this technique offers a better regularization method to decrease overfitting in neural networks and recover generalization error. This deep learning method gives better output on supervised learning tasks in computational biology, computer vision, speech recognition, document classification [88].

Max-Pooling: A predefined filter is applied across the non-overlapping regions (sub-regions) of the input layer and extracting the maximum values as the output. By using the max-pooling method, the computational cost for learning several parameters can also be reduced [89].

Batch Normalization: By reducing the covariate shift, the batch normalization technique increases the learning rate of the deep neural network. For each small batch, when the weights are adjusting during the training process, this method normalizes the input layer. The network stability can improve by normalizing the output from the last activation layer. Batch normalization methods also improve learning rates and reduce the training epochs [90].

Skip-gram: The skip-gram algorithms can be used for word embedding problems in the deep neural network. If two different sentences share similar contexts, then these sentences are also similar. E.g., the sentences “dogs are mammals” and “cats are mammals” are two different meaningful sentences. These sentences show the similar meaning “are mammals.” Skip-gram considers a context window containing terms, train the deep neural network by avoiding one term, and use the skip-gram model to predict skipped term [91]

Transfer learning: The trained model on a specific task is used to train another similar task in transfer learning. The extracted knowledge from solving one problem can be transferred to another related problem. While solving second related task, transfer learning gives rapid progress and improve performances [92].

Based on the above said deep learning models, some advantages and disadvantages of each technique are compared in Table 1.4 [93].

Table 1.4 Advantages and Disadvantages of Deep Learning Methods [93]

S.NO	Method	Description	Advantages	Disadvantages
1	Back Propagation	Used in the optimization problem	for calculation of the gradient	sensitive to noisy data
2	Stochastic Gradient Descent	to find optimal minimum in optimization problems	avoids trapping in local minima	Longer convergence time, computationally expensive
3	Learning Rate Decay	reduce learning rate gradually	increases performances reduce training time	Computationally expensive
4	Dropout	dropout units/connection during training	avoids overfitting	increases the number of iterations required
5	Max-pooling	applies a max filter	reduces dimension and computational cost	Considers only the maximum element which may lead to unacceptable result in some case
6	Batch Normalization	batch-wise normalization of input to a layer	Reduces covariant shift, Increases stability of the network, Network trains faster	Computational overhead during training
7	Skip-gram	used in word embedding algorithms	Can work on any raw text, Requires less memory	Soft-max function is computationally expensive, Training Time is high
8	Transfer learning	Knowledge of first model is transferred to the second problem	Enhances performance, Rapid progress in the training of the second problem	Works with similar problems only

1.6.2 DEEP LEARNING FRAMEWORK

Deep learning algorithms help in designing a neural network more quickly without going into details of original algorithms. Usually, each framework design for every different problem statement [94]. Few deep learning frameworks are summarized below in Table 1.5.

TensorFlow: TensorFlow supports many contemporary programming languages such as Python, C++, and R and developed by Google brain. It enables the deployment of deep learning models in CPUs and GPUs as well.

Keras: Keras API runs on top of TensorFlow and is written in Python. It enables fast experimentation that supports both Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Like TensorFlow, Keras also enables the deployment of deep learning models in CPUs and GPUs as well.

PyTorch: PyTorch can be used for building deep neural networks as well as executing tensor computations. PyTorch is a Python-based package that provides Tensor computations. PyTorch delivers a framework to create computational graphs.

Caffe Yangqing Jia developed Caffe, and it is open source as well. Caffe stands out from other frameworks in its speed of processing as well as learning from images. Caffe Model Zoo framework facilitates us to access pre-trained models, which enable us to solve various problems effortlessly.

Deeplearning4j Deeplearnig4j is implemented in Java, and hence, it is more efficient when compared to Python. The ND4J tensor library used by Deeplearning4j provides the capability to work with multi-dimensional arrays or tensors. This framework supports CPUs and GPUs. Deeplearnig4j works with images, CSV as well as plaintext.

Table 1.5 Comparison of Deep Learning Framework [94]

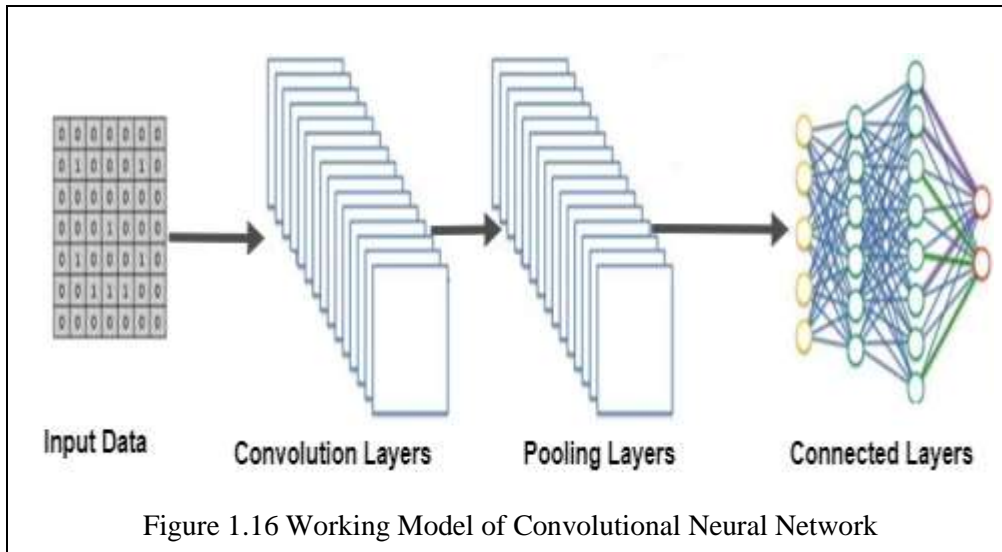
S. No	Deep Learning Framework	Release Year	Language Written in	CUDA Supported	Pre-trained Model
1	TensorFlow	2015	C++, Python	YES	YES
2	Keras	2015	Python	YES	YES
3	Pytorch	2016	C, Python	YES	YES
4	Caffe	2013	C++	YES	YES
5	Deeplearning4j	2014	C++, JAVA	YES	YES

1.6.3 COMMON DEEP LEARNING ALGORITHMS

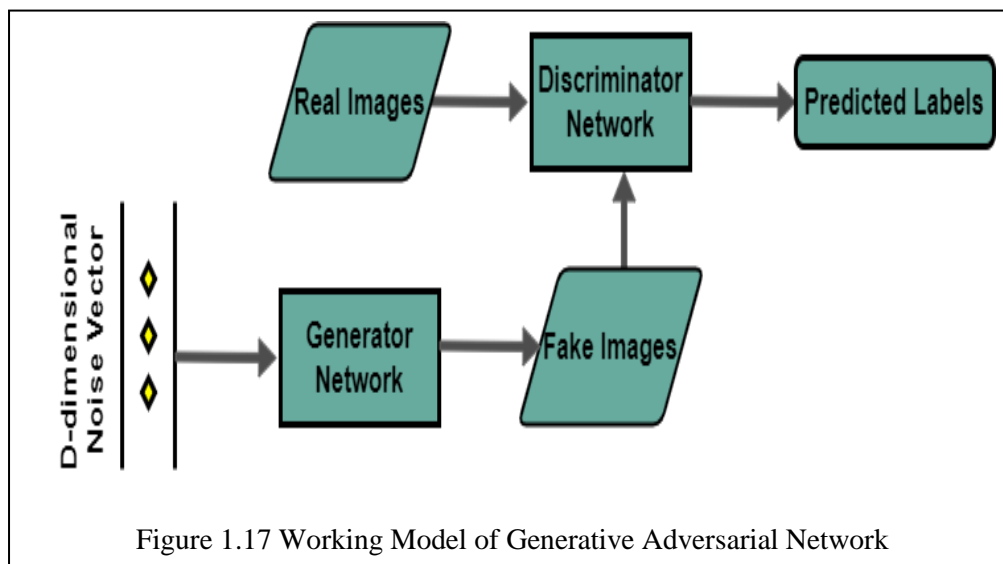
The most commonly used deep learning algorithms are CNN, RNN, and Generative Adversarial Networks (GAN). Along with these algorithms, VGGNet, ConvNets, LSTM, and DCGAN are few sub-category DL algorithms [95, 96, 97, 98, 99, 100]. Because these sub-categories can be derived from three basic DL algorithms (CNN, RNN, GAN), so, it is not important to include all sub-categories here [101]. As per the thesis direction, let us discuss CNN & GAN algorithms in brief and RNN with LSTM algorithms in detail.

1.6.3.1 CNN- GAN ALGORITHMS

Convolutional Neural Networks (CNN): In CNN, in their layers, a specialized convolutional neural network is used at least once instead of using general matrix multiplication. It comprises three layers that are convolutional layers, pooling layers, and connected layers [102]. Figure 1.16 shows that CNN uses convolutional and pooling layers for feature extraction and the connected layers as classifiers.



Generative Adversarial Networks (GAN): GAN is a machine learning technique that automatically discovers and learns patterns in input values. GAN is the combination of existing back-propagation algorithms. A class of noises takes as input for learning the real data distribution and new data generation. GAN can generate realistic models especially in image processing such as image translation of day to night, summer to winter, generation of photorealistic objects, etc.



Two basic GAN sub-models are the Generation Model (GM) and Discrimination Model (DM). Figure 1.17 shows that the GM is used to capture the real data distribution whereas DM is used as a binary classifier [103].

1.6.3.2 RECURRENT NEURAL NETWORKS WITH LONG SHORT-TERM MEMORY

Since the inception of LSTM, various practical experimental (theoretical as well) works published with RNN, where data is sequential, surprising results also obtained from various application domains like speech-to-text transcription, language modeling, machine translation, etc. The following section presents the research gap of IE in related to the agricultural corpus.

1.7 RESEARCH GAP AND DIRECTION

The existing Information extraction algorithms for mining the semantics from the agricultural corpus are not matured enough to meet the expected solution of domain problems [104, 105, 106][274, 275]. Moreover, the direct and indirect effects of various factors (weather, soil, pest, and fertilizers) continuously creating problems. Therefore, a textual information extraction system for the agriculture sector is highly appreciated for crop yielding.

1.7.1 PROBLEM DEFINITION

Traditional IE algorithms are not capable enough to extract the semantic information from an agricultural corpus, due to its complexity in terms of unstructured in nature. The following mathematical model represents the problem statement.

\vec{y} or a solution set $\vec{Y} = \{\vec{y}_1, \vec{y}_2, \dots \dots \dots, \vec{y}_n\}$ such that:

$$\begin{aligned} \vec{y}_1 & \text{ is to be maximized} \\ \vec{y}_2 & \text{ is to be minimized} \\ \vec{y}_3 & \rightarrow I \end{aligned}$$

Where \overline{y}_1 Accuracy, Precision, Recall & Sensitivity, \overline{y}_2 is Learning Rate, Error Rate, Loss Function, Gradient decedent (but not ≈ 0) and \overline{y}_3 Nash-Sutcliffe Efficiency Coefficient & F-measure.

1.7.2 RESEARCH OBJECTIVE

The objective of the present research program is to design & implement an algorithm using “*Deep Learning Technique to Extract Semantic Information from Agricultural Corpus*”.

Sub-objective

- Identify the issues in various existing information extraction techniques related to the agricultural domain.
- Devise an algorithm for information extraction in the agricultural corpus using the deep learning technique.
- Evaluating the performance of the newly devised algorithm against the standard information extraction algorithms.

1.8 THESIS NAVIGATION

The remainder of the present thesis stated as follows. Chapter 2 demonstrates the overview of different research areas of information extraction followed by state-of-the-art IE for NER and EE using deep learning and agricultural domain, provided in the literature with a detailed description of their related algorithms. Chapter 3 offers a framework for semantic information extraction using deep learning techniques for the agricultural domain. Chapter 4 deals with various data preprocessing algorithms and creating a knowledge base. Chapter 5, extraction of NER and EE, along with semantic extraction. Chapter 6 presents details about the experimental results of the proposed approach. Chapter 7 concludes the thesis with a summary and directions for future work.

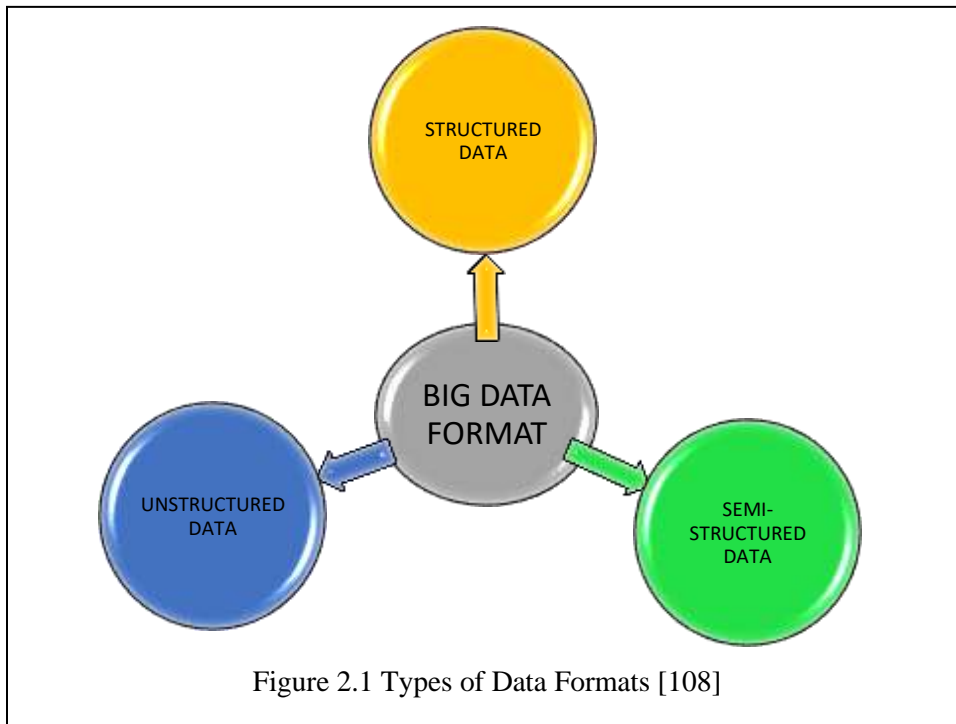
CHAPTER -2

LITERATURE SURVEY

Information extraction is one of the significant research areas of NLP. The present chapter pinpoints various existing research issues of IE tools & techniques with deep learning algorithms in different areas of NLP. Further, a comprehensive literature review presented through different IE tasks along with deep learning techniques and finally outlined various challenges for realizing IE in real essence.

2.1 DATA DEFINITION FRAMEWORK

Based on four major characteristics of big data i.e. volume, velocity, variety, and veracity, the data can categorize into three formats: structured data, semi-structured data, and unstructured data [107]. Figure 2.1 depicts the various data formats [108].



Structured Data: In a well-organized form of data containing a definite data model, structure, or format, so that it can easily be processed and analyzed. E.g., relational database, ERP, data warehousing, CRM, etc. [109].

Semi-structured Data: Partially structured data that can be organized by applying few desired operations. E.g., CSV, JSON, EDI, NoSQL, XML files and tab-delimited files etc. [110].

Unstructured Data: Data stored in different file formats and does not have a definite structure/pattern. A bit hard to process and analyze, so requires some advanced tools and techniques. E.g., streaming data, pdf, images, videos, etc. [111]. The following sections present the significance of IE.

2.2 SIGNIFICANCE OF INFORMATION EXTRACTION

The amount of text data over the web is increasing at a rampant rate. More than 85% of the data is unstructured nature due to various raw data sources such as web pages, text files, weblogs, sensor data, images, audio & video files, social media data, banking data, communication over the web (live chat, messages, and web meetings), survey responses, publications, etc. [112]. Now a day, almost all sectors/firms store and exchange data over the web in the form of digital libraries, blogs, social networks, etc. Internet is over-flooded with such textual information and it becomes a need for an organization to find out ways for managing and analyzing unstructured data to make crucial business decisions. As the manual analysis of this unstructured textual information is impractical, so it is a difficult task to find out the trends and patterns in free data to extract relevant information [113]. Thus, this large volume of data has raised the necessity of text mining techniques to extract significant information.

Text Mining is a part of Data Mining, where data is stored in the form of text and mostly unstructured [114]. Due to this amorphous data structure, text mining is generally managed through some search engines [115]. A search engine takes a

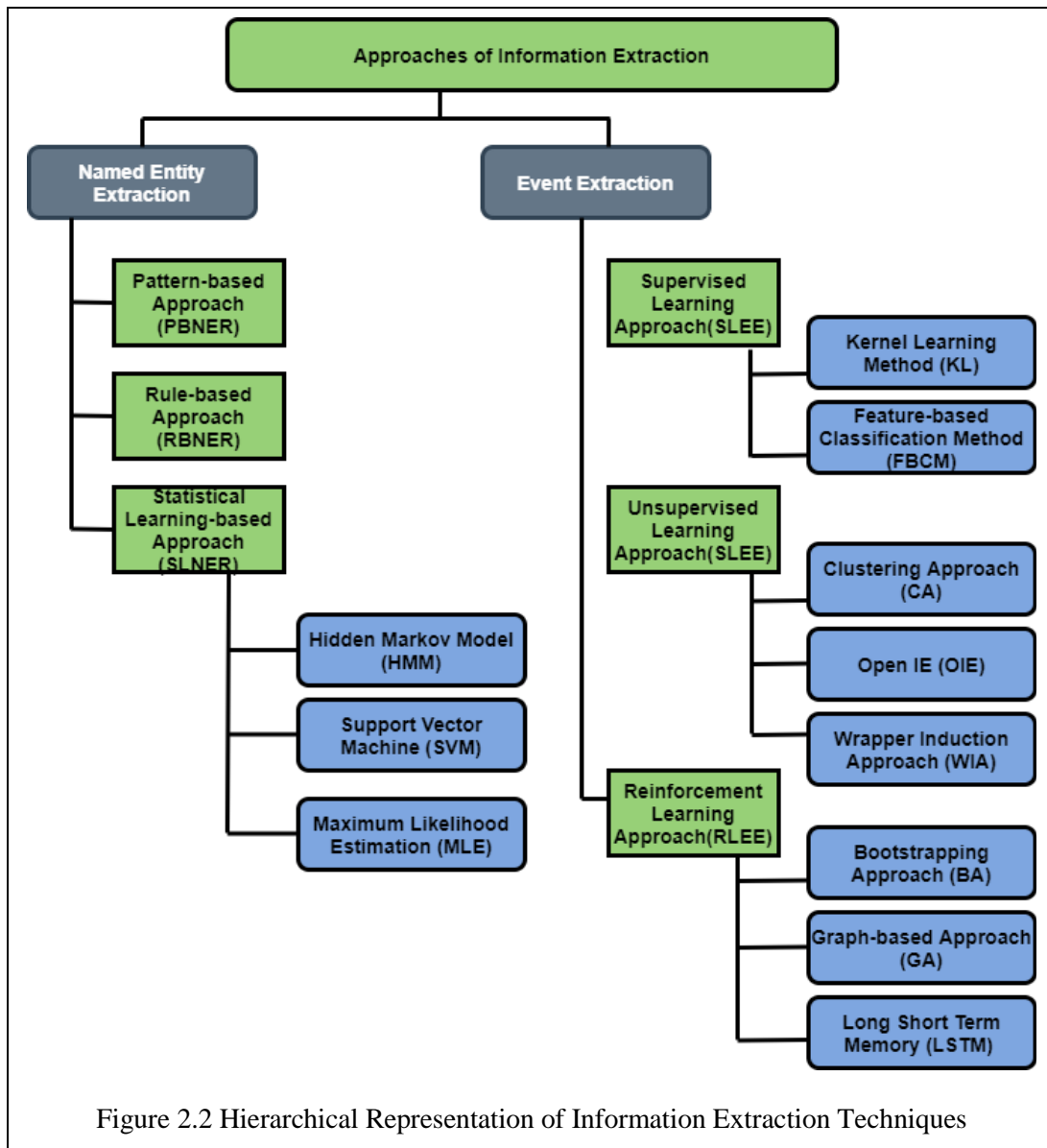
textual query as input and returns the useful information for a given textual query as output [116]. Traditionally research in the field of IR has focused more on information access [117, 118] rather than analyzing the patterns to discover information. However, the main goal of text mining is to focus on analyzing the pattern, trends of data stored and finally extract useful information from it [119].

To make the task easier, researchers have used the approach for first converting the unstructured data to semi-structured/structured and then use it for IE. Generally, some predefined formats like Named Entity Recognition (NER) and Event Extraction (EE) of IE tasks are most commonly used for this process (Chapter-1, Section 2).

Researchers are working on it for over two decades and concentrated more on the identification of named entities and related events among them using natural language text. Message Understanding Conference (MUC) [120] and Automatic Content Extraction [118] based on the extraction of structured entities. Literature surveys [121, 122, 123] are examples of research where the researchers have focused on the extraction of significant information from the entity named relationships like “CEO” of “Company”. The following section presents various existing methods of IE

2.3 VARIOUS METHODS OF INFORMATION EXTRACTION

To process the textual information, IE methods can classify into two fundamental parts i.e. NER and EE [124]. Figure 2.2 represents an overview of various methods of IE based on NER and EE. Each NER and EE category is further divided into sub-categories and has diverse research scope in various fields on NLP [125]. Section 2.4 discusses the various IE-based NER approaches and section 2.5 presents the sub-categories that fall under EE.



2.4 INFORMATION EXTRACTION-BASED ON NER APPROACHES

The NER classified into three sub-categories that are pattern-based approach, rule-based approach, and statistical learning-based approach. Each category of NER-based methods is explained below.

2.4.1 PATTERN-BASED APPROACH

A pattern-based NER (PBNER) used to extract certain patterns (entities and attributes) from an input unstructured corpus [126]. PBNER comprises three different steps first, a primary set of patterns are extracted based on keyword matching, second, a set of regular expressions is applied to identify information from extracted patterns, and finally, a set of rules are defined to acquire metastatic sites of identified information.

2.4.2 RULE-BASED APPROACH

The Rule-based NER (RBNER) methods, extraction based on particular grammar rules based on domain-specific linguistic knowledge [127]. It requires maximum human intervention for designing these rules. Typical RBNER systems have two major components Gazetteers (lexicon) and Local Grammars [128]. The gazetteers are a set of named entities, which known before designing rules. These gazetteers are further classified into some semantic classes for better rule designing. The local grammar is used to identify and classify named entities that are not present in gazetteers. This grammar is also responsible for the final classes of named entities. The semantics extraction of a single sentence may have a different meaning, so it is difficult to assign the NE category. Therefore, even for a domain expert, RBNER is a very complex and time taking process [129].

2.4.3 STATISTICAL LEARNING APPROACH

Because RBNER approaches are very time-consuming, error-prone, and require handcrafted rules for extracting named entities, Statistical Learning Approach based (SLNER) helps to extract entities automatically [130]. SLNER are very commonly used approaches and under the categories of NER, further divided into three sub-categories i.e. maximum likelihood estimation approach, hidden Markov model, and support vector machine [131].

Maximum Likelihood Estimation (MLE): Among three sub-categories of SLA, the MLE generally inappropriate for information extraction in NLP due to the data sparseness (Discounting or irregular). According to the parameter choice, the highest probability of training data can assign [132].

Hidden Markov Model (HMM): Markov process with hidden states used to design the system to extract named entity from free text corpus [133]. These hidden states are called unobserved states.

Support Vector Machine (SVM): The data used for regression and classification purposes, analyzed by SVM approaches. SVM used statistical learning-based algorithms [134]

2.5 INFORMATION EXTRACTION-BASED ON EVENT EXTRACTION

EE is classified into three fundamental categories supervised learning methods, unsupervised learning methods, and reinforcement learning methods. Let us have a brief overview of these three methods.

2.5.1 SUPERVISED LEARNING APPROACH

In the Supervised Learning approach for Event Extraction (SLEE), the inner relation between input data may or may not know however output (event) of the model is very well known [135]. Based on already labeled data, training of the machine will process. Until the input labeled data changed, data scientists should reconstruct the learning model to assure the given perceptions remain true. Figure 2.3 shows the internal process of a supervised learning algorithm [136].

SLEE is divided into two sub-categories i.e. Feature-based Classification Method (FBCM) and Kernel Learning Model (KLM). In FBCM, a set of features (segment) is used to train the model [137]. Whereas KLM uses a higher-dimension input feature space to design the raw data as possible linearly segregated [138].

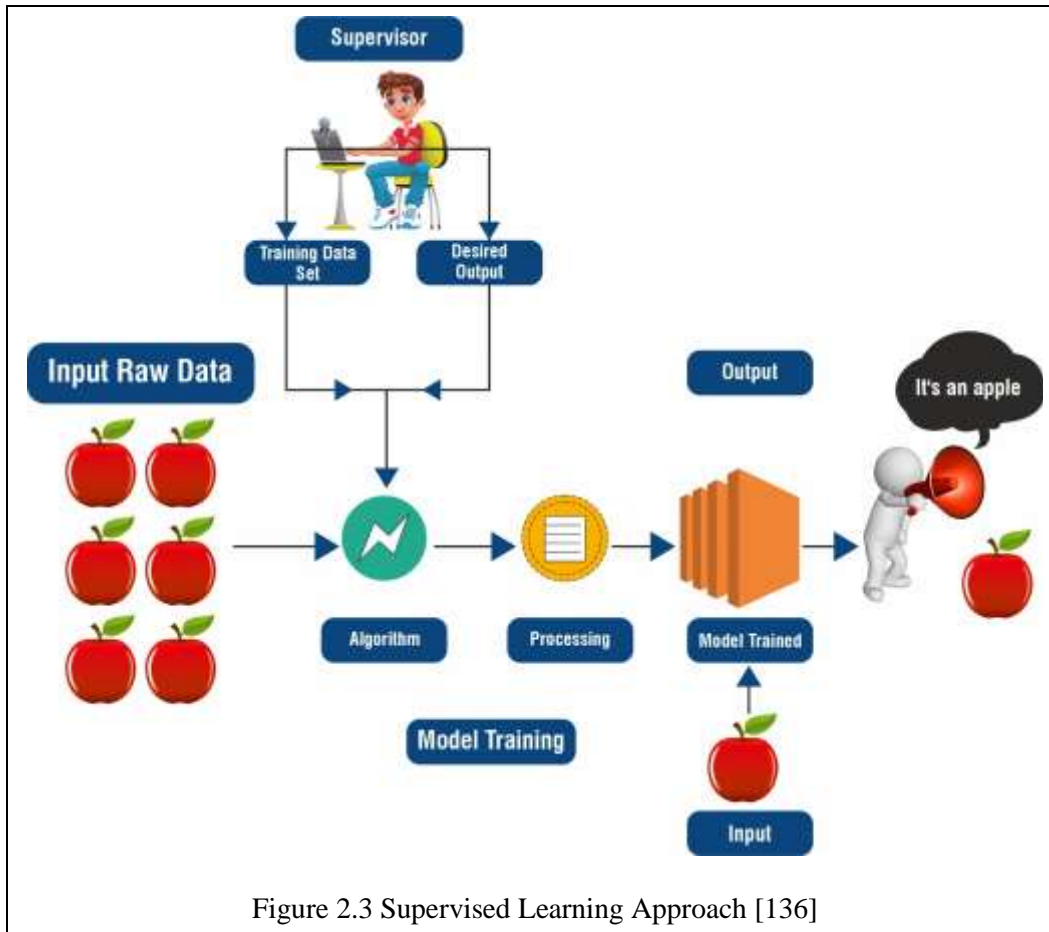


Figure 2.3 Supervised Learning Approach [136]

2.5.2 UNSUPERVISED LEARNING APPROACH

The Unsupervised Learning approach for Event Extraction (ULEE) is a self-organized EE approach. The purpose of ULEE is to find out the primary patterns and predicts an event. Raw input data provides a learning model that tries to explore the hidden features and clusters present in the corpus [139]. This learning approach is further divided into three subparts i.e. Clustering Approach (CA), Open IE (OIE), and Wrapper Induction Approach (WIA) [140]. CA uses the process of marking similar event groups together to train the model, OIE finds the relation between the data, and WIA generates the wrapper (set of rules) from existing

examples and counterexamples [141, 142, 143]. Figure 2.4 shows the fundamental architecture of the unsupervised learning approach [144].

