

Language Intelligence

Expanding Frontiers in Natural Language Processing

Akshi Kumar

Department of Computing
Goldsmiths, University of London
United Kingdom

 **IEEEPress**

WILEY

Contents

List of Figures *xiii*

List of Tables *xv*

About the Author *xvii*

Preface *xix*

Acknowledgements *xxi*

1	Foundations of Natural Language Processing	1
1.1	History of NLP	3
1.2	Approaches to NLP	5
1.3	Understanding NLP through NLU and NLG: Examples and Case Studies	9
1.3.1	Practical Case Studies	9
1.4	NLP Pipeline	10
1.5	NLP's Transformative Impact on Business and Society	12
2	Navigating the Data Landscape for NLP	15
2.1	Types of Data in NLP	16
2.2	Data Acquisition	17
2.2.1	Data Requirements and Considerations for NLP Tasks	19
2.2.2	Data Augmentation	19
2.3	Challenges in NLP Data Acquisition and Management	21
2.4	Data Quality Check in NLP	22
2.4.1	Checking for Biases	23
2.4.2	Addressing Data Imbalances	24
2.4.3	Identifying and Correcting Inconsistencies	24
2.4.4	Ensuring Data Accuracy and Reliability	25

2.4.5	Evaluating Data Quality Regularly	25
2.5	Ethical Considerations in NLP Data Management	25
3	Fundamental Text Processing	31
3.1	Text Cleaning	32
3.1.1	Techniques and Examples of Text Cleaning	32
3.1.2	Challenges in Text Cleaning	33
3.2	Sentence Splitting	34
3.3	Tokenization	36
3.3.1	Types of Tokenization	37
3.3.2	Challenges in Tokenization	42
3.3.3	Evolution of Tokenization	43
3.4	Lemmatization and Stemming	44
3.4.1	Stemming	44
3.4.2	Advantages and Disadvantages of Stemming	44
3.4.3	Lemmatization	45
3.4.4	Advantages and Disadvantages of Lemmatization	45
3.4.5	Stemming versus Lemmatization: When to Use Which?	46
3.5	Stop Word Removal	48
3.6	Part-of-Speech Tagging	49
3.7	Parsing and Syntactic Analysis	50
3.7.1	Parsing: An Overview	51
3.7.2	Contemporary Relevance of Parsing	52
3.7.3	Syntactic Analysis: Delving Deeper	52
3.7.4	Implications of Syntactic Analysis for Modern NLP	54
3.8	Tools and Libraries for Text Processing	56
3.8.1	Python	56
3.8.2	Java	57
3.8.3	C++ (with Python Bindings)	57
4	Linguistic Features in NLP	63
4.1	Levels of Linguistic Analysis	64
4.1.1	Phonetics and Phonology: Sounds of Language	64
4.1.2	Syntax: Sentence Structure and Grammatical Rules	67
4.1.3	Semantics: Meaning of Words and Sentences	67
4.1.4	Pragmatics: Language in Context and Usage	67
4.1.5	Discourse: Structure and Coherence in Texts	71
4.2	Features in NLP	73

4.3	Vector Space Representation in NLP	75
4.3.1	Bag of Words (BoW): Vector Space Models and Frequency Distributions	76
4.3.2	Term Frequency-Inverse Document Frequency: Weighing Terms in Documents	77
4.3.3	N-Gram Language Model	79
4.3.4	Conversion of Categorical Data to Numerical Data in NLP and One-Hot Encoding	80
4.4	Semantic Features in NLP	81
4.4.1	Lexicons	82
4.4.2	Ontologies	84
4.4.3	Knowledge Graphs	85
4.4.4	Word Embeddings	86
4.5	Feature Generation in NLP: Manual versus Automatic Approaches	89
4.5.1	Choosing Between Manual and Automatic Features	91
5	Computational and Cognitive Approaches in Natural Language Processing	95
5.1	Machine Learning for NLP	97
5.1.1	Tasks Accomplished by ML in NLP	97
5.1.2	Selecting the Right Machine Learning Model for NLP Tasks	99
5.1.3	Evolution and Impact of ML in NLP	100
5.2	Memory and Recall Models	100
5.2.1	Working with Long-Term Dependencies	101
5.2.2	Neural Turing Machines (NTMs)	103
5.2.3	Differentiable Neural Computers (DNCs)	104
5.3	Attention Mechanisms	105
5.3.1	Transformer Models	105
5.3.2	Multi-Head and Self-Attention	110
5.4	Human-Like Reasoning	112
5.4.1	Analyzing NLP Systems with Human Reasoning Capabilities	113
5.4.2	Commonsense Knowledge and NLP	114
5.5	Transfer Learning in NLP	114
5.5.1	Power of Pre-Training	116
5.5.2	Models that Generalize	118
5.6	Learning with Minimal Examples	122
5.7	Neuro-Symbolic Approaches	123
5.7.1	Techniques in Neuro-Symbolic NLP	124

