

ICT이노베이션스퀘어 AI복합교육 고급 언어과정

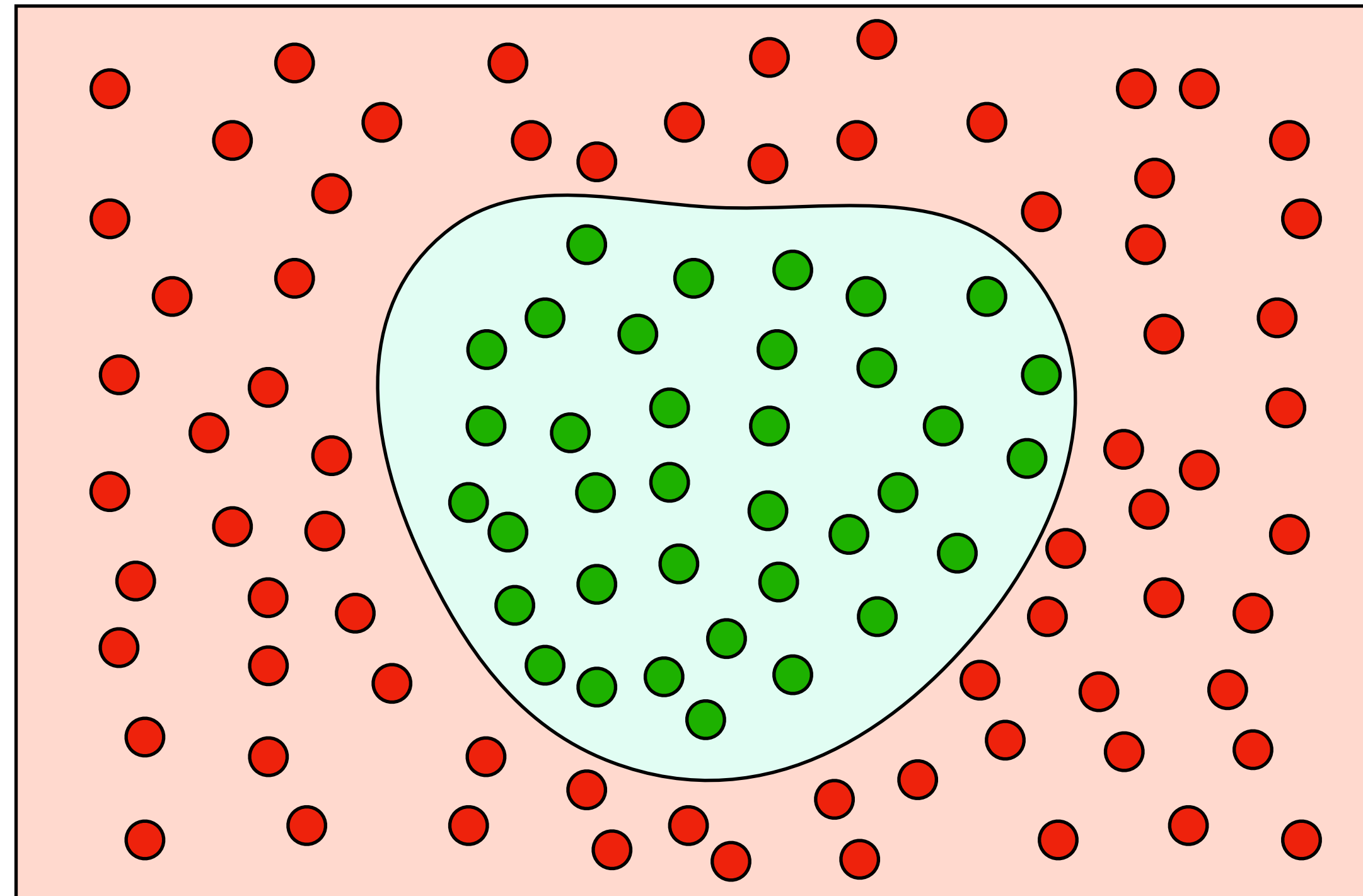
자연어처리를 위한 Text Classification

현청천

2021.04.19

What is Text Classification

- Text를 미리 정해진 Class로 분류하는 Task



What is Text Classification

- Text를 미리 정해진 Class로 분류하는 Task
 - 2 class: binary classification
 - N class: multi class classification
 - N label: multi label classification

