Setting Educational Objectives for the Book: Machine Learning Techniques for Text[1]

Nikos Tsourakis[2]

Abstract

The dominant trend in technology is moving towards crafting machines capable of learning from data to perform intelligent tasks. Preparing for this paradigm shift is crucial for individuals and organizations to remain competitive and innovative in the rapidly evolving digital landscape. The book "Machine Learning Techniques for Text" aims to equip learners with the relevant skills, focusing on text data and human language. The book covers topics like analyzing texts, embarking on machine learning, and effectively utilizing cutting-edge deep learning frameworks for different tasks. Incorporating Anderson and Krathwohl's Taxonomy enables the set of clear learning objectives, preparing students for real-world challenges and permitting measurable learning outcomes. In the current work, we provide a summary of the book along with an indicative list of relevant educational goals, learning outcomes, and assessment tasks.

Keywords: Anderson & Krathwohl's Taxonomy, Educational Goals, Machine Learning, NLP

1. Introduction

Despite recent advancements across various scientific fields, the language phenomenon is still shrouded in mystery (Binder & Smith, 2013). Astonishingly, it is Homo sapiens that have pioneered this complex system for information exchange, resulting in some of humanity's most remarkable achievements. While oral and gestural forms of language served as primary drivers over millennia, the written version played a pivotal role in disseminating knowledge globally. Today, text data is prolifically generated through activities such as social network interactions, scientific publications, and multimedia transcriptions, among others. There is an urgent need to equip professionals with the appropriate skill set to analyze this type of data effectively and make informed decisions. The "Machine Learning for Text" book (Tsourakis, 2022) assists learners in this endeavor.

In the new machine age era, delegating the effort of analyzing human language to a computer is an attractive option simply because it can process more data in a fraction of the time. However, the execution of this task is not merely quantitative. We can also teach a machine to perform it efficiently. The focus of the book is to present techniques in practical scenarios that allow a machine to extract meaningful insights from text data and act intelligently to solve a particular problem. It consists of ten chapters, each focusing on one specific case study using real-world datasets. For that reason, the book is solution-oriented, and it's accompanied by Python code in the form of Jupyter notebooks to help learners obtain first-hand experience.

Following an educational objectives framework in a textbook holds paramount importance for several reasons (Wiggins & McTighe, 2011). Firstly, it establishes a clear and organized path for both educators and learners, providing well-defined goals and outcomes that facilitate the learning journey. This structured approach ensures that the content is systematically organized,

[1] Tsourakis, N. (2022). *Machine learning techniques for text: Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation*. Packt Publishing Ltd.

[2] *Nikos Tsourakis is a Research Associate at the University of Geneva and Adjunct Professor at the International Institute in Geneva.*

making it easier for students to grasp complex concepts progressively. Moreover, it assists educators in crafting effective teaching strategies and assessments that align precisely with the desired learning outcomes, enhancing the efficiency of the learning process. Finally, it promotes consistency and coherence in education, ensuring all instructional material is aligned with recognized standards and objectives. This paper uses Anderson & Krathwohl's taxonomy of learning objectives (Anderson & Krathwohl, 2001) as a guide, offering insights into the educational goals and cognitive complexity of each chapter. We begin in the next section with a summary of the book's content before delving into its educational objectives.

2. Book Summary

The book utilizes a case study approach that enables active engagement instead of passively absorbing information. The problem statement is set in each chapter from the beginning, ensuring learners understand the challenge. Even when the discussion momentarily deviates from the primary objective, for instance, presenting some fundamental concept, readers quickly refocus on the problem under study. A recurring pattern is that we first try to gain some intuition on the data and then implement and contrast different solutions. In the following sections, we provide a high-level summary for each chapter.

2.1 Chapter 1: Introducing Machine Learning for Text

This introductory chapter sets the stage for the book. It highlights human language's distinct characteristics alongside its challenges, notably its ambiguity. It then covers key concepts in artificial intelligence, machine learning, deep learning, natural language processing, and big data, emphasizing the book's focus on the intersection of these fields. The chapter presents the main techniques for machine learning for text, the relevant terminology, and the challenges while using text corpora. Readers also familiarize themselves with the basic concepts behind text processing. There is a discussion on the notion of what a machine can learn, along with the taxonomy of different types of learning. The chapter concludes by stating the importance of visualization and evaluation techniques.