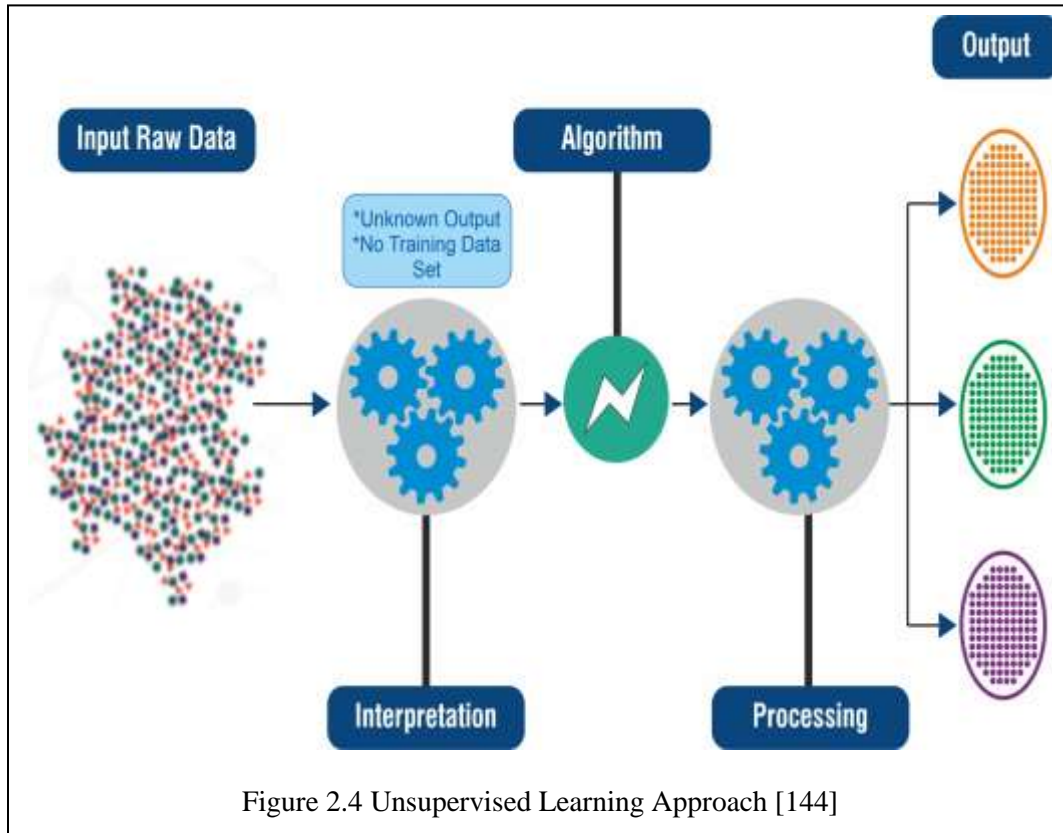
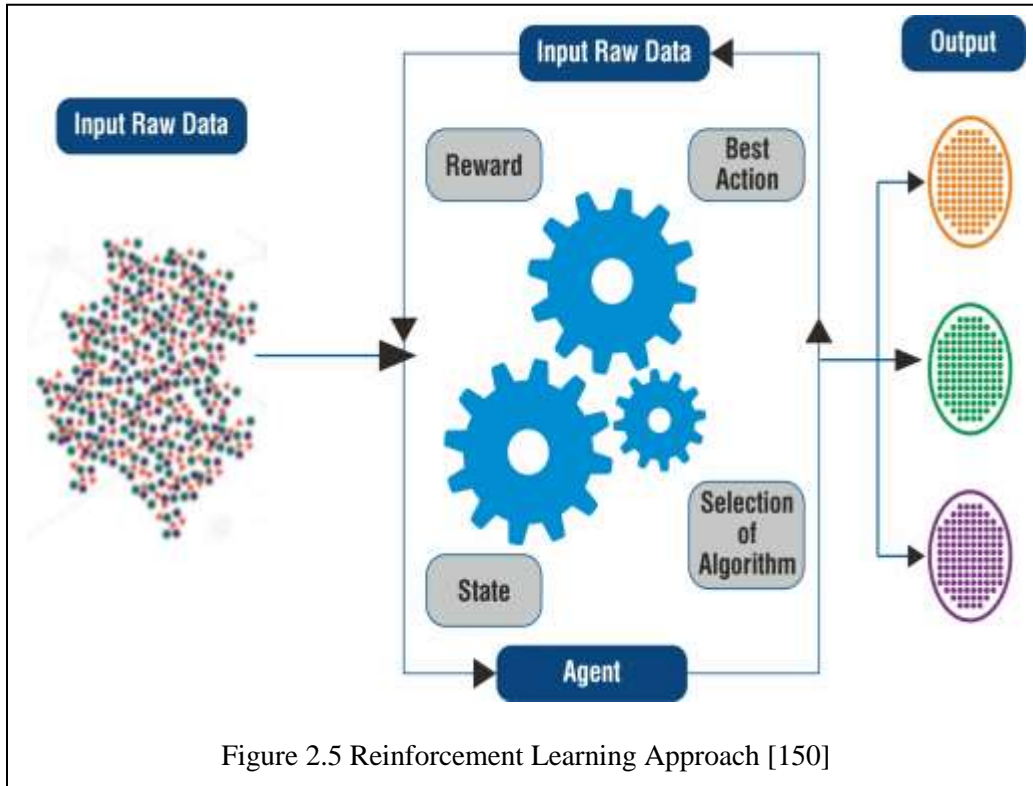


Figure 2.4 Unsupervised Learning Approach [144]

2.5.3 REINFORCEMENT LEARNING APPROACH

In the Reinforcement Learning for Event Extraction (RLEE), the Learning Agent (LA) has a start state, end state, and multiple paths in between. LA always manipulates the learning environment while traveling from one state to another state. For every successful finding, the LA gets appreciation (reward) [145, 146]. RLEE is classified into Bootstrapping Approach (BA), Graph-Based Approach (GBA), and LSTM. BA helps to prevent overfitting and improves the overall performance of learning algorithms [147], GBA finds sub-events within the input data & design a graph (subgraphs in a large corpus) to train the model [148] and LSTM based on time series data model to classify, process and make predictions

from input data [149] (detail in Chapter-3, section 3.3.2). Figure 2.5 shows the architecture of the reinforcement learning model [150].



Each category/sub-category of IE has a major role in various research fields. These techniques have their strength and weakness. Table 2.1 deeply elaborates on various pros and cons of each technique in a specific domain.

Table 2.1 Strength and Weakness of Information Extraction Sub-categories

IE Approach	Category	Domain	Ref	Strength and Weakness
NER	PBNER	Unstructured Data	[151]	The proposed method must require a dictionary.
		Social Media Data	[152]	Improvement required.

NER		BioMedical data	[153]	There are some cases, which involve false-negative results.
	RBNER	Social Media Data	[154]	The dataset is less and syntax features did not use.
		Biomedical Data	[155]	More than one species in a single document cannot recognize.
		Chemical Data	[156]	Data labeled manually which is not feasible when data is huge.
		Unstructured Data	[157]	The defined rules are less and do label all entities like time, percentage, and date.
	MLE	Social Media Data	[158]	This method is only useful for English, not any other languages
		Unstructured Data	[159]	The effectiveness of the POS tagger is low.
		Biomedical Data	[160]	Suicide-related genes present in databases do not equate to finding the appropriate Relations within a document.
	HMM	Social Media Data	[161]	Entity types such as products, movies, and songs have not been recognized.
		Biomedical Data	[162]	HMM has not evaluated that what happens if it is plugged into other existing systems.
		Chemical Data	[163]	The survey was taken for chemicals datasets only.
		Unstructured Data	[164]	The domain of study is just NERC research, which is too short.
	SVM	Social Media Data	[165]	Retweets do not consider for NER.

		Unstructured Data	[118]	Only three categories of NER recognized (which are too short.)
EVENT EXTRACTION	FBCM	Social Media Data	[166]	The combinations of relations, contributions to feature selection, and Hidden information in social media are not studied.
		Biomedical Data	[167]	Extracted relations are not on Gene-disease.
		Chemical Data	[168]	Did not attempt to explain the related discussions in the document texts, which might have yielded stronger features.
	KLM	Biomedical Data	[169]	The framework did not capture the semantic similarity between words as domain experts before relation extraction has manually tagged well as named entities.
		Unstructured Data	[153]	The processing speed is low when kernels are applied as well as complexity becomes high.
	CNN	Biomedical Data	[156]	Dictionary, as well as training samples, are not enormous.
		Unstructured Data	[170]	In the trained model with any order strategy, the choice of the evaluation strategy does not affect the performance.
	MIL	Biomedical Data	[171]	The whole database is created manually which is a very time-consuming task.

EVENT EXTRACTION		Chemical Data	[172]	The domain of the research is not wide, only a chemical database considered.
	BA	Social Media Data	[116]	The types of family relations are very short as well as the recall is low.
		Biomedical Data	[173]	The protein-protein interactions by using bootstrapping do not apply.
		Chemical Data	[174]	The value of recall is very low.
		Unstructured Data	[175]	Only two attributes are recognized which is too short.
		Social Media Data	[166]	This algorithm does not take into account the complex relationship within message threads.
	GBA	Biomedical Data	[114]	Richer features are not used and EDG does not use them for the combinations of features. This method is only tested when datasets are large
		Chemical Data	[176]	This algorithm is too time-consuming.
		Unstructured Data	[167]	Regular grammar may not use to convert simple natural language questions.
		Social Media Data	[177]	Unbounded Relations as well as some entities relations like height, GDP, etc. are not considered.
	OIE	Unstructured Data	[178]	Freebase does not use for the training set. Domain explicit features do not implement.

EVENT EXTRACTION	CA	Social Media Data	[167]	Analyzed a small number of social networks.
		Biomedical Data	[179]	Suicide-related genes present in databases do not equate to finding the appropriate relations within a document, which is one of the limitations of the evaluation approach.
		Unstructured Data	[180]	Some other constraints such as dependency and web search results can introduce to improve the extraction precision.
	WIA	Social Media Data	[156]	Additionally, analyses are required for checking the conduct of the framework in various setups.
		Unstructured Data	[181]	Lack of robustness in the learned wrappers.

2.6 EXISTING NER TECHNIQUES

The main objective of the NER is to extract the name of a person, organization, place, etc. from a large collection of documents. In the year 1991, F. Rau was the first one to present a paper in this field at the Seventh IEEE Conference on Artificial Intelligence Applications [178, 182]. He proposed an AI model for recognizing “company name from the text”. His approach used both heuristics and handcrafted rules. Further research in this area became more popular after 1996 [169] when a major event MUC-6 [177] was held and the research in this field has not declined until today. There have been numerous events in this area, HUB-4 [167], MET-2

and MUC-7 [176], IREX [183], CONLL [184], ACE, and HAREM [185] majorly contributed to the research in this field.

With the advancement of machine learning-based approaches, researchers are inclined to deep learning-based approaches and have applied them for NER in various fields, especially for Information Extraction. D.Huynh [186] has worked upon noisy and uncertain data available through social media (tweets) using NER and obtained an accuracy of 75%. K. Bontcheva et al., proposes an open-source Microblog. [187] For noisy and uncertain data in social media applied NER and obtained an accuracy of 76%. M.B. Habib et al. [188] applied NER using CNN and CRF on social media data and obtained an accuracy of 80%. NER has been widely used in the field of Medical data as well.

S.Verma et.al. [189], applied Rule-Based Open Information Extraction in the field of medical data from tweets during a mass emergency and obtained an accuracy of 75%. S. Vieweg et.al. [190], and applied framework for Information Extraction on data collected from Twitter on natural hazards emergency. They obtained an accuracy of 78%.

NER has gained much importance in the field of Agriculture as well. In the early 80's, the Food and Agriculture Organization of the UN and European Countries brought the AGROVOC concept server and Agropedia [191]. Initially developed in English Language but later on due to its popularity, further translated into other four languages: Chinese, Spanish, Arabic, and French. It was based on how controlled vocabularies are of limited semantics and how they can be improved for IE by doing reengineering. Data from soil and weather are used for IE for a type of crop. Researchers have word on the data of land use for different regions to extract the information about suitable crops for that region.

A. Nair, et al., [192] proposed the method to anticipate the Indian Summer Monsoon Rainfall esteems utilizing precipitation yields from Global Climate

Model (GCM). They applied Artificial Neural Network (ANN) in the GCM in India. The ANN procedure connected to different ensemble entities from the GCMs individual to get month-wise scale expectations for India and its sub-divisional region. In the present investigation, straightforward randomization and double folded approval method were merged and applied to minimize the over-fitting problem while ANN method training. The ANN anticipated rainfall is executed from GCMs individuals and decided by examining the contrast, box plots, absolute error, and percentile in linear error, in probability sample space. Experimental results proposed the critical changes after applying the ANN system of these GCMs individuals in forecast expertise. The datasets depend on the previous estimations of the primary variable, however not on relevant factors that can influence the framework or variable.

Estimation of crop yield production is studied by O. Satir, et al., [193] by using Vegetation Indices methods and Stepwise Linear Regression (SLR). By applying, object-based classification and multi-temporal land-sat data set; mapping formed on related crop patterns of an area. In this scenario, by applying real-time measurement methods like Mean Percent Error (MPE) prediction was estimated. MPE estimated for the cotton, corn & wheat and combined with different soil salinity degrees. The proposed method successfully forecast weather data with a single soil parameter but the prediction was reduced.

B. Das, et al., [194] investigated the hybrid algorithms such as ANN, SMLR, penalized regression models consists of the ENET, PCA, and LASSO for foreseeing the yield of rice productivity with the support of long-term weather data. The experimental outcomes stated that LASSO-ENET provided good performance because these methods reduced the model complexity and prevented the overfitting by using magnitude coefficients. The pairwise multiple comparison test found that the hybrid models utilized very well for the prediction of the crop on the west coast of India. The combination of feature selection methods and feature extraction with

neural networks includes PCA-SMLR provides poor performance because, during the alteration of input variables, PCA did not include the dependent variable.

A. Haidar and B. Verma, [195] developed a deep CNN for monthly rainfall prediction in a particular region of Australia. The proposed CNN model compared with the existing forecasting method like the Australian Community Climate (ACC) and Earth-System Simulator-Seasonal Prediction System (ACCESS) that released by the Bureau of Metrology. The results showed that CNN provided improved performance in nearly nine months with higher annual averages, whereas the existing method provided poor performance. In low yearly averages, the ACCESS methods lead to better performance to predict the monthly rainfall for the eastern region. Nevertheless, CNN was unable to learn the non-linear relationships for three months, such as April, June, and July. The rainfall variability is affected due to the need for extra data that not found in the generated dataset.

X. He, et al., [196] implemented HWNN method that included Particle Swarm Optimization (PSO), Mutual Information (MI), and Multi-resolution Analysis (MRA) into ANN for predicting rainfall from precursor climate indices and monthly rainfall dataset. The Maximal Overlap Discrete Wavelet Transform (MODWT) decomposed the large-scale climate indices and standardized monthly rainfall anomaly into subseries components with various time scales. The PSO algorithm was applied to find the optimal neuron numbers in ANN's layers (hidden), and the certain predictor predicted anomaly sub-series each rainfall. The HWNN method reduced over predictions and effectively forecasted the rainfall (monthly) from the large-scale weather signals and lagged rainfall anomaly data. HWNN method was more efficient for particular season rainfall prediction but took high prediction time in different season rainfall prediction.

C. Malarkodi, et al. [197], proposed Conditional Random Fields (CRFs) approach for named entity extraction from real-time heterogeneous agricultural data. Named entity tag-set for nineteen grained tags created and these tags could

cover almost all noticeable entities of the agricultural domain. Authors obtained 76% accuracy with three parameters (word, chunk information, and POS) for linguistically enriched and domain-independent features activities. To minimize the tags ambiguity, the post-processing heuristics approach implemented and improve precision & recall values are 83.24% and 83.13% respectively. Proposed methods also compared with CreateGazetteMED and improved unsupervised NEX technique. A relation extraction model is highlighted for the agricultural domain as a future scope.

S. Jiang, et al. [173] to support sustainability in the agricultural domain, proposed a model for integration of heterogeneous agricultural data and converted it into a database. This database may support the farmers for their day-to-day activity and researcher to work on real-time data. By reviewing contemporary semantic web applications for agricultural data, the authors also discuss the limitations and advantages of ontologies construction in the agricultural field.

Neha Kaushik and Niladri Chatterjee [198] worked on designing and automatic extraction of ontology for the agriculture domain. RelExOnt a rule-based reasoning method proposed that works into two parts, first part combines natural language processing techniques and regular expressions to extract ontologies from the agriculture domain. The second part identifies the semantic relationships between the phrases and extracted entities. The results (precision) of proposed algorithms was 86.89% that is higher as compared to human evaluation i.e. 75.7%. Proposed RelExOnt unable to extract relations like `grows_in_soil`, `grows_in_weather`, etc. due to less connected input data, so AGROVOC technique used to calculate recall values.

2.7 EXISTING EVENT EXTRACTION TECHNIQUES

The term ‘Event’ is related to an activity, which occurs at a specific time, place and involves one or more participants. The IE systems can be extended to check

the reported events on the various forum were really happened or not [199]. It is also used for extraction event arguments, identify their roles, to cluster and tract similar kinds of events. Event extraction has applications in many domains like market responses [200], trading suggestion [181], risk analysis [201], security events [202], real-time news [203] and botanical sciences [204].

O. Medeleyan et.al proposed a new algorithm for extracting index terms from agriculture-related documents and obtained an accuracy of 86% [175]. Information Extraction in the field of Agriculture plays a very important role because agriculture production depends a lot on various factors like weather, temperature, soil, etc. Therefore, gathering information from the aforementioned data and then performing event-based for analyzing the crops is a challenging task.

William H. et.al demonstrated the potential of Compound Specific Stable Isotope (CSSI) for soil resource management and protection of water resources for crop-specific sediment [171]. They worked on different crop regimes suited for different sediments. Ontology-based identification of diseases in crops is presented in [205]. A lot of work has been done in Extreme Weather Events (EWE) for the agriculture sector.

Cogato A. et.al have worked upon the EWE to interrelate EWE and the agriculture sector through a systematic and quantitative analysis [180]. They have analyzed nineteen major crops like horticulture, cereals, viticulture, legumes, and pastures across five continents. As per their survey, there lies a large research gap concerning EWE on major cash crops (grapevine and tomatoes).

Zurovac et.al discusses the challenges and opportunities in the field of agriculture in Bosnia and Herzegovina's (BH) [206]. BH is highly susceptible to climate change, which further causes a lot of problems to the agriculture sector. As per their suggestion, the government of BH should highly focus on event extraction techniques for weather prediction to save the crops.

In continuation to it, Falco C et.al is a survey report on “to which extent does the agriculture system depends upon the climate change”. Climate is one of the major factors that affect the agriculture system. They have discussed the three models to quantify the economic impact of climate change in the agriculture sector: agro-economic model, time-series model, and cross-sectional models [207].

Agro-economic models combine both crop and economic models. Based on a set of parameters, generally related to a particular crop whose growth depends on a change in environmental changes (temperature, precipitation, etc.). The agro-economic model focuses on the growth of Greenhouse Gas Emissions (GHS), the mapping of CHS into climate change, and the economic cost that occurred due to climatic change. The agro-economic model is applied to various crops and their results are then combined to stimulate the agricultural system [208]. The key advantage of this model lies that it estimates the change in crops required due to climate change and the farmer’s activities.

The cross-sectional model is based on the relationship between land rent and its corresponding productivity [209]. Based on Ricardian theory, where he has given a study for present and future rents of lands for maximizing the profits for farmers. The study was based on land rent, weather conditions, land productivity, etc. for different crops.

Time-Series Model also known as Panel-Data Model. It is an advancement in climate econometrics. It discusses the economic impact of climate change due to in the agriculture sector due to weather conditions [210]. They surveyed the effect agriculture caused due to fluctuations in weather, especially in US countries. In this model, all the other factors like soil quality, soil management can put to some fixed limit and the effects on the crops caused due to weather can consider individually. This model has thus motivated the researchers to think in this line of action, where all the other factors can fix and only the effect of weather can be used to IE regarding Crops.

Few researchers have worked on weather data extraction, which leads to several applications like weather suitable for different crops, to alarm the natural hazards w.r.t to weather, etc. Rossi et.al proposed the early detection of a flood using data from social media streams [211]. They have proposed automated services, which first collect the information about the weather forecasts from online social media. Based on extracted data, they have proposed qualitative feedback for metrological models, any sort of emergency event, and create awareness regarding upcoming natural hazards. The proposed approach acclaims to be highly useful to metrological offices and monitoring agencies, who act in the early warning phase.

AL Garrido et.al developed an NLP application that extracted the information from worldwide weather forecasts and further qualified the accuracy of weather forecasts [166]. This automatic task is used in verification task that reduces the typical human error while performing it manually. The accuracy obtained by them is very high and is very much liable to be used in real-time applications.

Table 2.2 Survey on Information Extraction Tasks in Various Application Areas

IE Task(s) [Ref]	Type of Data	Application Area	Technique Applied	Evaluation of Accuracy	Language Used	Single or Multilingual
NER [186]	Noisy & Uncertain Data	Social Media (Tweets)	SVM	Precision =75% Recall = 50%	English	Single
EE [119]	Noisy & Uncertain Data	Social Media (Tweets on earthquakes and typhoons)	SVM	Precision =80-90% Recall = 60-65%	English	Single

NER [187]	Noisy & Uncertain Data	Social Media (Tweets)	CRF & Labeled Latent Dirichlet Allocation	Precision =70% Recall = 60%	English	Single
NER-Segmentation & Classification [212]	Noisy & Uncertain Data	Social Media (Tweets)	Combine KNN & CRF	Precision =80% Recall = 70%	English	Single
EE [174]	Noisy & Uncertain Data	Social Media (Tweets)	CRF Based Approach	Precision = 85%	English	Single
NER & EE [213]	Noisy Data	Social Media (Tweets)	Rule-based IE	Not Mentioned	English	Single
NER & EE [170]	Unstructured data	Social Media	Framework for IE	Not Mentioned	English	Single
NER & EE [214]	Noisy & Uncertain Data	Social Media	Framework for IE	Not Mentioned	English	Single
NER [189]	Noisy Data	Medical	Rule-based OIE	Precision = 75%	English	Single
NER [190]	Noisy Data	Medical	Framework for IE	Not Mentioned	English	Single
NER-Propositions [215]	Unstructured Data	Domain-Independent	Rule-based OIE	Precision [English]= 75% Precision [Portuguese]= 53% Precision [Spanish] = 55%	English, Portuguese and Spanish	Multilingual
NER & EE [179]	Noisy Data	agricultural	Support Vector Machine (SVM)	Not Mentioned	English	Single

NER & EE [216]	Noisy data	Social Media (Tweets)	Rule-based IE	Not Mentioned	English	Single
NER & EE [217]	Semi-Structured Data	Research papers	HMM	92.9% (accuracy)	English	Single
NER & EE [218]	Structured Data	Social media	LSTM	95%	English	Single

Table 2.2 presents the summary of recent state-of-the-art approaches for Information Extraction in various fields focused on unstructured data. The present research work focuses on the extraction of information from the agricultural domain with deep learning techniques. After having a detailed discussion on various Information extraction fundamental tasks, Table 2.3 shows the detailed survey on various deep learning techniques.

Table 2.3 Survey on Agricultural Domain in related to Deep learning Techniques

Agri-Area	Problem Description	Data Used	DL	DL Arch.	Accuracy	Comparison	Hardware / Ref Real-Time?	Ref
Plant Disease Identification	Find suitable DL architecture for real-time tomato diseases and pest recognition	5000 images dataset created by the authors in several farms from Korea	Faster R-CNN, SDD, and R-FCN	VGG ResNet ResNeXt	86 % (R-FCN)	N/A	GPU / Yes	[276]
	Recognize cucumber diseases using leaf symptoms images	14208 images dataset created by the authors	CNN	Developed by authors	93.40%	AlexNet: 94% SVM: 81.9% RF: 84.8%	GPU (Nvidia Quadro) / N/A	[277]
	Detection of ESCA disease in Bordeaux Vineyards	6000 images dataset created by the authors in two	CNN	MobileNet	90.70%	SIFT Encoded: 87.9%	GPU (Nvidia GTX) / Yes	[278]

		Bordeaux Vineyards						
	Identification of plant/disease combination in 25 different species with 58 classes	Open database with 87484 photos of healthy and infected leaves	CNN	AlexNet, VGG, GoogleNet and Overfeat	99.48% (VGG)	N/A	GPU (Nvidia GTX) / Yes	[279]
	Identify 10 common rice diseases in leaves	500 images of healthy and diseased rice leave from database	CNN	Developed by authors	95.48%	BP: 92% SVM: 91% PSO: 88%	N/A / N/A	[280]
	Identification of 8 kinds of maize leaves diseases	500 images collected from public sources (Plant Village dataset and Google Websites)	CNN	Modified GoogleNet and Cifar10 (CAFFE FW)	98.9% (Google Net)	N/A	GPU / N/A	[281]
Plant Recognition	Identification of 16 plant species	DataSet collected by the governmental project over 1200 agro-stations	CNN	Developed by authors	97.47%	SVM : 89.94%	GPU (Nvidia Quadro) / N/A	[282]
	Identify 184 different species	LeafSnap dataset, FOLIAGE dataset, and FLAVIA dataset	CNN	LeafNet (Developed by authors)	LeafSnap : 86.3% FOLIAGE: 95.8% FLAVIA : 97.9%	SVM accuracy: LeafSnap : 79.66% FOLIAGE: 98.75% FLAVIA : 98.69%	GPU (Nvidia Gtx) / N/A	[283]
	Species and trait recognition from herbarium scans	Images of 1000 species on GBIF dataset (a subset of 170 species)	CNN	Modified ResNet model (TensorFlow FW)	82.40% (96.3% Top5)	Alternative DL approach : 90.3% (Top5)	GPU (Nvidia Titan) / N/A	[284]

						accuracy)		
	Identification of 1000 species	LifeClef 2015 dataset (90,000 images)	CNN	AlexNet, GoogleNet and VGGNet (CAFFE FW)	80%	N/A	GPU (Nvidia Tesla) / N/A	[285]
Land Cover Classification	Identify 13 different crop types	Multi-temporal Landsat enhanced Vegetation Index	RNN and CNN	LSTM and Conv1D	85.54% (Conv1D)	MLP accuracy: 83.81% XGBoost: 84.17% RF: 84.09% SVM: 83.09%	GPU (Nvidia Quadro) / N/A	[286]
	Crop mapping of 14 different species	Multi-temporal Sentinel-1 SAR images	CNN	Keras (TensorFlow FW)	91%	N/A	CPU (Intel Xeon) / N/A	[287]
	Identify 19 crop types	Sentinel-2A observations	RNN	LSTM	84.40%	Other RNN: 83.4% CNN: 76.8% SVM: 40.9%	N/A / N/A	[288]
Weed classification	Weed detection and classification in soybean crops	400 crop images captured by the authors with UAV	CNN	CafeNet (CAFFE FW)	98%	SVM: 98% AdaBoost: 98.2% Random Forest: 96%	GPU (Nvidia TITAN) / N/A	[289]
	Weed detection and classification by spectral band analysis	200 Hyperspectral images with 61 bands	CNN	MatConvnet	94.72%	HoG: 74.34%	CPU (intel i7) / N/A	[290]
	Accelerate a DL approach with FPGA for weed classification	18 000 weed images from the DeepWeedX dataset	CNN	VGG-16, DenseNet-128-10,	90.08% (DenseNet)	Other work approaches: ResNet: 95.7%	GPU and FPGA (intel DE1-0C) / Yes	[291]

Seed Classification	Estimate number of seeds into soybeans on pods**	Pods photography over a lightbox. The dataset created by authors	CNN	Developed by authors (Theano FW)	82.70%	SVM: 50.4 %	GPU (Nvidia GTX) / N/A	[292]
	Classify and sort 6 kinds of seeds	Pictures of hundreds of thousands of seeds (created by authors)	CNN	ResNet-18	99%	YOLO900: 86% SSD: 94 %	4 CPU and 1 GPU /Yes (500 fps)	[293]
Fruit counting	Detect mango fruits in trees canopies and estimate fruit load	1300 images acquired by authors at 5 mango orchards	CNN	Darknet (Mango Yolo - YOLO modified by authors)	98.30%	R-CNN: 95.3% SSD: 98.3 % YOLOv3 : 96.7% YOLOv2 : 95.9 %	GPU (Nvidia GTX) /Yes (14 fps)	[294]
Fruit classification	Combine DL, tracking & SFM for robust visible fruit counting on orange and apple orchards	N/A	FCN	Developed by authors	Orange count: 99 % Apple count: 97 %	N/A	N/A / N/A	[295]
	Classification of 18 types of fruit	3600 images acquired by authors and downloaded from public websites	CNN	Developed by authors	94.94%	PCA + kSVM: 89.11% WE+BB O: 89.47% FRFE + BPNN: 88.99%	GPU (Nvidia GTX) / Yes	[296]
Soil/Root segmentation	Segment Soil/ /root in X-ray tomography with CNN+SVM	Imagenet Dataset	CNN	N/A	0.57 (Specific quality metric)	N/A	CPU (intel Xeon) and GPU(A MD) / N/A	[297]

	Detect and quantify chicory roots	50 annotated Chicory root images created by authors	CNN	U-Net	99.70%	FrangiNet: 99.6%	GPU (Nvidia Titan) / N/A	[298]
Cattle detection	Detect and count cattle in UAV images	13520 images captured during UAV flight	CNN	Developed by authors	95.50%	N/A	CPU (Intel i7) / 3.2s/frame	[299]
Precision Irrigation	Inferring moisture conditions from images	Simulated large datasets of 1200 aerial images of Vineyards	CNN	CNNUP and CNNCF	0.034 (mean error)	SVM: 0.038 RFUP: 0.060 RFCF: 0.094 2 layer NN: .086 (mean error)	1 CPU (intel i7) and 3 GPU (Nvidia Titan) / N/A	[300]
Obstacle Detection	Detect humans, obstacles and traversable obstacle agricultural fields for safe machinery operation	ImageNet and COCO dataset	CNN and FCN	Darknet (improved YOLO), VGG and DeepAnomal (AlexNet)	70.81 % (accuracy of fusion of different approaches)	N/A	N/A / Yes	[301]
Measure Features	Stalk count and stalk width of crops	400 Stereo camera Images collected with the robotic platform	Faster RCNN and FCN	StalkNet (created by authors)	10% error for count; 2.76 mm error for measure	N/A	GPU (Nvidia GTX)/ 2 fps	[302]
Crop yield Estimation	Yield prognosis in vineyards by object counting	50 GB of Images from 2 vineyards at different crop stages	Faster R-CNN, R-FCN, SSD	TensorFlow FW	99.6% (F.R-CNN for object detection)	N/A	N/A / Yes	[303]

Automatic labeling of agricultural regulations	Translation of phytosanitary regulations into formal rules	N/A	CNN and BLSTM	Developed by authors	88.3% (F1 score)	N/A	GPU (Nvidia Quadro) / N/A	[304]
--	--	-----	---------------	----------------------	------------------	-----	---------------------------	-------

For weight manipulation in every backpropagation, deep learning techniques apply an activation function based on a well-suited optimizer. Heuristic algorithms aim to find a good/feasible solution to any optimization problem. They are based on “Trial and Error “methods, in a reasonable amount of computing time. Metaheuristic Algorithms are the advanced version of heuristic algorithms. These are the higher level of heuristic algorithms. Several other metaheuristic algorithms used nowadays, for instance, Evolutionary Algorithms (EA) including Genetic Algorithms (GA), Simulated Annealing (SA), Tabu Search (TS), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Bee Algorithms (BA), Firefly Algorithms (FA), and Harmony Search.

Table 2.4 shows the detailed solution-based survey on a single solution and population-based metaheuristic approaches.

There are different domains where IE is playing an important role. In the present work as well, four standard domains weather, soil, agricultural, and pest & fertilizers are used. The research target is to design a concluded corpus based on the above-said corpora. Various existing tools such as Protege-OWL ontology editor and knowledge-base framework, Cmap Tools, Oracle 11g database management for Kharif Database, Jena Adapter for Oracle Database, and Eclipse IDE applied on these corpora.