6	Fundamental Language Processing Techniques	129
6.1	Topic Modelling and Subject Identification	129
6.1.1	Topic Modelling Methods	130
6.1.2	Use Cases and Applications of Topic Modelling and Subject Identification Across Industries	131
6.1.3	Tools and Software Packages for Topic Modelling and Subject Identification	133
6.1.4	Challenges and Considerations in Topic Modelling and Subject Identification	135
6.2	Named Entity Recognition	136
6.2.1	Core Techniques in Named Entity Recognition	137
6.2.2	Applications of Named Entity Recognition	137
6.2.3	Tools and Best Practices in Named Entity Recognition	142
6.2.4	Challenges in Named Entity Recognition	143
6.3	Text Coherence and Cohesion	144
6.3.1	Text Summarization	146
6.4	Stylistic Analysis	151
6.4.1	Techniques for Stylistic Analysis	152
6.5	Semantic Role Labelling	154
6.5.1	Techniques for Semantic Role Labelling	154
7	Natural Language Processing for Affective, Psychological, and Content Analysis	159
7.1	Sentiment Analysis: <i>Dissecting Text for Opinion Mining</i>	159
7.1.1	Modern Tools and Techniques	161
7.1.2	Applications of Advanced Sentiment Analysis	163
7.1.3	Datasets for NLP-Based Sentiment Analysis	165
7.1.4	Challenges and Innovations in Sentiment Analysis	167
7.2	Emotion Recognition: <i>Beyond Polarity</i>	168
7.2.1	Emotion Recognition versus Sentiment Analysis	168
7.2.2	Complexity of Emotion Recognition	169
7.2.3	Applications of Emotion Recognition	171
7.2.4	Text-only Datasets for NLP-Based Emotion Recognition	174
7.2.5	Challenges in Emotion Recognition and Future Directions	174
7.3	Irony and Sarcasm Detection: <i>Between the Lines</i>	175
7.3.1	Challenges in NLP for Irony and Sarcasm Detection	176
7.3.2	Datasets for NLP-Based Irony and Sarcasm Detection	178
7.4	Humor Identification in Text: <i>Tapping into Textual Tickle</i>	180
7.4.1	Computational Linguistics of Humor	180
7.4.2	Applications of Humor Identification	180