2.2 Chapter 2: Detecting spam emails

Email, a pervasive Internet service for message exchange, faces the ongoing challenge of distinguishing and preventing unsolicited messages. Spam detectors play a crucial role in this task, striving to catch spam without impeding legitimate communication. The first problem of the book outlines a step-by-step process for crafting and assessing a typical spam detector. The discussion starts with the constraints of conventional programming for this task, followed by the fundamental text representation and preprocessing methods. Subsequently, an open-source dataset is employed to build and evaluate two classifiers using supervised learning and the Naïve Bayes and SVM algorithms. Finally, the performance of the two classifiers is contrasted using standard metrics.

2.3 Chapter 3: Classifying Topics of Newsgroup Posts

Businesses deal with various forms of unstructured text data daily, such as news posts, support tickets, and customer reviews, which, if not efficiently analyzed, can result in missed opportunities and customer dissatisfaction. Automated systems that handle large data volumes can offer a scalable solution for classifying this data. The time factor is also crucial, as a company can monitor in real time the different issues and react accordingly.

This chapter categorizes text documents into predefined topics, utilizing supervised and unsupervised machine learning techniques. It builds upon exploratory data analysis, enhancing visualizations through dimensionality reduction techniques like PCA and LDA. It also introduces word embeddings, an advanced word representation technique with unique properties. Finally, the analysis is based on the KNN and Random Forests algorithms.

2.4 Chapter 4: Extracting Sentiments from Product Reviews

Understanding the emotional context conveyed by a series of words holds significant value in analyzing survey responses, customer feedback, and product reviews. The rise of social networks has expanded opportunities for instantaneous expression of opinions on various subjects, leading companies, academia, and government entities to seek insights from public sentiment. By gauging the sentiment behind these expressions, organizations can make informed decisions, enhance customer satisfaction, tailor marketing strategies, and detect emerging trends, ultimately improving their products, services, and overall engagement with their audience.

This chapter addresses another common challenge in natural language processing: extracting sentiment from a piece of text. It incorporates an open-source dataset featuring customer reviews from Amazon. The exploratory data analysis phase is extended, and dimensionality reduction is used for feature selection. The focus is on deep learning techniques, and to facilitate their explanation, the chapter discusses linear and logistic regression. Concepts related to minimizing loss and gradient descent constitute part of this discussion. Readers learn how to construct, train, and test a deep neural network model in Keras for sentiment analysis.

2.5 Chapter 5: Recommending Music Titles

In today's vast digital landscape, customers often face overwhelming choices. Understanding their habits and preferences is highly valuable in fulfilling their needs and enabling companies to promote new products and services. However, given the predominantly online nature of most services, direct access to customers can be challenging. Their preferences, such as favorite music genres or authors, can be extracted from their purchase history and product reviews. Automatic systems can then use this data to recommend content or products, enhancing the user experience and extending their interaction with the service.

This chapter focuses on creating recommender systems for music titles by utilizing customer reviews on Amazon. The discussion begins with exploratory data analysis and rigorous data cleaning to address potential sample biases. The content-based and collaborative-filtering recommender types are introduced, contrasting their strengths and weaknesses. Using t-SNE for dimensionality reduction and RBM, we implement a system that suggests music titles. Finally, the topic of hyperparameter tuning is revisited, using a grid search to identify the optimal combination of the hyperparameters.

2.6. Chapter 6: Teaching Machines to Translate

Language barriers can hinder the spread of information and ideas in our culturally diverse world, and it's a significant challenge in effective human communication. The concept of a universal translator, often featured in science fiction literature, movies, and TV shows like Star Trek, envisions a device capable of seamlessly translating alien languages into the user's native tongue. The ongoing efforts to train machines to function as proficient translators have yielded remarkable progress in recent years despite the inherent challenges posed by language ambiguity and flexibility.