Table 2.4 Survey on Single-Solution and Population-based Metaheuristic Approaches

Metaheuristic	Optimizer Name	Method Used	Application	Ref
Single-solution Based Metaheuristics	Simulated Annealing	Statistical Mechanics	Discrete or continuous optimization problems	[219]
	Tabu Search	Escape from local minima	Embedded local minima	[220]
	Greedy Randomized Adaptive Search Procedure	Memory-less multi-start metaheuristic	Combinatorial optimization	[221]
	Variable Neighborhood Search	Variable neighborhood descent	Exploration of dynamically changing neighborhoods for a given solution	[222]
	Guided Local Search	Augmented objective function	Traveling salesman problem	[223]
	Iterated Local Search	Perturbation mechanism	Used for iterative problems	[224]
Population-based Metaheuristics	Evolutionary Computation	Darwinian principles	Combinatorial optimization	[225]
	Estimation of Distribution Algorithms	Probabilistic Model-Building Genetic Algorithms	Problems in domains such as Engineering, Biomedical Informatics, Robotics	[226]

	Differential Evolution	Global optimization	To solve single-objective optimization problems in continuous	[227]
	Coevolutionary algorithms	Coevolution	Game playing strategies, evolving better pattern recognizers, coevolve complex agent behaviors	[228]
	Cultural Algorithms	Computational models	Modeling the evolution of agriculture, job shop scheduling problem, re-engineering of Large-scale Semantic Networks, combinatorial optimization problems, multi-objective optimization problems, agent-based modeling systems, etc. Recently, many optimization methods have been combined with CAs, such as evolutionary programming, particle swarm optimization,	[229]
	Scatter Search and Path Relinking	Combining decision rules and problem constraints	Scatter Search	[230]

2.8 SUMMARY

The current chapter describes three data formats that are structured, semi-structured, and unstructured data. Various existing IE methods are discussed i.e. used to process and extract significant information from an unstructured corpus. Existing NER and EE techniques shows the strength and weakness of contemporary IE algorithms. A detailed survey on different fields of IE on various parameters shows the need to work on IE. Deep learning along with existing optimizers define good results in various application areas. The following chapter of the thesis aims to build the research framework for the extraction of semantic information from four corpora i.e. weather, soil, agriculture, and pest & fertilizer.

Chapter 3

PROPOSED FRAMEWORK FOR SEMANTIC INFORMATION EXTRACTION

3.1 INTRODUCTION

In the last two decades, several studies and experiments have been conducted to explore the research on IE [231]. Various statistical and machine learning techniques provided expected results when applied to numerous application areas. However, when the transformation of datasets to corpus happened then researchers decided to change their focus from simple machine learning algorithms to deep learning-based algorithms to get the semantic information extraction.

In the present chapter, a theoretical framework has been presented for semantic information extraction. The proposed framework comprises three layers: to merge various behavior corpora into a single unified corpus, to offer the structure for understanding the deep learning-based relation extraction from the agricultural corpus, and lastly post-processing of obtained results to improve the overall efficiency. The introduction of the presented chapter discussed in section 3.1, section 3.2 discussed the existing frameworks related to semantic IE. In section 3.3, discussion on various tools and techniques used for designing the proposed framework are preprocessing methods, LSTM, Adam optimizer along with various post-processing tools. Section 3.4 describes the theoretical aspects of the proposed framework and development using the deep learning technique; it also highlights the detailed explanation of philosophical viewpoints on conducting qualitative

research. Proposed frameworks have four components each tackling the issue as enumerated in this chapter. In section 3.4.1, the problem around the merging of multiple corpora into a single unit (detailed in Chapter-4). In section 3.4.2. Extraction of NER, EE, and RE through deep learning is described (detailed in Chapter-5). Section 3.4.2.1 demonstrates the techniques for semantic (relation) extraction. Section 3.4.2.2 discusses the significance of post-processing the relation extraction using heuristics and linguistics rules for better accuracy (detailed in Chapter-6) and section 3.5 summarizes the present chapter.

3.2 EXISTING FRAMEWORKS OF INFORMATION EXTRACTION

The structure to align research activities of theoretical and practical concepts on a specific area is called a research framework. Several frameworks are discussed below and present the practical and theoretical approaches for IE: [232] uses a framework for IE from biomedical literature tables, [233] presents a robust framework for IE from a different website, [234] employs a framework for automatic IE from research papers based on nanocrystal devices and [235] states an IE framework to build legislation network. Other practical approaches of IE frameworks are: [236] extracts semantic IE from unstructured data using RDF, [237] document spanners based relational framework for IE, and [238] shows the deep learning-based IE from the agricultural sector. The following section and subsections defined the tools and techniques i.e. used to design the proposed framework.

3.3 TOOLS & TECHNIQUES USED IN PROPOSED FRAMEWORK

This section is divided into four parts that are preprocessing methods, LSTM, Adam optimizer, and post-processing rules. The following subsections are detailed further.

3.3.1 PREPROCESSING METHODS

Data ambiguity generally arises while handling a large amount of data (corpus). Word Sense Disambiguation (WSD) is a research area of NLP and ML. WSD is a solution of ambiguous words, which arises due to distinct meaning words used in different contexts [239]. There are two types of WSD i.e. knowledge-based WSD and Corpus-based WSD.

3.3.1.1 KNOWLEDGE-BASED WSD

While working with a huge amount of lexical tokens like thesauri, dictionaries, and corpora, knowledge-based WSD becomes widely focused. By using the knowledge base from corpora, it mainly seeks to ignore training based on a large amount of data [240]. Generally, these WSD techniques use existing structured lexical knowledge base resources different from the following

- The use of lexical resources like monolingual or/and bilingual machine-readable dictionaries (MRD), thesauri, etc.
- The information is mentioned in the lexical resource.
- The property is used to find out the relation between words and senses.

Knowledge-based WSD techniques are recognized as ready-to-use tools for all words because this technique does not need sense-annotated data [241].

3.3.1.2 CORPUS-BASED WSD

Corpus-based learning is also called supervised learning in the area of NLP. The training of ML algorithms or statistical classification techniques prompted by using the semantically annotated corpus. Trained modules are enough cable to choose word sense from desired contexts. Commonly WordNet tools are applied for manually tagging by using semantic class (from specific lexical-semantic recourse) to corpora. Therefore, it requires maximum human intervention for training purposes [239].

3.3.1.3 TOOLS FOR WSD

There are several tools and web links also that can be directly used to automate the process of WSD. The following mentioned tools are commonly used in research areas of NLP or computer vision.

Resource Description Framework (RDF): RDF data model behaves similarly to ER or class diagram (classical conceptual modeling approaches) [242]. It works especially for web resources for making statements in triples expressions (subject–predicate–object). The subject, predicate, and object denote the resource, traits, or aspects of the resource and express the properties of the subject respectively. RDF also uses another approach i.e entity–attribute–value. Where object (entity) is used instead of the subject, attribute as predicates and value as an object e.g., this ink has color red. In some object-oriented approaches: an entity is ink, an attribute is a color and the value is red.

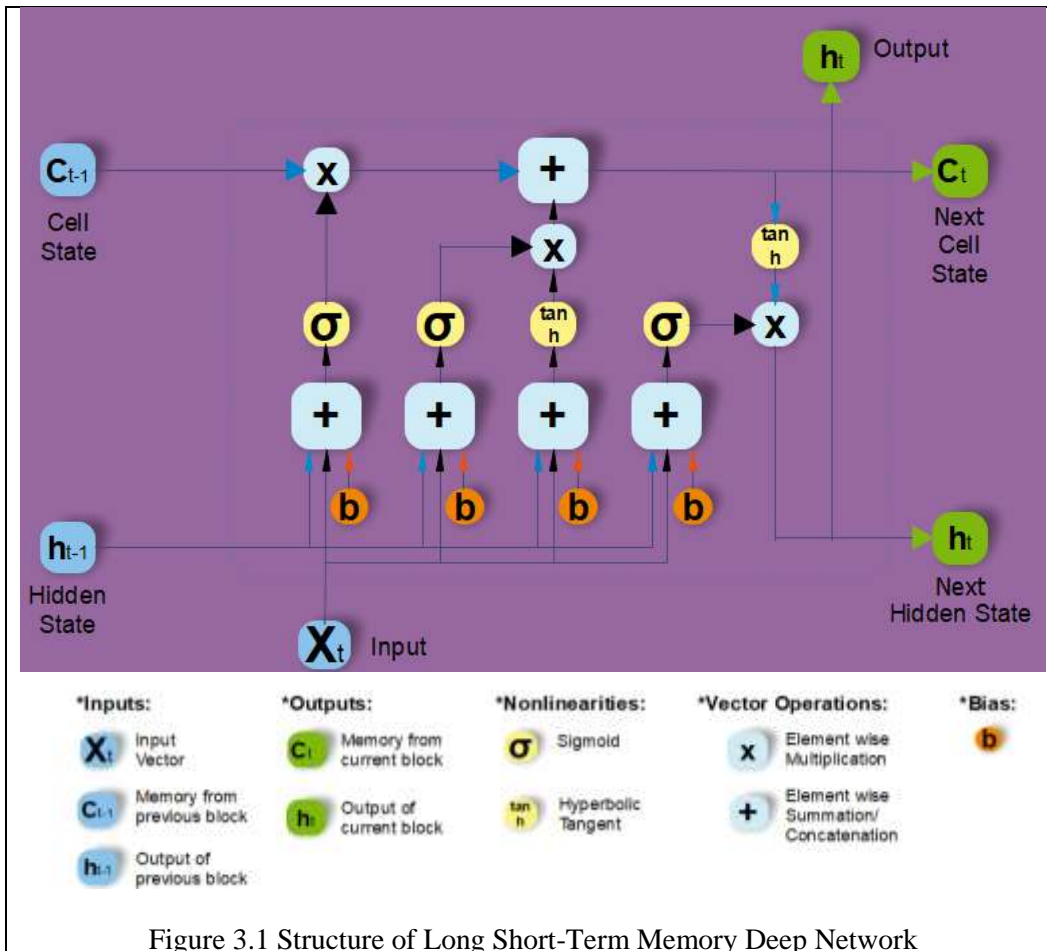
Ontology Language (OWL): OWL (semantic web language) is used to demonstrate the rich and complex information about an entity, group of entities, and relation between entities. The knowledge represented in OWL can be exploited by a computer program, so it is also called computational logic-based language. OWL documents (Ontologies) can be published on the Web and referred from (refer to) other OWL like Resource Description Framework (RDF), Resource Description Framework Schema (RDFS) and SPARQL Protocol and RDF Query Language (SPARQL), OWL is also the part of W3C’s Semantic Web technology [243, 244].

DARPA Agent Markup Language (DAML): Like RDF and OWL, DARPA is also used in the semantic web. It is a markup language based on RDF. It is used to define the sets of facts for making an ontology. DARPA had its roots in three main languages - DARPA Agent Markup Language (DAML), OIL (Ontology Inference Layer), and Simple HTML Ontology Extensions (SHOE) [245].

Natural Language Toolkit (NLTK): NLTK library is an open-source tool developed by Princeton University. The core functions of NLTK tools are to provide training data sets, taggers, stemmers, Wordnet corpus, various tokenizers, and lemmatizers [246].

3.3.2 LONG SHORT TERM MEMORY

The advanced version of RNN is known as LSTM. LSTM was introduced to resolve the gradient decedent problem of RNN by adding an extra memory cell per module. Figure 3.1 prescribed the architecture of LSTM [247].



LSTM is a special type of RNN by which remembering information and long-term dependencies can learn by the DL model for long periods. LSTM has four layers with a unique communication model.

LSTM uses the memory cell (gates) to handle the memorizing process, so it can use to design for preventing long-term dependency problems. While constructing the LSTM, the initial step is to identify the unessential information and delete it from the network using the memory gate. The sigmoid function is used to identify and exclude data from the network. This function takes current input X_t at time t and output of previous LSTM h_{t-1} at time $t-1$. The sigmoid function also determines which portion from the last output should omit from the network. This function is processed by the forget gate (f_t) in equation (1).

As per each number in the C_{t-1} (cell state), the value of vector f_t is ranging from 0 to 1,

$$f_t = \sigma (W_f [h_{t-1}, X_t] + b_f) \quad (1)$$

In the forget gate, weight, and bias are represented by W_f and b_f respectively. To decide, store, and update the cell state from input X_t , the following steps are used in two parts that are sigmoid layer and tanh layer. In equation (2), based on 0 or 1 sigmoid layer will decide that the new information should update or ignore and tanh function assign the value (-1 to 1) in equation (3) and decide the level of importance

After the multiplication of these two values, LSTM will update the state of the new cell. This updated memory added to the previous memory C_{t-1} resulting in a new C_t in equation (4)

$$i_t = \sigma (W_i [h_{t-1}, X_t] + b_i) \quad (2)$$

$$N_t = \tanh(W_n [h_{t-1}, X_t] + b_n) \quad (3)$$

$$C_t = [C_{t-1}f_t] + N_t i_t \quad (4)$$

Here, at time t-1 and t cell states are C_{t-1} and C_t , whereas weight matrices and bias of the cell state are denoted by W and b respectively. In the last step, the h_t (output values) is based on the O_t (output of cell state). Firstly in equation (5) sigmoid layer function selects which cell state part makes it to the output, then O_t (sigmoid gate output) multiplied by updated C_t (Cell state) values produced by the tanh layer i.e. ranging from -1 and 1 in equation (6).

$$O_t = \sigma (W_o[h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = O_t \tanh(C_t) \quad (6)$$

Here, weight matrices and bias of output gate are denoted by W_o and b_o respectively.

3.3.3 ADAM OPTIMIZER

The modification of SGD (stochastic gradient descent) called Adam (adaptive moment estimation) optimizer, which is broadly adopted by deep learning techniques especially in the area of NLP and computer vision [248, 249]. The SGD maintains the common alpha (linear learning rate) value for updating all weights and alpha does not change (or update) during the complete training process. Whereas Adam maintains the specific adaptive learning rate for the individual parameter of the first and second moment of the gradient. Adam is a combination of AdaGrad and RMSProp.

AdaGrad (Adaptive Gradient Algorithm) Maintain the individual parameter-learning rate to upgrade the performance of sparse gradient problems [250, 251].

RMSProp (Root Mean Square Propagation) Similar to AdaGrad, maintains the individual parameter-learning rates i.e. average weight of recent gradient magnitudes. It is well suited to the noisy online and non-stationary problems.

Like Adadelta and RMSprop, Adam also used to store an exponentially decaying average of the previous gradient m_t and squared gradient (v_t) in equations (7), (8), and (9). Getting gradients for the stochastic object at t time step

$$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1}) \quad (7)$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (8)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (9)$$

m_t and v_t are used to evaluate the first and second gradient moment i.e. mean and uncentered variance. The authors of Adam optimizer observes that m_t and v_t are moving towards biasing towards zero because they are initialized 0's vectors. This biasing towards zero noticed especially during the starting time steps and small decay rates (i.e. β_1 and β_2 close to 1). The biases offset defined by computing bias-corrected 1st and 2nd-moment estimates in equation (10) and (11):

$$m_t^{new} = \frac{m_t}{1 - \beta_1^t} \quad (10)$$

$$v_t^{new} = \frac{v_t}{1 - \beta_2^t} \quad (11)$$

The updated parameters in Adam with the combination of Adadelta and RMSprop optimizer in equation (12):

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t^{new}} + \epsilon} m_t^{new} \quad (12)$$

Table 3.1 is the recommendation of default parameter for Adam optimizer is especially for the deep learning algorithm's libraries defined below [248]. Where α , β_1 , β_2 and ϵ are learning rate, beta1, beta2 and epsilon respectively.

Table 3.1 Recommended Values of Default Parameters

DL Library	Learning Rate (α)	Beta 1 (β_1)	Beta 2 (β_2)	Epsilon (ϵ)	Decay Factor
TensorFlow, Lasagne, Caffe, MxNet and Torch	0.001	0.9	0.999	1.00E-08	NA
Keras	0.001	0.9	0.999	1.00E-08	0
Blocks	0.002	0.9	0.999	1.00E-08	1

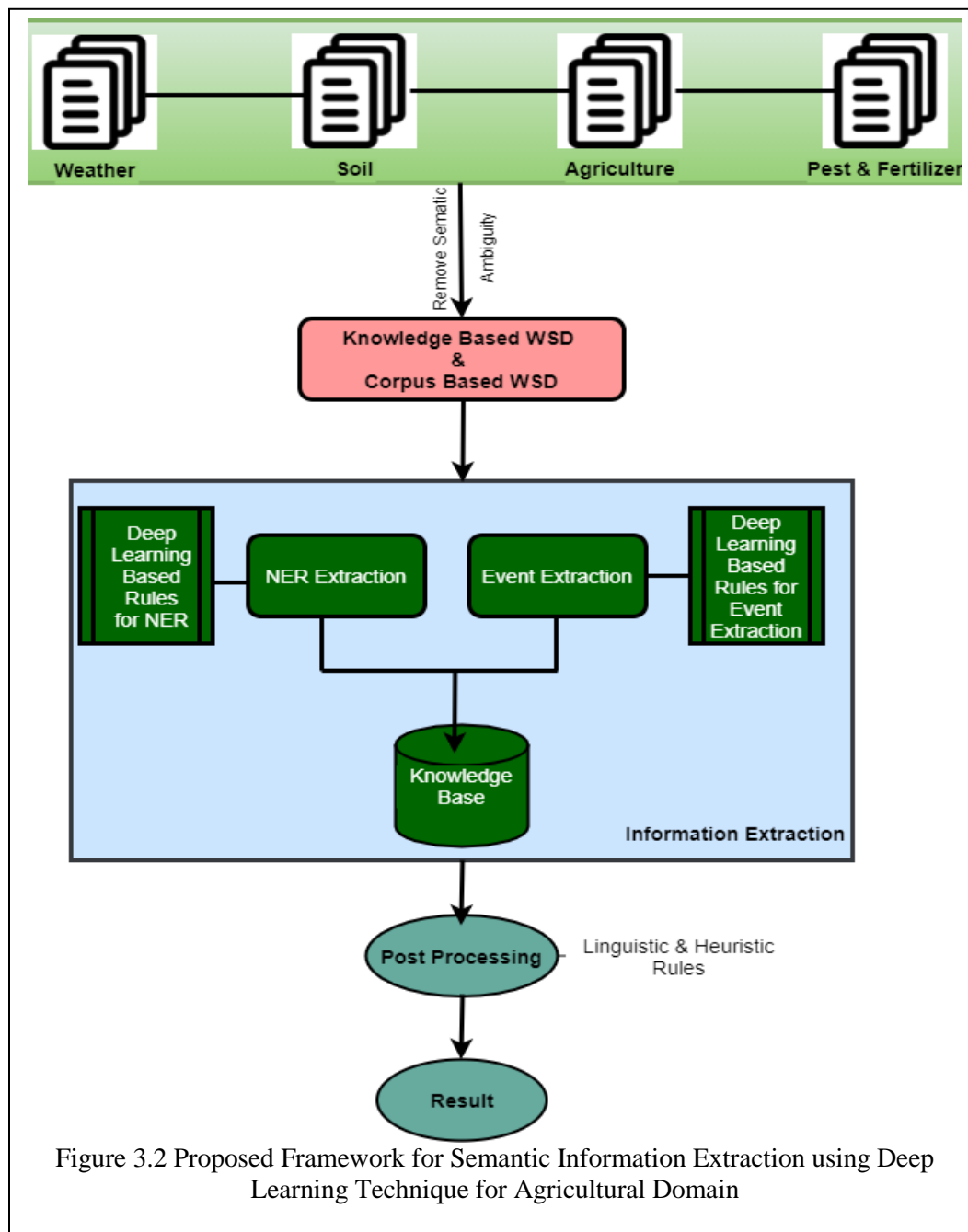
3.3.4 POST-PROCESSING RULES

Post-processing is not mandatory for all the obtained results, but to improve accuracy and remove inconsistencies few post-processing techniques can be beneficial. In few cases, precision and recall values are not enough to conclude that the algorithm behaves exceptionally in all conditions. The commonly used techniques are called linguistic and heuristic rules. Therefore, some boundary conditions have to be defined for getting better results. In an automatic IE system, a small slight error is very critical, so it is a good option to manually interrogate the system and fix the bugs (if required). e.g., extracting information about the distance value (such as 100km) is not a difficult task, but finding the distance between playground and hotel is always more difficult. So, it is recommended to define and maintain the boundary [252] (complete explanation in Chapter-6).

3.4 PROPOSED SEMANTIC EXTRACTION FRAMEWORK FOR AGRICULTURAL CORPUS

The research framework presents a theoretical and practical approach to extracting semantic information. Unlike few existing frameworks in literature, this

approach attempts to give a structure that highlights the fundamental concepts and components of semantic information extraction. Unlike few existing frameworks in literature, this approach attempts to give a structure that highlights the fundamental concepts and components of semantic information extraction.



The methodology followed in this study is composed of a collection of articles in the selected areas, collection of authenticating data in those relevant fields (mostly the benchmark datasets from repositories) selection of appropriate data mining tools, data storage tools (Excel, Oracle), and editing tools.

In this section, the operational framework is elaborated for presenting the complete flow of the research components carried out for this study. Since this study mainly spins around information gathering, data pre-processing, semantic extraction, and data post-processing. These four core or concentrated parts involve in practical implementations of this framework.

Figure 3.2 shows the overall view of the present research, framework divided into four different modules are Corpus concatenation, Deep network-based NER and EE, Semantic Extraction, and Post-processing. The following subsection presents a brief of each step, and subsequent chapters provide a detailed description.

3.4.1 CORPORA CONCATENATION

This module comprises two submodules i.e. corpora collection and Corpora Amalgamate. These two submodules are discussed in the next two subsections.

3.4.1.1 CORPORA COLLECTION

In the corpora collection, four unstructured nature corpora were collected from different data collection sources. Weather corpus collection from data sources India Meteorological Department, Meteorological Centre Dehradun. This dataset contains the values of the last twenty years' data based on six different parameters. The parameter for weather corpus is cloud coverage, minimum temperature, maximum temperature, average temperature, vapor pressure, and rainfall. For a reference of input corpus, Table 3.2 contains the sample data set for the weather corpus. Chapter-4 containing a detailed explanation of input corpora.

Table 3.2 Sample Data for Weather Corpus

Data Samples	Cloud Cover	Minimum Temp	Average Temp	Maximum Temp	Vapor Pressure	Rain fall
0	10.063	16.626	21.756	26.898	18.923	0
1	14.538	19.608	24.733	29.874	21.37	0
2	16.185	17.775	22.939	28.104	19.629	0.1
3	16.632	17.444	22.773	28.164	19.583	0.2
4	18.407	17.461	22.59	27.732	18.534	0
5	11.62	17.208	22.283	27.385	19.127	1.5
6	23.886	19.117	24.102	29.134	20.185	9.1
7	24.425	19.289	24.419	29.56	20.185	0.6
8	16.652	19.89	25.028	30.17	21.19	0.5
9	19.057	18.712	23.502	28.309	19.952	16
10	18.166	19.034	24.135	29.26	20.628	1.6
11	17.932	18.951	24.073	29.214	20.185	0.1
12	17.932	19.644	24.781	29.921	21.312	5.7

Soil corpus is also collected from various sources such as The National Bureau of Soil Survey and Land Use Planning (<https://www.nbsslup.in/>). In the last thirty years, this organization has developed a soil database based on laboratory and field studies. This has generated maps and soil information at different scales, showing the area and distribution of various soil groups in different agro-ecological sub-regions. Another soil corpus from the Department of Agricultural Research and Education, Ministry of Agriculture and Farmers Welfare (<http://dare.nic.in/>) also used as input corpus. Soil corpus parameter includes the primary, secondary and tertiary level in micronutrients. These nutrients play a vital role in agricultural productivity. Primary level macro-nutrients like nitrogen, phosphorus, and potassium; secondary level macronutrients are sulfur, calcium, and magnesium; and the third level of micronutrients are copper, iron, manganese, zinc, boron, chloride, nickel, and molybdenum in the soil. Quality of soil may include soil test results based on Cation Exchange Capacity (CEC), soluble salts (salinity), Organic

Matter (OM) content, and soil pH. Table 3.3 represents the example of Zinc (Zn) recommendation for Corn crop production utilizing the DPTA extractable.

Table 3.3 Zinc (Zn) Recommendation for Corn Crop Production

	Zn Soil Test	Zn application (ib/acre)	
	PPM	Broadcast	Band
Low	0.0-0.5	12	3
Marginal	0.6-0.9	7	2
Adequate	1.0+	3	1

The vital dataset for the current research is agricultural corpus i.e. freely downloadable from data.gov.in website https://data.gov.in/catalog/district-wise-season-wise-crop-production-statistics?filters%5Bfield_catalog_reference%5D=87631&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc+). The Krishi Vigyan Kendra, Dhakrani, District Dehradun, Uttarakhand also share the agricultural database based on season factor for rainfall data which can download from <http://agricoop.nic.in/sites/default/files/UKD7-Dehradun-10.07.14.pdf>. Table 3.4 shows the sample data for seasonally based crop productivity of the Dehradun region.

Table 3.4 Sample Data for Seasonal-based Crop Productivity

Season	Major Crops	Area ('000ha)	Productivity (kg/ha)
Rain (June-December)	Rice	11.4	19689.9
Winter (January-March)	Barley	0.7	19.57
Summer (April-May)	Potato	0.668	22140

Table 3.4 describes the sample data collected for major crop productivity like rice, barley, and potato. The rice gives more productivity like 19689.9kg per hectares (ha) during rainfall season, whereas barley gives nearly 20kg per ha for the winter season. Here, rice is considered the most important crop of its productivity. The potato can be cultivated during the summer season which produces 22140kg per ha.

The fourth and last component of data collection is the pest and fertilizer corpus. As similar to an agricultural database, this is also a free downloadable database. National Fertilizer Limited (NFL) is a government undertaking organization that is working in the area of pest and fertilizer. NFL is not only maintaining the database but also producing pest and fertilizer-based products (neem-coated urea, bio-fertilizer, bentonite Sulphur, etc.).

Table 3.5 Usage of Pesticide Classified by Crop Types

Objective	Types of Crop							
	Vegetable		Farm		Fruit		Rice	
	Fq	%	Fq	%	Fq	%	Fq	%
Herbicides	2	1.4	90	65.5	3	2.5	109	79.7
Acaricides	3	2.1	14	10.2	4	2.8	47	35.2
Fungicides	3	2.1	33	23.5	5	3.4	58	45.8
Insecticides	1	0.9	45	34.2	6	4.4	106	75.9
Others	2	1.4	5	3.3	0	0	4	2.8

Table 3.5 demonstrate that most of the farmers cultivated rice crop and majorly herbicides pesticide uses (79.7%), followed by insecticide (75.9%), fungicides (45.8%) and acaricides (35.2%). The farm crop is the second crop and the 65.5% pesticide use in this crop was herbicides, followed by insecticides, fungicides, and acaricides (32.4, 25.7%, and 9.6% respectively). Herbicides may be cheaper, more efficient, and more practical to use than weeding costs [253, 254]. After corpora

collection, their amalgamation becomes more important, so the next subsection describes the steps of corpora amalgamation.

3.4.1.2 CORPORA AMALGAMATE

After collection of these above said unstructured corpora, these shapeless corpora have to merge into one single corpus. As the nature of these corpora is different (details in chapter 4), so various preprocessing algorithms like data normalization, min-max algorithm, and data manipulation have been used. Knowledge-based WSD and corpus-based WSD are applied to this vast amount of data. The main purpose behind using these two popular algorithms is to remove semantic ambiguity and merge different natured corpora into a single entity. This single corpus is ready for the input of a deep network-based information extraction engine.

3.4.2 DEEP NETWORK-BASED NER AND EVENT EXTRACTION

In the second module, a complete IE engine plays the main role. Here input value is an unstructured corpus that is a combination of weather, soil, agricultural, and pest, and fertilizer database. A deep learning-based neural network designed to extract the named entity recognition and event extraction processes. This deep network uses the LSTM for NER and EE extraction. LSTM has an advantage over the RNN in that it has a memory cell (forget gate). This cell helps the network to store the last state in memory, resulting in that the deep network overcome will resolve the gradient decedent problem. RAO is also used to optimize the network model's shape and mold it into the most possible accurate form by weights futzing.

3.4.2.1 SEMANTIC EXTRACTION

Deep learning is used to automate the process of capturing significant information from unstructured texts. In the decision process, various states represent important information. After extraction of these two different information extraction components i.e. extracted Named entity recognition and information

about the entity extraction, the framework directed the research towards the semantic (relations) extraction between these two components.

3.4.2.2 POST-PROCESSING

The extracted semantic is further post-processed by using heuristics and linguistic rules to increase the overall accuracy of the proposed model. Post-processing helps the proposed model to improve the other matrix parameters (precision, recall, and f-score). The post-processed data are the expected research outcome that is by using the proposed techniques farmers may advise selecting the appropriate crop for farming; results may increase the efficiency of crop yielding.

3.5 SUMMARY

The present chapter demonstrates the approaches, methodologies, and other related details of the proposed framework for designing a deep learning-based model to extract the semantic information from agricultural corpus. Through various applications and existing frameworks of IE, the proposed framework guides to design of the research pathway in the IE area; and maintains a research structure, which helps to move forward in practical aspects of the theoretical framework. The proposed model intends to minimize the gap that exists between the practically experienced user and the corresponding engineering requirements. In the subsequent chapter, the proposed framework helps to perform data preprocessing, semantic extraction (between NER and EE), and post-processing of semantics. These three modules are divided into the next three chapters that experiments to conduct the investigation based on several parameters and discussed how the acquired results help to validate the framework. The next chapter focuses on the behavior of the target corpora, defined along with the corpora integration through WSD based techniques and proposed disambiguation algorithm.

CHAPTER -4

DATA PREPROCESSING

4.1 INTRODUCTION

To extract semantic information from an unstructured corpus is an important task in NLP. Various existing methods successfully mined the information from free text. The present research work proposes a framework for extracting semantic information from the agricultural domain. This chapter covers the preprocessing methods applied for the experiments carried out for the proposed study and the material's minimum needed to appreciate the same. Mathematical expressions and definitions are also present to support the materials needed. The implementation details of the various applied data preprocessing techniques and the reason behind choosing these techniques to handle data pre-processing are also discussed. The complete chapter is divided into five modules i.e. introduction, corpora collection, data normalization, corpora concatenation, and chapter summary. Section 4.1 shows the chapter's introductory part. Four different natured corpora collection approaches discuss in section 4.2. Section 4.3 describes various data normalization techniques to handle noises in targeted corpora. Section 4.4 is classified into two parts: the first part shows the combination of four said corpora into a single unit based on the integration of knowledge-based WSD into corpus-based WSD and the second part describes the implementation of the proposed disambiguation algorithm. Section 4.5 concluded the summary of the present chapter.

