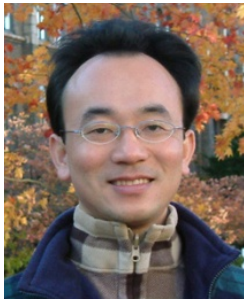
人工智慧文本分析
(Artificial Intelligence for Text Analytics)
人工智慧文本分析課程介紹
(Course Orientation on Artificial Intelligence for Text Analytics)

1082AITA01
MBA, IMTKU (M2455) (8410) (Spring 2020)
Wed 8, 9 (15:10-17:00) (B605)



Min-Yuh Day
戴敏育
Associate Professor
副教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

http://mail. tku.edu.tw/myday/

2020-03-04

人工智慧
文本分析
(Artificial Intelligence for Text Analytics)

淡江大學108學年度第2學期
課程教學計畫表
Spring 2020 (2020.03 - 2020.06)

- 課程名稱：**人工智慧文本分析**
  (Artificial Intelligence for Text Analytics)
- 授課教師：戴敏育 (Min-Yuh Day)
- 開課系級：資管所碩士班 (TLMXM1A)
- 開課資料：選修 單學期 2 學分 (2 Credits, Elective)
- 上課時間：週三 8, 9 (15:10-17:00)
- 上課教室： B605 (淡江大學淡水校園)

淡江大學

資訊管理學系

資訊創新

管理思維

IMTKU

淡江大學資訊管理系(所)教育目標

Tamkang University 淡江大學

- 致力於資訊科技與經營管理知識之科際整合研究發展，為國家與社會培育兼具資訊技術能力與現代管理知識的中高階人才。

淡江大學資訊管理系(所)核心能力

Tamkang University 淡江大學

- A. 現代管理知識應用。(10%)
- B. 邏輯思考。(10%)
- C. 關鍵分析。(10%)
- **D. 結合資訊技術與管理。(30%)**
- E. 研究與創新。(10%)
- **F. 資料分析與應用。(20%)**
- G. 資通安全管理。
- H. 言辭與文字表達。(10%)

課程簡介

- 本課程介紹人工智慧文本分析基本概念與研究議題。
- 課程內容包括
  - 文本分析的基礎：自然語言處理 (NLP)、
  - Python自然語言處理、
  - 處理和理解文本、
  - 文本表達特徵工 程、
  - 文本分類、
  - 文本摘要和主題模型、
  - 文本相似度和分群、
  - 語意分析與命名實體識別 (NER)、
  - 情感分析、
  - 深度學習和通用句子嵌入模型、
  - 問答系統與對話系 統、
  - 和文字探勘個案研究。

Course Introduction

- This course introduces the fundamental concepts and research issues of artificial intelligence for text analytics.
- Topics include
  - Foundations of Text Analytics: Natural Language Processing (NLP),
  - Python for NLP,
  - Processing and Understanding Text,
  - Feature Engineering for Text Representation,
  - Text Classification,
  - Text Summarization and Topic Models,
  - Text Similarity and Clustering,
  - Semantic Analysis and Named Entity Recognition,
  - Sentiment Analysis,
  - The Promise of Deep Learning and Universal Sentence-Embedding Models,
  - Question Answering and Dialogue Systems,
  - and Case Study on AI Text Analytics.

課程目標
(Objective)

- 瞭解及應用人工智慧文本分析基本概念與研究議題。
  Understand and apply the fundamental concepts and research issues of artificial intelligence for text analytics.

- 進行人工智慧文本分析相關之資訊管理研究。
  Conduct information systems research in the context of artificial intelligence for text analytics.

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2020/03/04	人工智慧文本分析課程介紹 (Course Orientation on Artificial Intelligence for Text Analytics)
2	2020/03/11	文本分析的基礎：自然語言處理 (Foundations of Text Analytics: Natural Language Processing; NLP)
3	2020/03/18	Python自然語言處理 (Python for Natural Language Processing)
4	2020/03/25	處理和理解文本 (Processing and Understanding Text)
5	2020/04/01	文本表達特徵工程 (Feature Engineering for Text Representation)
6	2020/04/08	人工智慧文本分析個案研究 I (Case Study on Artificial Intelligence for Text Analytics I)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2020/04/15	文本分類 (Text Classification)
8	2020/04/22	文本摘要和主題模型 (Text Summarization and Topic Models)
9	2020/04/29	期中報告 (Midterm Project Report)
10	2020/05/06	文本相似度和分群 (Text Similarity and Clustering)
11	2020/05/13	語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)
12	2020/05/20	情感分析 (Sentiment Analysis)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2020/05/27	人工智慧文本分析個案研究 II (Case Study on Artificial Intelligence for Text Analytics II)
14	2020/06/03	深度學習和通用句子嵌入模型 (Deep Learning and Universal Sentence-Embedding Models)
15	2020/06/10	問答系統與對話系統 (Question Answering and Dialogue Systems)
16	2020/06/17	期末報告 I (Final Project Presentation I)
17	2020/06/24	期末報告 II (Final Project Presentation II)
18	2020/07/01	教師彈性補充教學

教學方法與評量方法

- 教學方法

  –講述、討論、
  發表、實作

- 評量方法

  –討論、實作、報告

教材課本

- 教材課本
  - 講義 (Slides)
  - 人工智慧文本分析相關個案與論文 (Cases and Papers related to AI for Text Analytics)

參考書籍 (References)

1. Dipanjan Sarkar (2019),
   Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.
2. Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),
   Applied Text Analysis with Python:
   Enabling Language-Aware Data Products with Machine Learning, O'Reilly.
3. Charu C. Aggarwal (2018),
   Machine Learning for Text, Springer.
4. Gabe Ignatow and Rada F. Mihalcea (2017),
   An Introduction to Text Mining: Research Design, Data Collection, and Analysis, SAGE Publications.

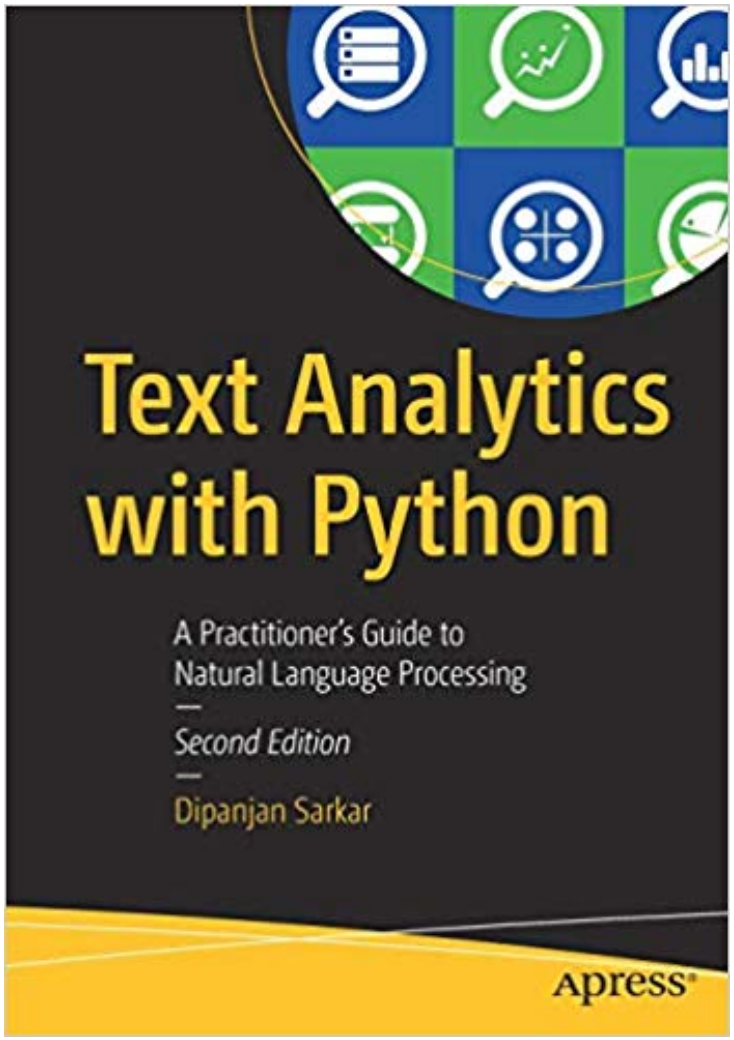
作業與學期成績計算方式

- 作業篇數
  - 3篇
- 學期成績計算方式
  - ☑期中評量：30 %
  - ☑期末評量：30 %
  - ☑其他（課堂參與及報告討論表現）：40 %

Dipanjan Sarkar (2019),

Text Analytics with Python:

A Practitioner's Guide to Natural Language Processing, Second Edition. APress.



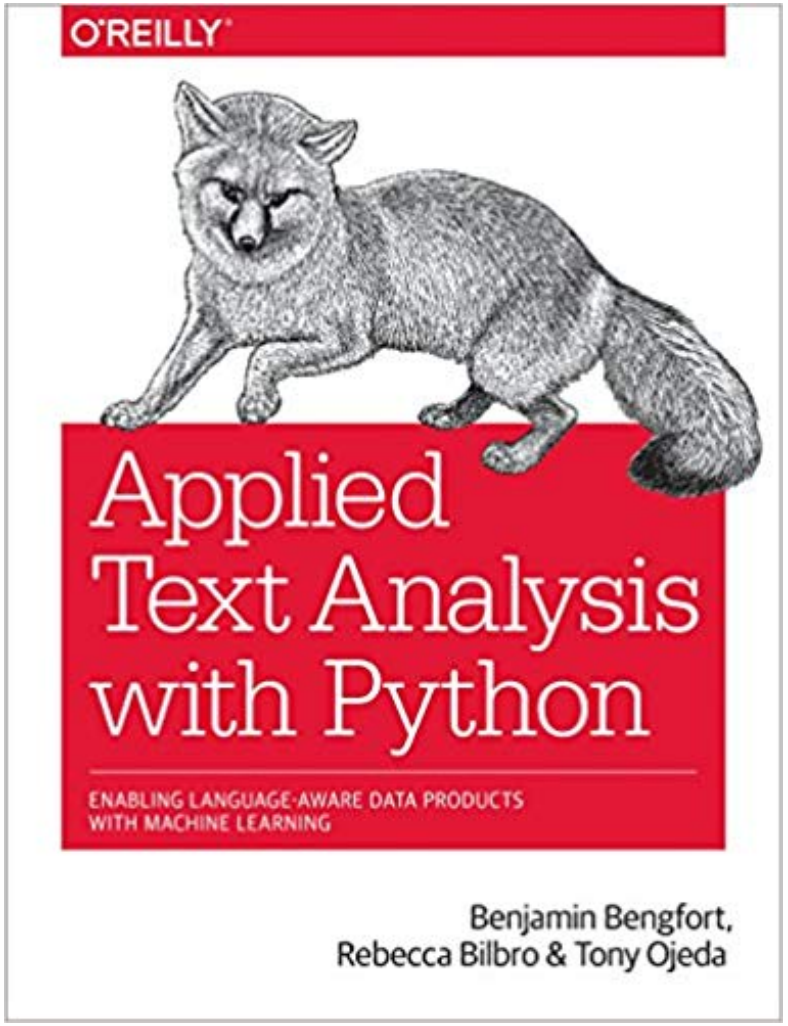
Source: https://www.amazon.com/Text-Analytics-Python-Practitioners-Processing/dp/1484243536

Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018),

Applied Text Analysis with Python:

Enabling Language-Aware Data Products with Machine Learning,

O'Reilly.

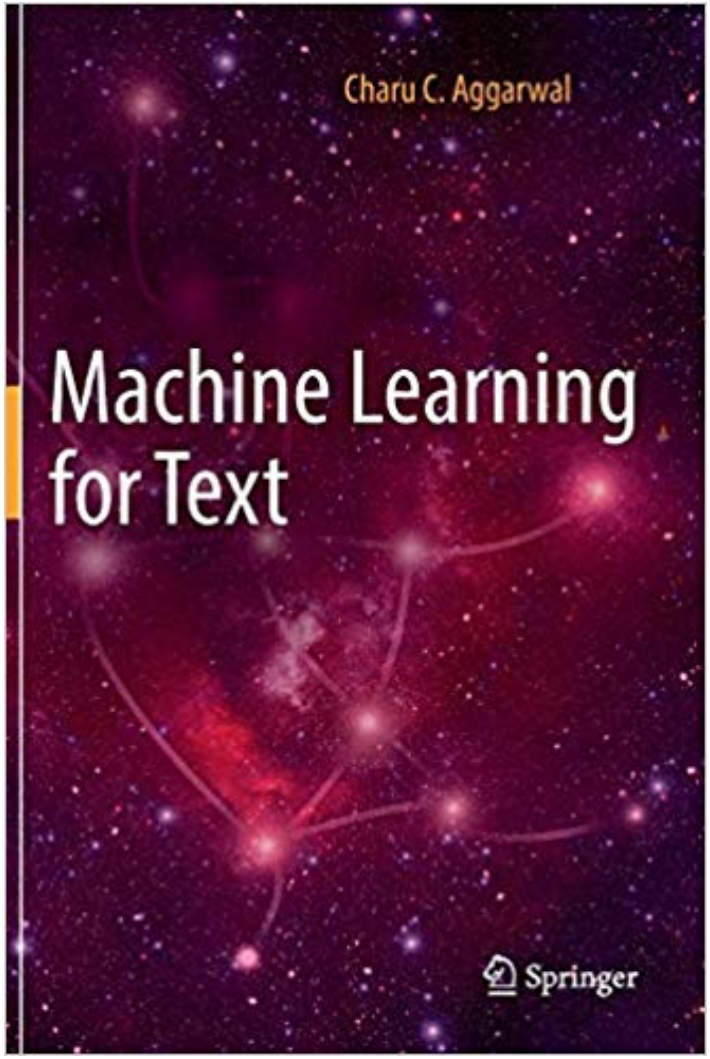


Source: https://www.amazon.com/Applied-Text-Analysis-Python-Language-Aware/dp/1491963042