This chapter introduces various machine translation techniques while enriching readers' knowledge of standard NLP methods. It provides an opportunity to compare top-down and bottom-up approaches in system design. Rule-based and statistical machine translation constitute an excellent way to introduce fundamental concepts on the topic. Readers become familiar with typical NLP methods such as POS tagging, parse trees, and NER. The discussion on deep learning models becomes more challenging as the focus is now on sequence-to-sequence learning. An extended section describes the famous encoder/decoder architectures using RNN and LSTM in detail. A seq2seq model is put into action to create an English-to-French translator, and the chapter ends with the typical evaluation of machine translation systems based on the BLEU score.

2.7. Chapter 7: Summarizing Wikipedia Articles

The rise of Web 2.0 transformed individuals from passive content consumers into active content creators, leading to an immense volume of online text data. However, the abundance of content has made it increasingly challenging to discover and consume valuable information efficiently. Consequently, there is a pressing need for automated systems that can accurately summarize longer texts and filter out irrelevant content. This task becomes even more complex when considering various constraints, such as the target audience and the type of content being summarized.

This chapter explores methods for creating effective summarization systems using web-derived data and covers techniques for automatically accessing and parsing web resources like Wikipedia. In addition to traditional text summarization methods, the chapter delves into the Transformer's cutting-edge architecture that offers outstanding performance across various real-world applications. This advanced architecture builds upon the previously discussed seq2seq models and incorporates key concepts highlighted throughout the book. The chapter concludes by discussing the ROUGE score for evaluating the performance of relevant systems.

2.8 Chapter 8: Detecting Hateful and Offensive Language

The proliferation of hate speech and the dissemination of fake news are significant negative consequences of the growing influence of social networks. Often shielded by anonymity, users on these platforms may find it easier to express hateful comments or share false information, as they are not immediately accountable for their actions. In order to combat these issues, major social networks invest substantial resources in machine learning algorithms to detect and remove inappropriate language.

The focus of this chapter revolves around efficiently utilizing and customizing pre-existing third-party models to identify hate speech and offensive language on Twitter. An open-source dataset containing inappropriate tweets is employed to construct a BERT language model for classification. The role of the validation set to fine-tune the model's hyperparameters and the strategies for dealing with imbalanced data are also examined. The classification tasks are based on boosting algorithms and CNN.

2.9 Chapter 9: Generating Text in Chatbots

State-of-the-art artificial intelligence applications can produce humanlike content, spanning written essays, music, and artwork. These applications hold great potential in the journey towards achieving artificial general intelligence, where systems can comprehend and perform any intellectual task that humans can. The generative aspects of natural language processing emphasize conversational agents, commonly known as chatbots. These chatbots find widespread application in various domains, including customer service for large organizations, language acquisition in education, and data collection in research. They can be engaged through various communication channels, such as speech, text, or even facial expressions and gestures.

This chapter covers a wide range of NLP techniques, starting with basic regular expressions and progressing towards more advanced approaches based on deep learning. We provide insights into crafting language models from scratch or fine-tuning pre-existing ones. Readers also gain familiarity with reinforcement learning techniques and how to develop graphical user interfaces for interacting with the chatbot. Additionally, we introduce perplexity as an evaluation metric and delve into TensorBoard, a tool that illuminates the inner workings of deep neural networks.

2.10 Chapter 10: Clustering Speech-to-Text Transcriptions
When working with real-world problems, it's frequently the case that their data come unlabeled. Take, for instance, grouping news topics, analyzing customer call transcriptions, evaluating user tweets, and more. In all these preceding cases, businesses aim to gain valuable insights by uncovering pertinent information within this data. Manually labeling each sample is often impractical due to the time and cost involved. Consequently, it becomes essential to incorporate techniques capable of handling such unlabeled datasets.

This chapter focuses on unsupervised learning to cluster similar data points into cohesive categories. Hard and soft clustering techniques are presented, offering an understanding of their inner workings and practical implementation. The hard clustering methods introduced are hierarchical clustering, k-means, and DBSCAN. There is also a relevant discussion on choosing the optimal number of clusters. Another distinctive aspect is the creation of the text corpus using speech-to-text technology and assessing performance using WER. The chapter concludes by applying soft clustering and LDA to identify the topics in a dataset.