Binary Classification

- Spam filtering (SPAM, HAM)

(광고) [라이지움] 누구나 노동부 교육비 지원 CISSP/CPPPG/PMP/CIA/감리사/CISA 개강안내

SPAM

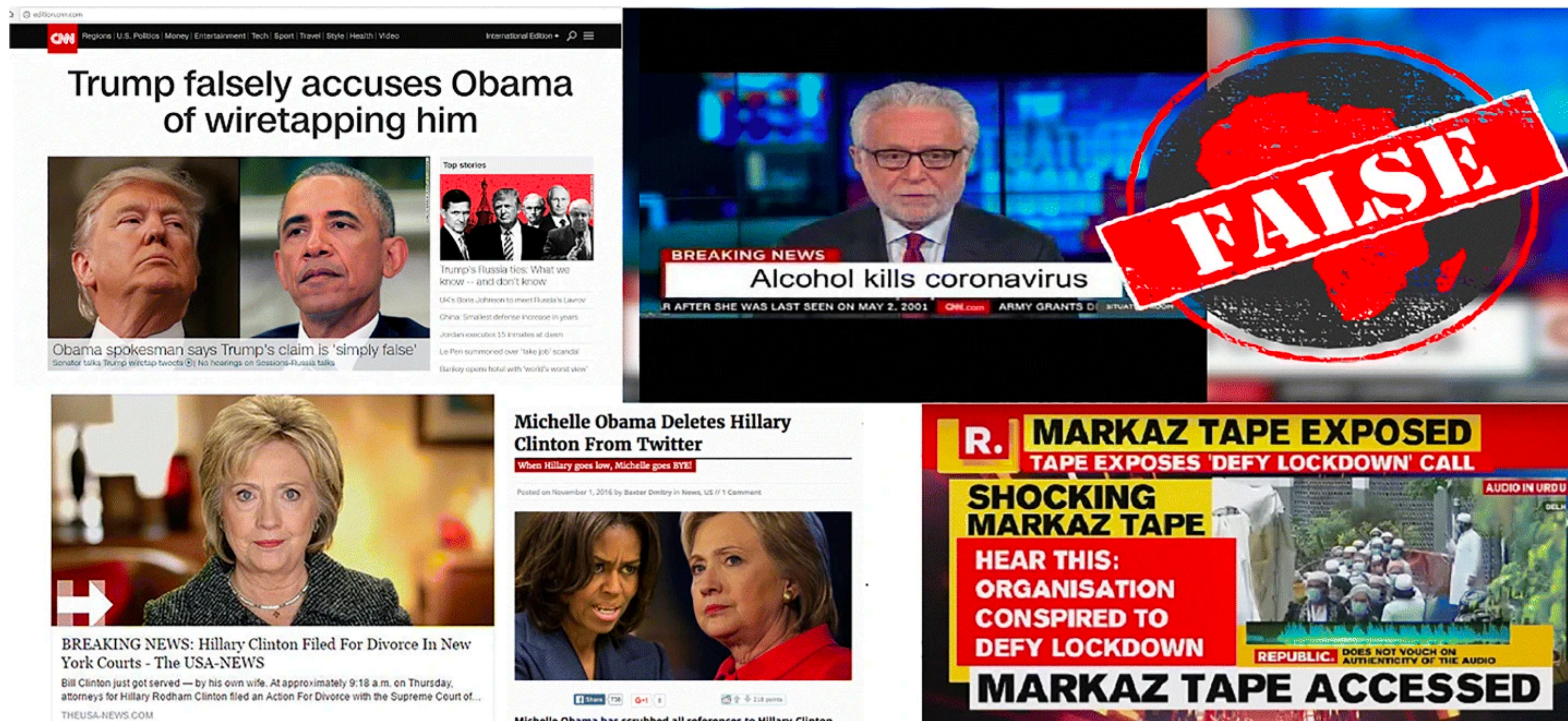
Binary Classification

- Spam filtering (SPAM, HAM)
- Sentiment Analysis (긍정, 부정)