Charu C. Aggarwal (2018),

Machine Learning for Text,

Springer



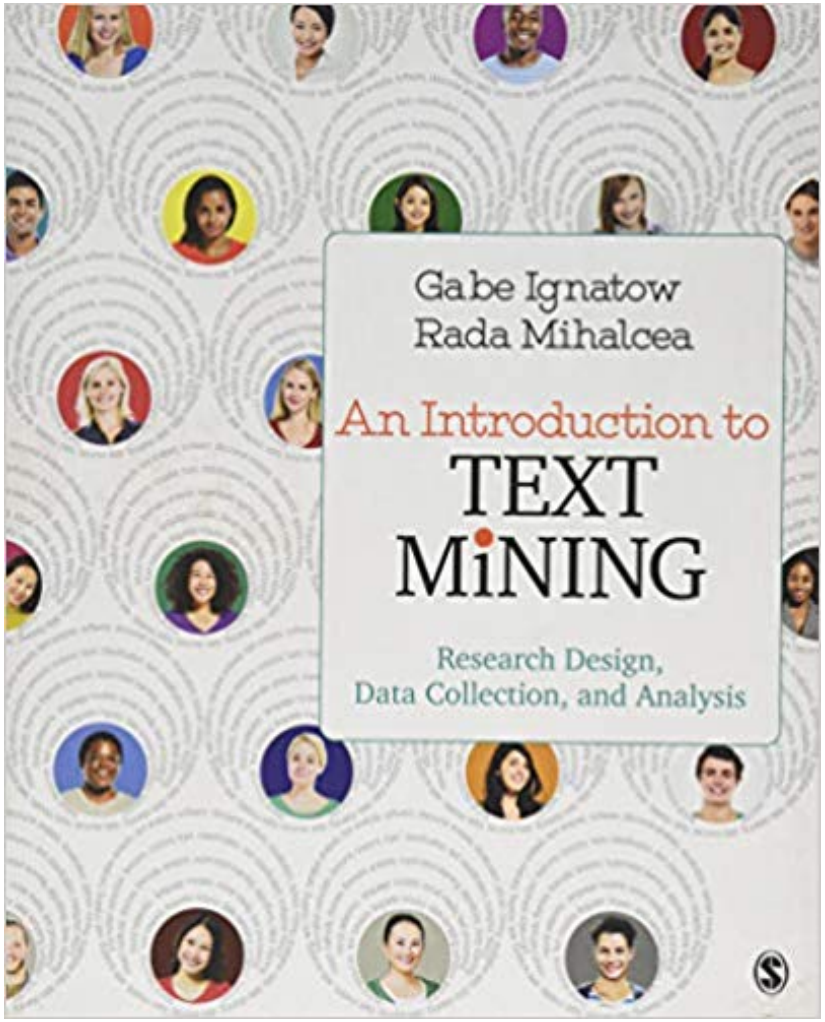
Source: https://www.amazon.com/Machine-Learning-Text-Charu-Aggarwal/dp/3319735306

Gabe Ignatow and Rada F. Mihalcea (2017),

An Introduction to Text Mining:

Research Design, Data Collection, and Analysis,

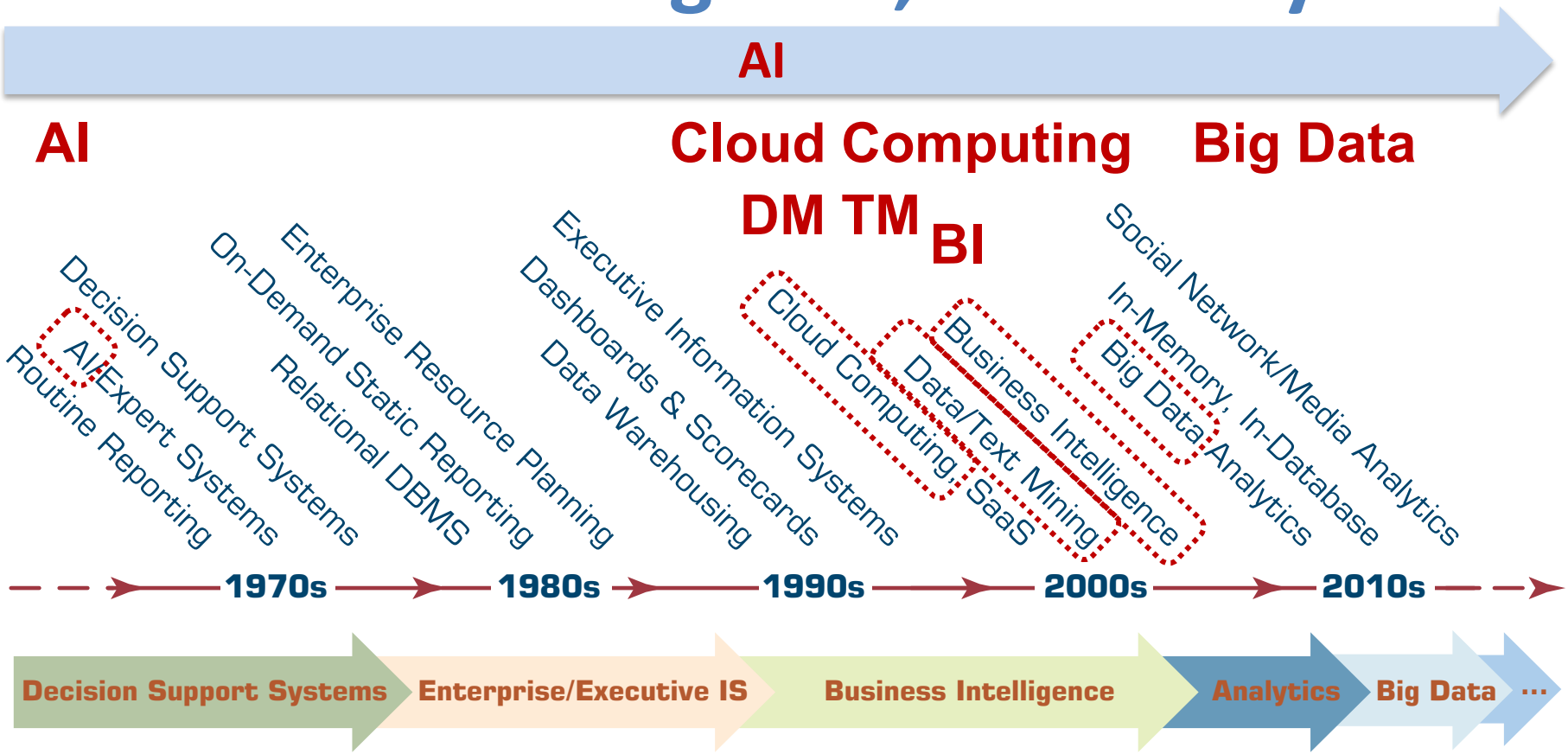
SAGE Publications.



Source: https://www.amazon.com/Introduction-Text-Mining-Research-Collection/dp/1506337007

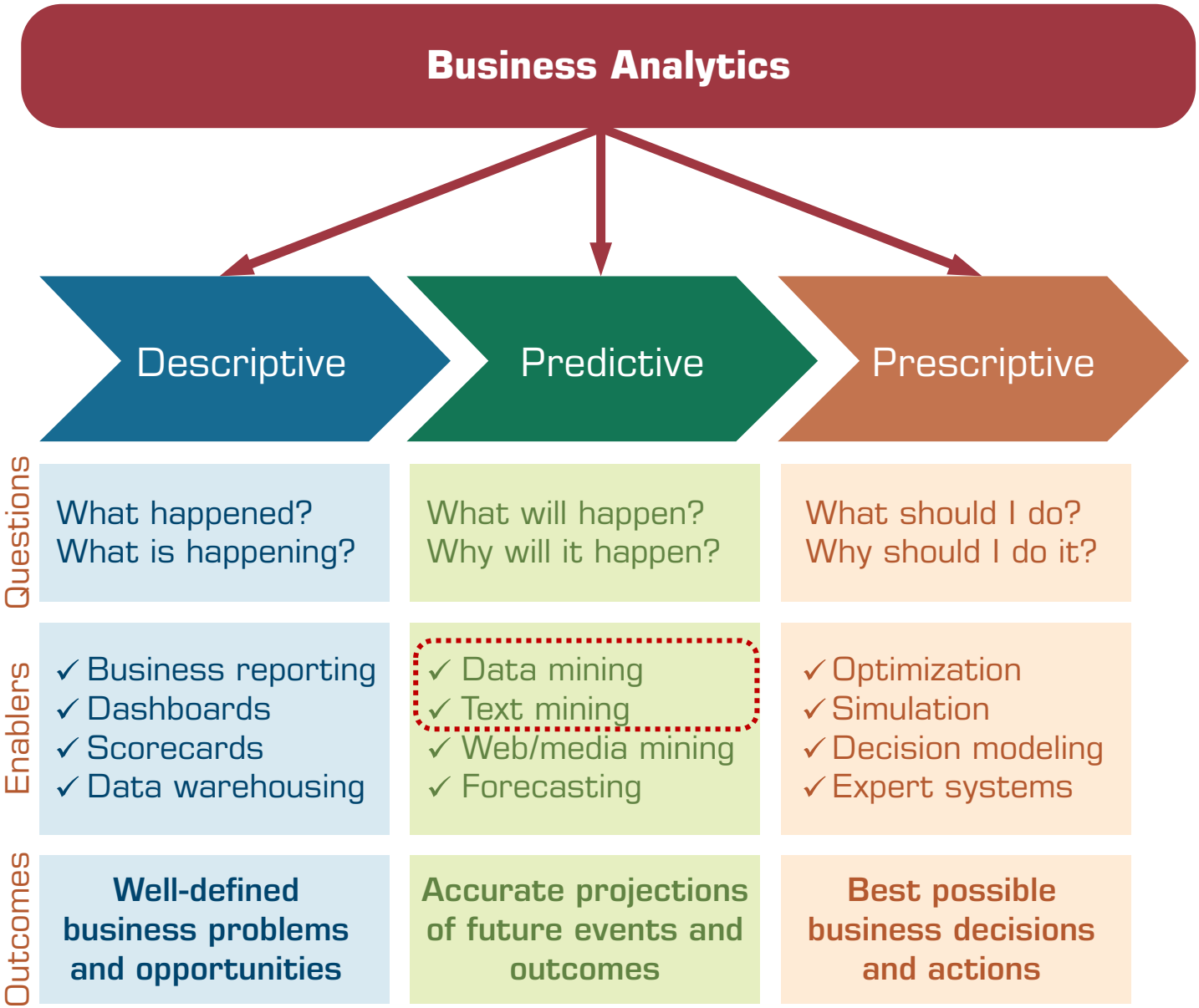
AI, Big Data, Cloud Computing

Evolution of Decision Support, Business Intelligence, and Analytics



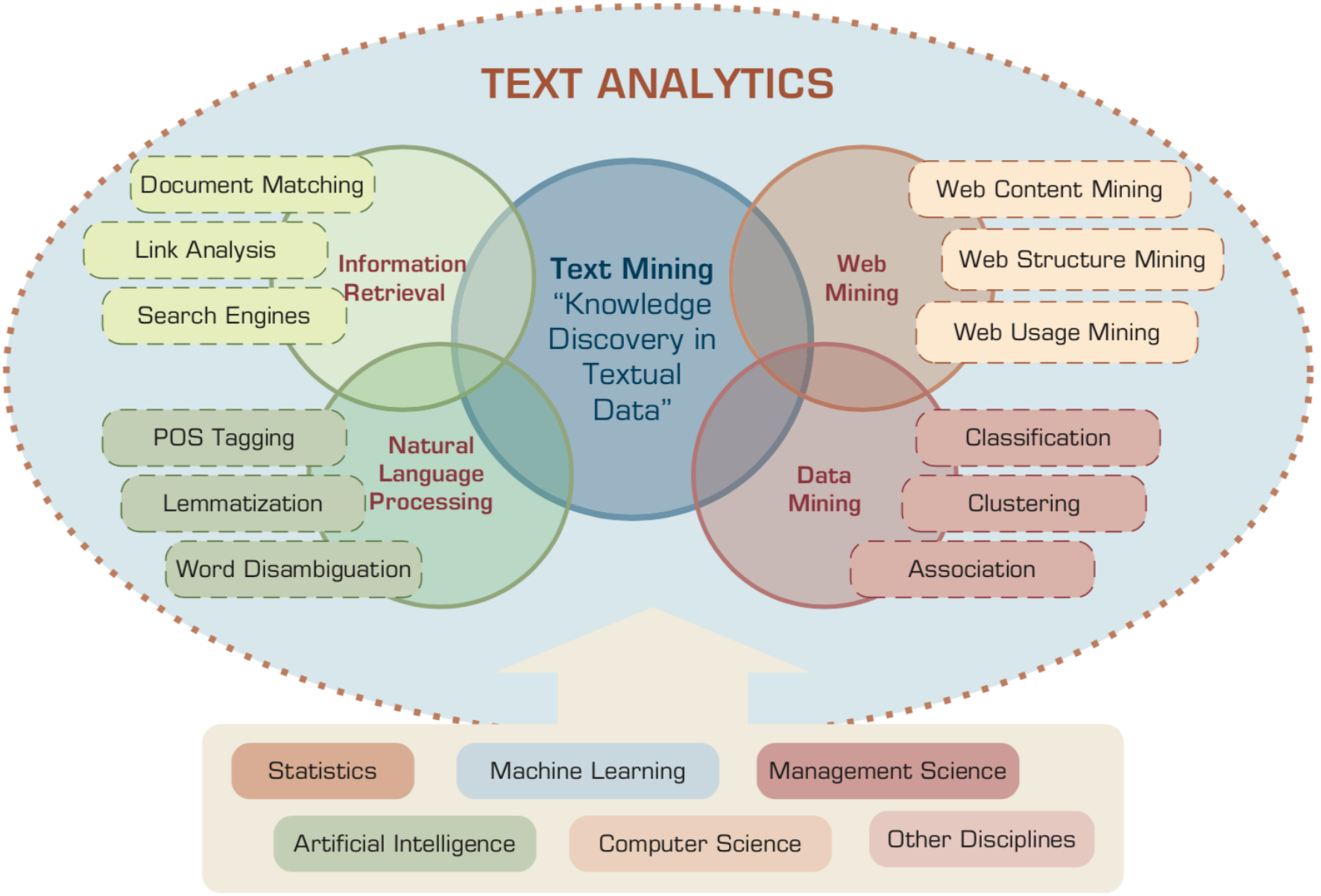
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Three Types of Analytics



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017),
Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Text Analytics and Text Mining



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Text Analytics

- **Text Analytics** =
  Information Retrieval +
  Information Extraction +
  Data Mining +
  Web Mining
- **Text Analytics** =
  Information Retrieval +
  Text Mining

Text mining

- Text Data Mining
- Knowledge Discovery in Textual Databases

Application Areas of Text Mining

- Information extraction
- Topic tracking
- Summarization
- Categorization
- Clustering
- Concept linking
- Question answering

Natural Language Processing (NLP)

- Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics.