3. The Taxonomy of Educational Objectives

Bloom's Taxonomy of Educational Objectives (Bloom, 1956) is an established framework for organizing and categorizing educational goals and learning outcomes. Since its introduction, this taxonomy has been pivotal in shaping curriculum development, instructional design, and assessment strategies in education. It provides a structured way to understand and address the diverse cognitive processes involved in learning. The taxonomy consists of six hierarchical levels, each representing a different level of cognitive complexity, from lower-order thinking skills to higher-order ones (Figure 1 - left).

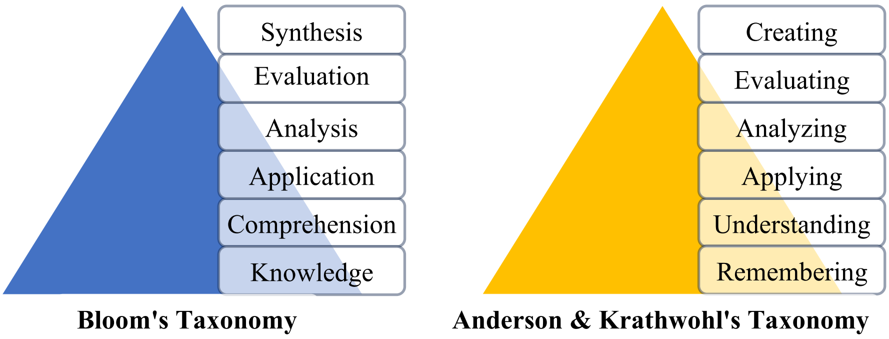

Figure 1: Bloom's versus Anderson & Krathwohl's Taxonomy

While the original taxonomy by Bloom is renowned for its foundational role in education, a revised version of this taxonomy was published in the early 2000s (Anderson & Krathwohl, 2001). The older version primarily emphasizes the cognitive domain, focusing on knowledge acquisition and cognitive skills. On the other hand, the revised version extends the taxonomy to include the affective and psychomotor domains, encompassing emotional and physical aspects of learning. It also offers a more contemporary perspective, emphasizing the importance of metacognition in today's diverse educational landscape and recognizing that learning is a dynamic process involving lower- and higher-order thinking skills. Overall, Anderson and Krathwohl's work provides a more comprehensive and adaptable framework for educators to address a broader range of learning objectives (Figure 1 - right).

The six educational objectives of Anderson and Krathwohl's taxonomy are:

1. **Remembering**: Recognize and recall facts, concepts, and information. Learners demonstrate their understanding by retrieving previously learned material from memory.
2. **Understanding**: Comprehend and interpret information. Learners can explain concepts and ideas in their own words, demonstrating a deeper understanding of the subject matter.
3. **Applying**: Apply knowledge and understanding in new situations or to solve problems. Learners can transfer their learning to practical scenarios and demonstrate the ability to use concepts and principles effectively.
4. **Analyzing**: Breakdown complex information into its components and understand the relationships between them. Learners can differentiate between different elements and assess their significance.
5. **Evaluating**: Make judgments and assess the value or quality of information, arguments, or solutions. Learners can critique, justify, and provide evidence for their evaluations.
6. **Creating**: Generate new ideas, solutions, or products based on current knowledge and skills. Learners can design, construct, and invent, demonstrating a high level of originality and innovation.

The next section provides an overview of the taxonomy used to categorize and classify the learning objectives for one of the book's chapters, along with relevant assessment tasks.

4. Learning Objectives and Assessment Tasks for Chapter 2

4.1 Anderson & Krathwohl's Taxonomy

This section presents the learning outcomes for the book's second chapter, thoughtfully crafted to align with Anderson and Krathwohl's taxonomy principles. As already explained, this educational framework provides a structured approach to learning and assessment, allowing the set of clear objectives and serving as a valuable roadmap to the chapter's content. The educational objectives and learning outcomes per cognitive level are shown in Table 1.