4.2 CORPORA COLLECTION

The direct or indirect impact on agricultural sectors by various areas such as weather, soil map, pest and fertilizer used, farmer's knowledge, climate change,

regional pollution, infrastructure, and many more, always considered to analyze and predict overall productivity in the agricultural domain [255, 256]. To extract the semantic from the agricultural domain, the present research work considers the weather, soil, and pert & fertilizer corpora taken along with agricultural corpus. The collection of said corpora discussed in the following subsections.

4.2.1 WEATHER CORPUS

The present research work focused on the Dehradun region. Weather corpus based on Dehradun region collected from India Meteorological Department (IMD), Meteorological Centre Dehradun. This knowledge base consists of seven parameters such as cloud coverage, minimum temperature, maximum temperature, average temperature, vapor pressure, rainfall, and average relative humidity of the last twenty years. The collected weather corpus contains the geographical area the dataset of Table 4.1 contains the sample of weather corpus.

Table 4.1 Sample Data of Weather Corpus-IMD, Dehradun

S. No	Cloud Cover (oktas)	Minimum Temp (°C)	Average Temp (°C)	Maximum Temp (°C)	Vapor Pressure (Pa)	Rainfall (mm)	Average Relative Humidity (g/m ³)
1	18.92300 0000000	7.04838 70968	13.1887096 8	19.3290322 58065	21.756000 000000	29.00000 0000000	92.50000 0000000
2	21.37000 0000000	10.3285 714286	16.8535714 28571	23.3785714 28571	24.733000 000000	23.00000 0000000	77.65666 6666667
3	19.62900 0000000	13.1000 000000	19.4806451 61290	25.8612903 22581	22.939000 000000	181.4000 0000000 0	86.15161 2903226
4	19.58300 0000000	16.9466 666667	23.6316666 66667	30.3166666 66667	22.773000 000000	60.90000 0000000	64.08000 0000000
5	18.53400 0000000	21.3387 096774	28.8483870 96774	36.3580645 16129	22.590000 000000	10.70000 0000000	91.87096 7741936
6	19.12700 0000000	23.1833 333333	28.8600000 00000	34.5366666 66667	22.283000 000000	144.9000 0000000 0	67.92100 0000000
7	20.18500 0000000	23.6741 935484	27.1761290 32258	30.6780645 16129	24.102000 000000	566.0000 0000000 0	96.30123 2258065

8	20.18500 000000	23.2161 290323	26.8580645 16129	30.5000000 00000	24.419000 000000	654.2000 0000000 0	62.37857 1428571
9	21.19000 000000	21.5466 666667	26.6016666 66667	31.6566666 66667	25.028000 000000	78.00000 0000000	94.86129 0322581
10	19.95200 000000	17.2322 580645	23.6919354 83871	30.1516129 03226	23.502000 000000	27.30000 0000000	64.31666 6666667
11	20.62800 000000	12.7200 000000	19.9000000 00000	27.0800000 00000	24.135000 000000	3.500000 0000000	82.35806 4516129
12	20.18500 000000	7.97741 93548	14.9241935 48387	21.8709677 41936	24.073000 000000	8.800000 0000000	61.53666 6666667
13	21.31200 000000	19.6440 000000	24.7825000 00000	29.9210000 00000	24.781000 000000	5.700000 0000000	79.67806 4516129
14	7.851612 903226	6.56451 61290	14.3209677 41936	22.0774193 54839	19.329032 258065	56.70000 0000000	56.93103 44828
15	9.946428 571429	9.56428 57143	15.8208525 34562	22.0774193 54839	22.200000 000000	56.70000 0000000	66.93103 44828
16	12.38064 5161290	11.6000 000000	18.4500000 00000	25.3000000 00000	24.534482 758621	66.20000 0000000	78.10000 00000
17	18.54000 000000	18.0000 000000	24.6500000 00000	31.3000000 00000	21.400000 000000	39.70000 0000000	95.03666 66667
18	21.16129 0322581	20.3000 000000	26.8333333 33333	33.3666666 66667	23.378571 428571	29.00000 0000000	90.51935 48387
19	22.64000 000000	23.9838 709677	30.2919354 83871	36.6000000 00000	25.428571 428571	12.50000 0000000	58.70666 66667
20	24.17096 7741936	23.9838 709677	29.3919354 83871	34.8000000 00000	21.325806 451613	402.4000 0000000 0	66.72903 22581
21	24.03225 8064516	24.2290 322581	28.2445161 29032	32.2600000 00000	25.300000 000000	402.4000 0000000 0	54.80000 00000
22	22.23000 000000	23.1774 193548	26.8487096 77419	30.5200000 00000	25.200000 000000	395.6000 0000000 0	81.55666 66667
23	17.16129 0322581	17.0419 354839	24.1976344 08602	31.3533333 33333	22.077419 354839	490.2000 0000000 0	52.39677 41935
24	11.30000 000000	12.7000 000000	21.7000000 00000	30.7000000 00000	22.096428 571429	22.30000 0000000	95.73666 66667
25	9.006451 612903	7.40000 00000	16.9733333 33333	26.5466666 66667	27.100000 000000	8.700000 0000000	58.46451 61290
26	18.03666 6666667	9.90000 00000	16.2048387 09677	22.5096774 19355	21.077419 354839	28.60000 0000000	54.50000 00000
27	21.51935 4838710	13.7000 000000	20.4000000 00000	27.1000000 00000	25.861290 322581	25.20000 0000000	65.20000 00000
28	23.70666 6666667	17.6225 806452	25.5112903 22581	33.4000000 00000	29.500000 000000	7.900000 0000000	87.40000 00000

29	23.72903 2258065	20.3000 000000	28.4000000 00000	36.5000000 00000	28.587096 774194	14.20000 0000000	83.98387 09677
30	23.80000 0000000	24.0000 000000	27.9983870 96774	31.9967741 93548	25.428571 428571	24.60000 0000000	90.29193 5483871

Table 4.1 contains the weather information of the Dehradun region, as the values mentioned in the above table are noisy, so before merging with other corpora, data normalization techniques must be applied to it (see section 4.3).

4.2.2 SOIL CORPUS

Soil health is an important component for better crop yielding. The analysis of the soil map provides the necessary information to set nutrient values in the soil. This information further used in later stages for calculating the pest & fertilizer applied. For evaluating soil nutrients, Nitrate-Nitrogen (Dryland), Nitrate-Nitrogen (Irrigated) (lb/ac), Phosphorus (lb/ac), Potassium (lb/ac), Sulphur (lb/ac), Copper (ppm), Manganese (ppm), Iron (ppm), Zinc (ppm), Boron and Chloride (ppm) are commonly used term. Soil corpus gathered from various research organizations such as District Soil Testing Laboratory, Dehradun/ Soil Testing Laboratories located at Nanda ki Chowki, Premises of Directorate of Agriculture, Premnagar, Dehradun. Based on lab and field experiments, this organization has developed a soil database for the last 30+ years. Another soil corpus from the Department of Agriculture, Cooperation & Farmers Welfare (DACFW) under the Ministry of Agriculture and Farmers Welfare (MAFW) (<http://agricoop.nic.in/hi>) also used. The reports from this organization categorized the soil corpus into three levels i.e. primary (nitrogen, phosphorus, and potassium), secondary (sulfur, calcium, and magnesium), and third (copper, iron, manganese, zinc, boron, chloride, nickel, and molybdenum) level in the soil. Table 4.2 shows the soil health card results based on Dehradun agricultural terrain for these three micronutrient levels.

Table 4.2 Soil Health Card Issued by DACFW

S. No.	Parameter	Soil Result	Observation	Unit
1	pH	7.12	Neutral	mEq/L
2	EC	0.25	Normal	mEq/L
3	OC	0.5	Low	mEq/L
4	Nitrogen (N)	280	Low	mEq/L
5	Phosphorus (P)	24.82	High	mEq/L
6	Potassium (K)	147.44	High	mEq/L
7	Sulphur (S)	33.52	Very High	PPM
8	Zinc(Zn)	0.3	Low	PPM
9	Boron (B)	3.45	Normal	PPM
10	Iron (Fe)	13.42	Very High	PPM
11	Manganese (Mn)	31.25	Very High	PPM
12	Copper (Cu)	2.31	Very High	PPM

The above table shows various micronutrients based soil results along with the soil level (low, neutral, high, and very high). These values are useful for experimenting with the proposed model. This database contains additional information also like details of the farmer, soil sample, expected yielding with and without fertilizers, recommendation table for secondary micronutrients level, and many more. However, the information mentioned in Table 4.2 along with the recommendation table for secondary level micronutrients (Table 4.3) is significant for the proposed model.

Other extracted information from the soil corpus used to recommend zinc (secondary level micronutrients) is described in Table 4.3. It is recommendation information of Zinc (Zn) for Corn crop production utilizing the DPTA extractable in micronutrients

Table 4.3 Using DPTA Extractable Zn Extraction Method, Zn Recommendation for Corn Crop Production

	Zinc Soil Test	Zinc application (ib/acre)	
	PPM	Broadcast	Band
Low	0.0-0.5	12	3
Marginal	0.6-0.9	7	2
Adequate	1.0+	3	1

4.2.3 AGRICULTURAL CORPUS

The third vital corpora for the proposed model is the agricultural knowledge base. This is a freely downloadable corpus from the data.gov.in website. The Krishi Vigyan Kendra (KVK), Dhakrani, Dehradun, Uttarakhand also share the agricultural corpus based on seasonal crop yielding data which can download from <http://agricoop.nic.in/sites/default/files/UKD7-Dehradun-10.07.14.pdf>. This corpus contains unstructured information related to geographical information, rainfall, and agricultural products yielding for the Dehradun region. The below-mentioned details in the area (000'ha) and geographical area, net sown area, cultivable area, forest area, gross crop area, land under non-agricultural use, cultivable wasteland, land under misc. (tree crops, groves) are 308.8, 45.5, 48.9, 20.5, 66.3, 21.4, 63.4 and 14.7 respectively. Table 4.4 shows the normal rainfall in Dehradun based on the four below-mentioned seasons.

Table 4.4 Sample Data of Seasonal-based Normal Rainfall

S. No.	Season	Normal Rainfall (mm)
1	SW Monsoon (June-Sep)	1767.6
2	NW Monsoon (Oct-Dec)	86.6
3	Winter (Jan-March)	147.6
4	Summer (Apr-May)	63.7
5	Annual	2065.7

Table 4.5 Sample Data for Seasonal-based Crop Productivity

S No.	Season	Major Crops	Area ('000ha)	Productivity (kg/ha)
1	Rain (June-December)	Rice	11.4	19689.9
2	Winter (January-March)	Barley	0.7	19.57
3	Summer (April-May)	Potato	0.668	22140
4	Spring (Feb-March)	Maize	9.3	9115
5	Summer (April-May)	Amaranth	1.2	5638

Table 4.5 shows the sample data of seasonally based crop productivity of the Dehradun region. It describes the sample data collected for major crop productivity like rice, barley, potato, maize, and amaranth. The rice gives more productivity like 19689.9kg per hectares (ha) during rainfall season, whereas barley gives nearly 20kg per ha for the winter season. Here, rice is considered the most important crop of its productivity. Maize gives 9115 kg per ha in the spring season. The potato and amaranth can be cultivated during the summer season which produces 22140kg and 5638 kg per ha.

4.2.4 PEST AND FERTILIZER CORPUS

Pest and fertilizer (PAF) corpus is the last component of corpora collection. As similar to the agricultural database, the PAF corpus is also freely downloadable. Three corpora collected from the following mentioned resource areas that are State Fertilizer Quality Control Laboratory, Dehradun, Fertilizer Quality Control Laboratory, Nanda ki Chowki, Dehradun, and National Fertilizer Limited (NFL). NFL is a government undertaking organization that is working in the area of pest and fertilizer. NFL not only maintaining the database, but NFL also producing pest and fertilizer-based products (neem-coated urea, bio-fertilizer, bentonite Sulphur, etc.). Table 4.6 shows the requirements, availability, and sale of various fertilizers

likes Urea, DAP, MOP, and NPKS. It is the comparison between requirement, availability, and sales of Kharif crop.

Table 4.6 Fertilizer-based Comparison between Requirement, Availability, and Sales of Kharif Crop [257]

S. No	Fertilizer	Requirement		Availability		DBT Sale	
		2018	2019	2018	2019	2018	2019
1	Urea	148.9	156.22	212.43	212.38	154.39	153.69
2	DAP	49.18	51.22	73.35	89.77	37.75	35.71
3	MOP	20.25	20.39	23.18	27.73	14.15	11.71
4	NPKS	49.73	52.97	86.49	92.97	50.12	46.35

Table 4.7 Usage of Pesticide Classified by Crop Types

Objective	Types of Crop							
	Vegetable		Farm		Fruit		Rice	
	Fq	%	Fq	%	Fq	%	Fq	%
Herbicides	2	1.4	90	65.5	3	2.5	109	79.7
Acaricides	3	2.1	14	10.2	4	2.8	47	35.2
Fungicides	3	2.1	33	23.5	5	3.4	58	45.8
Insecticides	1	0.9	45	34.2	6	4.4	106	75.9
Others	2	1.4	5	3.3	0	0	4	2.8

Table 4.7 demonstrate that most of the farmers cultivated rice crop and majorly herbicides pesticide uses (79.7%), followed by insecticide (75.9%), fungicides (45.8%) and acaricides (35.2%). The farm crop is the second crop and the 65.5% pesticide use in this crop was herbicides, followed by insecticides, fungicides, and acaricides (32.4, 25.7%, and 9.6% respectively). Herbicides may be cheaper, more efficient, and more practical to use than weeding costs [253, 254].

After collection of various corpora, there is some issue in corpora like missing values, blank values, and imbalance data. So some data preprocessing techniques

are required to plugin these issues. The following section discussed the applied data normalization techniques on the above-said corpora.

4.3 DATA PREPROCESSING

Deep learning techniques depend on the quality of data. Raw data collection generally has some noises like missing values, blank values, imbalanced data, inconsistent data, and outlier data. Therefore, there is an essential need to preprocessing these data before taking them as input. Several data preprocessing techniques available such as data cleaning, data integration, data transformation, data normalization, and data reduction. Considered corpora contained imbalanced values (Table 4.1-4.7), so min-max algorithm was used for normalizing the collected corpora. The normalization method is applied to reduce the difference in the original data that affects the performance of the proposed model. In equation (13), input data is provided as $x_{i,n}$, where i denotes the elements and n denotes the number of instance. The output of normalized data denoted as $x'_{i,n}$

$$x'_{i,n} = \frac{x_{i,n} - \min(x_i)}{\max(x_i) - \min(x_i)} (nMax - nMin) + nMin \quad (13)$$

Table 4.8 Output of Min-Max Algorithm Applied on Weather Corpus

S.No	Cloud Cover	Minimum Temp	Average Temp	Maximum Temp	Vapour Pressure	Rainfall (MM)	Average Relative	Actual	Predicted	Crop_expert	Crop_predict
1	18.92	7.05	13.19	19.33	21.76	29.00	92.50	0	0	Rice	Rice
2	21.37	10.33	16.85	23.38	24.73	23.00	77.66	0	0	Rice	Rice
3	19.63	13.10	19.48	25.86	22.94	181.40	86.15	0	1	Rice	Maize
4	19.58	16.95	23.63	30.32	22.77	60.90	64.08	0	0	Rice	Rice

5	18.5 3	21.3 4	28.8 5	36.3 6	22.5 9	10.70	91.8 7	0	0	Ric e	Rice
6	19.1 3	23.1 8	28.8 6	34.5 4	22.2 8	144.9 0	67.9 2	0	1	Ric e	Maiz e
7	20.1 9	23.6 7	27.1 8	30.6 8	24.1 0	566.0 0	96.3 0	0	1	Ric e	Maiz e
8	20.1 9	23.2 2	26.8 6	30.5 0	24.4 2	654.2 0	62.3 8	0	1	Ric e	Maiz e
9	21.1 9	21.5 5	26.6 0	31.6 6	25.0 3	78.00	94.8 6	0	0	Ric e	Rice
10	19.9 5	17.2 3	23.6 9	30.1 5	23.5 0	27.30	64.3 2	0	0	Ric e	Rice

Data normalization (min-max algorithm) is applied to all the selected corpora. The results of this algorithm were significant. The following sample table contains the output of the min-max algorithm applied to the weather corpus.

The min-max algorithm applied to the weather data and Table 4.8 shows the output of the said algorithm. The rainfall values can calculate by using predicted values from the sample table. The values 0 in the predicted column indicate the low rainfall, 1 indicates medium rainfall, and 2 represents the high rainfall values.

4.4 CORPORA CONCATENATION

Data pre-processing technique (min-max algorithm) applied on the collected corpora to minimize the noisy data. The unstructured corpora have to unite into a single corpus. As the nature of these corpora is different (details in chapter 4), merging the corpora required maximum human intervention. Knowledge-based WSD (KB-WSD) and corpus-based WSD (CB-WSD) are two basic methods used for combining two or more corpora into a single entity. The main purpose behind the integration of these two popular algorithms is to remove semantic ambiguity and merge different natured corpora into a single entity. The integration of CB-WSD into a KB-WSD successfully has a low improvement rate in many cases. So,

for the current research work, the integration of KB-WSD into a CB-WSD used in the proposed model and this model will discuss in the next section.

4.4.1 INTEGRATION OF KB-WSD SYSTEM INTO CB-WSD SYSTEM

In this technique, the domain heuristic is used to improve the overall efficiency of the system, because domain features directly added the domain heuristic information. Due to the domain knowledge, this method reduces the word polysemy. A cross-validation testing, if more examples were available, could be appropriate to perform a domain tuning for each word to determine which words must use this preprocess and which not. However, integration of KB-WSD System into a CB-WSD System. KB-WSD method, like domain heuristic, can also integrate successfully into a corpus-based system, to obtain a small improvement.

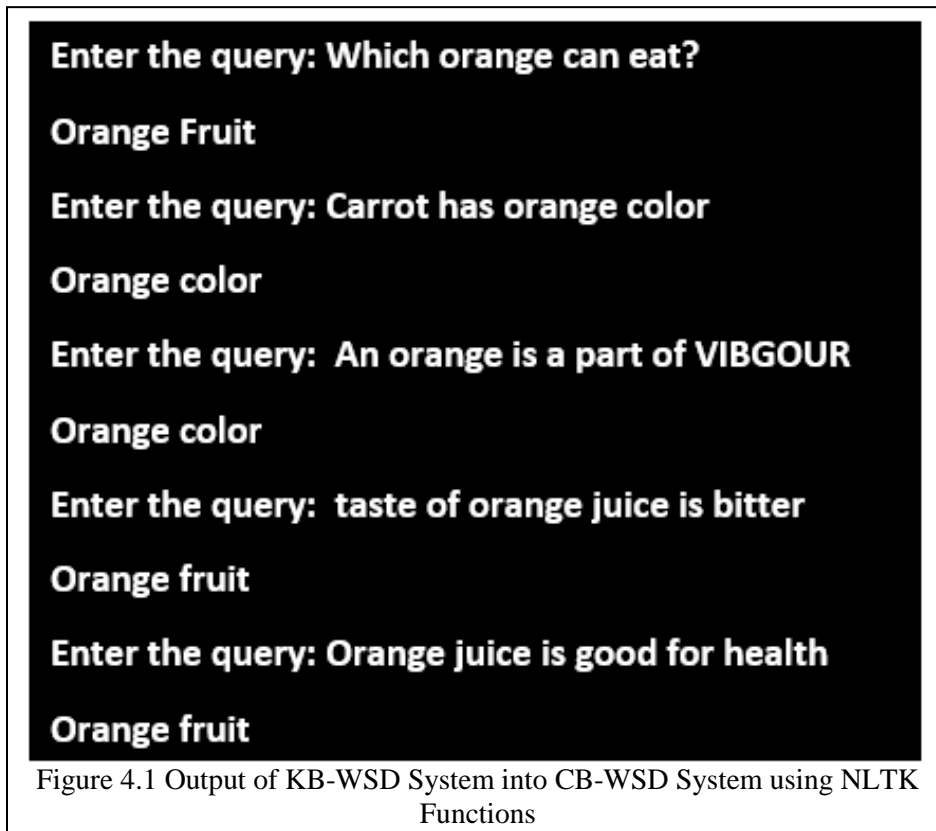


Figure 4.1 Output of KB-WSD System into CB-WSD System using NLTK Functions

To integrate KB-WSD System into a CB-WSD System, the NLTK library functions are used. For testing purposes, this library has given two input data files i.e. color.txt and fruit.txt. The first input file (color.txt) contains few sentences that refer the orange as a color and the second file (fruit.txt) consists of some sentences where orange is a fruit. The sentences of these two input files combined into one file, filtered, and checked similarity index using NLTK function. The comparison sentences normalized and Figure 4.1 output show that whether the query related to an orange fruit or orange color.

4.4.2 PROPOSED DISAMBIGUATION ALGORITHM

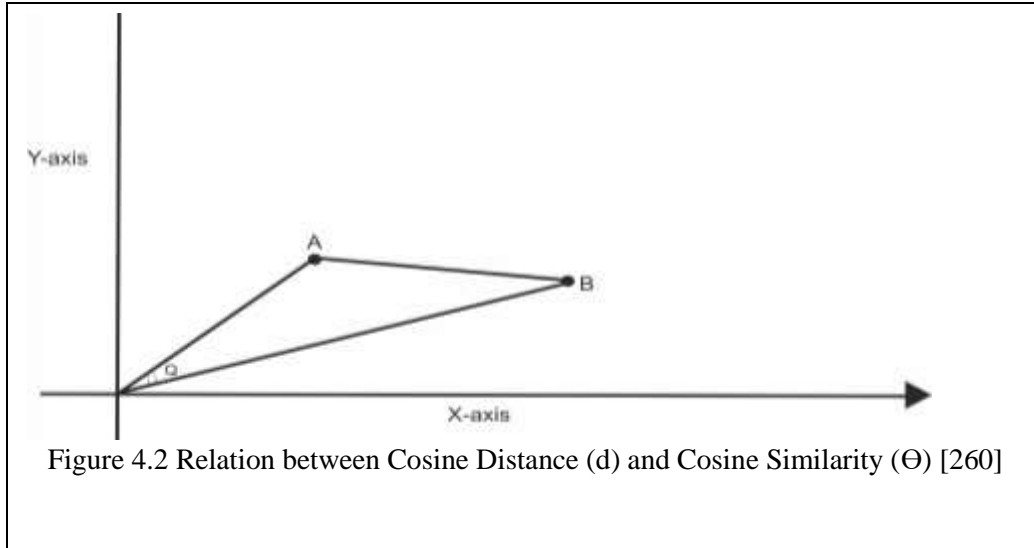
The term ambiguous, defined as similar words with a different meaning. These ambiguous words normally encountered while processing two same or different natured data. So, before applying any natural language processing technique, data should process under any disambiguation algorithm. Few existing methods can use to extract the sense of ambiguous words from an unstructured text [258]. The proposed algorithm is used to extract the sense of ambiguous words present in the corpus collected for current research.

Cosine similarity uses to measure the similarity between two words and cosine distance uses to find the similarity distance between two words [259].

$$Sim(W_i, S_i) = \frac{\sum_{i=1}^n W_i \cdot S_i}{\sqrt{\sum_{i=1}^n W_i^2} * \sqrt{\sum_{i=1}^n S_i^2}} \quad (14)$$

$$D_{amb}(W_i, S_i) = 1 - Cosine(W_i, S_i) \quad (15)$$

From equations (14) and (15), Cosine similarity and cosine distance between two Words (W_i , and S_i) are defined by $Sim(W_i, S_i)$ and $D_{amb}(W_i, S_i)$ respectively. Range of cosine dis between 0 to 1, where 1 represents W_i and S_i are different in nature and 0 (≈ 0) represents W_i associated with S_i . Figure 4.2 shows the relationship between cosine similarity and cosine distance [260].



Algorithm 1: Extract Senses from Ambiguous Word

1. \forall word $W_j \exists$ a sense S_i
 - a. do:
 - b. Calculate $Sim(W_i, S_i) = \frac{\sum_{i=1}^n W_i \cdot S_i}{\sqrt{\sum_{i=1}^n W_i^2} * \sqrt{\sum_{i=1}^n S_i^2}}$
 - c. Calculate $D_{amb}(W_i, S_i) = 1 - Cosine(W_i, S_i)$
 - d. If $D_{amb} \approx 0$ then assign similarity S to W
 - e. End
 2. Sense: = Sen1
 3. \forall ambiguous word calculate sense
 - a. do:
 - b. If $D_{amb}(W_j, S_i) > D_{amb}(W_j, Sen1)$
 - c. Assign Sense = S_i
 - d. End
 4. Assign calculate sense S_i to ambiguous word W_j .
-

The above-said algorithm applied to the following small paragraph (next complete paragraph). For a single word, there are various meanings (sense). To demonstrate the proposed algorithm, the following paragraph used as input, and

Table 4.9 represented the output in tabular form. The term “session” is related to the period of activity, a serious meeting, and weather session. By applying the disambiguation algorithm, the word ‘session’ is related to a weather session only. The output of the proposed algorithm is presented in tabular form.

Ginger is a medicinal plant. There is not a particular period to sow this plant but the pre-monsoon shower session is considered a better period. It is considered a Kharif crop. One month of dry weather before harvesting ginger gives better results.

Table 4.9 Output of Proposed Disambiguation Algorithm

S. No	Ambiguous Word	Existed Senses	Extracted Sense
1	Plant	Factory	Tree
		Tree	
2	Sow	Female Pig	Seed
		Seed	
3	Crop	Grain	Grain
		No of People	
4	Dry	Boing	Rain
		Not Sweet	
		Rain	
5	Session	Period of Activity	Weather
		Weather	
		A Series of Meeting	

4.5 SUMMARY

This chapter includes various corpora selection methods. These corpora consist of Weather, soil map, agricultural corpus, and pest & fertilizer data. Firstly, the nature of each corpus describes along with its significant parameter (some tables and screenshots). The min-max algorithm applied to remove noises from the targeted corpora and integrated CB-WSD and KB-WSD system make research in a forward direction by merging four different natured corpora into a single unified corpus. The proposed disambiguation algorithm successfully extracts the senses from ambiguous words. The next chapter discusses the deep learning algorithm (LSTM and RAO) to extract AGNER, AGEE, and Sematic IE from deep learning techniques.

CHAPTER 5

NAMED ENTITY RECOGNITION AND EVENT EXTRACTION

5.1 INTRODUCTION

Semantic information extraction is used to extract and categorize the meaningful organized information from a free text (unstructured and/or semi-structured data). Named Entity Recognition (NER) and Event Extraction (EE) are two essential tasks while extracting semantic from an agricultural-based corpus. Substantial results were obtained by applying machine learning techniques in the general domain of NER but not in the agriculture domain. NER is a task to identify and extract proper noun phrases from unstructured text and classify them into a small-predefined set of semantic classes. General NER consists of the name of a person, actor, scientist, novel, city, state, river, organization, place, date, etc. Where an Agricultural Named Entity Recognition (AGNER) is the process to extract and recognize the entities from the agricultural domain such as the name of a crop, a fertilizer, a pest, pesticides, summer, winter, black soil, etc. Similarly, Agricultural Event Extraction (AGEE) is used to extract the events from the agricultural corpus.

The previous chapter demonstrated the corpora collection, preprocessing algorithm, integration of KB-WSD into CB-WSD System, and proposed disambiguation algorithms. The present chapter proposes a deep learning-based algorithm to extract the AGNER, AGEE, and relationships between entities from the agricultural corpus. LSTM and RAO are used to define the agricultural NER and EE. The chapter organization is as follows: Section 5.1 discussed the introduction of the present chapter. In section 5.2, the existing rules related to agricultural NER, EE, and relation extraction are highlighted. In section 5.3, the proposed deep learning-based algorithm discussed the significance of LSTM with

RAO to extract AGNER, AGEE and identify the relationship (semantics) between them. Section 5.4 discussed the summary of the present chapter.

5.2 EXISTING RULES FOR AGNER AND AGEE

The fundamental division of NER and EE are supervised, semi-supervised and unsupervised extraction. Several methods were identified that can apply to the agricultural corpus to extract the AGNER and AGEE. This section analyzes the existing extraction patterns used only to process plain agricultural documents. These extraction rules are based on semantic and syntactic constraints. These constraints help to identify the significant information within a text document. Before applying these extraction rules, pre-process methods as syntactic analyzer and semantic tagger should be applied to it.

AutoSlog: It builds extraction patterns dictionaries called concepts (concept nodes). A concept node has a conceptual anchor and linguistic pattern. The conceptual anchor used to activate the concept whereas the linguistic pattern, with a set of enabling conditions, guarantees the applicability of the concept node [261].

LIEP: It represents the multi-slot extraction rules. In a sentence, LIEP generates a rule for all the interesting items rather than a single extraction rule for every individual interested item [262].

PALKA: The patterns extraction rules consist of a phrasal pattern and a meaning frame known as FP (frame-phrasal) structures. Phrasal patterns emphasizes on the order of lexical entities and the meaning frame represent a semantic constraint-based extraction item [263].

Crystal: It produces multi-slot concepts (concept nodes), which allow the exact word and semantic constraints on a component phrase [264]. Because PALKA patterns do not accept semantic constraints, so CRYSTAL rules are more expressive.

WHISK: It generates regular expressions-based extraction rules that consist of two components, the first component describes context-based rules from unstructured text, and the second specifies the exact delimiters from structured text [265].

RAPIER: It learns single slot extraction rules that use precise syntactic knowledge (part-of-speech tagger's output) and semantic class knowledge (hypernym links from WordNet). A typical RAPIER extraction task consists of three components: an input document, information that needs extraction, and the extraction patterns [266].

5.3 INFORMATION EXTRACTION USING DEEP LEARNING

This section comprises three sub-models, first sub-section contains the hybrid deep learning-based algorithm used for extracting meaningful information from agricultural text, the second sub-section consists of AGNER extraction rules based on deep learning (LSTM and RAO) techniques and the last sub-section contains the AGEE rules for extracting events from an agricultural corpus based on deep learning techniques (LSTM+RAO).