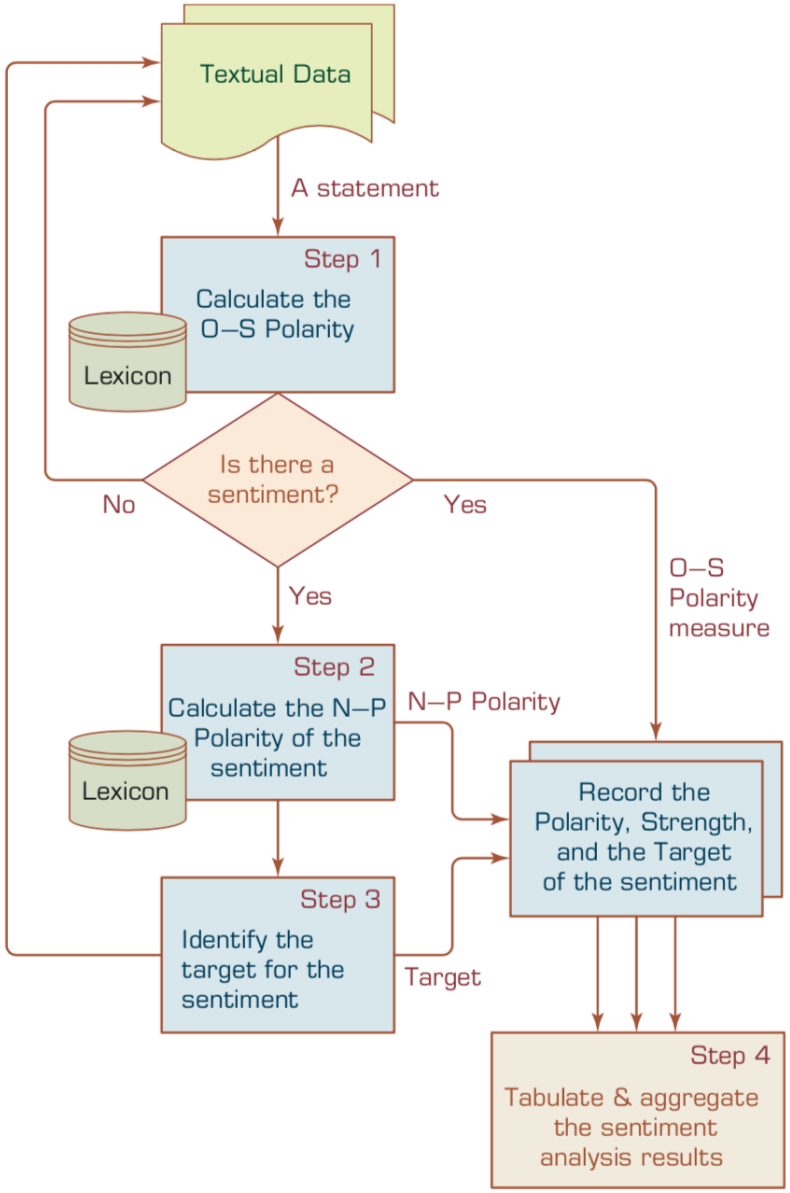
Natural Language Processing (NLP)

- Part-of-speech tagging
- Text segmentation
- Word sense disambiguation
- Syntactic ambiguity
- Imperfect or irregular input
- Speech acts

NLP Tasks

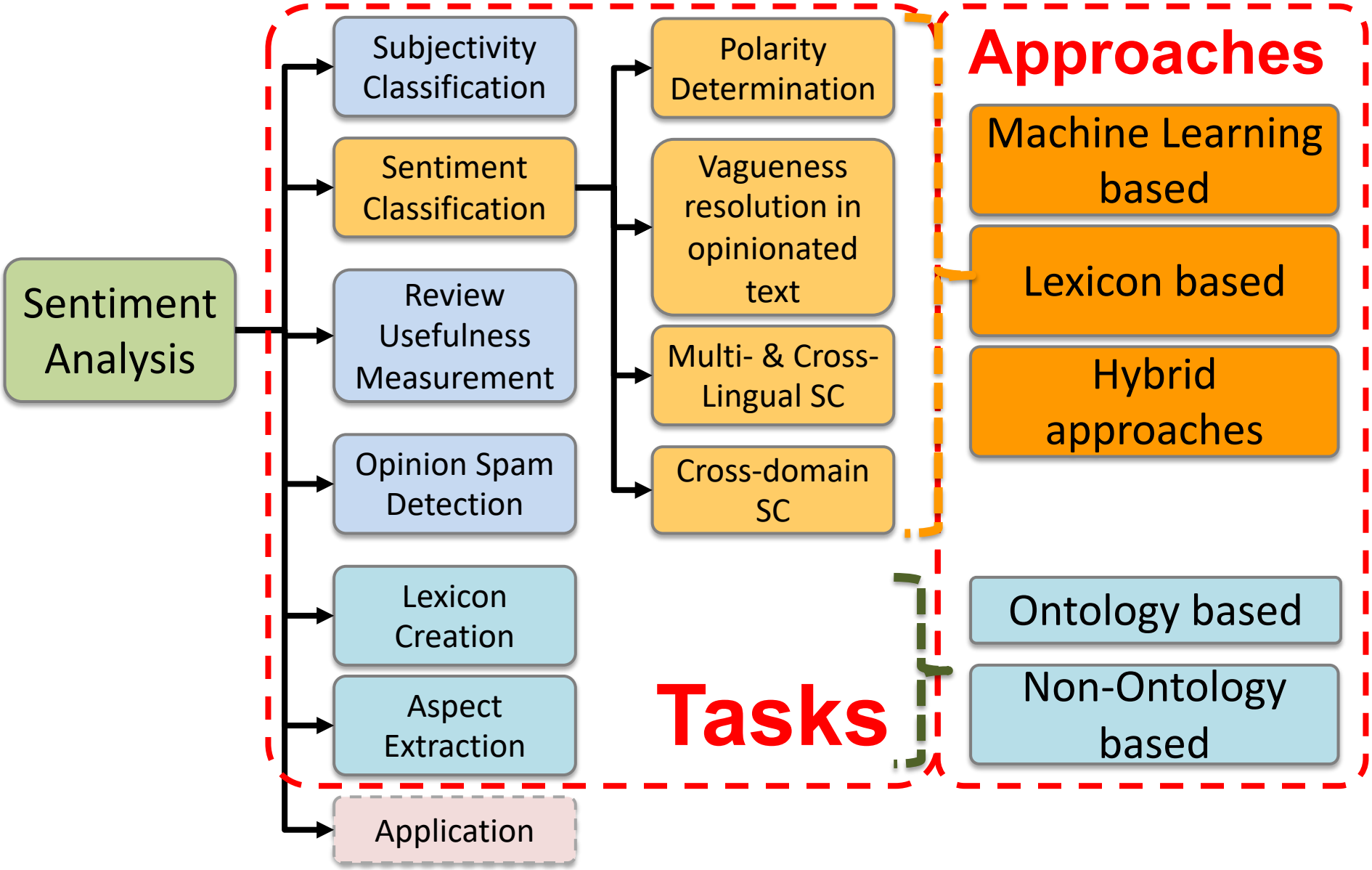
- Question answering
- Automatic summarization
- Natural language generation
- Natural language understanding
- Machine translation
- Foreign language reading
- Foreign language writing.
- Speech recognition
- Text-to-speech
- Text proofing
- Optical character recognition

A Multistep Process to Sentiment Analysis



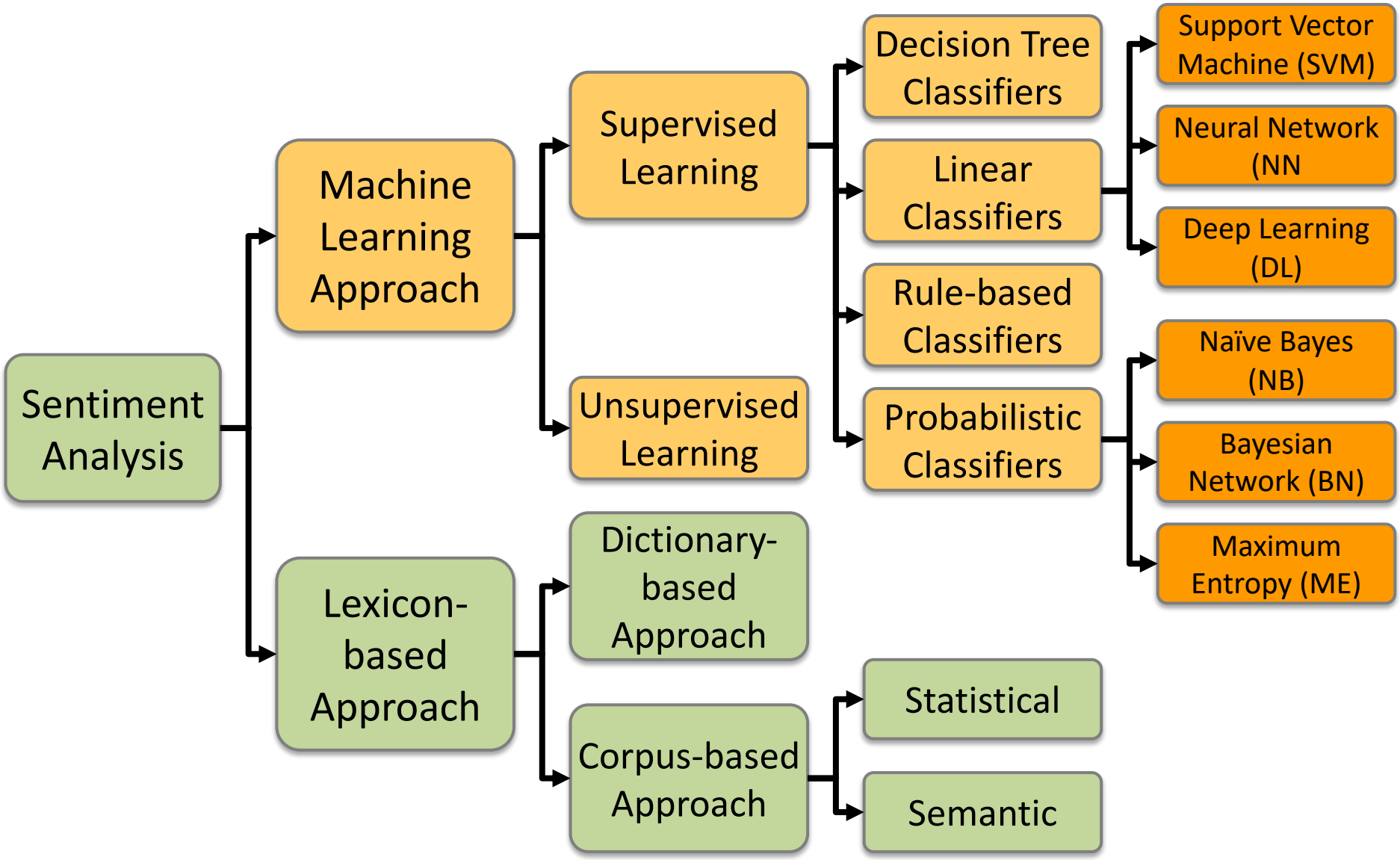
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Sentiment Analysis



Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Techniques



Source: Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma (2015), "Sentiment analysis: A review and comparative analysis of web services," Information Sciences, 311, pp. 18-38.