Remembering
Define the typical steps for training a classifier.
Identify when machine learning is suitable for a given task.
Describe the basic steps of the exploratory data analysis.
List the basic methods for word representation.
Recall the basic data preprocessing steps.
Understanding
Summarize the role and purpose of training and test sets.
Understand the limitations of traditional programming for solving complex problems.
Explain why specific metrics are not appropriate in the context of an ML problem.
Demonstrate how to extract pertinent features for a given problem.
Applying
Construct a basic data preprocessing pipeline.
Apply the basic techniques for text representation and preprocessing.

Implement classifiers using an open-source dataset.

Analyzing

Analyze the tradeoffs between training accuracy and generalization capacity of a model.

Examine the performance metrics used to evaluate spam detectors.

Contrast the strengths and weaknesses of the implemented classifiers.

Evaluating

Assess the effectiveness of machine learning models based on standard metrics.

Critically evaluate the tradeoffs associated with different algorithms and techniques used in machine learning.

Creating

Create and design new strategies for improving spam detection beyond what is discussed in the chapter.

Develop a comprehensive understanding of the entire spam detection pipeline and potentially modify it to suit different scenarios.

Table 1: Anderson & Krathwohl's taxonomy of the learning objectives for Chapter 2.
Later in the paper, we will also provide an indicative list of the other chapters. Before that, however, we present different tasks that assess these objectives.

4.2 Assessment tasks
Assessment has been signified early on as an integral part of the educational process, helping to evaluate learning outcomes, improve teaching practices, and ensure accountability and quality in education (Tyler, 1949). In this respect, we propose in Table 2 sample tasks for each cognitive level that can assist in the assessment.

Remembering

Students are expected to demonstrate their ability to retrieve information, identify key terms, and remember essential details from the learning materials. This level involves tasks like creating flashcards, completing multiple-choice quizzes, listing fundamental concepts, matching terms to their definitions, and recalling specific facts without reference.

Understanding

Learners are expected to explain ideas in their own words, create concept maps to visualize connections between key ideas, summarize content with a focus on cause-and-effect relationships, and compare and contrast different concepts.

Applying

Assessment tasks involve solving problems related to learned concepts, demonstrating the application of tools or techniques in Python, and offering step-by-step instructions for applying the learned principles in a different way.

Analyzing

Students at this level are expected to analyze case studies, identify key issues and solutions, deconstruct intricate ideas to explain relationships, compare different approaches or theories, and identify patterns, trends, or recurring themes within the content.

Evaluating

Learners are expected to critically assess the validity and reliability of data, construct persuasive arguments for or against specific concepts or theories, develop criteria for evaluating the effectiveness of solutions, compare and contrast ideas from various sources, and determine the significance of content within a broader context.
Creating
Students are challenged to design original projects, create presentations to synthesize key concepts, develop new models, theories, or frameworks, write essays proposing innovative solutions, and generate interactive learning materials to teach others.

Table 2: Proposed assessment tasks for Anderson & Krathwohl's taxonomy.

The next section concludes the discussion with more learning objectives for the entire book.

5. Indicative Learning Objectives for all chapters

Excluding Chapter 2, we provide in Table 3 an indicative list of learning outcomes for each chapter and taxonomy level. These can assist as a reference to other tutors who wish to use the book in their teaching activities.

Remembering	
ch1	Remember the basic taxonomy of machine learning algorithms.
ch3	State the purpose of exploratory data analysis.
ch4	Recall the concept of overfitting and how it can be addressed using regularization.
ch5	Memorize the concept of hyperparameter tuning in machine learning.
ch6	Name the distinction between top-down and bottom-up techniques in machine learning and natural language processing.
ch7	List the typical interchange formats for organizing data.
ch8	Define the benefits of transfer learning and fine-tuning of large language models.
ch9	Provide the definitions for autoregressive and auto-encoding models.
ch10	Tell the differences between hard and soft clustering methods.
Understanding	
ch1	Understand how machines learn versus traditional programming.
ch3	Make sense of word embeddings and their relevance in text representation.
ch4	Explain how optimization techniques can minimize the loss function.
ch5	Comprehend the importance of data cleaning in preparing datasets for analysis.
ch6	Describe the components and functioning of the transformer model.
ch7	Clarify the challenges associated with a single context vector in the encoder-decoder model and the role of attention mechanisms.
ch8	Outline the concept of ensemble learning and the rationale behind combining multiple machine learning models.
ch9	Interpret the relevance and appropriateness of perplexity as an evaluation metric for language models.
ch10	Illustrate the ways to visualize hierarchical relationships between objects in a dataset.
Applying	
ch1	Implement the suitable machine learning method for a given problem.