5.3.1 PROPOSED ALGORITHM FOR SEMANTIC INFORMATION EXTRACTION USING LSTM-RAO

The following algorithm uses the min-max algorithm for data normalization, having the advantages of the LSTM techniques over RNN methods. For each iteration of backpropagation, an RAO is applied to modify the weights in a deep network. This optimizer inherited the properties of RMSProp and AdaGard optimizer.

Algorithm 2 : LSTM based Algorithm with Rectification in Adam Optimizer

1. Class LSTM-RNN [_cropdata, _weights, _biases]
2. Categorize the input data by state-wise
3. Crop data processed by season and state name
4. $D = \{ x_{i,n}, y_n \mid i \in f \text{ and } n \in N \}$
//Database with N instances and f features
5. $x'_{i,n} = \frac{x_{i,n} - \min(x_i)}{\max(x_i) - \min(x_i)} (nMax - nMin) + nMin$
//Min-Max normalization within the range of -1 to 1, or 0 to 1
6. $X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow \text{train_test_split}(\text{crop_data_processed}, \text{test_size}=0.3)$
// spitting the data into training data and testing data for the LSTM-RAO model
7. def model: while θ_t is not converged do
8. $m_0, v_0 \leftarrow 0, 0$ (1st and 2nd moment moving initialize)
9. $\rho_\infty \leftarrow \frac{2}{(1 - \beta)} - 1$
10. while $t = \{1, \dots, T\}$ do
 - a. $g_t \leftarrow \Delta_{\theta} f_t(\theta - 1)$
 - b. $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
//Update exponential moving 2nd moment
 - c. $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$
//Update exponential moving 1st moment
 - d. $\hat{m}_t \leftarrow \frac{m_t}{1 - \beta_1^t}$
//Compute bias-corrected moving average
 - e. $\rho_t \leftarrow \rho_\infty - 2t \frac{\beta_2^t}{(1 - \beta_2^t)}$

- //Compute the length of the approximated SMA*
- f. if the variance is tractable, i.e., $\rho_t > 4$ then
- i. $\hat{v}_t \leftarrow \sqrt{\frac{v_t}{(1 - \beta_2^t)}}$
- //Compute bias-corrected moving 2nd moment*
- ii. $r_t \leftarrow \sqrt{\frac{((\rho_{(t-4)})(\rho_{(t-2)})(\rho_\infty))}{((\rho_{(\infty-4)})(\rho_{(\infty-2)})(\rho_t)}}$
- iii. $\theta_t \leftarrow \theta_{t-1} - \alpha_t r_t \frac{\hat{m}_t}{\hat{v}_t}$ *//Update parameters with adaptive momentum*
- g. else
- h. $\theta_t \leftarrow \theta_{t-1} - \alpha_t \hat{m}_t$

End while

Return; Resulting parameter θ_t

5.3.2 DEEP LEARNING TECHNIQUE FOR AGNER

AGNER is the first step of information extraction from the structured, semi-structured, and unstructured agricultural-based knowledge base. AGNER seeks to extract and classify NER into various existing categories. For the agricultural domain, before extracting NER, it is better to design a Tagset. There is no agricultural domain-based Tagset exists, so a Tagset based on 15 fine-grained tags created to perform the proposed methodology.

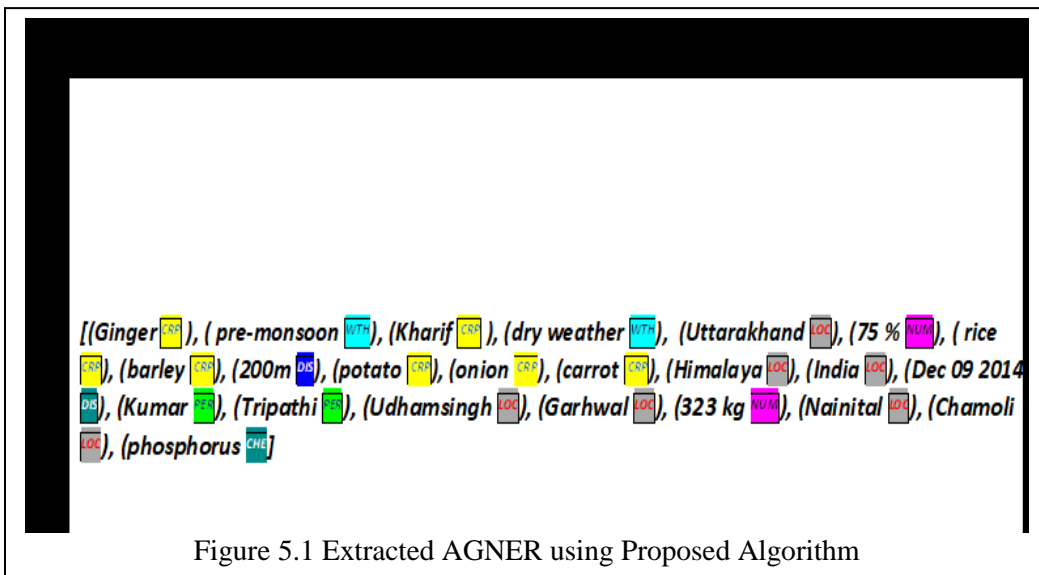
Table 5.1 explains various values that categories into three attributes Tag type, description of tags, and related examples of agricultural-based named entities.

Table 5.1 Types of Tags, Tags Descriptions, and Suitable NER Examples

S.N	Types of Tags	Tag Abbreviation	Tag Description	Examples of NER
1	Person	PER	Person name	Ramesh, John
2	Location	LOC	city, country, continents, waterbodies, place	Dehradun, India, Punjab
3	Organization	ORG	Name related to any institution, company, industry	Krishi Vigyan Kendra, Central Institute of Agricultural Engineering, Cotton Corporation of India
4	Chemicals	CHE	Fungicide, Insecticide, Pesticide and Fertilizer	Endosulfan, Mancozeb, Nitrogen, Methyl Parathion
5	Crop	CRP	Name related to grain, cereal, fruit, and vegetable	Orange, Banana, Radish, Ginger, Rice
6	Policy	POL	Agricultural policies or aids	Common agricultural Schemes
7	Food Products	FOP	Product extracted from Plant or animal	milk, curd, Cheese, bread
8	Natural Disaster	NTD	crop production affected by disaster	Kashmir Floods disaster, Uttarakhand Flash Floods, Tsunami, Super Cyclone
9	Events	ENT	Meeting, workshops, summits, conferences, and exhibition	Advanced Agricultural Biotechnology and Fertilizer Systems Conference
10	Number	NUM	Total number of any items	12 medicinal plants

11	Distance	DIS	Measurement of agricultural area	15 acres
12	Quantity	QUN	Measurement of agricultural-based product quantity	120 tones, 35 kg
13	Money	MON	Country currency	Rs. 5000, \$820 and 11300 euro
14	Temp.	TEM	Numerical value of climatic condition	34C
15	Date	DAT	Specify day month year (any format)	12 Dec, 12/12/2001

The proposed algorithm is applied on a sample of input corpus (collected and merged four corpora) and before extracting AGNER, the proposed model helps to assign the tags on the input sample. The subtask, tag assignment is executed based on the above-mentioned fifteen tag types. Once the tags assignment task is completed, the AGNER can be extracted with their frequency value. Figure 5.1 shows the output of AGNER from the input corpus.



5.3.3 DEEP LEARNING TECHNIQUE FOR AGEE

The proposed deep learning-based agricultural event extraction method has a different perspective as compared to existing techniques. First, the automated process of the proposed algorithm learns features from input sentences to reduce the dependency over the supervised toolkits. Second, for event triggering, the proposed method uses the word-embedding technique to make a better event representation. The AGEE is a kind of classification problem. For every AGNER in the input sentence, the proposed method predicts it as an event trigger, along with a comparison with the existing event set. For training purposes, the proposed model was limited with a fixed sample size by minimizing the longer sentences and used padding (if necessary) with a specially named entity.

Let the fixed sample size is $2n+1$ and $X = \{x_0, x_1, x_2, x_3, \dots, x_{2n-1}, x_{2n}\}$ is a set of entity trigger candidate values, also contain AGNE in X . Before entering into Step-6 of the proposed algorithm, each AGNE X_i changed into a real-valued vector to find the hidden semantic and syntactic properties of AGNE. As AGNE and its type already extracted and labeled, so it is easier to find the information related to each entity. Therefore, the initial event trigger X transformed into the following matrix W

$$W = \{w_0, w_1, w_2, w_3, \dots, w_{2n-1}, w_{2n}\} \quad (16)$$

The proposed method vanishing the gradient descent problem, by using the LSTM technique. The network training purpose, along with fundamental properties of AdaDelta and RMSprop optimizer, the rectification in Adam Optimizer used to achieve the optimal weights. Finally, the optimized weights apply during each iteration of the backtracking method of the proposed model.

5.3.4 APACHE SOLR FOR AGNER AND AGEE

Apache Solr is an open-source tool, written in JAVA, used for handling large corpus. This tool is used for full-text and faceted searching, dynamic clustering, real-time indexing, database integration, and corpus handling, etc. Solr was designed to increase scalability and fault tolerance by facilitating distributed searching and replication indexing [267]. Apache Solr individually runs as a full-text searching server, Apache Lucene for full-text indexing and JSON helps it to make usable from programming languages. Figure 5.2 represents the technical architecture used by the Apache Solr system [267].

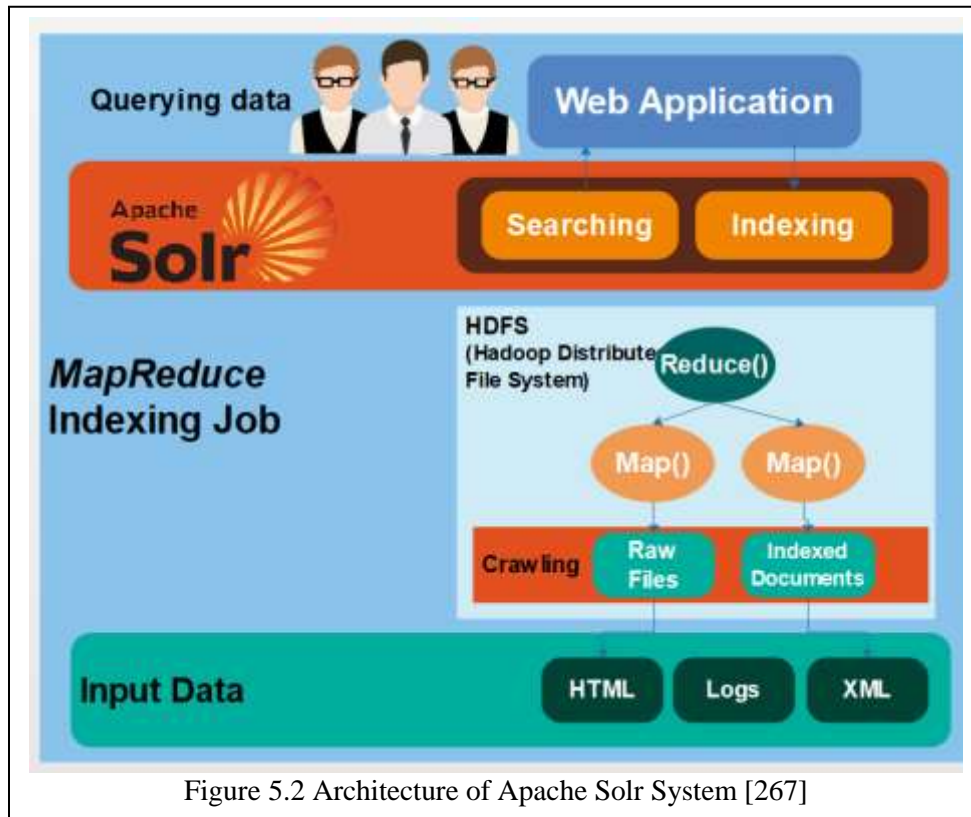


Figure 5.2 Architecture of Apache Solr System [267]

As described in Figure 5.2, the Apache Solr system takes input in the form of unstructured data (HTML/Web, XML Documents, Logs, etc.), intercepted by Hadoop Distributed File Storage (HDFS) layer, and finally searching and indexing

by Apache Lucene indexes [268]. As Solr accepts only HTTP requests, user can use their web browser to connect with Solr. Users can communicate with the Solr administration console by <http://localhost:8983/solr/admin> at the browser if the Solr server is installed on the localhost [268].

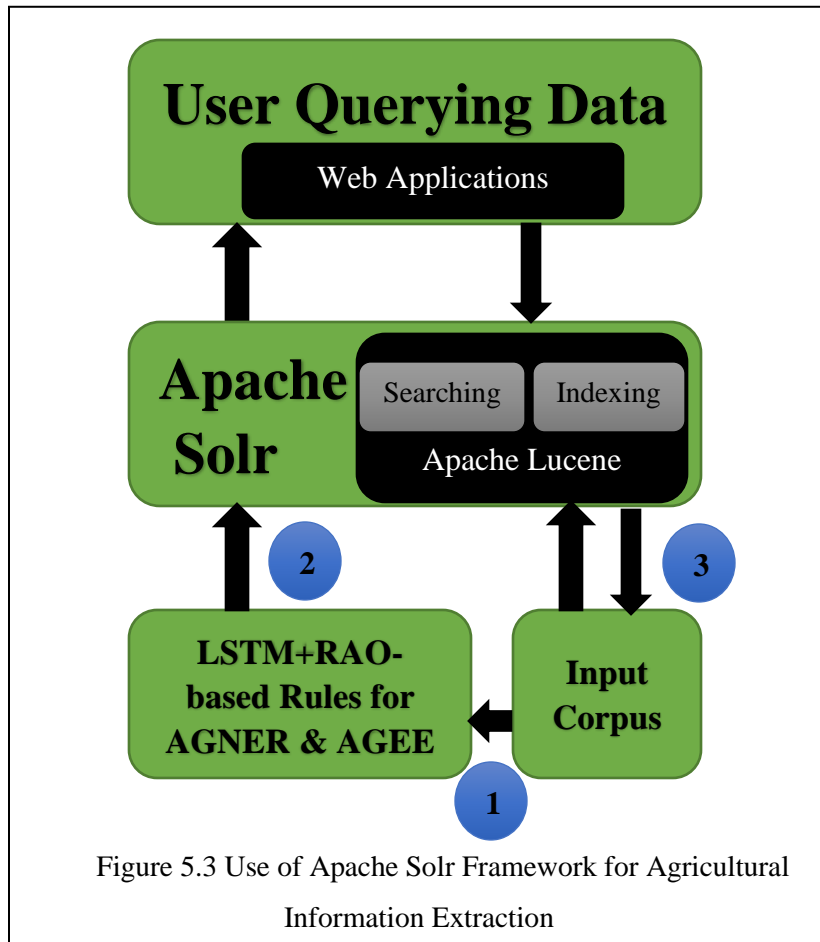


Figure 5.3 Use of Apache Solr Framework for Agricultural Information Extraction

As shown in

Figure 5.3, Apache Solr uses to improve the overall efficiency of the proposed system due to its full-text search capability and real-time indexing feature. The system takes input as an agricultural-based unified corpus (details in Chapter 4). LSTM with RAO uses to extract the AGNER and AGEE from the input corpus. After extracting the first AGNER and AGEE, the Apache Lucene layer reads

extracted AGNER & AGEE and stores extracted records using Lucene Scheme. It saves this data in an index as Lucene documents. Once indexing is completed at the Lucene layer, the user can perform queries against them.

Apache Solr platform is used to search the whole corpus for similar named entities and events. It also helps in categorize various extracted AGNER into four different categories i.e. input corpora based (Weather, soil, Agricultural, and Pest & Fertilizer). At Apache Solr layer, the designed schema matches the index value generated from the Lucene layer. Based on the query type, indexed data provides a faster response through the Request method on HTTP. Additionally, query results are instantly listed on the console, so output can also print out in any format such as XML, JSON, CSV, etc. formats. As discussed earlier, the Solr system accepts only HTTP requests, so the Apache Solr administrator console uses for running queries.

The extracted information can be used for multiple purposes like event extraction, event summarization, information retrieval, relation extraction, etc. For the present research work, Apache the overall efficiency of the proposed algorithm.

5.3.5 SEMANTIC INFORMATION EXTRACTION USING DEEP LEARNING

After extracting AGNER and AGEE, extraction of semantic information is an essential part of IE. NLTK module is used to extract the semantics between the AGNER and AGEE. Figure 5.4 represents the various relations and named entities related to the agricultural domain. Here, the crop segment consists of fiber crops, fruit crops, vegetable crops, agricultural crops, horticultural crops, and cereal crops. Various named entities exist in an agricultural corpus such as cashew, mango, apple, grapes, cucumber, onion, tomato, chilies, corn, wheat, and paddy. Here Figure 5.4 shows a few AGNER that are cashew, mango, apple, and grapes. Similarly, the relationship between AGNER are “is a” and “is type of”. Along with

said AGNER and AGEE, there exist many other NER and EE, chapter 6 highlighted the complete detail.

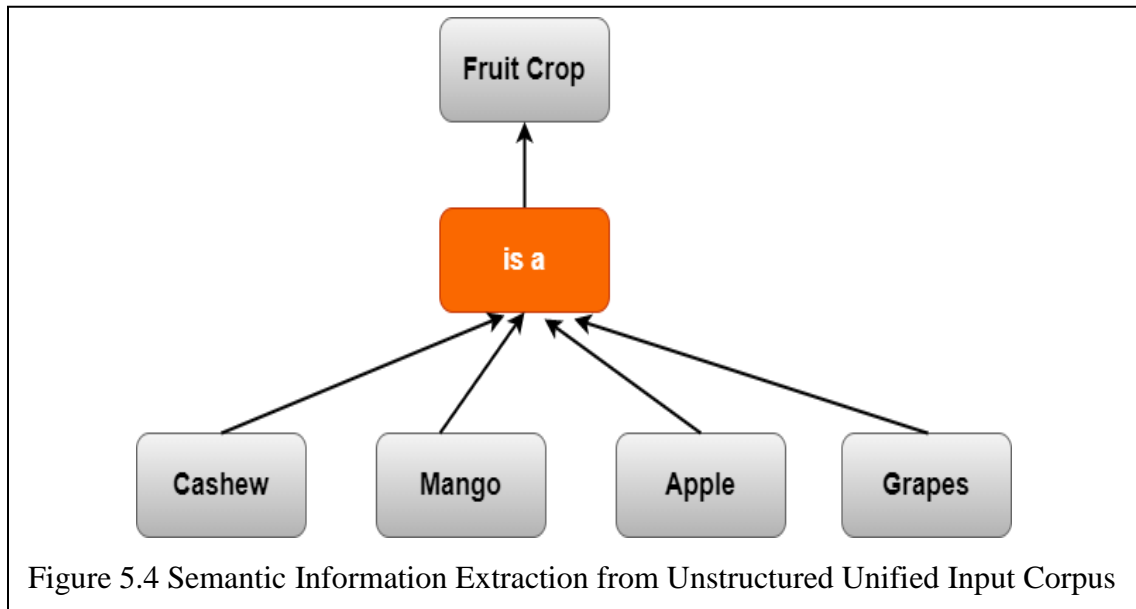


Figure 5.4 Semantic Information Extraction from Unstructured Unified Input Corpus

5.4 SUMMARY

The present chapter briefly presented various information extraction tools such as AutoSlog, LIEP, PALKA, Crystal, WHISK, and RAPIER. The proposed algorithm successfully demonstrated the procedure to find out the semantics of extracted information from the agricultural domain using deep learning. The min-max algorithm is used to normalize the input corpora. LSTM techniques are significantly used to forget and retain the last memory cell. Hyperbolic tangent and sigmoid activation function used to optimize the weights for every iteration. LSTM with RAO also enhances the overall performance of the proposed model. The proposed method used to extract the agricultural named entity extraction along with event extraction. After extraction of AGNER and AGEE, to find the semantics between these terms, heuristics, and linguistics approaches were also applied to the extracted output. In the next chapter, the experimental setup along with research outcomes will discuss.

CHAPTER 6

RESULTS AND DISCUSSION

The present chapter presents the evaluation of the proposed semantic information extraction approach for the agricultural domain. The chapter consists of four different stages: section 6.1 demonstrates the process of experimental setup and design. Graphical representation of obtained results and analyzed the same in section 6.2. Section 6.3 compares the results of the present research work with some existing standard methods. Section 6.4 summarizes the present chapter.

6.1 EXPERIMENTAL SETUP AND DESIGN

Table 6.1 presents the detailed requirements to implement the proposed algorithm for AGNER, AGEE, and semantic extraction.

Table 6.1 Basic Requirements for Experimental Setup and Design

S.No	Content	Description
1	Input Corpora	IMD Weather Data Set
		Soil Corpus
		Agricultural Corpus
		Pest & Fertilizer Dataset
2	Deep Learning Technique	LSTM with 32 hidden layer
3	Optimizer used	RAO
4	Cross Validation (%)	80-20, 70-30 and 60-40
4	Evaluation parameter	Accuracy, Precision, Recall, F-Score
5	No of Iteration	150
6	Language, Tools Used	Python, NLTK
7	Libraries Used	numpy, tensorflow, keras, pandas, scikit-learn, matplotlib
8	Efficiency Coefficient	Nash-Sutcliffe Efficiency Coefficient
9	Hardware Requirements	Windows 10/ RHEL 6/7, 64-bit (almost all libraries also work in Ubuntu), Minimum 4GB, Intel or AMD x86-64 processor, and 10GB free disk space
10	Software Requirements	Anaconda Navigator 3.5.2.0 (64-bit). Python 3.7

6.2 RESULTS OF PROPOSED MODEL

Anaconda is open-source software and provides a user-friendly environment for deep learning techniques. We have downloaded the Anaconda Python package from its official website (<https://www.anaconda.com/products/individual>) and installed the required packages.

In the next step, the proposed disambiguation algorithm was implemented and the obtained result converted into a tabular form so that it can easily be demonstrated.

Table 6.2 Result of Proposed Disambiguation Algorithm

S. No	Ambiguous Word	Existed Senses	Extracted Sense
1	plant	Factory	Tree
		Tree	
2	sow	Female Pig	seed
		Seed	
3	crop	Grain	grain
		No of People	
4	dry	Boing	rain
		Not Sweet	
		Rain	
5	session	Period of Activity	weather
		Weather	
		A Series of Meetings	

Table 6.2 described the results of the proposed disambiguation algorithm in a tabular manner so that it can be easily understood. It contained five ambiguous words that are plant, sow, crop, dry, and session. A plant can be a factory or a tree but according to extracted sense, it should be a tree here. Similarly, a session can be a period of activity, weather, or a series of meetings but the output of the proposed algorithm shows that a session is related to a weather session.

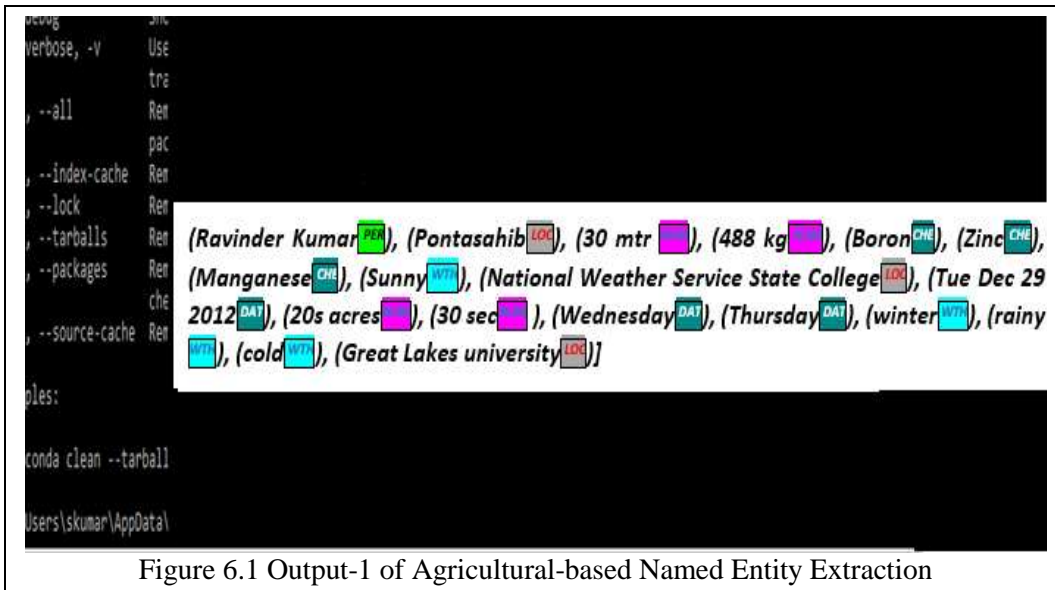


Figure 6.1 Output-1 of Agricultural-based Named Entity Extraction

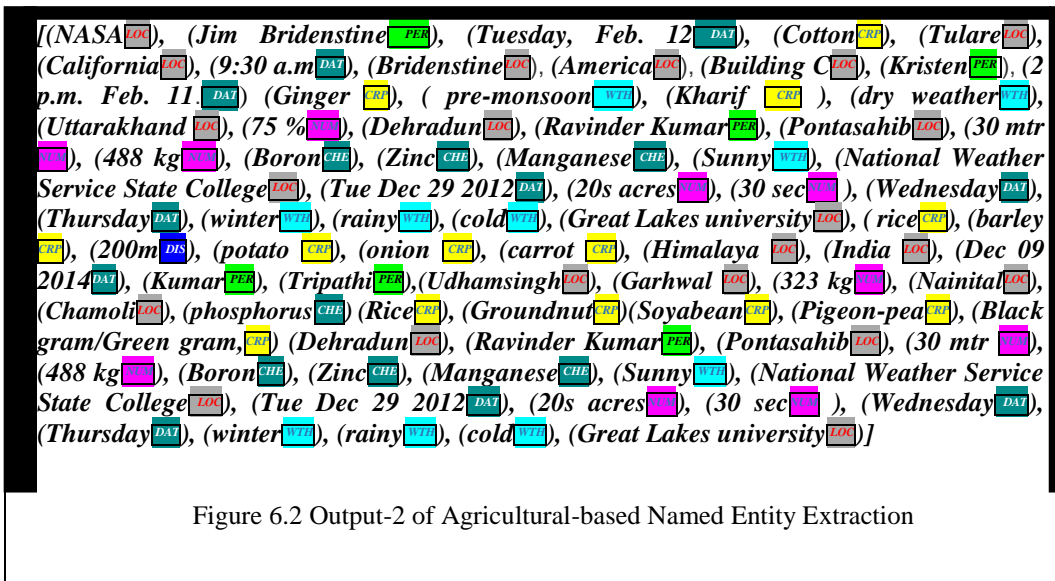


Figure 6.2 Output-2 of Agricultural-based Named Entity Extraction

The above two figures 6.1 and 6.2 depict the output of the proposed deep learning algorithm for NER. Each extracted named entity, a specific tag attached with it. These tags will help us to extract the events that are related to these named entities. Figure 6.3 shows the output of agricultural-based event extraction

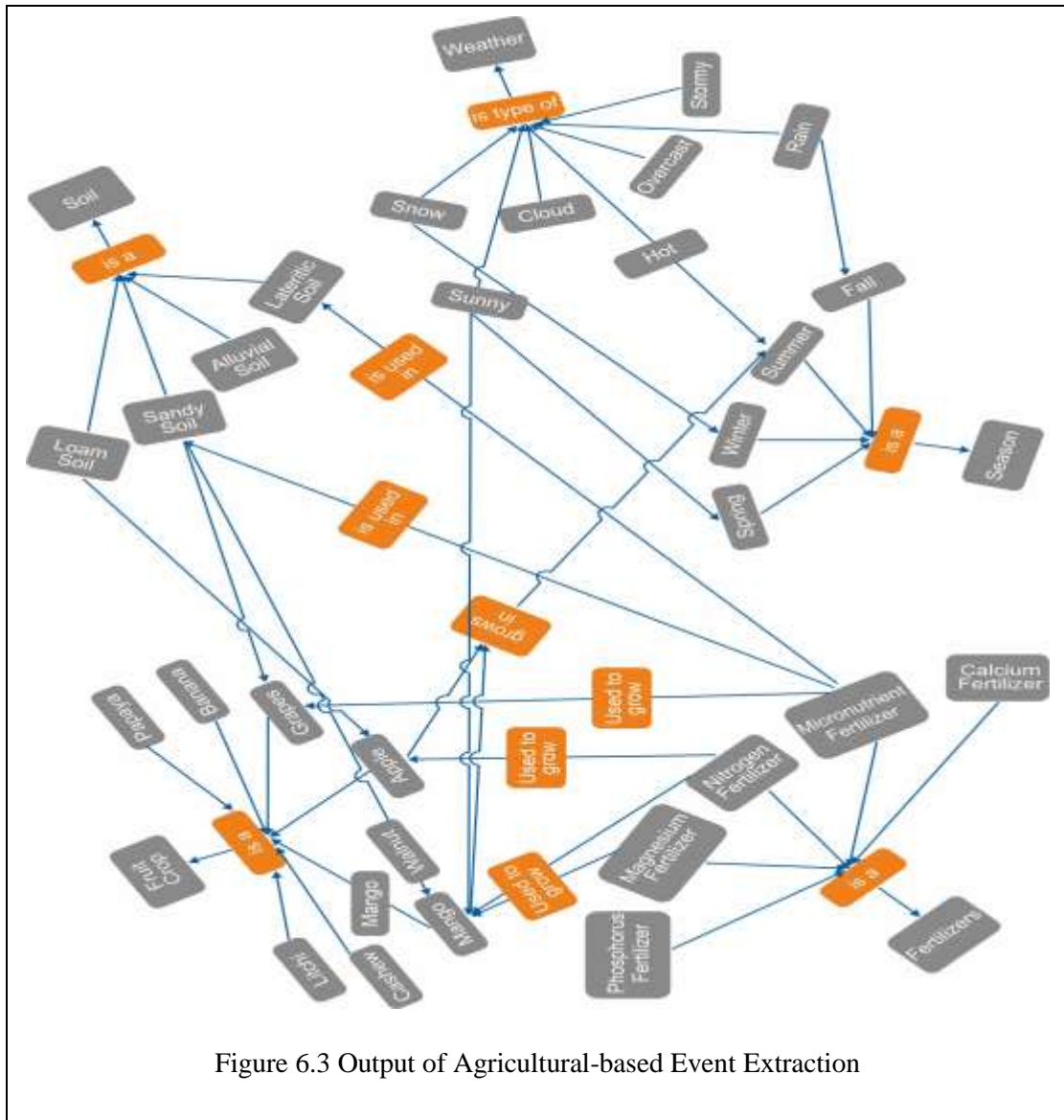


Figure 6.3 Output of Agricultural-based Event Extraction

6.3 COMPARISON WITH EXISTING METHODS

6.3.1 PARAMETER METRICS

In this study, the proposed method's performance was assessed and used the standard statistical performance evaluation criteria that include the accuracy, sensitivity, specificity, and F-Measures.