Example of Opinion: review segment on iPhone

"I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. … "

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

+Positive Opinion

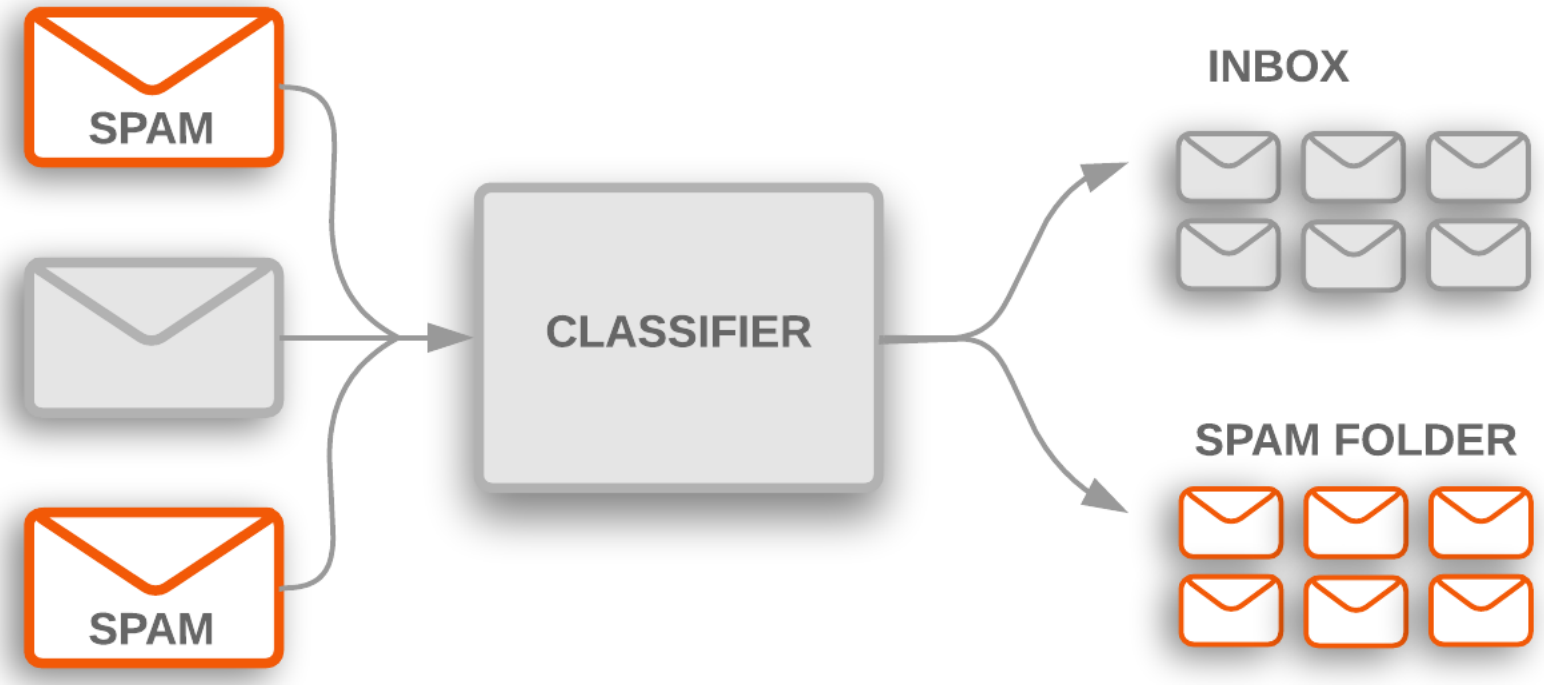
(4) The voice quality was **clear** too.

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too **expensive**, and wanted me to return it to the shop. … "

-Negative Opinion

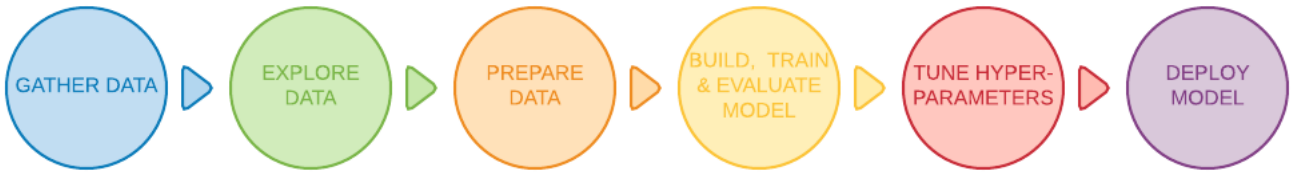
Text Classification



Source: https://developers.google.com/machine-learning/guides/text-classification/

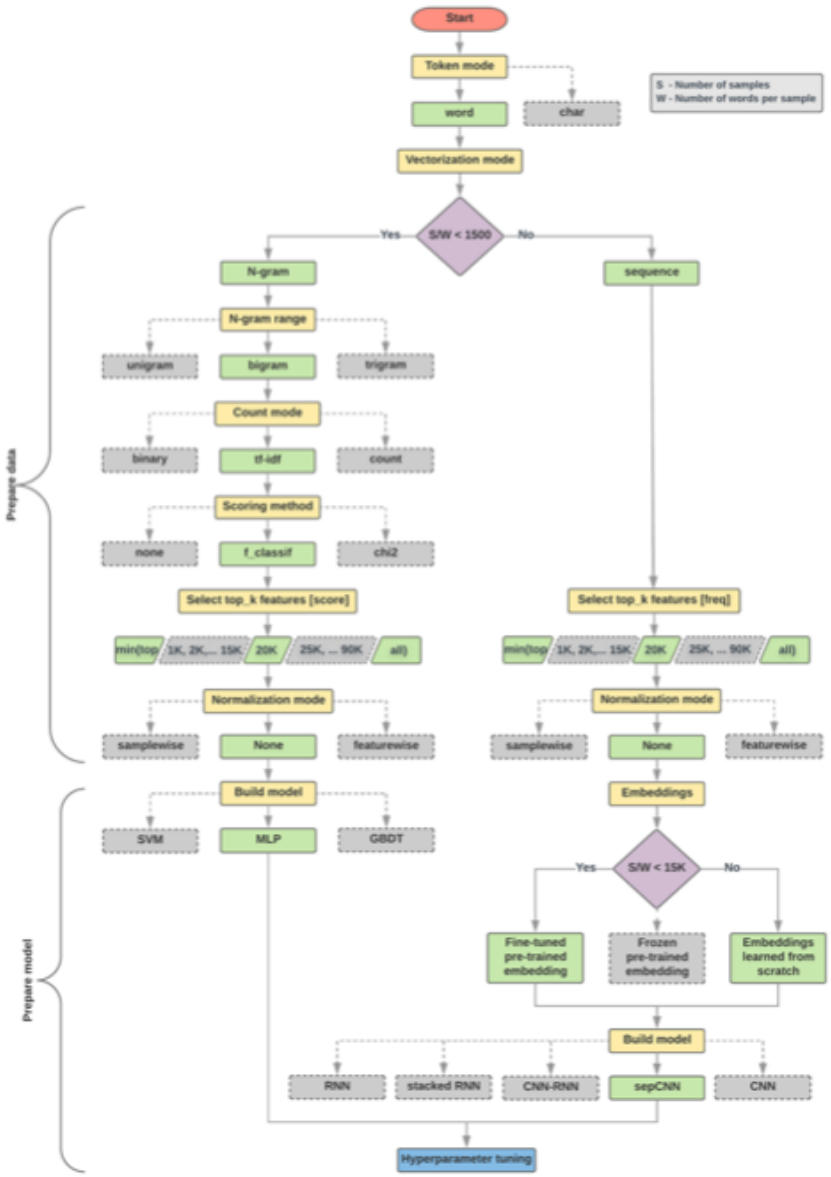
Text Classification Workflow

- Step 1: Gather Data
- Step 2: Explore Your Data
- Step 2.5: Choose a Model*
- Step 3: Prepare Your Data
- Step 4: Build, Train, and Evaluate Your Model
- Step 5: Tune Hyperparameters
- Step 6: Deploy Your Model



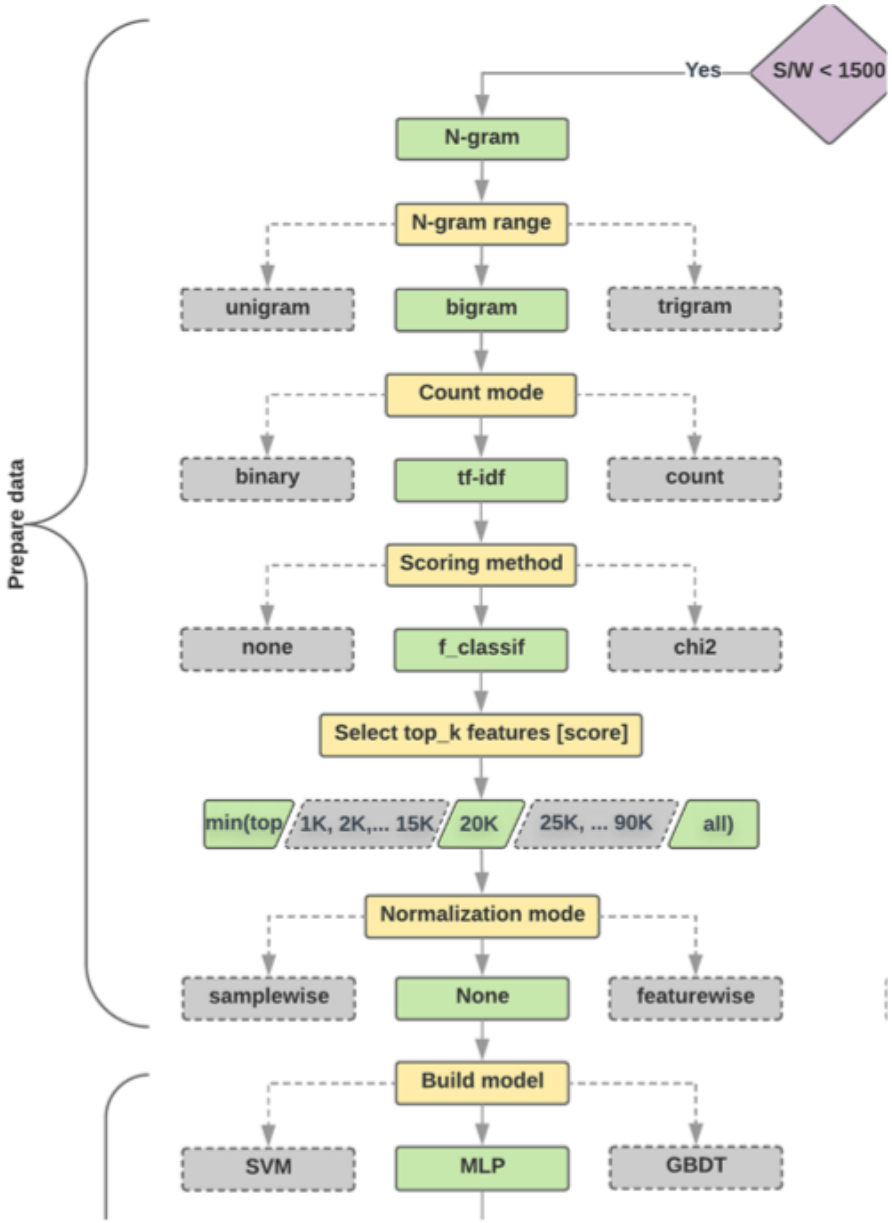
Source: https://developers.google.com/machine-learning/guides/text-classification/

Text Classification Flowchart



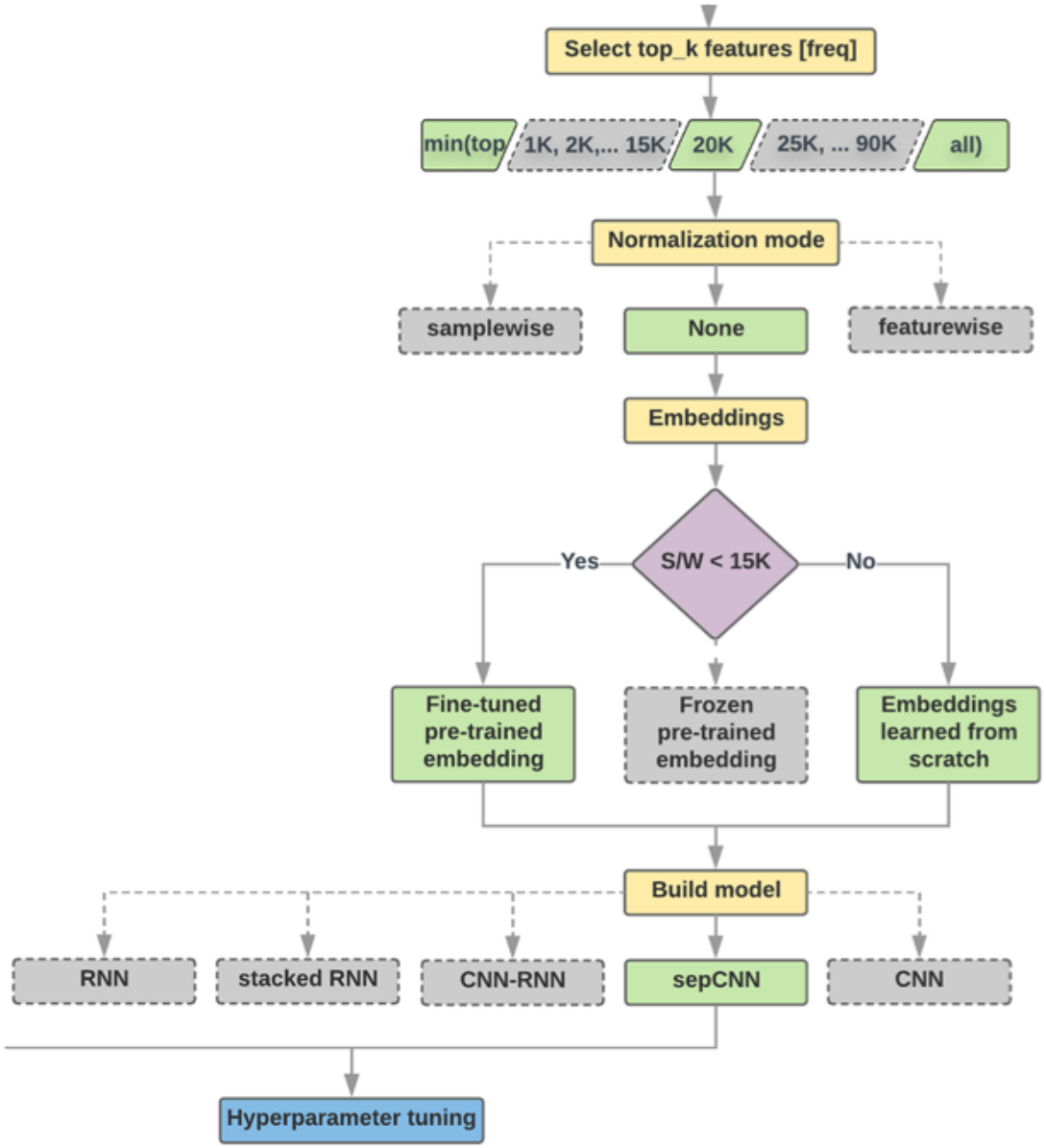
Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Text Classification S/W<1500: N-gram



Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Text Classification S/W>=1500: Sequence

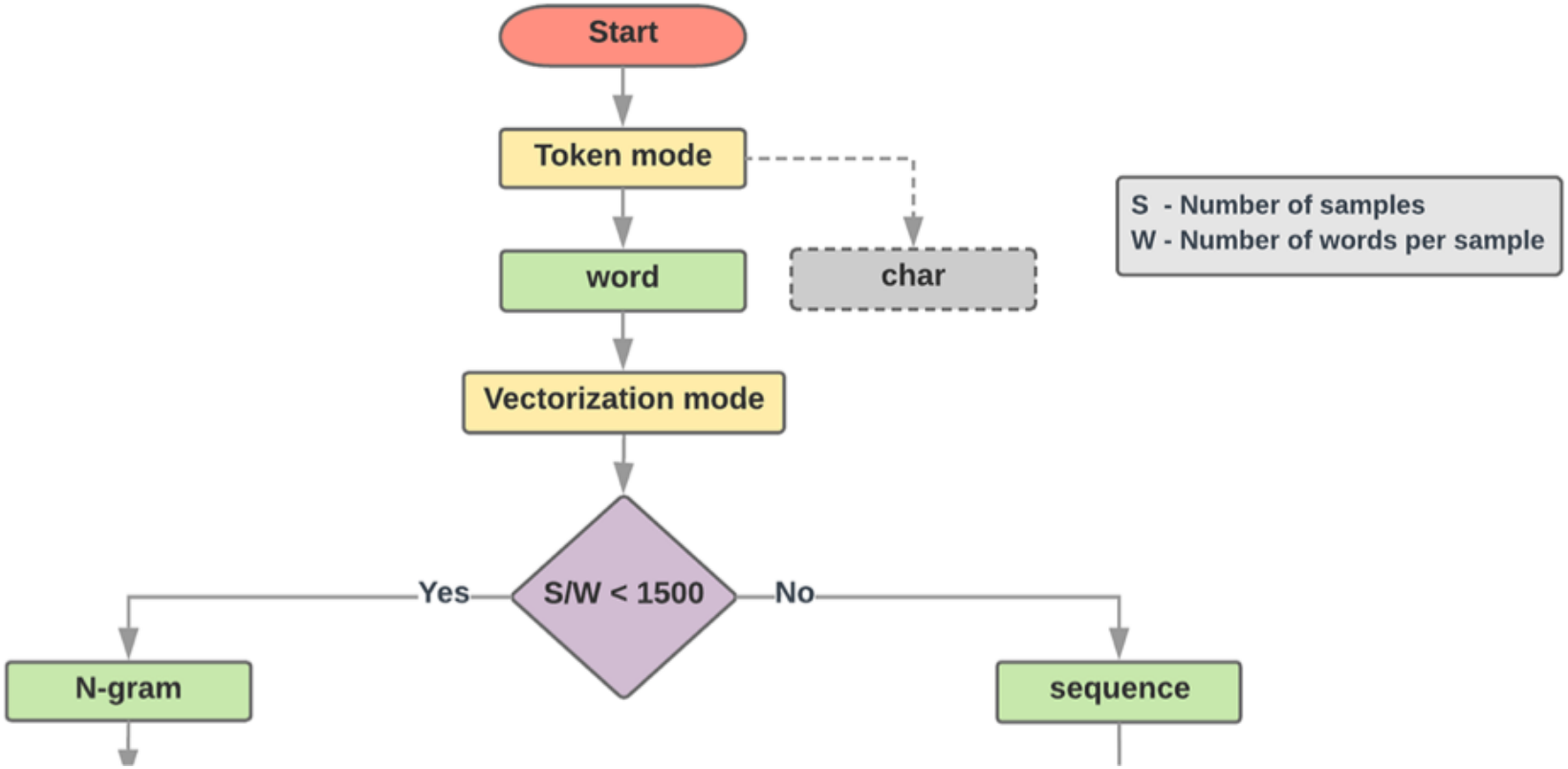


Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 2.5: Choose a Model

Samples/Words < 1500

150,000/100 = 1500



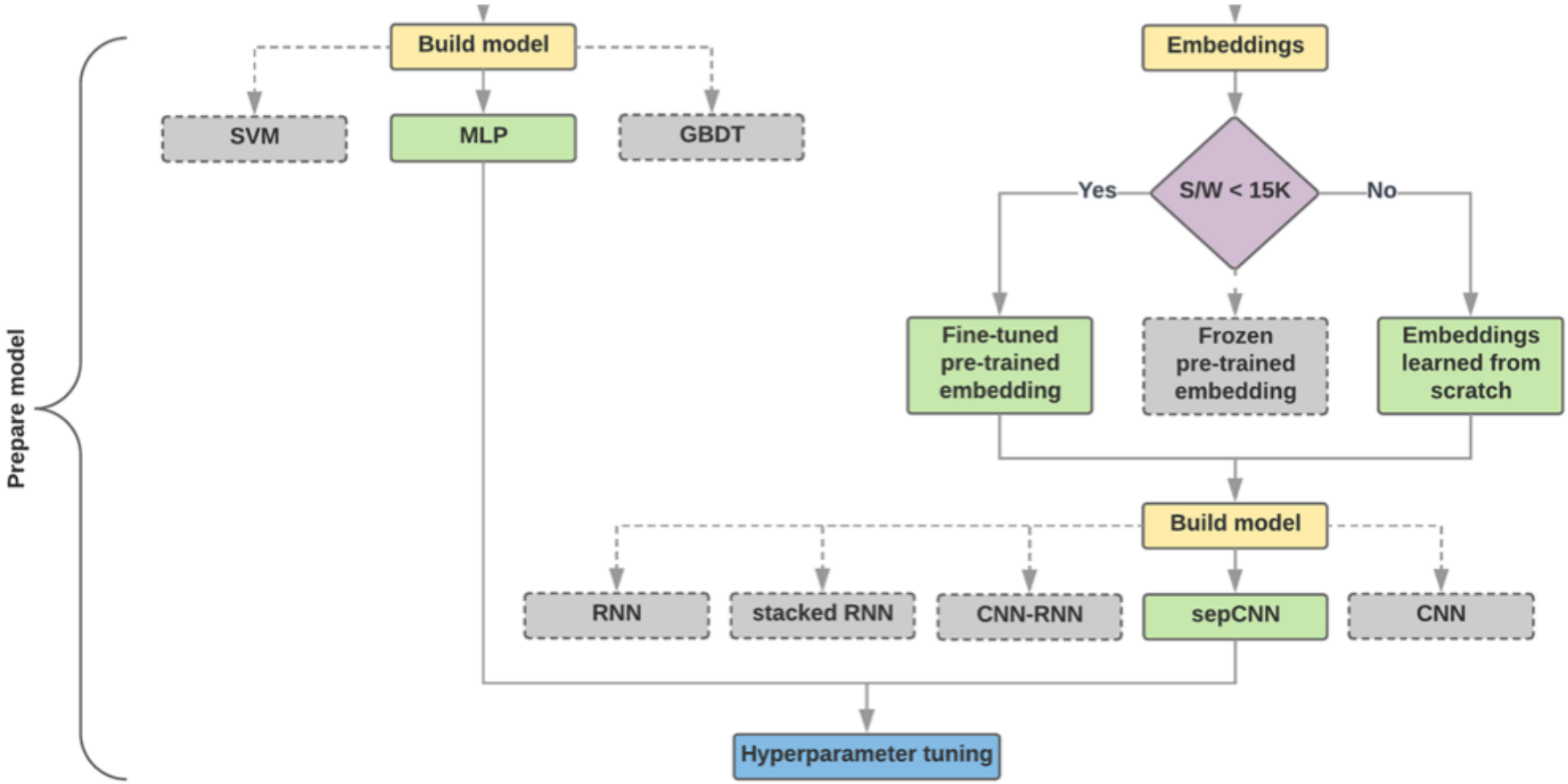
IMDb review dataset,
the samples/words-per-sample ratio is ~ 144

Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 2.5: Choose a Model

Samples/Words < 15,000

1,500,000/100 = 15,000



Source: https://developers.google.com/machine-learning/guides/text-classification/step-2-5

Step 3: Prepare Your Data

```
Texts:
T1: 'The mouse ran up the clock'
T2: 'The mouse ran down'

Token Index:
{'the': 1, 'mouse': 2, 'ran': 3, 'up': 4, 'clock': 5, 'down': 6,}.
    NOTE: 'the' occurs most frequently,
          so the index value of 1 is assigned to it.
          Some libraries reserve index 0 for unknown tokens,
          as is the case here.

Sequence of token indexes:
T1: 'The mouse ran up the clock' =
      [1,  2,   3, 4,  1,  5]
T1: 'The mouse ran down' =
      [1,   2,   3,  6]
```

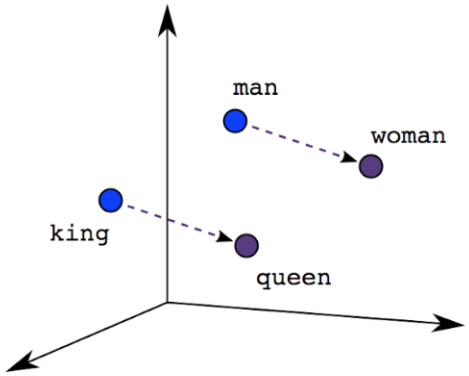
Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

One-hot encoding

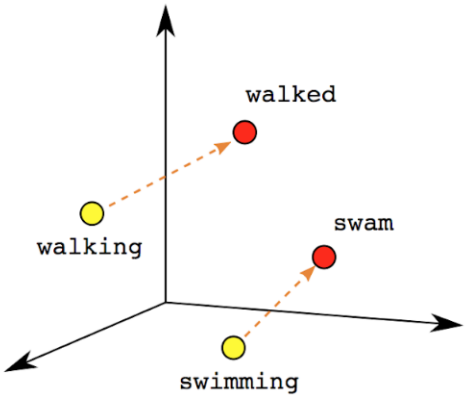
```
                'The mouse ran up the clock' =

The       1   [ [0, 1, 0, 0, 0, 0, 0],
mouse     2     [0, 0, 1, 0, 0, 0, 0],
ran       3     [0, 0, 0, 1, 0, 0, 0],
up        4     [0, 0, 0, 0, 1, 0, 0],
the       1     [0, 1, 0, 0, 0, 0, 0],
clock     5     [0, 0, 0, 0, 0, 1, 0] ]

                [0, 1, 2, 3, 4, 5, 6]
```

Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

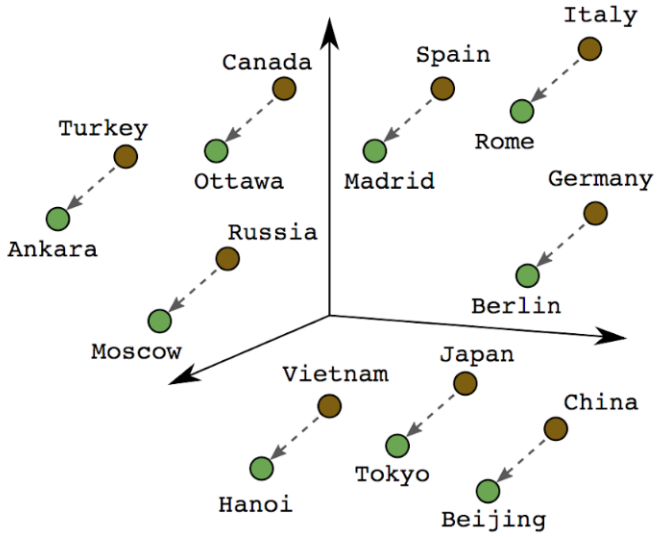
Word embeddings



Male-Female



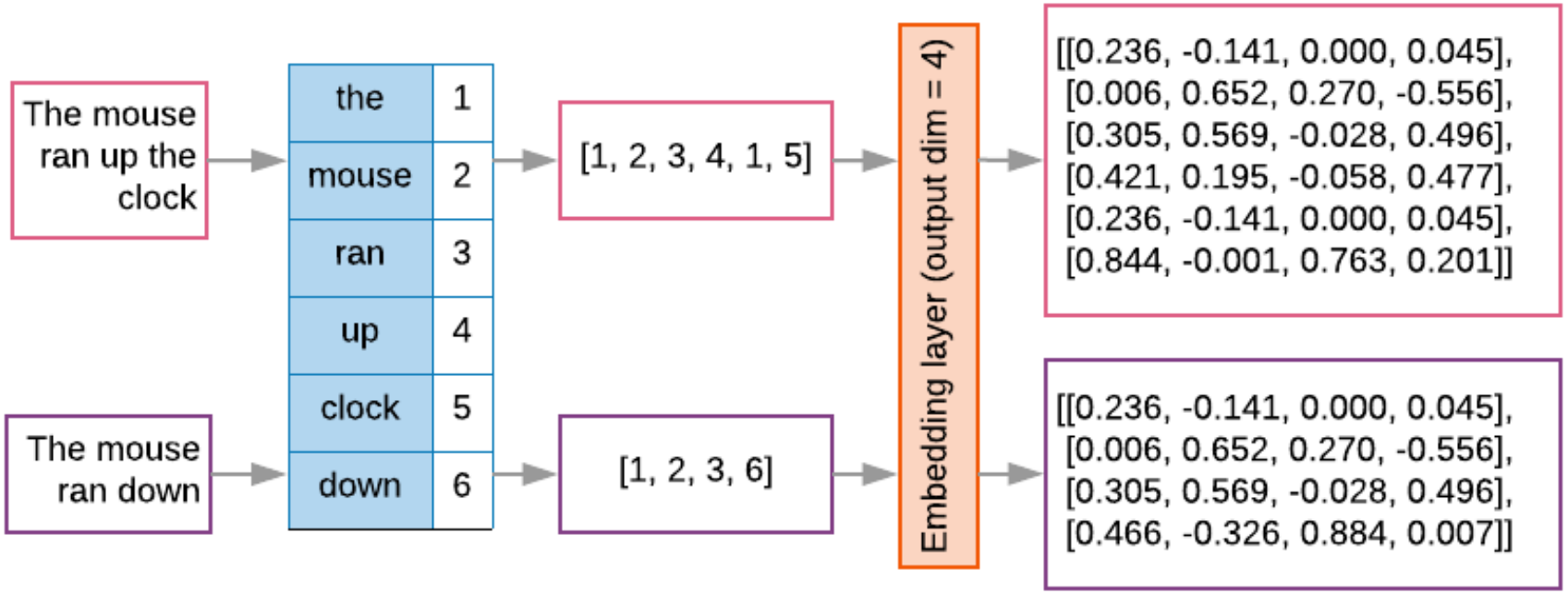
Verb Tense



Country-Capital

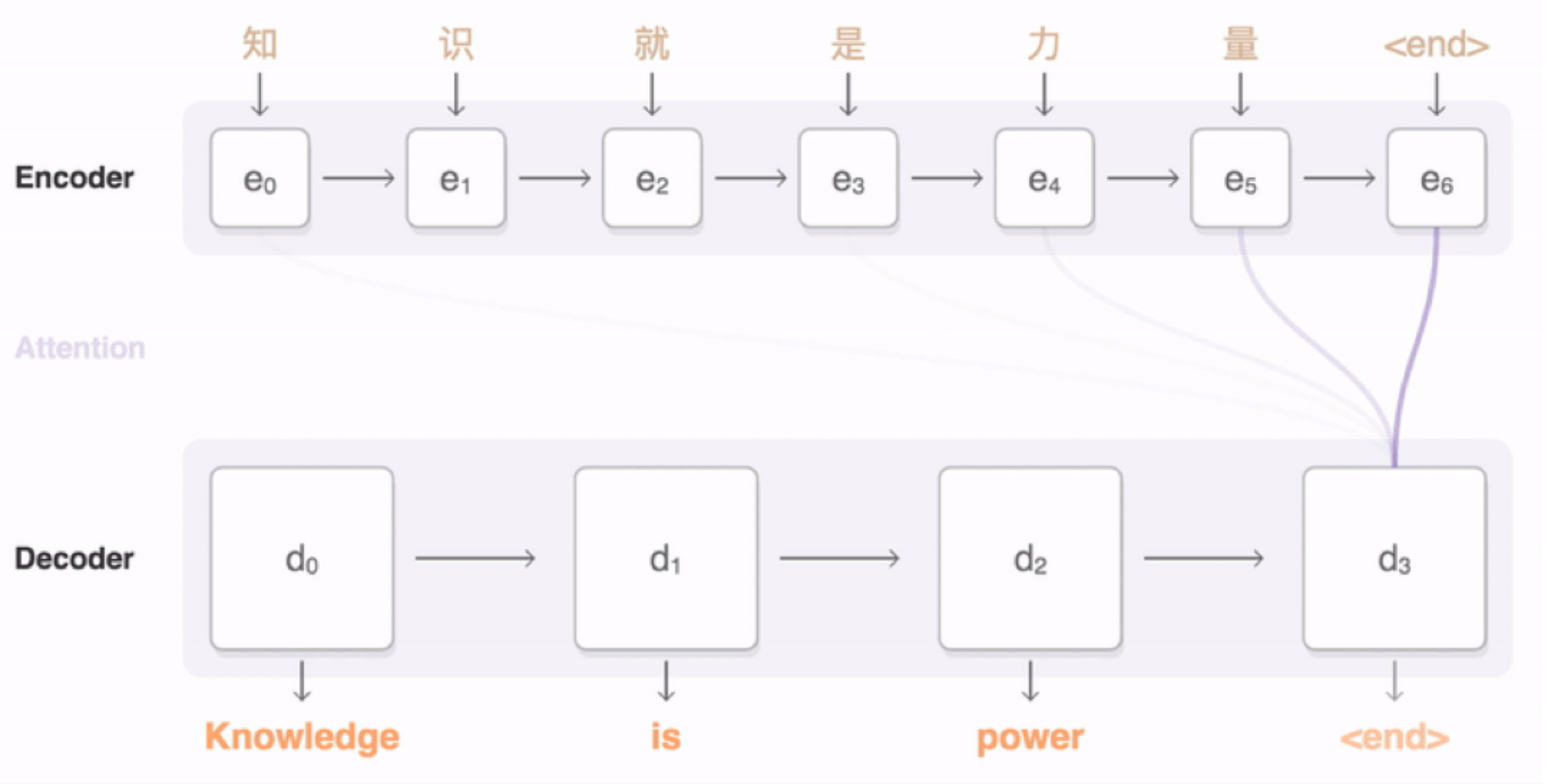
Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

Word embeddings



Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

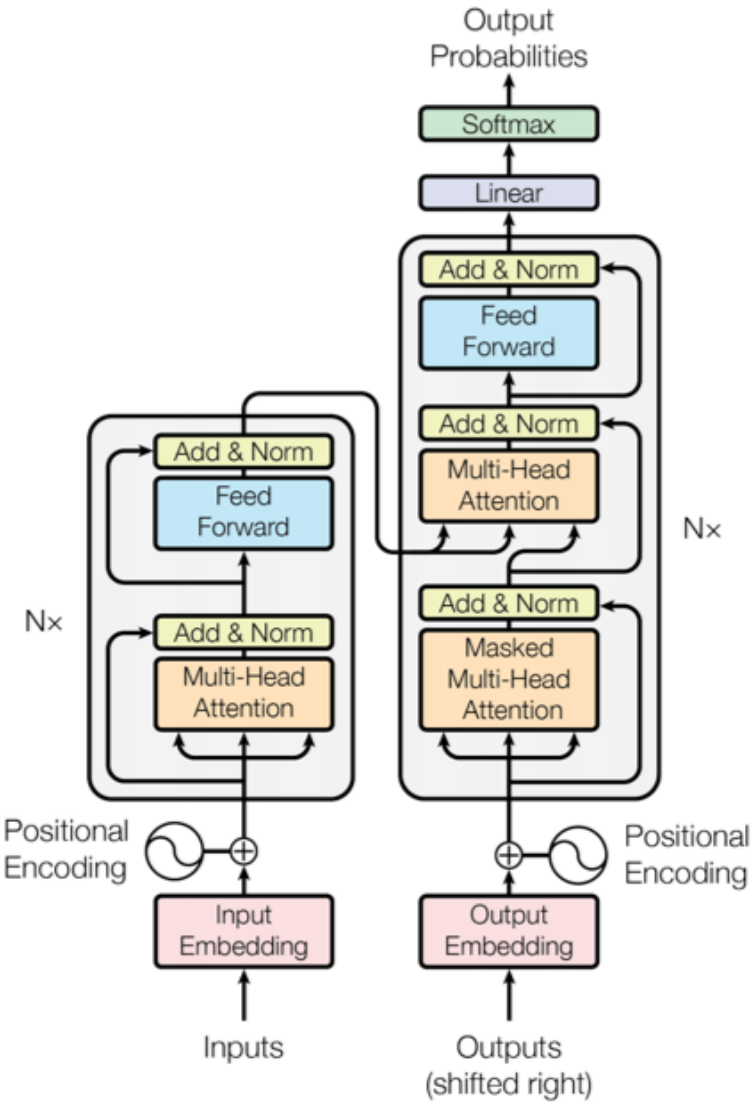
Sequence to Sequence (Seq2Seq)



Source: https://google.github.io/seq2seq/

Transformer (Attention is All You Need)

(Vaswani et al., 2017)

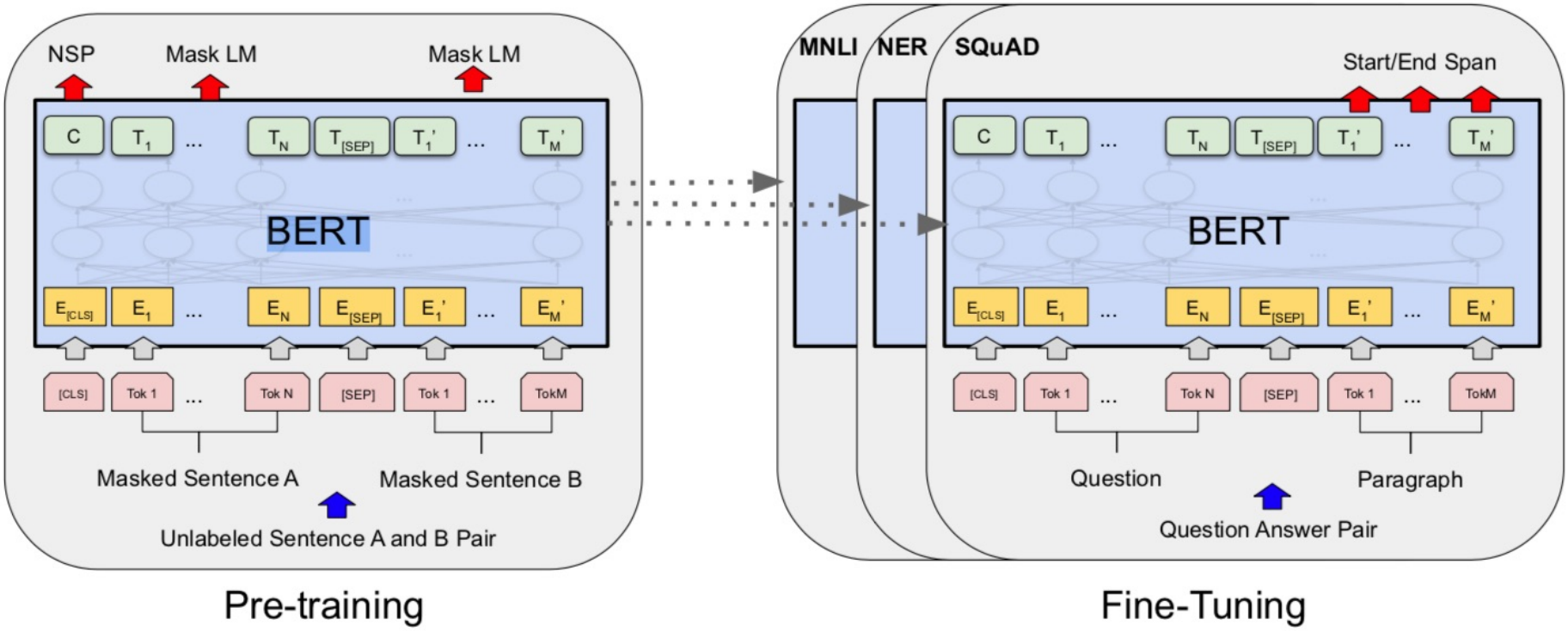


Source: Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

BERT (Bidirectional Encoder Representations from Transformers)