ch3 Utilize dimensionality reduction methods for improved data analysis and visualization.

ch4 Apply machine learning algorithms to predict continuous values.

ch5 Perform hyperparameter tuning efficiently using grid search.

ch6 Construct standard NLP processes for tagging and information extraction.

ch7 Develop web crawling and data scraping techniques for gathering data from online resources.

ch8 Make use of strategies for dealing with imbalanced data sets.

ch9 Experiment with the concept of fine-tuning language models using reinforcement learning.

ch10 Identify the optimal number of clusters using the appropriate methods.

Analyzing

ch1 Scrutinize possible model errors based on their severity on the problem under study.

ch3 Examine the gender biases often found in word embeddings.

ch4 Investigate the impact of regularization on model performance and overfitting prevention.

ch5 Predict the potential consequences and implications of misinterpreting correlation as causation.

ch6 Compare the appropriateness of different grammar formalisms for parsing and generation tasks.

ch7 Inspect how entities relate to each other using knowledge graphs.

ch8 Contrast training versus validation loss to apply early stopping.

ch9 Break down the steps to avoid an exploding gradient.

ch10 Analyze clustering results using domain knowledge.

Evaluating

ch1 Defend the use of computer- and human-centered metrics.

ch3 Compare the performance of the created models with the baseline one.

ch4 Judge the impact of dropout on model generalization.

ch5 Determine the performance of machine learning models on generalizing to unseen data using cross-validation.

ch6 Assess the quality of sequence-to-sequence model output using the BLEU score.

ch7 Estimate how model performance is affected by adapting the learning rate.

ch8 Rate the addition of pooling layers in convolutional neural networks.

ch9 Decide where to use an intrinsic or an extrinsic evaluation approach.

ch10 Evaluate the quality of topics generated by LDA in terms of relevance and coherence.

Creating

ch1 Produce innovative visualizations to present complex information with clarity and aesthetics.

ch3 Propose new ways to create simpler models for the same classification task.

ch4 Design custom solutions for addressing overfitting challenges in machine learning.

ch5	Create efficient strategies for hyperparameter tuning in different machine-learning applications.
ch6	Develop custom machine learning applications using knowledge-driven or data-driven approaches.
ch7	Invent a customized transformer-based architecture for sequence-to-sequence learning.
ch8	Elaborate novel chained models to advance ensemble learning techniques.
ch9	Construct custom language models using fine-tuning strategies for specific text generation tasks.
ch10	Integrate LDA with other NLP techniques to create innovative solutions for topic modeling in various domains.

Table 3: Indicative Anderson & Krathwohl's taxonomy of the learning objectives for all chapters.

6. Conclusions

Building machines that learn from observations is becoming the dominant paradigm due to the ever-increasing amount of data that cannot be processed with traditional methods. These resources pose fewer challenges in access and storage, which have become relatively inexpensive. Conversely, we need techniques to extract, visualize, and analyze text data to leverage this massive amount of unstructured information.

By the book's conclusion, learners acquire a versatile skill set for text preprocessing, representation, dimensionality reduction, machine learning, language modeling, visualization, and evaluation in Python, equipping them to tackle similar problems effectively.

We juxtaposed the contents of the book with specific levels of the Anderson & Krathwohl taxonomy, showcasing the diversity of cognitive processes and learning outcomes addressed throughout the book. Finally, we provided possible assessment tasks for each level in the taxonomy to assist other tutors in their teaching efforts.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson.

Binder, P.-M & Smith, K. (2013). *The Language Phenomenon: Human Communication from Milliseconds to Millennia*. Springer.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The Classification of Educational Goals*. Longmans.

Tsourakis, N. (2022). *Machine learning techniques for text: Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation*. Packt Publishing Ltd.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

Wiggins, G. P., & McTighe, J. (2011). *The understanding by design guide to creating high-quality units.* ASCD.