Specificity: Specificity is defined as the ratio of actual positives recognized accurately and the capacity to test and acknowledge positive results, which is explained in equation (17)

$$specificity = \left(\frac{TP}{TP * FN} \right) * 100 \quad (17)$$

Where TP =True Positive and FN =False Negative. The computation of the precision is the number of real positives that are precisely predictable. It connects with the capability of the experiment to identify positive outcomes (TP).

Sensitivity: Sensitivity is calculated as the ratio of negatives predicted correctly to the capability of testing to recognize negative outcomes, explained in equation (18)

$$Sensitivity = \left(\frac{TN}{TN * FN} \right) * 100 \quad (18)$$

Where TN = True Negative and FP =False Positive.

Accuracy: Accuracy is defined as the ratio of the sum of the number of TP and TN to the total number of objects in the dataset. Equation (19) describes the mathematical equation for accuracy,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (19)$$

F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure that is represented in equation (20)

$$F - measure = 2 * \left(\frac{precision * recall}{precision + recall} \right) \quad (20)$$

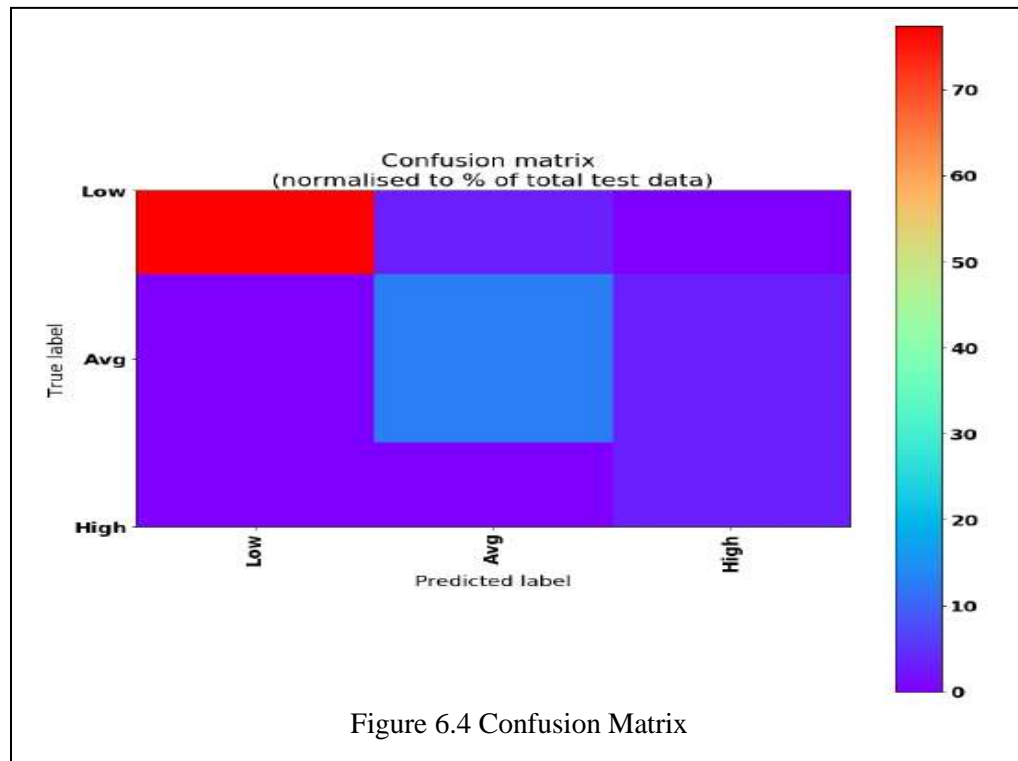
Nash–Sutcliffe Efficiency (NSE) Coefficient: The NSE is used to find out the analytical skill of hydrological models [269]. As shown in equation (21)

$$NSE = 1 - \frac{\sum_{T=1}^n (X_{Obj}^T - X_{Sim}^T)^2}{\sum_{T=1}^n (X_{Obj}^T - \overline{X_{Obj}})^2} \quad (21)$$

In equation (21), at time T, X_{Obj}^T is the observed value, X_{Sim}^T is the simulated value and $\overline{X_{Obj}}$ is the mean of all observed values which is dimensionless. The NSE ≈ 1 if the model is perfect with zero estimation error variance. NSE ≈ 0 , if a model shows that the estimation error variance and observed variance are equal.

6.3.2 CONFUSION MATRIX AND OTHER OBSERVATIONS

The pictorial representation of the confusion matrix is shown in Figure 6.4.



Some other observations like Mean Squared Error is 0.065, Root Mean Squared Error is 0.25, MAE is 0.065, and Nash-Sutcliffe Efficiency Coefficient is 0.99 (shown in Figure 6.5).

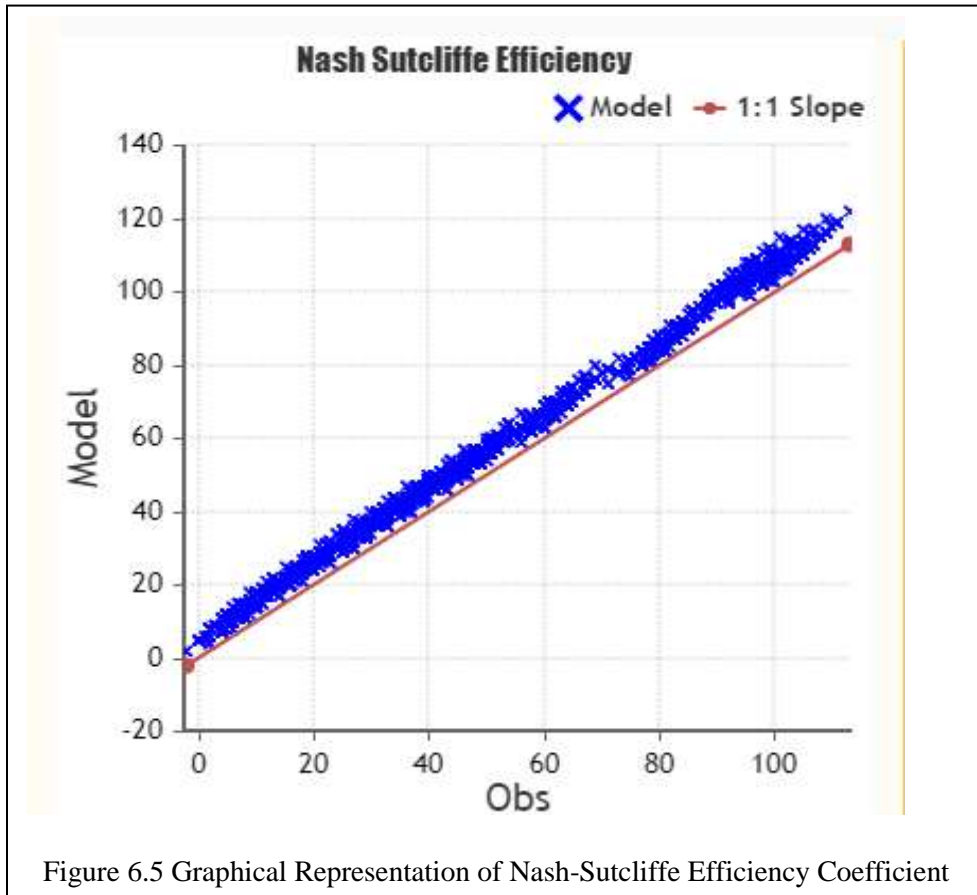
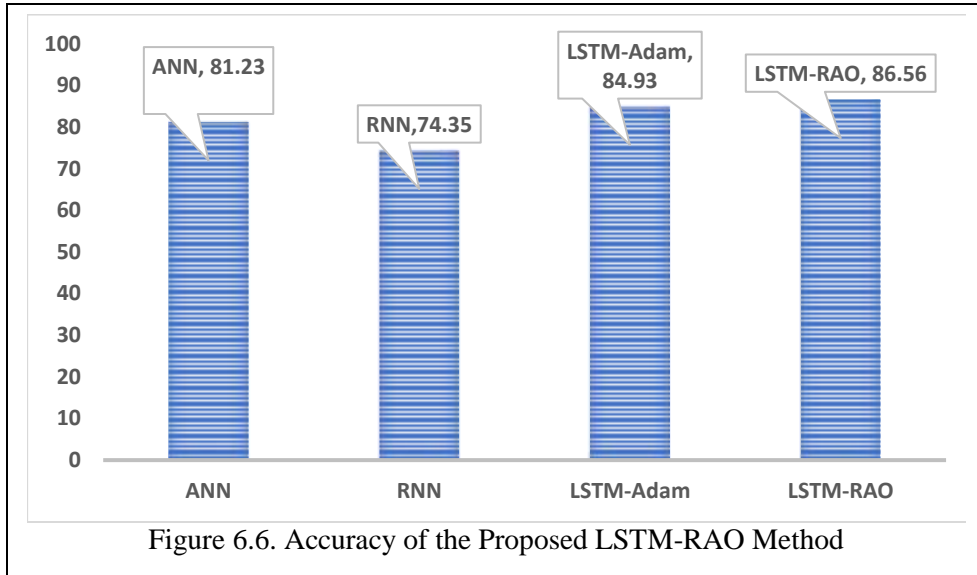
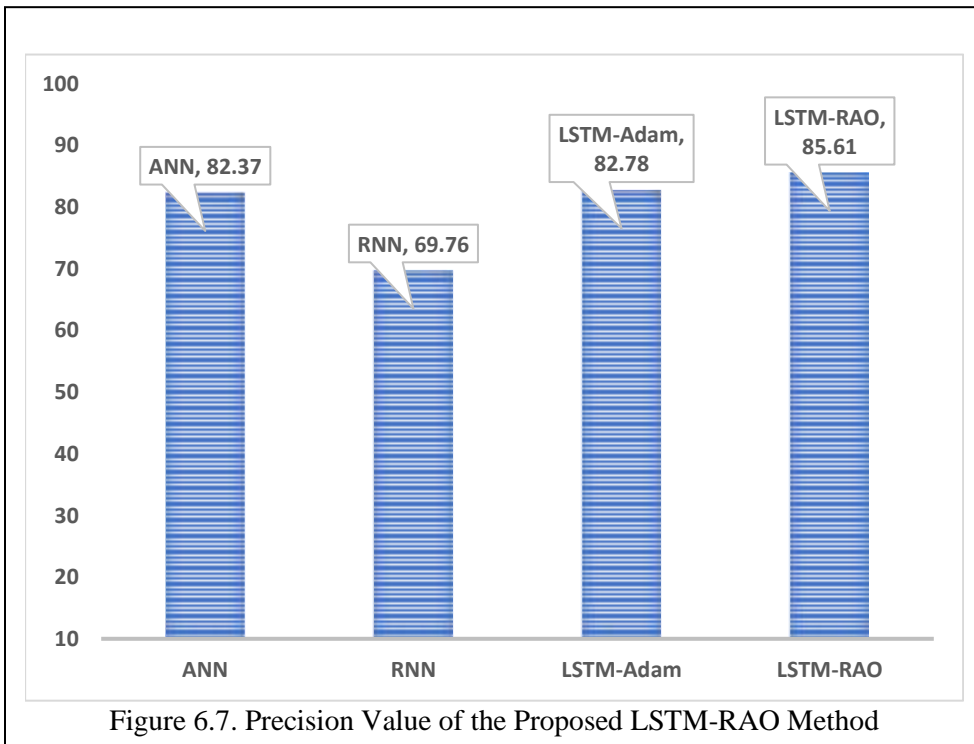


Figure 6.5 Graphical Representation of Nash-Sutcliffe Efficiency Coefficient

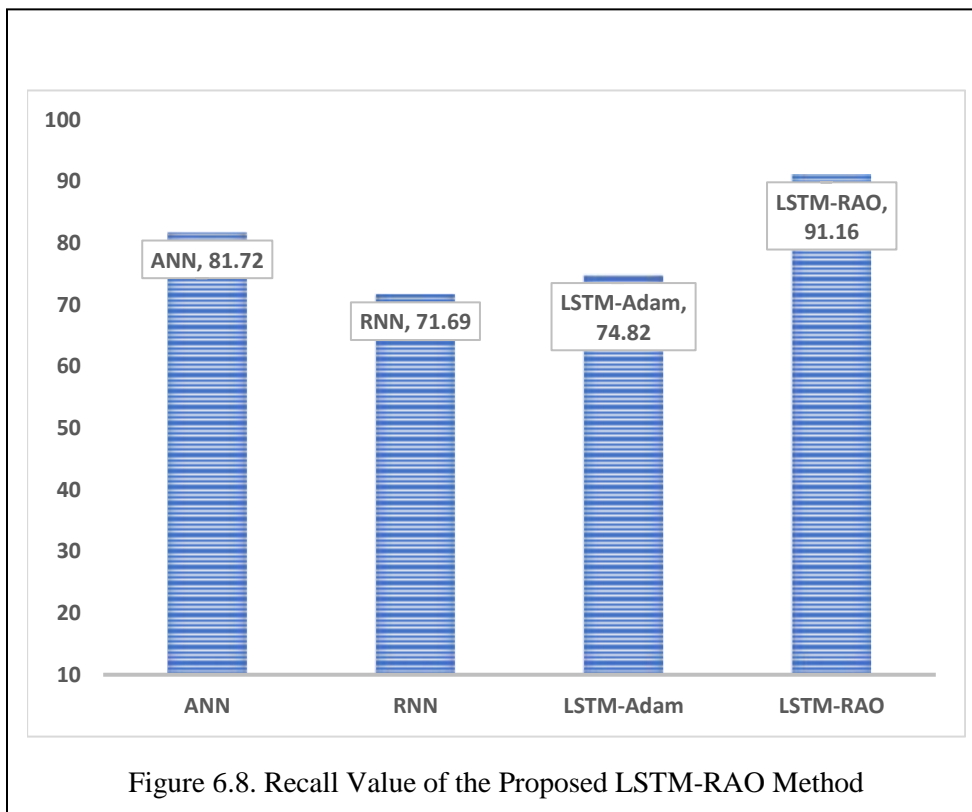
The IE is a domain-based system, so extraction of AGNER and AGEЕ cannot be easily compared with other domain-based IE systems. So it is compared with existing machine learning techniques only. This section provides a detailed description of the performance of the proposed LSTM-RAO method. The comparison of the proposed LSTM-RAO method is applied with cross-validation of 80% training and 20% testing. The cross-validation of the proposed LSTM-RAO method is also analyzed for 70-30% and 60-40% training-testing data.



The accuracy of the proposed LSTM-RAO method for the Uttarakhand Kharif dataset is shown in Figure 6.6. The proposed LSTM-RAO method has a higher accuracy of 86.56% and the LSTM-Adam optimization method has 84.93% accuracy.



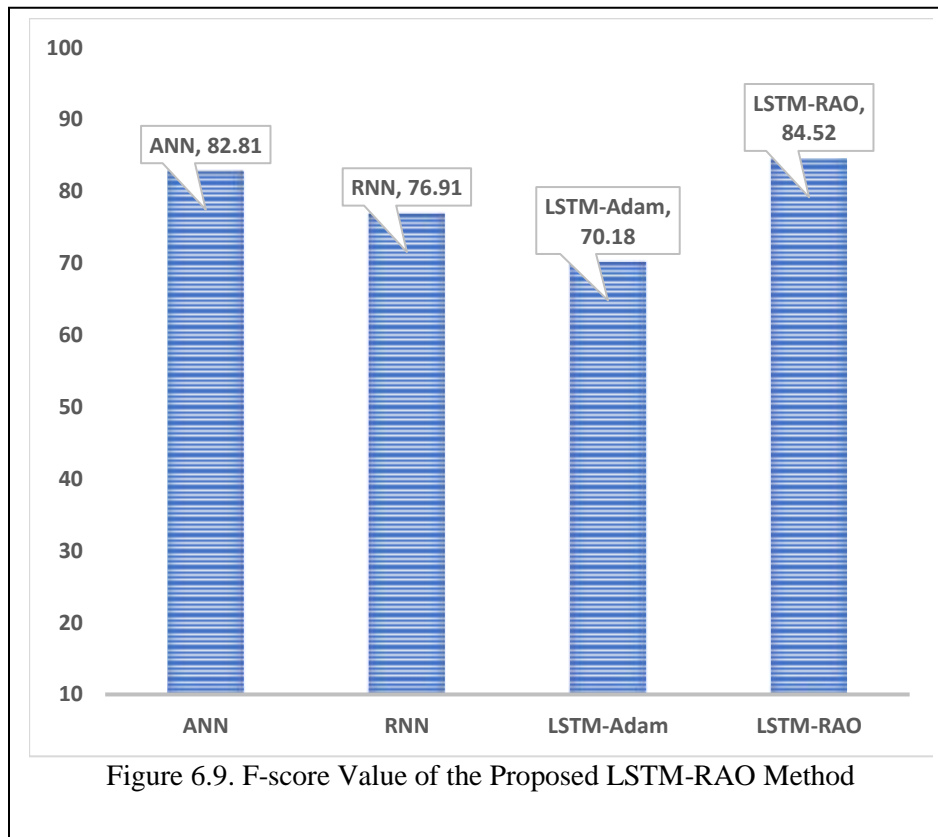
The proposed LSTM-RAO method precision value is measured in the IMD dataset, as shown in Figure 6.7. The proposed LSTM-RAO method has a higher precision value compared to other methods such as ANN, RNN, and LSTM-Adam optimization. The proposed LSTM-RAO method has the advantage of the selection of parameter settings for the LSTM network. The proposed LSTM-RAO method has a precision value of 85.61% and LSTM-Adam optimization has an 82.78% precision value.



The recall value of the proposed LSTM-RAO method is compared with ANN and LSTM, as shown in Figure 6.8. The comparison shows that the LSTM-RAO method has a higher recall value compared to the ANN, RNN, and LSTM-Adam optimization. The proposed LSTM-RAO method has the advantage of proper selection of the parameter for the LSTM network. The proposed method has a

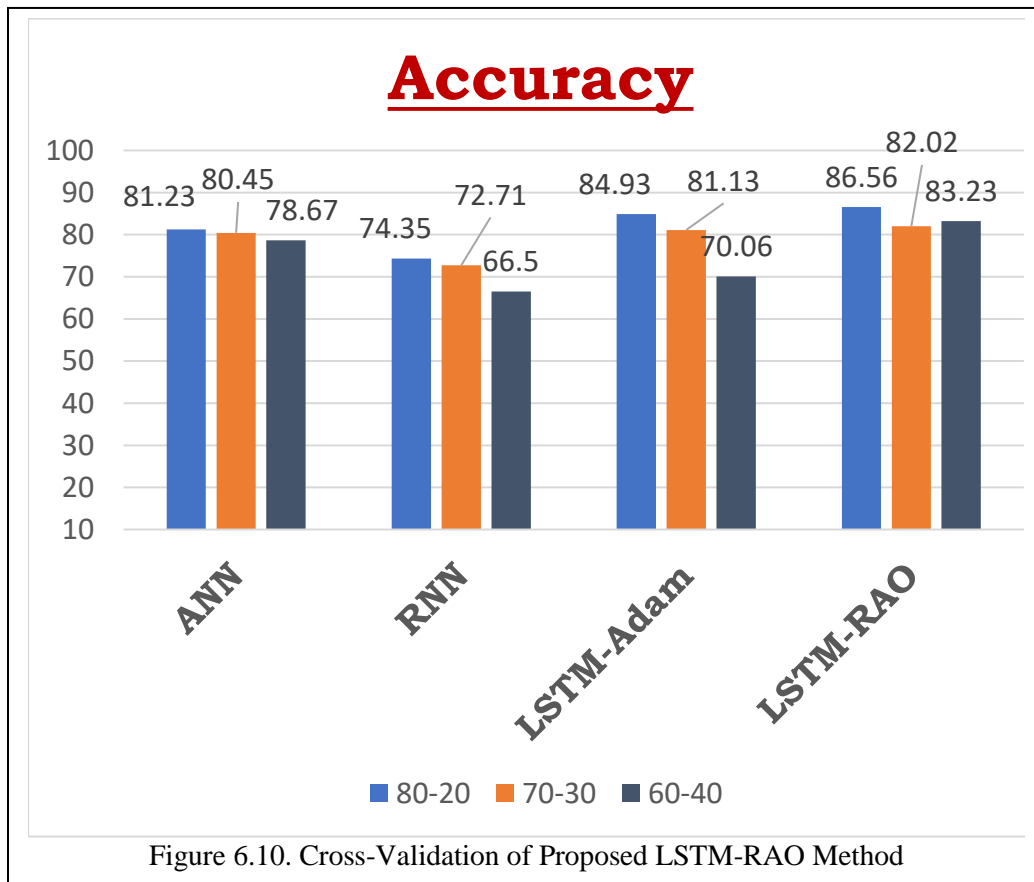
recall value of 91.16% and the LSTM-Adam optimization has a 74.82% recall value.

The F-Score value of the proposed LSTM-RAO method is measured and shown in Figure 6.9. The figure shows that the LSTM-RAO method has a higher F-Score value compared with ANN, RNN, and LSTM-Adam optimization. The proposed LSTM-RAO method has the advantage of the proper selection of parameters for the LSTM network. The proposed LSTM-RAO method has 84.52% F-Score and LSTM-Adam optimization method has 70.18% F-Score.

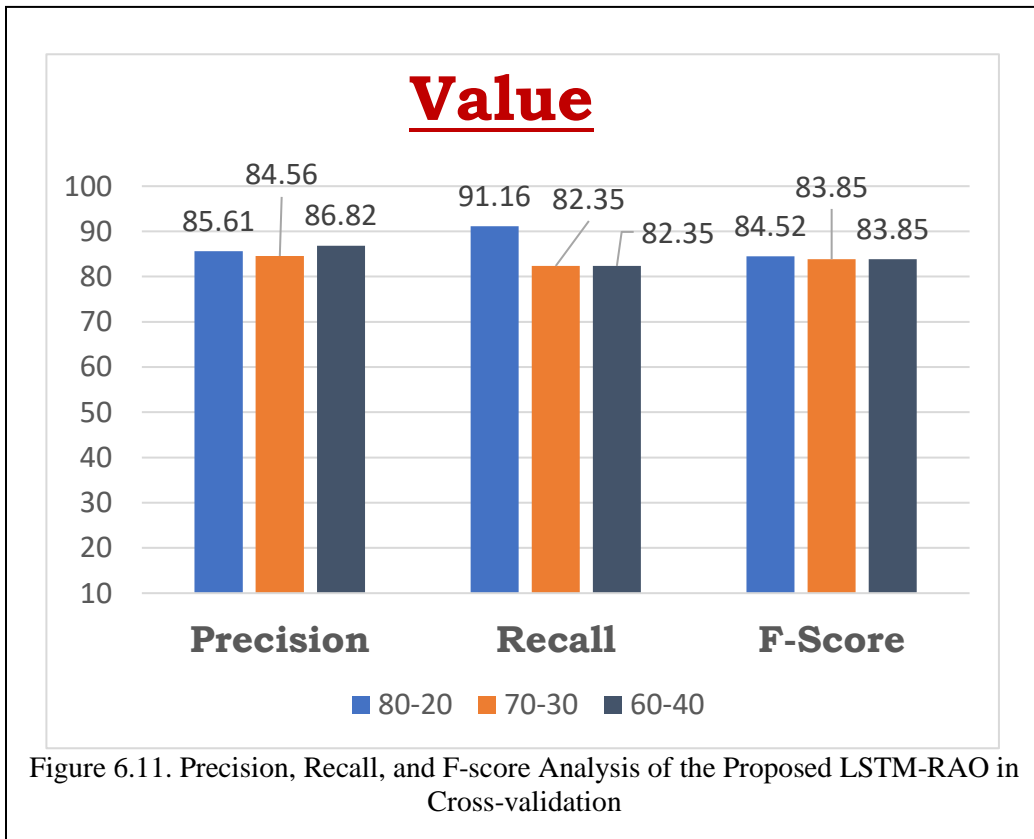


6.3.3 CROSS-VALIDATION

The Cross-validation of the proposed LSTM-RAO method for 80-20, 70-30, and 60-40 training and testing is analyzed, as shown in Figure 6.10. The proposed LSTM-RAO method has higher accuracy compared with other standard methods. The LSTM-RAO method has the advantage of the proper selection of LSTM network parameters. The proposed LSTM-RAO method has an accuracy of 86.56% and the LSTM-Adam optimization method 84.93% accuracy. The proposed LSTM-RAO method has a higher accuracy in the cross-validation analysis that shows that the LSTM-RAO method has able to provide higher accuracy for lower training data.



The Cross-validation analysis is applied for the proposed LSTM-RAO method and measured precision, recall, and F-Score, as shown in Figure 6.11. The proposed LSTM-RAO method has higher precision, recall and F-Score is high in cross-validation. The proposed LSTM-RAO method has a precision value of 85.61% and the LSTM-Adam method has 82.78% precision. The proposed LSTM-RAO method has the advantage of the proper selection of parameters for the LSTM network.



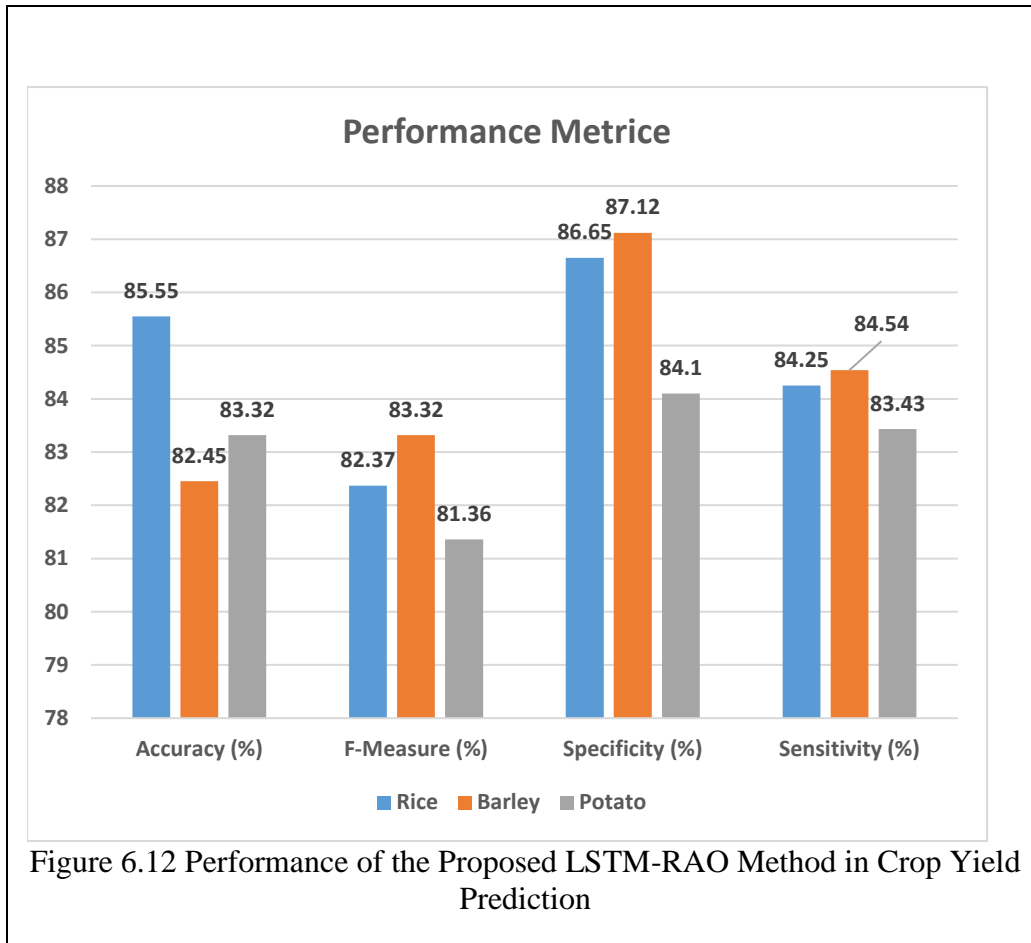
The performance of the proposed LSTM-RAO method is compared with ANN, RNN, and LSTM-Adam Optimization methods, as shown in Table 6.3. The table shows that the proposed LSTM-Adam Optimization has a higher efficiency

compared with other methods in crop yield prediction. The graphical representation is shown in Figure 6.12.

Table 6.3. Performance of the Proposed LSTM-RAO Method in Crop Yield Prediction

Cross-Validation (%)	Parameter	ANN	RNN	LSTM-Adam	LSTM-RAO
80-20	Accuracy	81.23	74.35	84.93	86.56
	Precision	82.37	69.76	82.78	85.61
	Recall	81.72	71.69	74.82	91.16
	F-Score	82.81	76.91	70.18	84.52
70-30	Accuracy	80.45	72.71	81.13	82.02
	Precision	80.11	80.22	81.6	84.56
	Recall	81.72	74.17	78.51	82.35
	F-Score	79.81	76.91	79.89	83.85
60-40	Accuracy	78.67	66.5	70.06	83.23
	Precision	81.96	56.92	73.64	86.82
	Recall	79.58	74.67	71.87	82.35
	F-Score	80.69	60.32	76.72	83.85

The proposed LSTM-RAO method has the benefits of the proper selection of weight parameters for LSTM networks. The proposed LSTM-RAO method provides high performance in 60-40 training-testing validation. This shows that the proposed LSTM-RAO method has able to perform with less training data.



6.3.4 COMPARATIVE ANALYSIS

Table 6.4 shows the performance criteria of different existing methods for IE. The proposed deep learning-based LSTM-RAO information extraction methods are compared with existing techniques such as Deep Learning-based Weighted SOM and Ensemble-Neural Network (ENN) were evaluated in the combinations of testing and training percentage like 80% training and 20% testing of collected data.