7.4.3	Challenges of Humor Identification	182
7.4.4	Datasets for NLP-Based Humor Detection	183
7.5	Psychometric NLP	184
7.5.1	Personality Trait Estimation	184
7.5.2	Theoretical Frameworks in Personality Psychology and Psychometric Assessment	185
7.5.3	Linguistic Indicators of Personality Traits	187
7.5.4	NLP Methodologies for Personality Estimation	188
7.6	Learning Disabilities Detection	190
7.6.1	Traditional versus NLP-Based Detection Methods	191
7.6.2	Harnessing NLP in the Detection of Learning Disabilities	192
7.6.3	Challenges and Limitations of NLP in Learning Disability Detection	193
7.7	Textual Indicators of Distress: <i>Addressing Depression, Anxiety, and Beyond</i>	194
7.7.1	Understanding Distress in Textual Form	195
7.7.2	Utilizing NLP to Identify Indicators of Mental Distress in Online Communications	196
7.8	Digital Content Moderation using NLP	198
7.8.1	Hate Speech and Cyberbullying Monitoring	199
7.8.2	Misinformation and Fake News Analysis	205
7.8.3	Extremism and Radicalization Identification	212
8	Multilingual Natural Language Processing	223
8.1	Translation and Transliteration	223
8.1.1	Challenges with Translation and Transliteration	226
8.1.2	Evaluation of Translation and Transliteration Quality	227
8.2	Cross-Lingual Models and Embeddings	228
8.2.1	Cross-Lingual Models	229
8.2.2	Cross-Lingual Embeddings: A Shared Semantic Space	230
8.3	Low-Resource Language Processing	235
8.3.1	Challenges in Low-Resource Language Processing	235
8.3.2	Techniques for Low-Resource Language Processing	237
8.3.3	Case Study: Natural Language Processing for Hindi	237
8.4	Cultural Nuance and Idiom Recognition in Natural Language Processing	240
9	Domain-Specific Natural Language Processing	243
9.1	Healthcare Natural Language Processing	243
9.1.1	Predictive Analytics in Healthcare NLP	245

9.1.2	Domain-Specific Models in Healthcare	247
9.1.3	Challenges in Implementing Healthcare NLP	249
9.2	Legal Natural Language Processing	250
9.2.1	Applications of Legal NLP: Automation and Predictive Analytics	251
9.2.2	Domain-Specific Models in Legal	251
9.2.3	Challenges in Legal Natural Language Processing	253
9.3	Finance Natural Language Processing	255
9.3.1	Core Applications of NLP in Finance	256
9.3.2	NLP in Financial Analysis	258
9.3.3	Domain-Specific NLP Models	260
9.3.4	Challenges Within Financial NLP and Future Scope	260
9.4	NLP in Education	262
9.4.1	Facilitating Learning Through NLP	262
9.4.2	NLP Techniques and Their Application in Education	264
9.4.3	Domain-Specific Models in Education	265
9.4.4	Challenges and Considerations in Educational NLP	266

10 Measuring Success in Natural Language Processing Evaluation and Metrics 269

10.1	Intrinsic versus Extrinsic Evaluation Techniques	269
10.1.1	Intrinsic Evaluation Techniques	269
10.2	Extrinsic Evaluation Techniques	270
10.2.1	Combining Intrinsic and Extrinsic Evaluations	270
10.3	Metrics for Text Classification	272
10.3.1	Confusion Matrix	272
10.3.2	Accuracy	273
10.3.3	Precision	273
10.3.4	Recall	274
10.3.5	F1 Score	274
10.3.6	ROC Curve and AUC	275
10.4	Evaluating Machine Translation and Text Summarization	276
10.4.1	Machine Translation Evaluation	276
10.4.2	Text Summarization Evaluation	277
10.5	Metrics for Question-Answering and Conversational AI	278
10.5.1	Evaluation in Ranking and Information Retrieval	279
10.6	Metrics for Text-Based Forecasting and Prediction	280
10.6.1	Root Mean Square Error (RMSE)	281
10.6.2	Mean Absolute Percentage Error (MAPE)	281