Overall pre-training and fine-tuning procedures for BERT

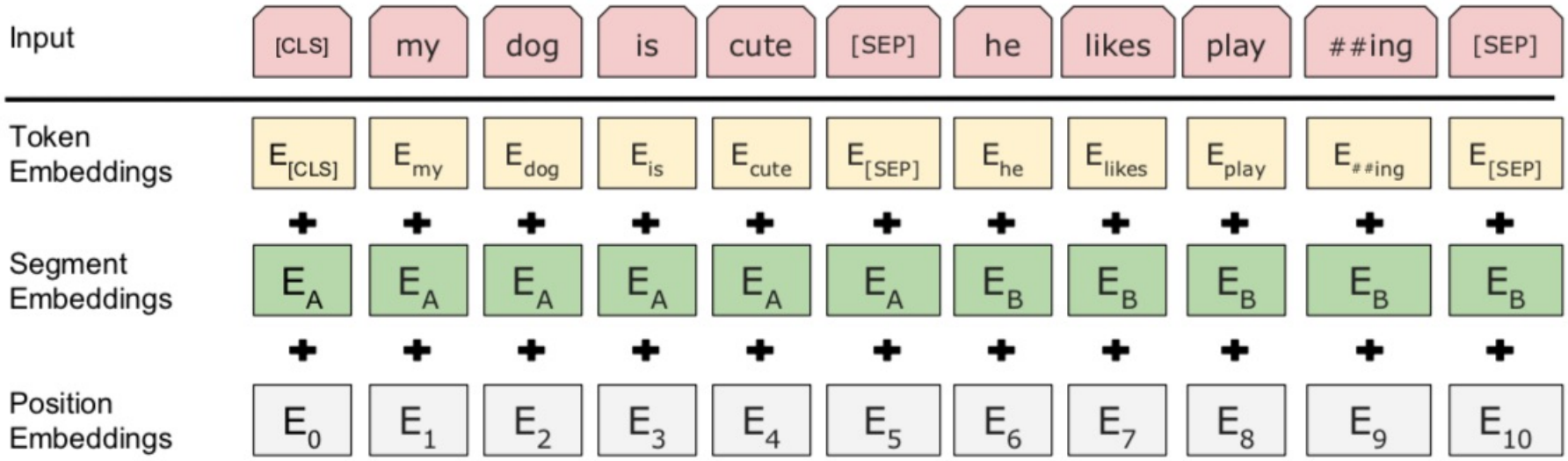


Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

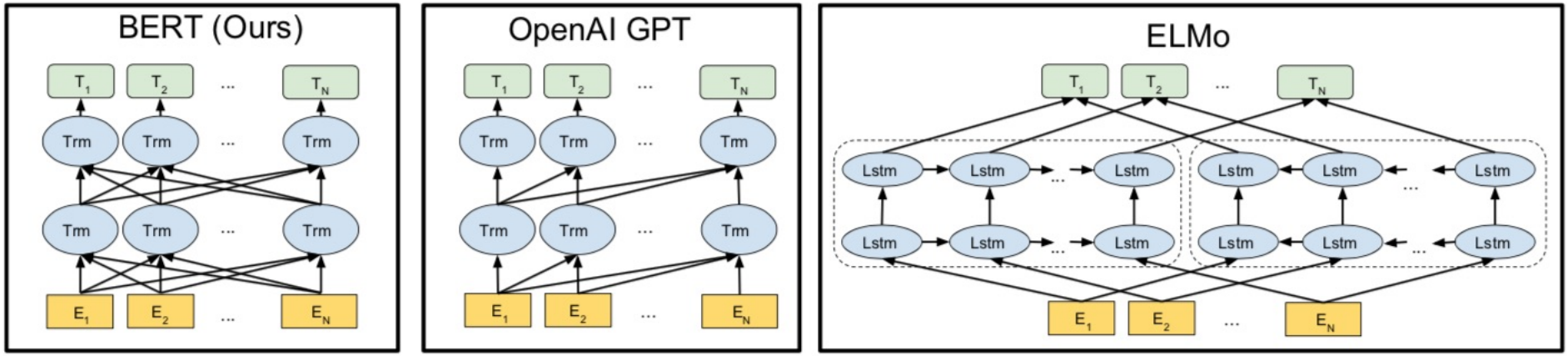
BERT (Bidirectional Encoder Representations from Transformers)

BERT input representation



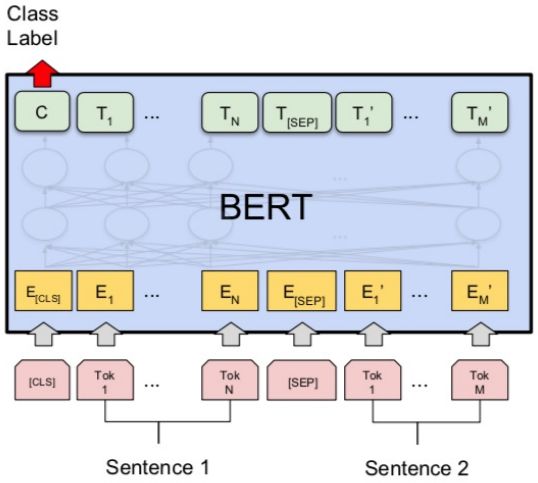
Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

BERT, OpenAI GPT, ELMo

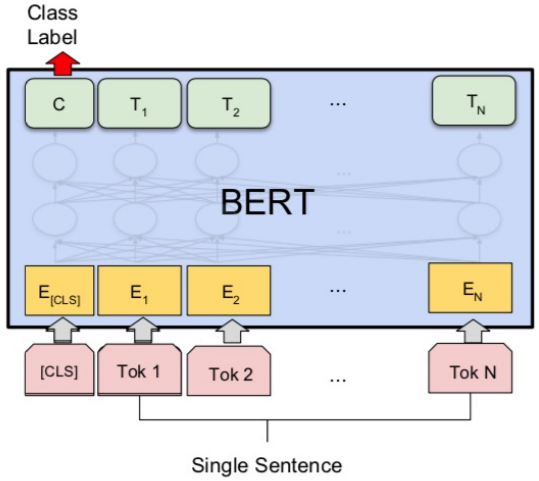


Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

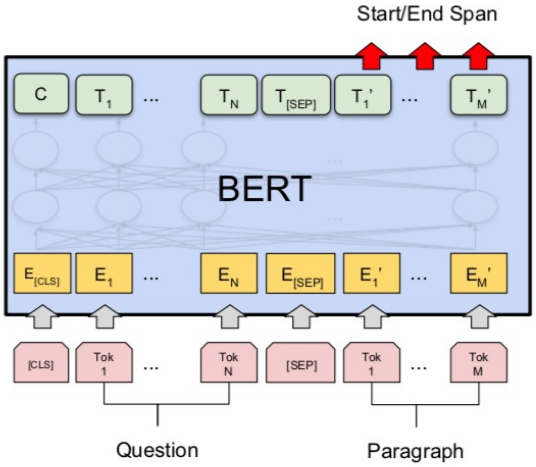
Fine-tuning BERT on Different Tasks



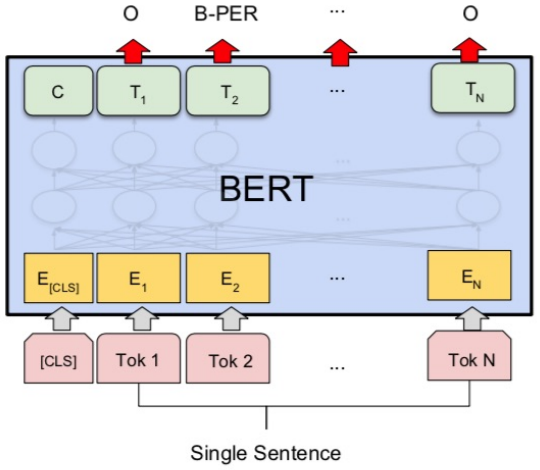
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



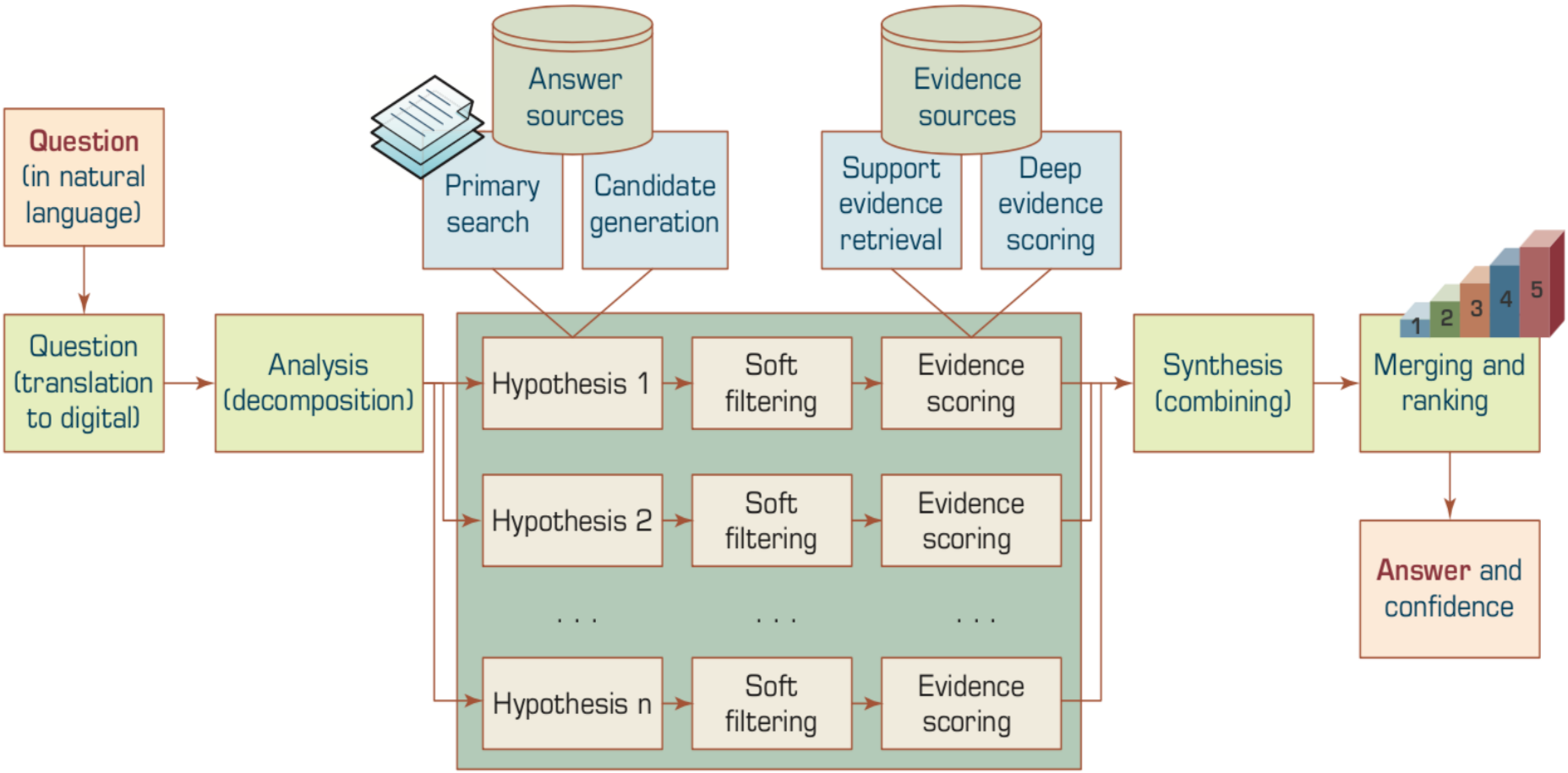
(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

A High-Level Depiction of DeepQA Architecture

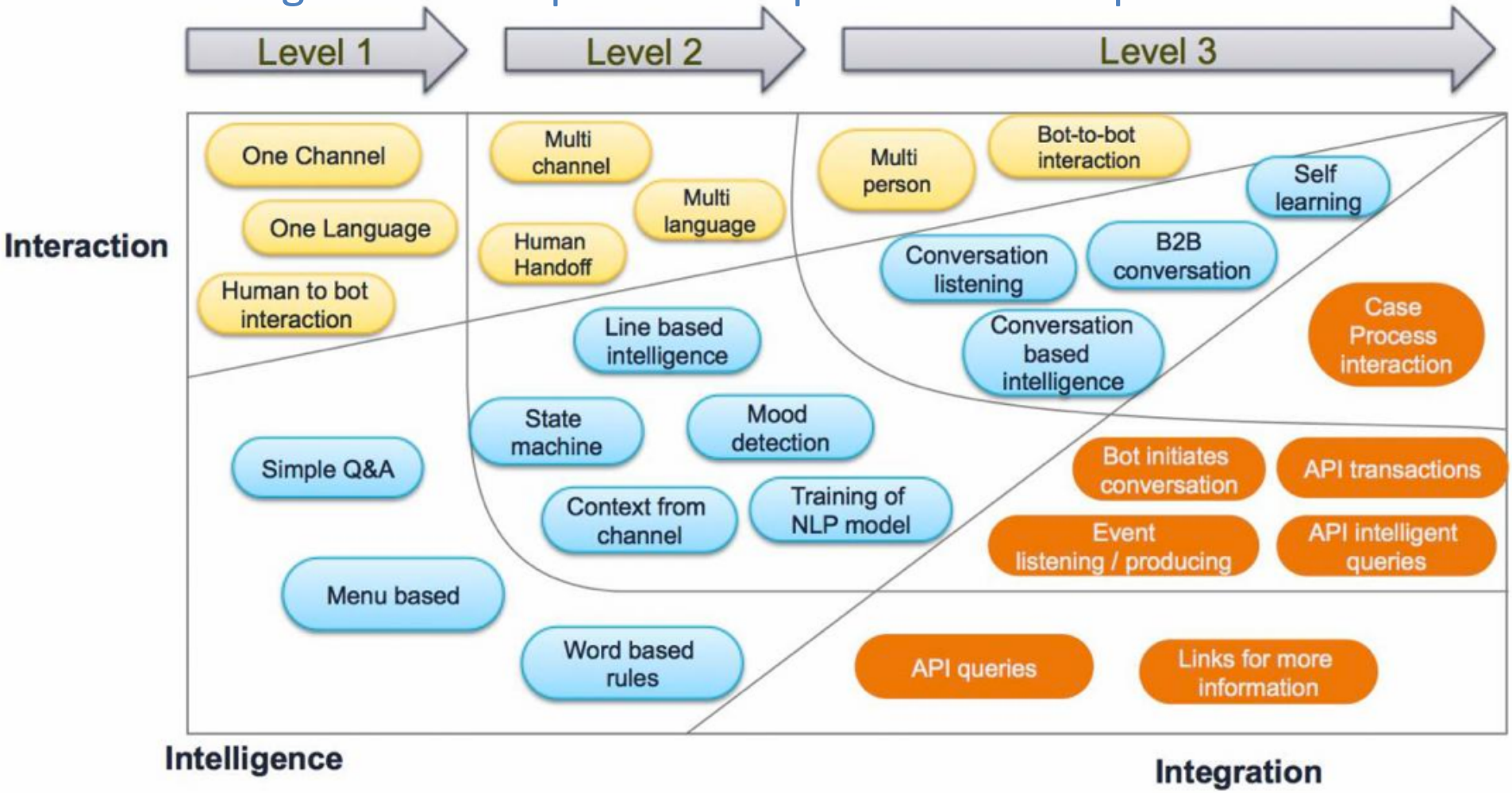


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Chatbots

Bot Maturity Model

Customers want to have simpler means to interact with businesses and get faster response to a question or complaint.