딘따 잼없다!!! 영화를 만들어야지 다큐를 만드냐?!	➡	부정
강추합니다 남산의 부장들 재밌네요 속이 다 시원함	➡	긍정
중후반은 시간이 어찌 흘렀는지...모를정도로 완전 집중몰입!!!!	➡	긍정


Binary Classification

- Spam filtering (SPAM, HAM)
- Sentiment Analysis (긍정, 부정)
- 가짜 뉴스 검출 (Real, Fake)



Multi Class Classification

- Language Detection (한국어, 영어, 일본어, 중국어, ...)

**Language Classifier**
English

Detect language in text. New languages were added for a total of 49 different languages arranged in language families.

Afrikaans-af

Arabic-ar

Armenian-hy

Azerbaijani-az

Basque-eu

Belarusian-be

Bengali-bn

Bulgarian-bg

Cantonese-yue

Catalan-ca

[View 40 more](#)

Test with your own text

Simia (nomen Latinum classicum), vel rarius simius, est animal Romanis notum quod ad ordinem primatum.

Classify Text

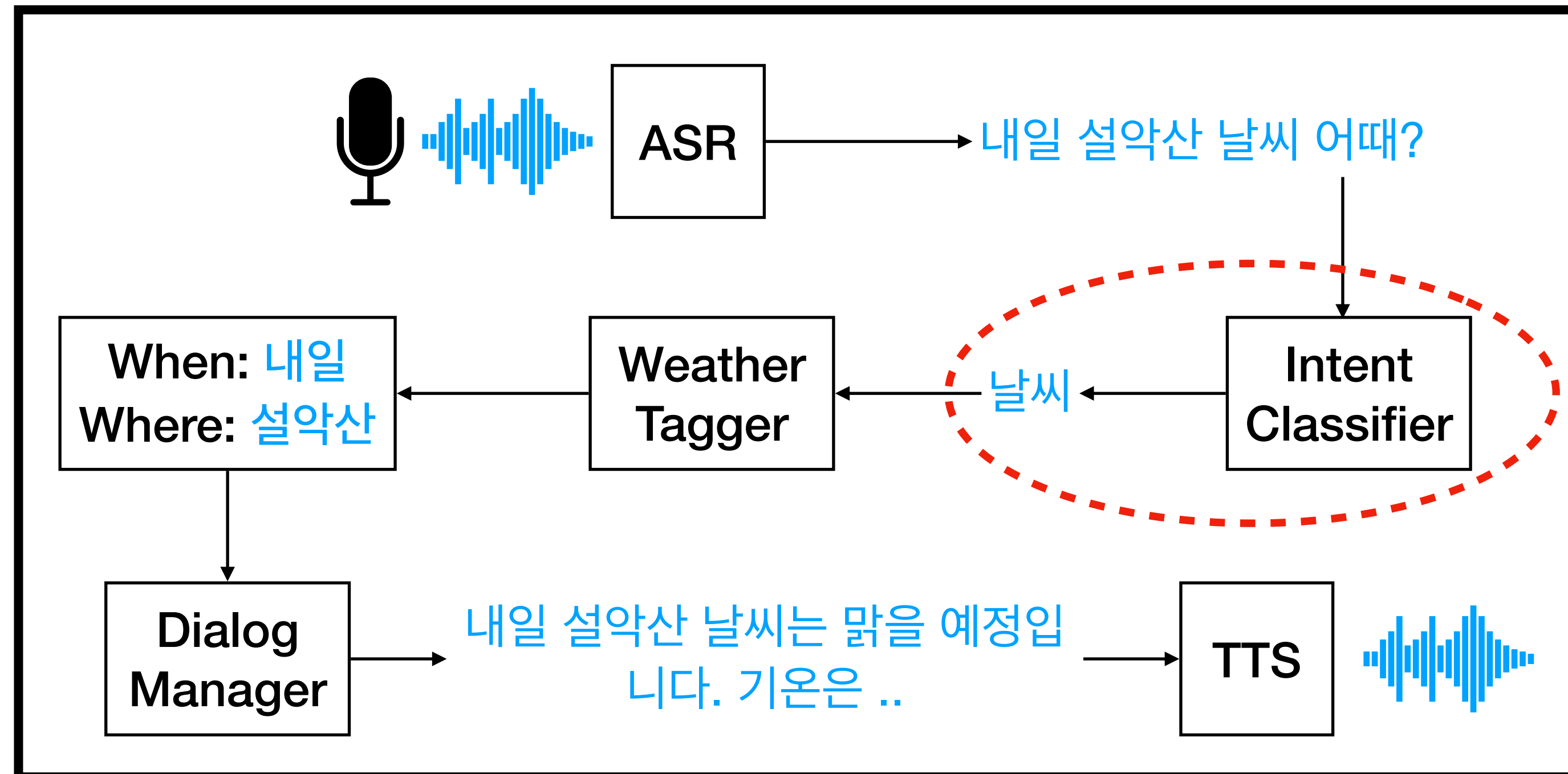
[LIST](#) [JSON](#)