Table 6.4 Comparison of Existing Techniques with Proposed Algorithm

Methodology	Accuracy	Sensitivity	Specificity
Weighted SOM [270]	78.98%	83.05%	81.45%
ENN [271]	65.12%	90.0%	90.6%
Proposed Algorithm (LSTM+RAO)	86.56%	91.16%	90.56%

The proposed deep learning-based method focused to extract the important features for agricultural corpus like AGNER, AGEE and semantic in between. Results show that the proposed method is better than the existing methods. From the above results, the proposed DL method achieved nearly 87% accuracy when compared with ENN and weighted SOM. The ENN method achieved less accuracy when compared with SOM because of consuming more time, but achieved better performance in both sensitivity and specificity. So, this tool can be used to recommend the farmers for crop selection that may improve the overall crop productivity.

6.4 SUMMARY

The present chapter summarizes with experimental setup and design for the proposed algorithm. Various results for AGNER and AGEE extracted using the deep learning-based LSTM-RAO model. Comparison parameters like precision, recall, accuracy, and f-measure helped the comparison between proposed methods and existing machine learning algorithms such as ANN, RNN, and LSTM-Adam. The proposed methods also cross-validated based on 60-40%, 70-30%, and 80-20% training and test data. Comparison between the proposed method with existing deep learning methods shows the dominance of proposed methods. In the next chapter, the conclusion of the present research work along with the future research directions will discuss.

CHAPTER - 7

CONCLUSION AND FUTURE DIRECTION

7.1 SUMMARY AND CONTRIBUTIONS

Information extraction is a crucial component of natural language processing and extracting semantic from IE makes it more meaningful. From a semi-structured or/and unstructured corpus, IE tasks mainly consider as retrieval of significant information like entities, events, and their relationship; finally stored it into a structured format (database), so that it can be easily processed in the future. While processing a huge amount of amorphous information, deep learning techniques play an important role to mine it. In the agricultural domain, deep learning methods used to extract the AGNER, AGEE, and the relationship between them, also realize the predominance of semantic information extraction.

The present research work emphasizes the implementation of deep learning-based information extraction techniques for the agricultural domain and provides meaningful structured information. The objective of the present research program is “Semantic Information Extraction using Deep Learning Techniques for Agricultural Domain”.

In the present research program, Chapter 1 started with a detailed introduction to Information Extraction, related tasks & tools, and the importance of Information Extraction techniques in related to the agricultural domain while mining significant information. Further, that chapter also presented contemporary deep learning algorithms for processing a large amount of data and the chapter concluded with the research objective of the present research program. Chapter 2 provided a detailed literature review on various data formats, information extraction tasks, deep learning techniques, and existing optimizers with special importance in the

agricultural field. The survey highlighted the functionality of statistical-based and machine learning-based information extraction techniques. Some ongoing projects along with few government initiatives are also described at the end of the chapter. Chapter 3 to chapter 6 described the research contribution of semantic information extraction for the agricultural domain and complete detailed research contributions mentioned in the following subsection. Finally, the thesis concluded in chapter 7 by recommending future research directions.

7.2 RESEARCH CONTRIBUTIONS

(Contribution 1): Proposed Framework for Semantic Information Extraction

Chapter 3 proposed a framework for semantic information extraction. The proposed framework discussed various tools and techniques required for IE. Initially, it accepted four corpora (weather, soil, agricultural, and pest & fertilizer) as input and converted them into a single unified corpus. Data preprocessing techniques are applied to normalize the input corpus. Deep learning-based rules are used to extract AGNER and AGEE. Further, semantic information was extracted based on mined AGNER and AGEE.

(Contribution 2): An Approach for data preprocessing

Chapter 4 described the data preprocessing stage. The chapter started with a detailed explanation of the four input corpora mentioned in this chapter. To remove noises such as missing values, blank values, imbalanced data, inconsistent data, etc. from input corpora, a min-max algorithm was applied. For merging these corpora and converted into a unified corpus, two different classes are used. Firstly, the integration of a Knowledge-based word sense disambiguation system into a corpus-based word sense disambiguation system through the NLTK library tools was proposed. Secondly, proposed a disambiguation algorithm to assign sense(s) of two same words with a different meaning in a document. Cosine similarity was used to

find the similarity between two words and cosine distance used to measure the distance between two words.

(Contribution 3): Extraction of semantic information from the agricultural corpus

Chapter 5 proposed a deep learning-based algorithm for semantic information extraction using LSTM and RAO. LSTM with RAO method-based deep learning rules well demonstrated the extraction of AGNER and AGEE. The Apache Solr and Lucene applied in proposed model to improve the overall efficiency of the system. Finally, the Semantic (relationship) between extracted AGNER and AGEE was extracted. The NLTK tools were used for the extraction of semantic relations between NER and EE of the agricultural domain.

Chapter 6 included the experimental design of the proposed model. Upon identification of the needs and the available technologies, the system model is presented along with the obtained results. These results were also compared with existing standard information extraction techniques. Further, the chapter graphically represented the obtained results and also analyzed them statistically. In most of the cases, it was observed that the proposed algorithm significantly performed than the compared standard information extraction algorithms.

7.3 FUTURE RESEARCH DIRECTIONS

Semantic information extraction from unstructured knowledgebase is one of the extremely important research areas of Natural language processing. For further improvement in the proposed model, the current research program can be explored in multiple directions.

Firstly, the research work can be extended by applying post-processing algorithms like heuristic and linguistic approaches to improve various components of the confusion matrix. At the data preprocessing stage, data transformation and data reduction techniques along with data normalization methods can be used in

future research. While creating a unified corpus, more parameters can be considered to make a more robust system. Agricultural diseases database can also be used to improve crop yielding. Real-time weather forecast data can also help in designing a robust system for agricultural information extraction. This enhanced system can use as a tool for farmers to improve overall yielding.

Leveraging the benefits of contemporary deep learning and development in existing techniques regularly opening new research areas along with an open call for improvement in ongoing research practices. Advanced deep learning methods such as mLSTM methods can reduce avoidable calculations and makes the training process speed faster. The present research program used fifteen tagset of AGNER and AGEE, these tags can be divided into subtags or few new tags can be added. The addition of POS tags may also improve the extraction of semantic information from AGNER and AGEE. As optimizers play a vital role in information processing, so the proposed Adam optimizer can be replaced by other existing or improved versions of optimizers. Stochastic Gradient Descent, Mini-Batch Gradient Descent, Momentum, Nesterov Accelerated Gradient also shown better results in various domains, so a hybrid optimizer approach can be fruitful.

Smart agriculture development with emergent IoT devices produced a huge amount of real-time data for data processing techniques. The IE in the agricultural domain is still at the evolving stage to cater to its actual purpose, and a lot of research is in progress.

REFERENCES

- [1] D. Sawant, A. Jaiswal, J. Singh and P. Shah, "AgriBot-An intelligent interactive interface to assist farmers in agricultural activities," in *IEEE Bombay Section Signature Conference (IBSSC)*, Mumbai, 2019.
- [2] K. M. Arjun, "Indian agriculture-status, importance and role in Indian economy," *International Journal of Agriculture and Food Science Technology*, vol. 4, no. 4, pp. 343-346, 2013.
- [3] A. Kant and A. Roy, "Discussion Paper National Strategy for Artificial Intelligence," NITI Aayog, 2018.
- [4] R. Sumithra and S. Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery.," in *IEEE*, 2010.
- [5] S. Patwardhan, *Widening the field of view of information extraction through sentential event recognition*, Doctoral dissertation, School of Computing, University of Utah, 2010.
- [6] DBMSInternals, "Unstructured data processing: Relational database Vs Hadoop," DBMS Internals, [Online]. Available: <http://www.dbmsinternals.com/blog/unstructured-data-processing-relational-database-vs-hadoop/>. [Accessed 22 Oct 2020].
- [7] A. P. Saygin, I. Cicekli and V. Akman, "Turing test: 50 years later," *Minds and machines*, vol. 10, no. 4, pp. 463-518, 2000.

- [8] A. P. Saygin, I. Cicekli and V. Akman, "Turing test: 50 years later," in *Minds and Machines*, Kluwer Academic Publishers, 2010, pp. 463-518.
- [9] R. Jiménez-Peris, C. Pareja-Flores, M. Patiño-Martínez and J. Á. Velázquez-Iturbide, "New technologies in computer science education," in *Computer Science education in the 21st century*, New York, Springer, 2000, pp. 113-136.
- [10] P. M. Nadkarni, L. Ohno-Machado and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association* , vol. 18, no. 5, pp. 544-551, 2011.
- [11] J. K. Gill, "Evolution and Future of Natural Language Processing (NLP)," Xenonstack A Stack Innovator, 5 Dec 2019. [Online]. Available: <https://www.xenonstack.com/blog/evolution-of-nlp/>. [Accessed 18 Nov 2020].
- [12] R. Kibble, "Introduction to natural language processing," University of London, 2013.
- [13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research* , pp. 2493-2537, 2011.
- [14] K. Verspoor and K. Bretonnel Cohen, "Natural Language Processing," in *Encyclopedia of Systems Biology*, New York, NY, Springer, 2013, pp. 1495-1498.
- [15] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard and N. Aswani, "Twitie: An open-source information extraction pipeline for

microblog text," in *international conference recent advances in natural language processing RANLP 2013*, 2013.

- [16] J. Piskorski and R. Yangarber, Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, Berlin, Heidelberg: Springer,, 2013, pp. 29-43.
- [17] D. E. Johnson and T. Hampp-Bahnmueeller, "Architecture of a framework for information extraction from natural language documents". U.S. Patent 6,553,385, 22 April 2003.
- [18] J. Cowie and W. Lehnert, *Information Extraction*, 1 ed., vol. 39, 1996, pp. 80-91.
- [19] E. Riloff and J. Lorenzen, "Extraction-based text categorization: Generating domain-specific role relationships automatically," *Natural language information retrieval*, pp. 167-196, 1999.
- [20] M. F. Moens, "Information Extraction: Algorithms and Prospects in a Retrieval Context," *Science & Business Media*, 2006.
- [21] E. M. Bender, "Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax," in *Synthesis lectures on human language technologies*, vol. 6, 2013, pp. 1-184..
- [22] K. Aberer, "Information Retrieval and Data Mining," EPFL-IC, Laboratoire de systèmes d'informations répartis, 2006-07.
- [23] F. Kiyoumars, "Evaluation of automatic text summarizations based on human summaries," *Procedia-Social and Behavioral Sciences*, vol. 192, pp. 83-91, 2015.

- [24] Y. K. Meena and D. Gopalani, "Domain independent framework for automatic text summarization," *Procedia Computer Science*, vol. 48, pp. 722-727, 2015.
- [25] S. D. Gheware, A. S. Kejkar and S. M. Tondare, "Data mining: Task, tools, techniques and applications," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 10, pp. 8095-8098, 2014.
- [26] S. B. Aher and L. M. R. J. Lobo, "Data Mining in Educational System using Weka," in *International Conference on Emerging Technology Trends (ICETT)*, Kollam, Kerala, 2011.
- [27] "Data Mining – Working, Characteristics, Types, Applications & Advantages," *Electricalfundablog.com*, 2015. [Online]. Available: <https://electricalfundablog.com/data-mining-working-characteristics-types-applications-advantages/>. [Accessed 14 Sep 2020].
- [28] A. J. Chamatkar and P. K. Butey, "Importance of data mining with different types of data applications and challenging areas," *Journal of Engineering Research and Applications*, vol. 4, no. 5, pp. 38-41, 2014.
- [29] S. P. Deshpande and V. M. Thakare, "Data mining system and applications: A review," *International Journal of Distributed and Parallel systems (IJDPS)*, vol. 1, no. 1, pp. 32-44, 2010.
- [30] J. H. Kroeze, M. C. Mathee and T. J. Bothma, "Differentiating between data-mining and text-mining terminology," *SA Journal of Information Management*, vol. 6, no. 4, 2004.

- [31] S. Dang and P. H. Ahmad, "Text mining: Techniques and its application," *International Journal of Engineering & Technology Innovations*, vol. 1, no. 4, pp. 22-25, 2014.
- [32] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang and J. Starren, "Automating content extraction of html documents," *World Wide Web*, vol. 8, no. 2, pp. 179-224, 2005.
- [33] M. T. Pazienza, M. Pennacchiotti and F. M. Zanzotto, "Terminology extraction: an analysis of linguistic and statistical approaches," in *Knowledge Mining*, Berlin, Heidelberg, Springer, 2005, pp. 255-279.
- [34] E. Lefever, L. Macken and V. Hoste, "Language-independent bilingual terminology extraction from a multilingual parallel corpus," in *12th Conference of the European Chapter of the ACL (EACL 2009)*, 2009.
- [35] A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72-79, 2001.
- [36] D. Jurafsky and J. H. Martin, "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition," in *Computational Linguistics and Speech Recognition*, Upper Saddle River, NJ: Prentice Hall PTR, 2000, pp. 638-641.
- [37] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge university press, 2007.
- [38] S. G. Small and L. Medsker, "Review of information extraction technologies and applications," *Neural computing and applications*, vol. 25, no. 3-4, pp. 533-548, 2014.

- [39] D. Jurafsky and J. H. Martin, "Chapter 21: Information Extraction," in *Speech and language processing*, 2019, pp. 1-30.
- [40] S. Zheng, J. J. Lu, N. Ghasemzadeh, S. S. Hayek, A. A. Quyyumi and F. Wang, "Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies," *JIMR Medical Informatics*, vol. 5, no. 2, pp. 1-11, 2017.
- [41] W3C, "internet usage and social media statistics," World wide web foundation, [Online]. Available: <http://www.internetlivestats.com>. [Accessed 27 OCT 2020].
- [42] J. Bingel and T. Haider, "Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers," in *9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.
- [43] R. Stern, "Identification automatique d'entités pour l'enrichissement de contenus textuels," PhD thesis, Université Paris-Diderot, Paris, 2013.
- [44] S. Sekine, K. Sudo and C. Nobata, "Extended Named Entity Hierarchy," in *3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, 2002.
- [45] R. Grishman and B. Sundheim, "Message Understanding Conference–6: A Brief History," in *16th International Conference on Computational Linguistics (COLING)*, 1996.
- [46] E. F. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *6th Conference on*

Computational Natural Language Learning (CoNLL), Taipei, Taiwan, 2002.

- [47] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel and R. Weischedel, "The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation.," in *4th International Conference on Language Resources and Evaluation (LREC)*, 2004.
- [48] N. Oostdijk, M. Reynaert, P. Monachesi, G. V. Noord, R. Ordelman, I. Schuurman and V. Vandeghinste, "From D-Coi to SoNaR: A Reference Corpus for Dutch.," in *6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.
- [49] *Name Recognition: NER and WEPS (unpublished lecture slides)*, 2009.
- [50] L. F. Rau, "Extracting Company Names from Text," in *7th IEEE Conference on Artificial Intelligence Applications*, Miami Beach, FL, USA, 1991.
- [51] P. Watrin, Une approche hybride de l'extraction d'information sous-langages et lexicque-grammaire, PhD thesis, Université Catholique de Louvain, 2006.
- [52] R. Florian, "Named Entity Recognition as a House of Cards: Classifier Stacking," in *6th Conference on Natural Language Learning*, 2002.
- [53] Z. Kozareva, Ó. Ferrández, A. Montoyo, R. Muñoz, A. Suárez and J. Gómez, "Combining Data-Driven Systems for Improving Named Entity Recognition," *Data & Knowledge Engineering*, vol. 61, no. 3, p. 449–466, 2007.

- [54] C. Brun, M. Ehrmann and G. Jacquet, "XRCE-M: A Hybrid System for Named Entity Metonymy Resolution," in *4th International Workshop on Semantic Evaluations*, Prague, 2007.
- [55] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard and N. Aswani, "TwitIE: An open-source information extraction pipeline for microblog text," in *International Conference Recent Advances in Natural Language Processing*, BULGARIA, 2013.
- [56] C. Aone and M. Ramos-Santacruz, "REES: A Large-Scale Relation and Event Extraction System," in *6th Conference on Applied Natural Language Processing*, Seattle, WA, USA, 2000, 2000.
- [57] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz and D. R. Radev, "TimeML: Robust Specification of Event and Temporal Expressions in text," in *New Directions in Question Answering*, 2003, p. 28–34.
- [58] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer and J. Pustejovsky, TimeML annotation guidelines, 1 ed., vol. 1, 2006, p. 31.
- [59] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel and R. Weischedel, "The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation," in *4th International Conference on Language Resources and Evaluation (LREC)*, 2004.
- [60] T. guardian, "Indian election 2014: your interactive guide to the world's biggest vote," Guardian News & Media Limited, 7 April 2014. [Online]. Available: <https://www.theguardian.com/world/2014/apr/07/-sp-indian->

election-2014-interactive-guide-narendra-modi-rahul-gandhi. [Accessed 20 Nov 2020].

- [61] M. B. Habib and M. v. Keulen, "Information Extraction for Social Media," Third Workshop on Semantic Web and Information Extraction, Dublin, Ireland, 2014.
- [62] G. K. Palshikar and R. Srivastava, *Information Extraction for Effective Knowledge Management*, TCS Design Services , 2015.
- [63] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015.
- [64] A. Kawtraku, "Knowledge as a Service for Agriculture Domain," *Information and Communications Technology Services*, 2011.
- [65] L. Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory," in *Computer Application and System Modeling (ICCASM)*, Taiyuan, China., 2010.
- [66] S. Verma, S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram and K. M. Anderson, "Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency," in *ICWSM*, Atlanta, Georgia, U.S., 2011.
- [67] S. Vieweg, A. L. Hughes, K. Starbird and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness.," in *SIGCHI conference on human factors in computing systems*, Vancouver BC, Canada, 2011.

- [68] X. Zhang, L. Jing, X. Hu, M. Ng, J. X. Jiangxi and X. Zhou, "Medical document clustering using ontology-based term similarity measures," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 4, no. 1, pp. 62-73, 2008.
- [69] P. Gamallo and M. Garcia, "Dependency-Based Open Information Extraction," in *Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012.
- [70] H. Wang, *Semantic deep learning*, University of Oregon, 2015, pp. 1-42.
- [71] F. Wu and D. S. Weld, "Automatically refining the wikipedia infobox ontology," in *17th international conference on World Wide Web*, 2008.
- [72] R. Calandra, T. Raiko, M. P. Deisenroth and F. M. Pouzols, "Learning deep belief networks from non-stationary streams," in *International Conference on Artificial Neural Networks*, Berlin, Heidelberg,, 2012.
- [73] J. Zhu, Z. Nie, X. Liu, B. Zhang and J.-R. Wen, "StatSnowball: A statistical approach to extracting entity relationships," in *International Conference on World Wide Web*, Madrid, Spain, 2009.
- [74] P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *25th International Conference on Machine Learning*, 2008.
- [75] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015.

- [76] Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798-1828, 2013.
- [77] Y. Bengio, Learning deep architectures for AI, Now Publishers Inc, 2009.
- [78] B. Y, "Deep learning of representations: Looking forward.," in *1st International Conference on Statistical Language and Speech Processing. SLSP'13.*, Tarragona, Spain, 2013.
- [79] M. Chen, X. ZE, W. KQ and S. F, "Marginalized denoising autoencoders for domain adaptation," in *29th International Conference in Machine Learning*, Edingburgh, Scotland, 2012.
- [80] X. Zhang, L. Jing, X. Hu, M. Ng, J. X. Jiangxi and X. Zhou, "Medical document clustering using ontology-based term similarity measures," *International Journal of Data Warehousing and Mining*, vol. 4, no. 1, pp. 62-73, 2008.
- [81] D. M. Dimiduk, E. A. Holm and S. R. Niezgod, "Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering," *Integrating Materials and Manufacturing Innovation*, vol. 7, no. 3, pp. 157-172, 2018.
- [82] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *ICML*, 2011.
- [83] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580, 2012.

- [84] J. Ahmad, H. Farman and Z. Jan, "Deep learning methods and applications," in *Deep Learning: Convergence to Big Data Analytics*, Singapore, Springer, 2019, pp. 31-42.
- [85] A. Panigrahi, Y. Chen and C.-C. J. Kuo, Analysis on Gradient Propagation in Batch Normalized Residual Networks, arXiv preprint arXiv:1812.00342, 2018.
- [86] J. Lorraine and D. Duvenaud, Stochastic hyperparameter optimization through hypernetworks, arXiv preprint arXiv:1802.09419, 2018.
- [87] A. Vieira and B. Ribeiro, Introduction to Deep Learning Business Applications for Developers, Apress, 2018.
- [88] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2897-2905., 2018.
- [89] T. Takahashi, "Statistical max pooling with deep learning". US Patent Patent 10,013,644., 3 July 2018.
- [90] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.
- [91] C. Liu, Y. Li, H. Fei and P. Li, "Deep skip-gram networks for text classification," in *International Conference on Data Mining*, 2019.
- [92] S. J. Pan and Q. Yang, "A survey on transfer learning.," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, p. 1345–1359, 2009.

- [93] A. Mathew, P. Amudha and S. Sivakumari, "Deep Learning Techniques: An Overview," in *International Conference on Advanced Machine Learning Technologies and Applications*, Singapore, 2020.
- [94] P. Sharma, Top 5 deep learning frameworks, their applications, and comparisons!, 2019.
- [95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Computer Vision and Pattern Recognition*, arXiv preprint arXiv:1409.1556, 2014.
- [96] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 1, no. 4, pp. 568-576, 2014.
- [97] Y. Fan, Y. Qian, F.-L. Xie and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [98] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," in *Neural networks*, vol. 18, 2005, pp. 602-610.
- [99] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Machine Learning*, arXiv preprint arXiv:1511.06434, 2015.
- [100] P. L. Suárez, A. D. Sappa and B. X. Vintimilla, "Infrared image colorization based on a triplet dcgan architecture," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

- [101] J. Jin, K. Fu and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991-2000, 2014.
- [102] S. R. Ali, H. Azizpour, J. Sullivan and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE conference on computer vision and pattern recognition workshops*, 2014.
- [103] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Machine Learning*, arXiv preprint arXiv:1511.06434, 2015.
- [104] K. G. Liakos, P. Busato, D. Moshou, S. Pearson and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.
- [105] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and electronics in agriculture*, vol. 147, pp. 70-90, 2018.
- [106] L. Santos, F. N. Santos, P. M. Oliveira and P. Shinde, "Deep learning applications in agriculture: A short review.," in *Iberian Robotics conference*, Cham, 2019.
- [107] N. S. Punn, S. Agarwal, M. Syafrullah and K. Adiyarta, "Testing Big Data Application," in *6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2019.
- [108] R. Sint, S. Schaffert, S. Stroka and R. Ferstl, "Combining unstructured, fully structured and semi-structured information in semantic wikis," in *4th Semantic Wiki Workshop (SemWiki 2009) at the 6th European Semantic Web Conference (ESWC 2009)*, , Hersonissos, Greece, 2009.

- [109] Y. Chen, W. Wang, Z. Liu and X. Lin, "Keyword search on structured and semi-structured data," in *In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 2009.
- [110] A. C. Eberendu, "Unstructured Data: an overview of the data of Big Data," *International Journal of Computer Trends and Technology*, vol. 3, no. 1, pp. 46-50, 2016.
- [111] M. Tsumura, Y. Wang, J. Burbank, J. Sam and G. Sun, "Obtaining data from unstructured data for a structured data collection". U.S. Patent 9,299,041, 29 March 2016.
- [112] S. S. Tandel, A. Jamadar and S. Dudugu, "A Survey on Text Mining Techniques," in *th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE*,, 2019.
- [113] N. Padhy, D. Mishra and R. Panigrahi, "The survey of data mining applications and feature scope," *International Journal of Computer Science, Engineering and Information Technology*, 2012.
- [114] J. Han and M. Kamber, *Data Mining: Concepts and Techniques 2nd Edition*, Morgan Kaufmann, 2005.
- [115] W. B. Croft, D. Metzler and T. Strohman, *Search Engines - Information Retrieval in Practice*, Pearson Education, 2009.
- [116] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval - the concepts and technology behind search* Second Edition, Harlow, England: Pearson Education Ltd, 2011.

- [117] K. Willett and J. P. Sparck, *Readings in Information Retrieval*, Morgan Kaufmann Publishers, 1997.
- [118] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation," *Lrec*, vol. 12, no. 1, pp. 837-840, 2004.
- [119] S. Sarawagi, "Information Extraction," *Foundations and Trends in Databases*, pp. 261-377, 2007.
- [120] R. Sundheim and B. Grishman, "Message understanding conference-6: A brief," in *6th Conference on Computational Linguistics*, USA, Morristown, 1997.
- [121] J. Turmo, A. Ageno and N. Catal`a, "Adaptive Information Extraction," *ACM Computer Services*, 2008.
- [122] H. Cunningham, *Information Extraction*, Encyclopedia of Language Second Edition, 2005.
- [123] R. Grishman, "Information extraction: Techniques and challenges," *SCIE*, 1997.
- [124] C. Siefkes and P. Siniakov, "An overview and classification of adaptive approaches to information extraction," *Journal on Data Semantics IV*, pp. 172-212, 2005.
- [125] M. Shaker, H. Ibrahim, A. Mustapha and L. N. Abdullah, "A strategy for extracting information from semi-structured web pages," *International Journal of Web Information Systems*, vol. 6, no. 4, pp. 304-318, 2010.

- [126] Z. Liang, J. Chen, Z. Xu, Y. Chen and T. Hao, "A pattern-based method for medical entity recognition from chinese diagnostic imaging text," *Frontiers in Artificial Intelligence*, 14 May 2019.
- [127] S. Abdallah, K. Shaalan and M. Shoaib, "Integrating rule-based system with classification for arabic named entity recognition," in *International Conference on Intelligent Text Processing and Computational Linguistics*, Berlin, Heidelberg, 2012.
- [128] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis and C. D. Spyropoulos, "Using machine learning to maintain rule-based named-entity recognition and classification systems," in *39th Annual Meeting of the Association for Computational Linguistics*, 2001.
- [129] H.-J. Song, B.-C. Jo, C.-Y. Park, J.-D. Kim and Y.-S. Kim, "Comparison of named entity recognition methodologies in biomedical documents," *BioMedical Engineering OnLine*, vol. 17, no. 2, p. 158, 6 November 2018.
- [130] C. Zhang, "Combining statistical machine learning models to extract keywords from chinese documents," in *In International Conference on Advanced Data Mining and Applications*, Berlin, Heidelberg, 2009.
- [131] A. Téllez, M. Montés and L. Villaseñor, "A machine learning approach to information extraction," in *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, 2005.
- [132] R. B. Millar, *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*, vol. 111, John Wiley & Sons, 2011.

- [133] S. Morwal, N. Jahan and D. Chopra, "Named entity recognition using hidden Markov model (HMM)," *International Journal on Natural Language Computing (IJNLC)*, vol. 1, no. 4, pp. 15-23, 2012.
- [134] A. Ekbal and S. Bandyopadhyay, "Named entity recognition using support vector machine: A language independent approach," *International Journal of Electrical, Computer, and Systems Engineering*, vol. 4, no. 2, pp. 155-170, 2010.
- [135] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *23rd international conference on Machine learning*, 2006.
- [136] TechVidvan, "Supervised Learning Algorithm in Machine Learning," TechVidvan, [Online]. Available: <https://techvidvan.com/tutorials/supervised-learning/>. [Accessed 22 Oct 2020].
- [137] J. Wu, L. Yao and B. Liu, "An overview on feature-based classification algorithms for multivariate time series," 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA).
- [138] Z. Wang, S. Chen and T. Sun, "MultiK-MHKS: a novel multiple kernel learning algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 348-353, 2007.
- [139] H. Singh, "Performance analysis of unsupervised machine learning techniques for network traffic classification," in *15th International Conference on Advanced Computing & Communication Technologies*, 2015.

- [140] M. Khanum, T. Mahboob, W. Imtiaz, H. A. Ghafoor and R. Sehar, "A survey on unsupervised machine learning algorithms for automation, classification and maintenance," *International Journal of Computer Applications*, vol. 119, no. 13, 2015.
- [141] P. Rai and S. Singh, "A survey of clustering techniques," *International Journal of Computer Applications*, vol. 7, no. 12, pp. 1-5, 2020.
- [142] G. Stanovsky and I. Dagan, "Open ie as an intermediate structure for semantic tasks," in *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015.
- [143] N. Kushmerick, D. S. Weld and R. Doorenbos, *Wrapper induction for information extraction*, Washington: University of Washington, 2007.
- [144] A. Mezcic, "Why Unsupervised Machine Learning is the Future of Cybersecurity," *TECHNATIVE*, 28 January 2020. [Online]. Available: <https://www.technative.io/why-unsupervised-machine-learning-is-the-future-of-cybersecurity/>. [Accessed 20 March 2020].
- [145] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.
- [146] L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237-285, 1996.
- [147] E. Riloff and R. Jones, "Learning dictionaries for information extraction by multi-level bootstrapping," *AAAI-99 Proceedings*, pp. 474-479, 1999.

- [148] X. Li, S. Mabu and K. Hirasawa, "A novel graph-based estimation of the distribution algorithm and its extension using reinforcement learning," *IEEE transactions on evolutionary computation*, vol. 18, no. 1, pp. 98-113, 2013.
- [149] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, pp. 802-810, 2015.
- [150] A. Gokhale, "Introduction to Reinforcement Learning : In 2(or a bit more) Minutes," the Society of Robotics And Automation, [Online]. Available: <http://www.sra.vjti.info/blog/machine-learning/introduction-to-reinforcement-learning-in-2-minutes>. [Accessed 23 November 2020].
- [151] J. Ang, M. Hong, D. L. Zhang and J. Li, "Information Extraction: Methodologies and Applications," in *Emerging Technologies of Text Mining: Techniques and Applications*, IGI Global, 2008.
- [152] V. Solovyev, R. Gareev, V. Ivanov, S. Serebryakov and N. Vassilieva, "Dictionary and pattern-based recognition of organization names in Russian news texts," *Information Technology and Computer Science*..
- [153] K. Amita, S. V. Shankar, M. Sanjay and B. M. Sarvesh, "Effectiveness of the Pattern-Based Approach in the Cytodiagnosis of Salivary Gland Lesions. Acta cytologica, 60(2).," *Acta cytologica*, vol. 60, no. 2, pp. 107-117, 2016.
- [154] B. Li, X. Si, M. R. Lyu, I. King and E. Y. Chang, "Question identification on twitter," in *Proceedings of the 20th ACM international conference on Information and knowledge management ACM.*, 2011.