Source: https://www.capgemini.com/2017/04/how-can-chatbots-meet-expectations-introducing-the-bot-maturity/

Dialogue on Airline Travel Information System (ATIS)

The ATIS (Airline Travel Information System) Dataset

https://www.kaggle.com/siddhadev/atis-dataset-from-ms-cntk

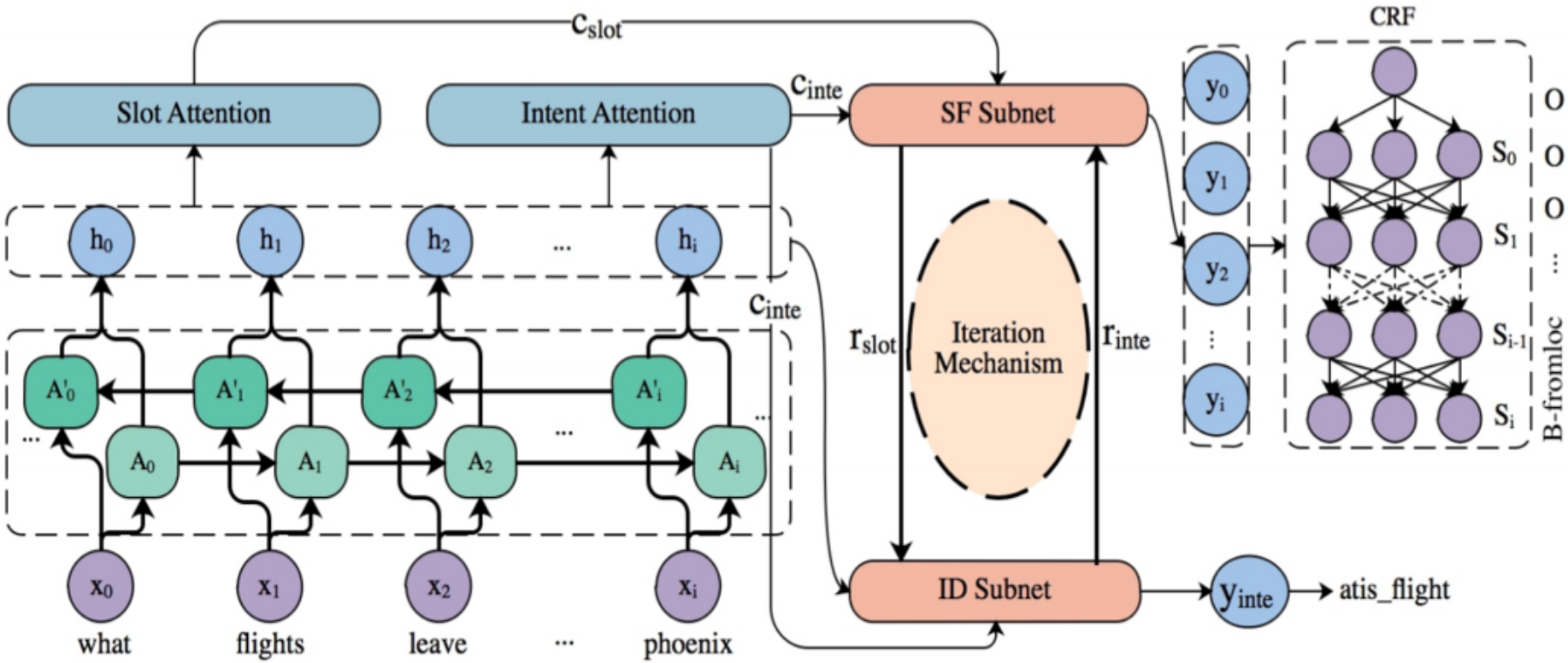
Sentence	what	flights	leave	from	phoenix
Slots	O	O	O	O	B-fromloc
Intent	atis_flight				

Training samples: 4978
Testing samples: 893
Vocab size: 943
Slot count: 129
Intent count: 26

Source: Haihong, E., Peiqing Niu, Zhongfu Chen, and Meina Song. "A novel bi-directional interrelated model for joint intent detection and slot filling." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5467-5471. 2019.

SF-ID Network (E et al., 2019)
Slot Filling (SF)
Intent Detection (ID)

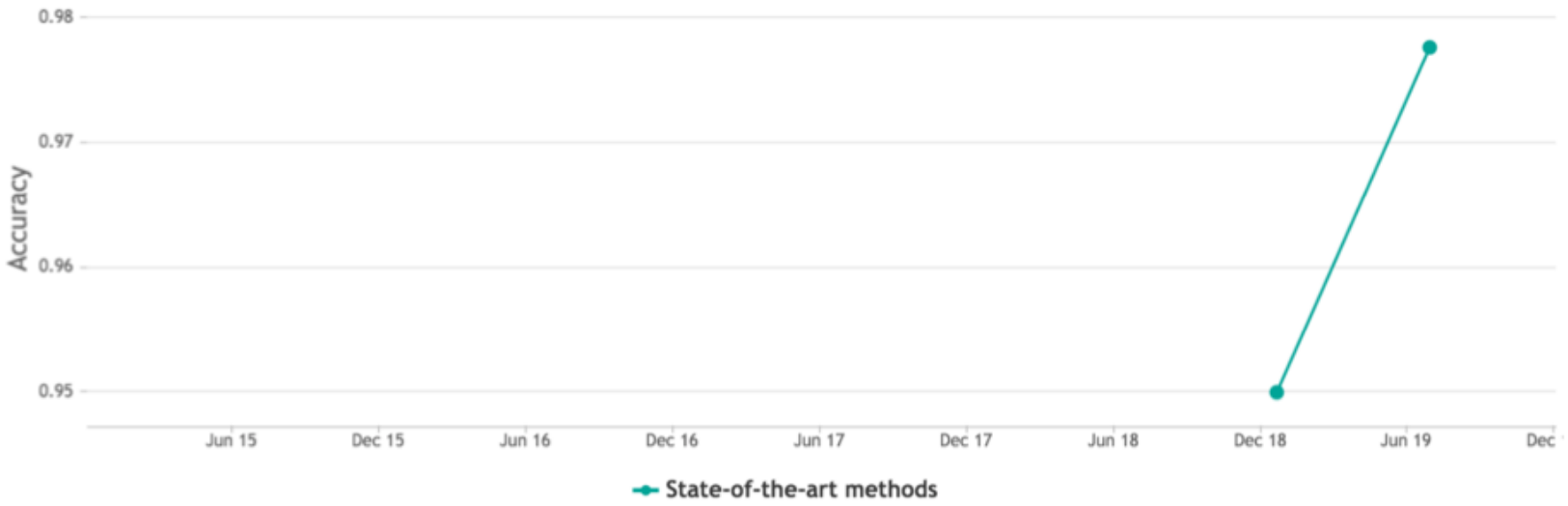
A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling



Source: Haihong, E., Peiqing Niu, Zhongfu Chen, and Meina Song. "A novel bi-directional interrelated model for joint intent detection and slot filling." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5467-5471. 2019.

Intent Detection on ATIS State-of-the-art

Intent Detection on ATIS



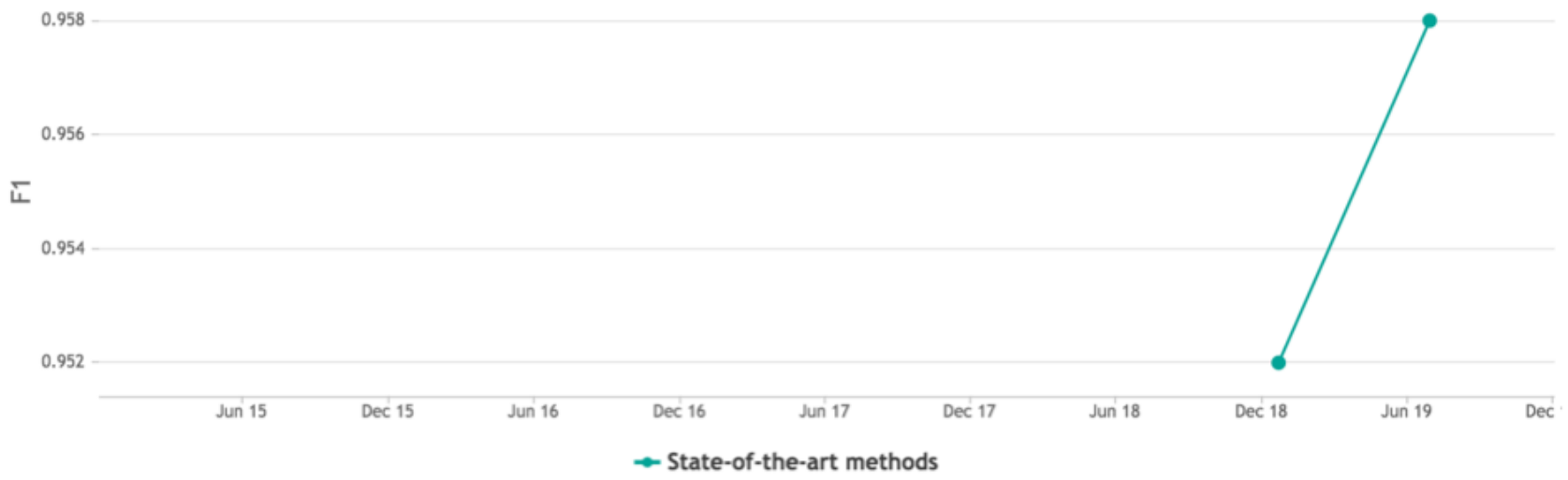
Edit

RANK	METHOD	ACCURACY	PAPER TITLE	YEAR	PAPER	CODE
1	**SF-ID**	0.9776	A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling	2019		
2	**Capsule-NLU**	0.950	Joint Slot Filling and Intent Detection via Capsule Neural Networks	2018		

Source: https://paperswithcode.com/sota/intent-detection-on-atis

Slot Filling on ATIS State-of-the-art

Slot Filling on ATIS



Edit

RANK	METHOD	F1	PAPER TITLE	YEAR	PAPER	CODE
1	SF-ID	0.958	A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling	2019		
2	Capsule-NLU	0.952	Joint Slot Filling and Intent Detection via Capsule Neural Networks	2018		

Source: https://paperswithcode.com/sota/slot-filling-on-atis

Restaurants Dialogue Datasets

- MIT Restaurant Corpus
  - https://groups.csail.mit.edu/sls/downloads/restaurant/
- CamRest676
  (Cambridge restaurant dialogue domain dataset)
  - https://www.repository.cam.ac.uk/handle/1810/260970
- DSTC2 (Dialog State Tracking Challenge 2 & 3)
  - http://camdial.org/~mh521/dstc/

任務型對話系統

The Evaluation of Chinese Human-Computer Dialogue Technology, SMP2019-ECDT

- 自然語言理解
  Natural Language Understanding (NLU)
- 對話管理
  Dialog Management (DM)
- 自然語言生成
  Natural Language Generation (NLG)

Source: http://conference.cipsc.org.cn/smp2019/evaluation.html
https://github.com/OnionWang/SMP2019-ECDT-NLU

Summary

- This course introduces the fundamental concepts and research issues of artificial intelligence for text analytics.
- Topics include
  - Foundations of Text Analytics: Natural Language Processing (NLP),
  - Python for NLP,
  - Processing and Understanding Text,
  - Feature Engineering for Text Representation,
  - Text Classification,
  - Text Summarization and Topic Models,
  - Text Similarity and Clustering,
  - Semantic Analysis and Named Entity Recognition,
  - Sentiment Analysis,
  - The Promise of Deep Learning and Universal Sentence-Embedding Models,
  - Question Answering and Dialogue Systems,
  - and Case Study on AI Text Analytics.

人工智慧文本分析
Artificial Intelligence for Text Analytics
Contact Information

戴敏育 博士 (Min-Yuh Day, Ph.D.)

副教授 (Associate Professor)

淡江大學 資訊管理學系

Department of Information Management, Tamkang University

電話：02-26215656 #2846

傳真：02-26209737

研究室：B929

地址： 25137 新北市淡水區英專路151號

Email： myday@mail.tku.edu.tw

網址：http://mail.tku.edu.tw/myday/