Knowledge Checkpoint Answers 285

A Text Representation Techniques: A Unified Overview 305

**B Step-by-Step Guide to NLP Processing on E-Commerce Customer
Feedback** 307

C Harnessing Python Libraries for NLP 311

Further Reading 313

Index 315

List of Figures

- Figure 1.1** NLP Pipeline. 10
- Figure 2.1** Data Acquisition. 18
- Figure 3.1** Sentence-Splitting Example. 35
- Figure 3.2** Exploring Tokenization Techniques in NLP. 37
- Figure 3.3** Syntactic Tree Diagram. 50
- Figure 3.4** N-gram Models in Text Analysis. 51
- Figure 3.5** Syntactic Structure: “*The Cat Sat on the Mat*”. 53
- Figure 3.6** Semantic Relationship Mapping: “*She Likes Chocolate*”. 54
- Figure 4.1** Levels of Linguistic Analysis. 65
- Figure 4.2** Example Knowledge Graph. 85
- Figure 4.3** Understanding Word Embedding. 87
- Figure 4.4** Word Embedding Models. 88
- Figure 5.1** Architecture of LSTM. 101
- Figure 5.2** Architecture of GRU. 102
- Figure 5.3** Architecture of NTM. 103
- Figure 5.4** Architecture of DNC. 104
- Figure 5.5** Architecture of Transformer Model. 106
- Figure 6.1** Schematic Representation of Topic Modelling Process. 130
- Figure 6.2** Example of Named Entity Recognition (NER) Annotations in a Text Passage. 137
- Figure 6.3** Extractive Summarization. 148
- Figure 6.4** Abstractive Summarization. 150
- Figure 7.1** Example of Sentiment Analysis. 160
- Figure 7.2** Overview of Sentiment Analysis Techniques. 161

- Figure 7.3** Computational Approaches for Sarcasm and Irony Detection. 178
- Figure 7.4** Spectrum of Online Information Disorders. 205
- Figure 7.5** Aarhus Model Showing How Violent Extremism Can Be Mitigated. 214
- Figure 8.1** Translation versus Transliteration: What's the Difference? 226
- Figure 8.2** Cross-Lingual XLM-R Transformer Model. 240
- Figure 9.1** Driving Factors Behind NLP in Healthcare. 244
- Figure 9.2** Overview of Legal NLP. 252
- Figure 9.3** Benefits of Using Natural Language Processing in Education. 264
- Figure 10.1** Binary Classification Problem (2×2 matrix). 272
- Figure 10.2** ROC Curve. 275

List of Tables

Table 4.1	Overview of Phonetics and Phonology in Speech Processing: Methodologies, Techniques, Applications, and Examples. 66
Table 4.2	Overview of Morphology in NLP: Methodologies, Techniques, and Applications. 68
Table 4.3	Overview of Syntax in NLP: Methodologies, Techniques, and Real-World Applications. 69
Table 4.4	Overview of Semantics in NLP: Methodologies, Techniques, and Real-World Applications. 70
Table 4.5	Overview of Pragmatics in NLP: Methodologies, Techniques, and Real-World Applications. 72
Table 5.1	Overview of Transformer Models. 108
Table 5.2	Summary of Methodologies and Use Cases for Embedding Commonsense Knowledge in NLP Systems. 115
Table 5.3	Key Benefits of Pre-Training in NLP Models. 119
Table 5.4	Prominent Datasets Used for Pre-Training NLP Models. 120
Table 6.1	A Snapshot of Topic Modelling Applications Across Various Sectors. 132
Table 6.2	Applications of Subject Identification Across Different Sectors. 134
Table 6.3	Comparative Overview of Core Techniques in Named Entity Recognition (NER). 138
Table 6.4	Applications of NER: Diverse Domains and Functional Utilization. 139
Table 6.5	Types of Text Summarization. 147
Table 7.1	Overview of Sentiment Analysis Techniques and Their Description. 162

Table 7.2	Summary of Emotion Recognition Techniques and Their Descriptions. 172
Table 7.3	Overview of Computational Linguistics Techniques for Humor Identification. 181
Table 7.4	NLP Indicators for Personality Trait Assessment. 189
Table 7.5	Applications of NLP Techniques in the Detection of Learning Disabilities. 192
Table 7.6	Comparative Analysis of NLP Techniques for Detecting Linguistic Indicators of Depression, Anxiety, and Broader Mental Health Issues. 197
Table 7.7	NLP Strategies for Fake News and Misinformation Detection. 208
Table 7.8	Advanced NLP Mechanisms in the Identification of Misinformation. 208
Table 7.9	Challenges in Identifying Extremist Content Online. 218
Table 8.1	Comparison of Foundational Translation Techniques. 224
Table 8.2	Methods Used for Creating Cross-Lingual Embeddings. 231
Table 8.3	Summary of Techniques for Low-Resource Language Processing. 238
Table 9.1	Overview of Domain-Specific NLP Models in Finance. 261