- [155] D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer and J. Fluck, "ProMiner: rule-based protein and gene entity recognition.," *BMC bioinformatics*, vol. 6, no. 1, 2005.
- [156] M. Dieb, "Ensemble approach to extract chemical named entity by using results of multiple cner systems with different characteristic.," *BioCreative Challenge Evaluation Workshop*, vol. 2, p. 162, 2013.
- [157] R. Alfred, L. C. Leong, C. K. On and P. Anthony, "Malay named entity recognition based on rule-based approach.," *International Journal of Machine Learning and Computing*, vol. 4, p. 300, 2014.
- [158] K. Sarkar, "A hidden markov model based system for entity extraction from social media english text," 2015.
- [159] J. Sun, J. Gao, L. Zhang, M. Zhou and C. Huang, "Chinese named entity identification using class-based language model.," *In Proceedings of the 19th international conference on Computational linguistics*, vol. 1, 2002.
- [160] C. Quan, M. Wang and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PloS one*, vol. 9, no. 7, 2014.
- [161] C.-K. Hsu, B.-J. P. Wang and C. E. K. Ming-Wei, "Simple and Knowledge-intensive Generative Model for Named Entity Recognition," 2013.
- [162] S. Zhao, "Named entity recognition in biomedical texts using an HMM model.," *In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications Association for Computational Linguistics*, pp. 84-87, 2004.
- [163] S. P. Umare and D. N. A. Deshpande, "A Survey on Machine Learning Techniques to Extract Chemical Names from Text Documents.,"

International Journal of Computer Science and Information Technology,
vol. 4, pp. 1263-1266.

- [164] S. Amarappa and S. V. Sathyanarayana, "Named entity recognition and classification in kannada language," *International Journal of Electronics and Computer Science Engineering*, vol. 2, no. 1, pp. 281-289, 2013.
- [165] K. B. Cohen and L. Hunte, "Natural language processing and systems biology," *Artificial intelligence and systems biology*, pp. 147-173, 2005.
- [166] A. L. Garrido, M. G. Buey, G. Muñoz and J.-L. Casado-Rubio, "Information extraction on weather forecasts with semantic technologies," in *International Conference on Applications of Natural Language to Information Systems*, Cham, 2016.
- [167] N. Chinchor, P. Robinson and E. Brown, *Hub-4 Named Entity Task Definition.*, 1998.
- [168] G. Ge, Z. Shi, Y. Zhu, X. Yang and Y. Hao, "Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four machine learning algorithms," *Global Ecology and Conservation*, vol. 22, p. 00971, 2020.
- [169] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, pp. 3-6, 2007.
- [170] A. Kawtraku, "Knowledge as a Service for Agriculture Domain," *Information and Communications Technology Services*, 2011.

- [171] W. H. Blake, K. J. Ficken, P. Taylor, M. A. Russell and D. E. Walling, "Tracing crop-specific sediment sources in agricultural catchments," *Geomorphology*, Vols. 139-140, pp. 322-329, 2012.
- [172] A. McCallum, D. Freitag and F. C. Pereira, "Maximum entropy Markov models for information extraction and segmentation.," in *ICML*, 2000.
- [173] S. Jiang, R. Angarita, R. Chiky, S. Cormier and F. Rousseaux, "Towards the Integration of Agricultural Data from Heterogeneous Sources: Perspectives for the French Agricultural Context Using Semantic Technologies," in *International Conference on Advanced Information Systems Engineering*, Cham, 2020.
- [174] G. Palshikar and R. Srivastava, "Information Extraction for Effective Knowledge Management," TCS Design Services, Mumbai, 2015.
- [175] O. Medelyan and I. H. Witten, "Thesaurus-based index term extraction for agricultural documents," *EFITA/WICCA*, pp. 1122-1129, 2005.
- [176] N. Chinchor, "Overview of MUC-7/MET-2," in *In Proc. Message Understanding Conference MUC-7*, 1999.
- [177] R. Grishman and B. Sundheim, "Message Understanding Conference-6: A Brief History.," in *International Conference on Computational Linguistics.*, 1996.
- [178] L. F. Rau, "Extracting Company Names from Text," in *Conference on Artificial Intelligence Applications of IEEE.*, 1991.
- [179] H. Liu, Z.-Z. Hu, M. Torii, C. Wu and C. Friedman, "Quantitative assessment of dictionary-based protein named entity tagging," *Journal of the*

American Medical Informatics Association, vol. 13, no. 5, pp. 497-507, 2006.

- [180] A. Cogato, F. Meggio, M. De Antoni Migliorati and F. Marinello, "Extreme Weather Events in Agriculture: A Systematic Review," *Sustainability*, vol. 11, 2019.
- [181] W. Nuij, V. Milea, F. Hogenboom, F. Frasincar and U. Kaymak, "An automated framework for incorporating news into stock trading strategies," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 4, pp. 823-825, 2014.
- [182] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel and R. Weischedel, "The Automatic Content Extraction (ACE) Program — Tasks, Data, and Evaluation.," in *Conference on Language Resources and Evaluation*, 2004.
- [183] E. F. Tjong Kim Sang, "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition.," in *Conference on Natural Language Learning.*, 2002.
- [184] E. F. Tjong Kim Sang and De Meulder, " Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.," in *Conference on Natural Lan-guage Learning.*, 2003.
- [185] D. Santos, N. Seco, N. Cardoso and R. Vilela, "HAREM: An Advanced NER Evaluation Contest for Portuguese," in *International Conference on Language Resources and Evalu-ation.*, 2006.
- [186] D. T. Huynh, "Entity Extraction from Unstructured Data on the Web," The University of Queensland, Australia., 2014.

- [187] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard and N. Aswani, "Twitie: An open-source information extraction pipeline for microblog text," in *In Proceedings of the international conference recent advances in natural language processing RANLP*, Bulgaria, 2013.
- [188] M. B. H. Morgan and M. V. Keulen, "Information extraction for social media," in *Workshop on Semantic Web and Information Extraction*, 2014.
- [189] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. Martin, M. Palmer, A. Schram and K. Anderson, "Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency," in *5th International AAI Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, 2011.
- [190] S. Vieweg, A. L. Hughes, K. Starbird and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *28th International Conference on Human Factors in Computing Systems, ACM*, Newyork, 2010.
- [191] M. Sini, S. Rajbhandari, J. Singh, J. Keizer, T. V. Prabhakar and A. Kawtrakul, "Smart Organization of Agricultural Knowledge The example of the AGROVOC Concept Server and Agropedia. Advances in Knowledge Organization," *Advances in Knowledge Organizations*, vol. 12, pp. 322-326, 2010.
- [192] G. S. Nair and U. C. Mohanty, "Prediction of Monthly Summer Monsoon Rainfall Using Global Climate Models Through Artificial Neural Network Technique,," *Pure and Applied Geophysics*, vol. 175, no. 1, pp. 403-419, 2018.

- [193] O. Satir and S. Berberoglu, "Crop yield prediction under soil salinity using satellite derived vegetation indices," *Field Crops Research*, vol. 192, pp. 134-143, 2016.
- [194] B. Das, B. Nair, V. K. Reddy and P. Venkatesh, "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," *International journal of biometeorology*, vol. 62, no. 10, pp. 1809-1822, 2018.
- [195] H. Ali and B. Verma, "Monthly Rainfall Forecasting using One-Dimensional Deep Convolutional Neural Network," *IEEE Access*, 2018..
- [196] X. He, H. Guan and J. Qin, "A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall," *Journal of Hydrology*, vol. 527, pp. 88-100, 2015.
- [197] E. L. Malarkodi and S. Devi, "Named Entity Recognition for the Agricultural Domain," *Research in Computing Science*, vol. 117, pp. 121-132, 2016.
- [198] N. Kaushik and N. Chatterjee, "Automatic relationship extraction from agricultural text for ontology construction," *Information processing in agriculture*, vol. 5, no. 1, pp. 60-73, 2018.
- [199] W. Xiang and B. Wang, "A Survey of Event Extraction From Text," *IEEE Access*, pp. 173111-173137, 2019.
- [200] S. Abrahams, "Developing and executing electronic commerce applications with occurrences,," Ph.D. dissertation, Dept. Comput. Sci. Tech-nol., Univ. Cambridge, , Cambridge, , 2002.

- [201] P. Capet, T. Delavallade, T. Nakamura, A. Sandor, C. Tarsitano and S. Voyatzi, "A risk assessment system with automatic extraction of event types," in *International Conference on Intelligent Information Processing*, 2008.
- [202] M. Atkinson, J. Piskorski, H. Tanev, E. v. d. Goot, R. Yangarber and V. Zavarella, "Automated event extraction in the domain of bordersecurity," in *International Conference on User Centric Media*, Heidelberg, 2009.
- [203] H. Tanev, J. Piskorski and M. Atkinson, "Real-time news event extraction for global crisis monitoring," in *International Conference on Application of Natural Language to Information Systems*, Berlin, Heidelberg, 2008.
- [204] S. Prasad and P. P. Singh, "Medicinal plant leaf information extraction using deep features," in *IEEE Region 10 Conference*, Penag, 2017.
- [205] B. Jiang, M.-x. Zhu and J. Wang, "Ontology-Based Information Extraction of Crop Diseases on Chinese Web Pages," *JOURNAL OF COMPUTERS*, vol. 8, no. 1, pp. 85-90, 2013.
- [206] O. Zurovec, P. Vedeld and B. Sitaula, "Agricultural Sector of Bosnia and Herzegovina and Climate Change—Challenges and Opportunities," *Agriculture*, 2015.
- [207] C. Falco, F. Donzelli and A. Olper, "A. Climate Change, Agriculture and Migration: A Survey," *Sustainability*, vol. 10, 2018.
- [208] G. Nelson, D. van der Mensbrugge, H. Ahammad, E. Blanc, K. Calvin, T. Hasegawa and P. Havlik, "Agriculture and climate change in global scenarios: Why don't the models agree," *Agric. Econ*, vol. 24, pp. 85-101, 2014.

- [209] R. Mendelsohn, W. Nordhaus and D. Shaw, "The impact of global warming on agriculture: A Ricardian," *Am. Econ. Rev*, vol. 84, p. 753–771, 1994.
- [210] O. Deschenes and M. Greenstone, "The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather.," *Am. Econ.* , p. 354–385, 2007.
- [211] C. Rossi, F. S. Acerbo, K. Ylinen, I. Juga, P. Nurmi, A. Bosca, F. Tarasconi, M. Cristoforetti and A. Alikadic, "Early detection and information extraction for weather-induced floods using social media streams," *International journal of disaster risk reduction*, vol. 30, pp. 145-157, 2018.
- [212] M. B. H. Morgan and M. V. Keulen, "Information Extraction for Social Media," in *Proceedings of Third Workshop on Semantic Web and Information Extraction*, August 2014..
- [213] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, pp. 1-21, 2015.
- [214] L. Chengcheng, "Automatic Text Summarization Based On Rhetorical Structure Theory,," in *Computer Application and System Modeling (ICCASM), 2010*, 2010.
- [215] D. Jurafsky and J. H. Martin, "Speech and Language Processing, T. E. draft,," *Ed., Stanford: Pearson International Edition*, 2017.
- [216] R. Galliers, "Towards a flexible information architecture: integrating business strategies,," *information systems strategies and business process redesign. Information Systems Journal*, vol. 3, no. 3, pp. 199-213, 1993.

- [217] C. Cunningham, "Can three incongruence tests predict when data should be combined?," *Molecular Biology and Evolution*, vol. 14, no. 7, pp. 733-740., 1997.
- [218] D. Soutner and L. Müller, "Application of LSTM neural networks in language modelling," in *International Conference on Text, Speech and Dialogue*. Springer, Berlin, Heidelberg, 2015.
- [219] W. A. Bennage and A. K. Dhingra, "Single and multiobjective structural optimization in discrete-continuous variables using simulated annealing," *International Journal for Numerical Methods in Engineering*, vol. 38, no. 16, pp. 2753-2773, 1995.
- [220] K. C. Tan, E. F. Khor, T. H. Lee and Y. J. Yang, "A tabu-based exploratory evolutionary algorithm for multiobjective optimization," *Artificial Intelligence Review*, vol. 19, no. 3, pp. 231-260, 2003.
- [221] L. S. Pitsoulis and M. G. Resende, "Greedy randomized adaptive search procedures," in *Handbook of applied optimization*, 2002, pp. 168-183.
- [222] C. C. Ribeiro and M. C. Souza, "Variable neighborhood search for the degree-constrained minimum spanning tree problem," *Discrete Applied Mathematics*, vol. 118, no. 1-2, pp. 43-54, 2002.
- [223] C. Voudouris and E. Tsang, "Guided local search and its application to the traveling salesman problem," *European journal of operational research*, vol. 113, no. 2, pp. 469-499, 1999.
- [224] T. Stützle, "Iterated local search for the quadratic assignment problem," *European Journal of Operational Research*, vol. 174, no. 3, pp. 1519-1539, 2006.

- [225] B. K. Ambati, J. Ambati and M. M. Mokhtar, "Heuristic combinatorial optimization by simulated Darwinian evolution: a polynomial time algorithm for the traveling salesman problem," *Biological Cybernetics*, vol. 65, no. 1, pp. 31-35, 1991.
- [226] C. Echegoyen, A. Mendiburu, R. Santana and J. A. Lozano, "A quantitative analysis of estimation of distribution algorithms based on Bayesian networks," Department of Computer Science and Artificial Intelligence, Technical Report EHU-KZAA-TR-2-2009 , 2009.
- [227] Y. E. Yildiz and A. O. Topal, "Large scale continuous global optimization based on micro differential evolution with local directional search," *Information Sciences* , vol. 477, pp. 533-544, 2019.
- [228] K. Krawiec and M. Heywood, "Solving complex problems with coevolutionary algorithms," in *the Genetic and Evolutionary Computation Conference Companion*, 2019.
- [229] C.-J. Chung and R. G. Reynolds, "A Testbed for Solving Optimization Problems Using Cultural Algorithms," *Evolutionary programming*, pp. 225-236, 1996.
- [230] F. Glover, "A template for scatter search and path relinking," in *European Conference on Artificial Evolution*, Berlin, Heidelberg, 1997.
- [231] J. Schöpfel, H. Prost and V. Rebouillat, "Research data in current research information systems," *Computer Science*, vol. 106, pp. 305-320, 2017.
- [232] N. Milosevic, C. Gregson, R. Hernandez and G. Nenadic, "A framework for information extraction from tables in biomedical literature," *International*

Journal on Document Analysis and Recognition (IJ DAR), vol. 22, no. 1, pp. 55-78, 2019.

- [233] R. Alfred, K. S. Gan, K. O. Chin and P. Anthony, "A robust framework for web information extraction and retrieval," *International Journal of Machine Learning and Computing*, vol. 4, no. 2, pp. 146-150, 2014.
- [234] T. M. Dieb, M. Yoshioka, S. Hara and M. C. Newton, "Framework for automatic information extraction from research papers on nanocrystal devices," *Beilstein journal of nanotechnology*, vol. 6, no. 1, pp. 1872-1882, 2015.
- [235] N. Sakhaee and M. C. Wilson, "Information extraction framework to build legislation network," *Artificial Intelligence and Law*, pp. 1-24, 2020.
- [236] K. Gandhi and N. Madia, "Information extraction from unstructured data using RDF," in *International Conference on ICT in Business Industry & Government (ICTBIG), IEEE*, 2016.
- [237] R. Fagin, B. Kimelfeld, F. Reiss and S. Vansummeren, "A relational framework for information extraction," *ACM SIGMOD Record*, vol. 44, no. 4, pp. 5-16, 2016.
- [238] N. Zhu, X. Liu, Z. Liu, K. Hu, Y. Wang, J. Tan, M. Huang, Q. Zhu, X. Ji, Y. Jiang and Y. Guo, "Deep learning for smart agriculture: Concepts, tools, applications, and opportunities," *International Journal of Agricultural and Biological Engineering*, vol. 11, no. 4, pp. 32-44, 2018.
- [239] A. Montoyo, A. Suárez, G. Rigau and M. Palomar, "Combining knowledge- and corpus-based word-sense-disambiguation methods," *Journal of Artificial Intelligence Research*, vol. 23, pp. 299-330, 2005.

- [240] E. Agirre and D. Martinez, "Knowledge sources for word sense disambiguation," in *International Conference on Text, Speech and Dialogue*, Berlin, Heidelberg, 2001.
- [241] A. Montoyo and A. Suárez, "The University of Alicante word sense disambiguation system," *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 131-134, 2001.
- [242] D. Brickley, R. V. Guha and A. Layman, "Resource description framework (RDF) schema specification," W3C, 5 January 1999. [Online]. Available: <https://www.w3.org/TR/PR-rdf-syntax/Overview.html>. [Accessed 23 March 2020].
- [243] G. Antoniou and F. V. Harmelen, *Web ontology language (owl) (Handbook on ontologies)*, Berlin, Heidelberg: Springer, 2004, pp. 67-92.
- [244] I. Gallo and E. Binaghi, "Information extraction and classification from free text using a neural approach," in *Iberoamerican Congress on Pattern Recognition*, Berlin, Heidelberg, 2007.
- [245] T. Finin, J. Mayfield and B. Grosz, "DARPA Agent Markup Language (DAML) Tools for Supporting Intelligent Annotation, Sharing and Retrieval," Maryland Univ Baltimore, Baltimore, 2007.
- [246] S. Bird, E. Klein and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, O'Reilly Media, Inc., 2009.
- [247] J. Aungiers, "Time Series Prediction using LSTM Deep Neural Networks," *Altum Intelligence*, 1 September 2018. [Online]. Available:

<https://www.altumintelligence.com/articles/a/Time-Series-Prediction-Using-LSTM-Deep-Neural-Networks>. [Accessed 10 May 2019].

- [248] J. Brownlee, "Gentle introduction to the adam optimization algorithm for deep learning," Machine Learning Mastery, 20 August 2020. [Online]. Available: <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>. [Accessed 25 Nov 2020].
- [249] D. P. Kingma and J. Ba., Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
- [250] S. Ruder, An overview of gradient descent optimization algorithms, arXiv preprint arXiv:1609.04747, 2016.
- [251] M. Devin, Q. V. Le, M. Z. Mao, M. A. Ranzato, A. Senior, P. Tucker, K. Yang and A. Y. Ng., "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223-1231.
- [252] S. Sarawagi, Information extraction, Now Publishers Inc, 2008.
- [253] A. W. Abu-Qare and H. J. Duncan, "Herbicide safeners: uses, limitations, metabolism, and mechanisms of action," in *Chemosphere*, vol. 48, 2009, pp. 965-974.
- [254] A. V. Waichman, E. Eve and N. C. d. S. Nina, "Do farmers understand the information displayed on pesticide product labels? A key question to reduce pesticides exposure and risk of poisoning in the Brazilian Amazon," in *Crop Protection*, vol. 26, 2007, pp. 576-583.

- [255] T. Chapagain and M. N. Raizada, "Impacts of natural disasters on smallholder farmers: gaps and recommendations," *Agriculture & Food Security*, vol. 6, no. 1, p. 39, 10 May 2017.
- [256] J. Burney and V. Ramanathan, "Recent climate and air pollution impacts on Indian agriculture," *The National Academy of Sciences*, vol. 111, no. 46, pp. 16319-16324, 2014.
- [257] "Annual report (2019-2020)," Department of Fertilizers, Ministry of Chemicals and Fertilizers, Government of India, 2019-2020.
- [258] I. Ghanimi, E. Benlahmar, A. Tragha and F. Ghanimi, "A Word Embedding based Approach for Word Sense Disambiguation," *International Journal of Advanced Science and Technology*, vol. 28, no. 16, pp. 144-153, 2019.
- [259] L. Muflikhah and B. Baharudin, "Document clustering using concept space and cosine similarity meaInternational conference on computer technology and development," in *International conference on computer technology and development*, 2009.
- [260] P. Xia, L. Zhang and F. Li, "Learning similarity with cosine similarity ensemble," in *Information Sciences*, 2015, pp. 39-52.
- [261] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *11th National Conference on Artificial Intelligence*, 1993.
- [262] S. B. Huffman, "Learning information extraction patterns from examples," in *International Joint Conference on Artificial Intelligence*, Berlin, Heidelberg, 1995.

- [263] J.-T. Kim and D. I. Moldovan, "Acquisition of linguistic patterns for knowledge-based information extraction," *IEEE transactions on knowledge and data engineering*, vol. 7, no. 5, pp. 713-724, 1995.
- [264] S. Soderland, D. Fisher, J. Aseltine and W. Lehnert, "CRYSTAL: inducing a conceptual dictionary," in *14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, 1995.
- [265] S. Soderland, "Learning information extraction rules for semi-structured and free text," *Journal of Machine learning*, vol. 34, no. 1-3, pp. 233-272, 1999.
- [266] R. Mooney, "Relational learning of pattern-match rules for information extraction," in *16th national conference on artificial intelligence*, 1999.
- [267] A. Serafini, *Apache Solr Beginner's Guide*, 1st ed., Packt Publishing, 2013, p. 324.
- [268] K. A. ABDELOUARIT, B. SBIHI and N. AKNIN, "Spark and Solr: a powerful and ergonomic combination for online search in the Big Data environment (case of the UAE)," in *International work-conference on Time Series*, Spain, Granada, 2017.
- [269] K. H. Cho, S. M. Cha, J.-H. Kang, S. W. Lee, Y. Park, J.-W. Kim and J. H. Kim, "Meteorological effects on the levels of fecal indicator bacteria in an urban stream: a modeling approach," *Water research*, vol. 44, no. 7, pp. 2189-2202, 2010.
- [270] P. Mohan and K. K. Patil, "Deep Learning Based Weighted SOM to Forecast Weather and Crop Prediction for Agriculture Application," *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 4, pp. 167-176, 2018.

- [271] H.-Y. Kung, T.-H. Kuo, C.-H. Chen and P.-Y. Tsai, "Accuracy analysis mechanism for agriculture data using the ensemble neural network method," in *Sustainability*, vol. 8, 2016, p. 735.
- [272] U. Y. Nahm and R. J. Mooney, "Text mining with information extraction," *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pp. 60-67, 2002.
- [273] V. Tourism-Dehradun, "Uttarakhand Map," [Online]. Available: <https://www.uttarakhand-tourism.com/map/uttarakhand-map.php>. [Accessed 20 Nov 2020].
- [274] G. Kaur, "Usage of regular expressions in NLP," *International Journal of Research in Engineering and Technology (IJERT)*, vol. 3, no. 1, 2014.
- [275] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan and H. Zhu, "SystemT: a system for declarative information extraction," *ACM SIGMOD Record*, vol. 37, no. 4, pp. 7-13, 2009.
- [276] A. Fuentes, S. Yoon, S. C. Kim and D. S. Park, "A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition," 9 ed., vol. 17, *Sensors*, 2017, p. 2022.
- [277] J. Ma, K. Du, F. Zheng, L. Zhang, Z. Gong and Z. Sun, "A recognition method for cucumber diseases using leaf symptom images based on deep convolutional neural network," *Computers and electronics in agriculture*, vol. 154, pp. 18-24, 2018.
- [278] F. Raçon, L. Bombrun, B. Keresztes and C. Germain, "Comparison of sift encoded and deep learning features for the classification and detection of Esca disease in bordeaux vineyards," *Remote Sensing*, vol. 11, no. 1, 2019.

- [279] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture* , vol. 145, pp. 311-318, 2018.
- [280] Y. Lu, S. Yi, N. Zeng, Y. Liu and Y. Zhang, "Identification of rice diseases using deep convolutional neural networks," *Neurocomputing* , vol. 267, pp. 378-384, 2017.
- [281] X. Zhang, Y. Qiao, F. Meng, C. Fan and M. Zhang, "Identification of maize leaf diseases using improved deep convolutional neural networks," *IEEE Access* , vol. 6, pp. 30370-30377, 2018.
- [282] H. Yalcin and S. Razavi, "Plant classification using convolutional neural networks," in *Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, 2016.
- [283] P. Barré, B. C. Stöver, K. F. Müller and V. Steinhage, "LeafNet: A computer vision system for automatic plant species identification," *Ecological Informatics* , vol. 40, pp. 50-56, 2017.
- [284] S. Younis, C. Weiland, R. Hoehndorf, S. Dressler, T. Hickler, B. Seeger and M. Schmidt, "Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks," *Botany Letters* , vol. 165, no. 3-4, pp. 377-383, 2018.
- [285] M. M. Ghazi, B. Yanikoglu and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing* , vol. 235, pp. 228-235, 2017.
- [286] L. Zhong, L. Hu and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote sensing of environment* , vol. 221, pp. 430-443, 2019.

- [287] M. Lavreniuk, N. Kussul and A. Novikov, "Deep learning crop classification approach based on coding input satellite data into the unified hyperspace," in *38th International Conference on Electronics and Nanotechnology (ELNANO)*, 2018.
- [288] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with long short-term memory neural networks," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 551, 2017.
- [289] d. S. A. Ferreira, D. M. Freitas, G. G. d. Silva, H. Pistori and M. T. Folhes, "Weed detection in soybean crops using ConvNets," *Computers and Electronics in Agriculture* , vol. 143, pp. 314-324, 2017.
- [290] A. Farooq, J. Hu and X. Jia, "Analysis of spectral bands and spatial resolutions for weed classification via deep convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 183-187, 2018.
- [291] C. Lammie, A. Olsen, T. Carrick and M. R. Azghadi, "Low-Power and High-Speed Deep FPGA Inference Engines for Weed Classification at the Edge," *IEEE Access* , vol. 7, pp. 51171-51184, 2019.
- [292] L. C. Uzal, G. L. Grinblat, R. Namías, M. G. Larese, J. S. Bianchi, E. N. Morandi and P. M. Granitto, "Seed-per-pod estimation for plant breeding using deep learning," *Computers and electronics in agriculture*, vol. 150, pp. 196-204, 2018.

- [293] Y. J. Heo, S. J. Kim, D. Kim, K. Lee and W. K. Chung, "Super-high-purity seed sorter using low-latency image-recognition based on deep learning," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3035-3042, 2018.
- [294] A. Koirala, K. B. Walsh, Z. Wang and C. McCarthy, "Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'," *Precision Agriculture*, vol. 20, no. 6, pp. 1107-1135, 2019.
- [295] X. Liu, S. W. Chen, S. Aditya, N. Sivakumar, S. Dcunha, C. Qu, C. J. Taylor, J. Das and V. Kumar, "Robust fruit counting: Combining deep learning, tracking, and structure from motion," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.
- [296] Y.-D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad and S.-H. Wang, "Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3613-3632, 2019.
- [297] C. Douarre, R. Schielein, C. Frindel, S. Gerth and D. Rousseau, "Transfer learning from synthetic data applied to soil-root segmentation in x-ray tomography images," *Journal of Imaging*, vol. 4, no. 5, p. 65, 2018.
- [298] A. G. Smith, J. Petersen, R. Selvan and C. R. Rasmussen, "Segmentation of roots in soil with U-Net," *Plant Methods*, vol. 16, no. 1, pp. 1-15, 2020.
- [299] A. Rivas, P. Chamoso, A. González-Briones and J. M. Corchado, "Detection of cattle using drones and convolutional neural networks.," *Sensors*, vol. 18, no. 7, p. 2048, 2018.

- [300] D. Tseng, D. Wang, C. Chen, L. Miller, W. Song, J. Viers, S. Vougioukas, S. Carpin, J. A. Ojea and K. Goldberg, "Towards automating precision irrigation: Deep learning to infer local soil moisture conditions from synthetic aerial agricultural images," in *IEEE 14th International Conference on Automation Science and Engineering (CASE)*, 2018.
- [301] D. Tseng, D. Wang, C. Chen, L. Miller, W. Song, J. Viers, S. Vougioukas, S. Carpin, J. A. Ojea and K. Goldberg, "Towards automating precision irrigation: Deep learning to infer local soil moisture conditions from synthetic aerial agricultural images," in *14th International Conference on Automation Science and Engineering (CASE)*, 2018.
- [302] H. S. P. T. Baweja, O. Mirbod and S. Nuske, Stalknet: A deep learning pipeline for high-throughput measurement of plant stalk count and stalk width." In *Field and Service Robotics*, Cham: Springer, 2018, pp. 271-284.
- [303] K. Heinrich, A. Roth, I. Breithaupt, B. Möller and J. Maresch, Yield prognosis for the agrarian management of vineyards using deep learning for object counting, 2019.
- [304] B. Espejo-Garcia, F. J. Lopez-Pellicer, J. Lacasta, R. P. Moreno and F. J. Zarazaga-Soria, "End-to-end sequence labeling via deep learning for automatic extraction of agricultural regulations," *Computers and Electronics*, vol. 162, pp. 106-111, 2019.

2224



Document Information

Analyzed document	Sunil_Final thesis 4.0.pdf (D104462665)
Submitted	5/10/2021 10:59:00 PM
Submitted by	Sunil Kumar
Submitter email	skumar@ddn.upes.ac.in
Similarity	5%
Analysis address	skumar.upes@analysis.orkund.com

Sources included in the report

W	URL: https://www.researchgate.net/publication/271297181_A_Robust_Framework_for_Web_Info ... Fetched: 5/10/2021 11:01:00 PM		1
W	URL: http://dare.nic.in/ Fetched: 5/10/2021 11:01:00 PM		1
W	URL: https://www.researchgate.net/publication/339990379_A_Survey_on_Deep_Learning_for_N ... Fetched: 7/7/2020 9:41:23 AM		1
W	URL: http://static.tongtianta.site/paper_pdf/402be354-80eb-11e9-995f-00163e08bb86.pdf Fetched: 3/1/2021 11:45:28 AM		10
W	URL: https://tel.archives-ouvertes.fr/tel-01943841/document Fetched: 5/10/2021 11:01:00 PM		3
W	URL: https://openrepository.aut.ac.nz/bitstream/handle/10292/12936/YuH.pdf?sequence=3&i ... Fetched: 2/24/2020 2:54:48 PM		6
W	URL: https://arxiv.org/pdf/1812.09449 Fetched: 10/22/2019 1:39:40 PM		6
W	URL: https://arxiv.org/pdf/1803.05667 Fetched: 2/2/2020 1:12:53 PM		2
W	URL: https://www.booktopia.com.au/information-extraction-marie-francine-moens/book/9789 ... Fetched: 5/10/2021 11:01:00 PM		1

2