TAG	CONFIDENCE
Latin-la	65.8%

https://app.monkeylearn.com/main/classifiers/cl_Vay9jh28/

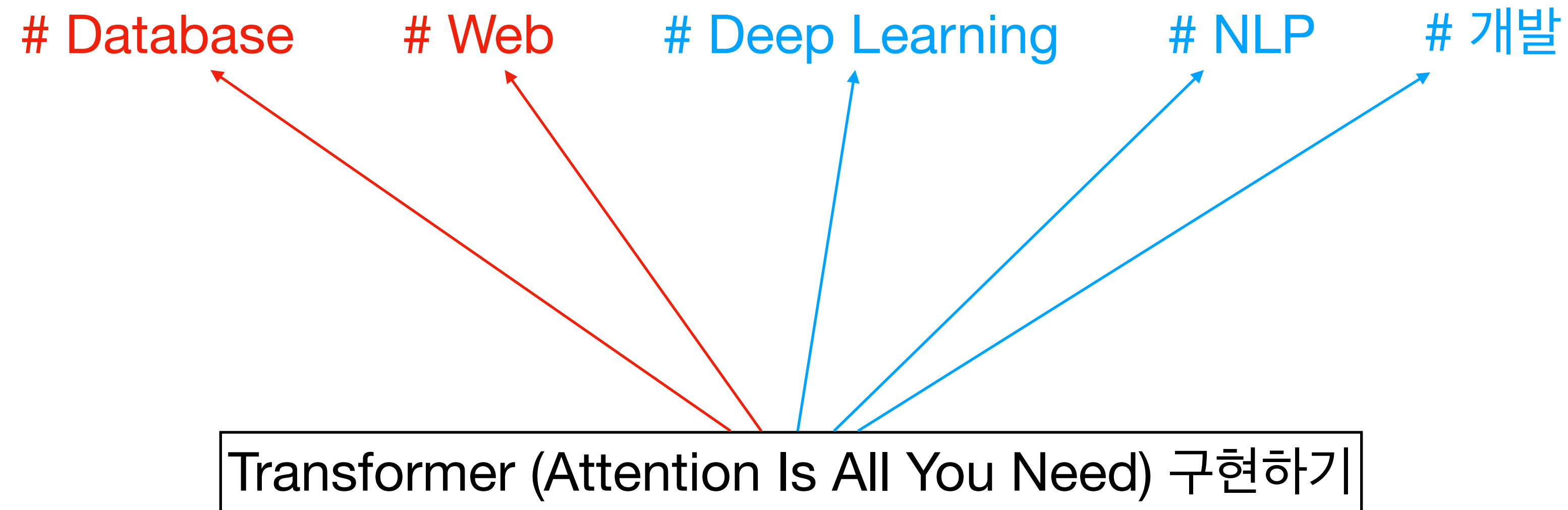
Multi Class Classification

- Language Detection (한국어, 영어, 일본어, 중국어, ...)
- Intent classification in dialog system



Multi Label Classification

- Blog hash tag classification



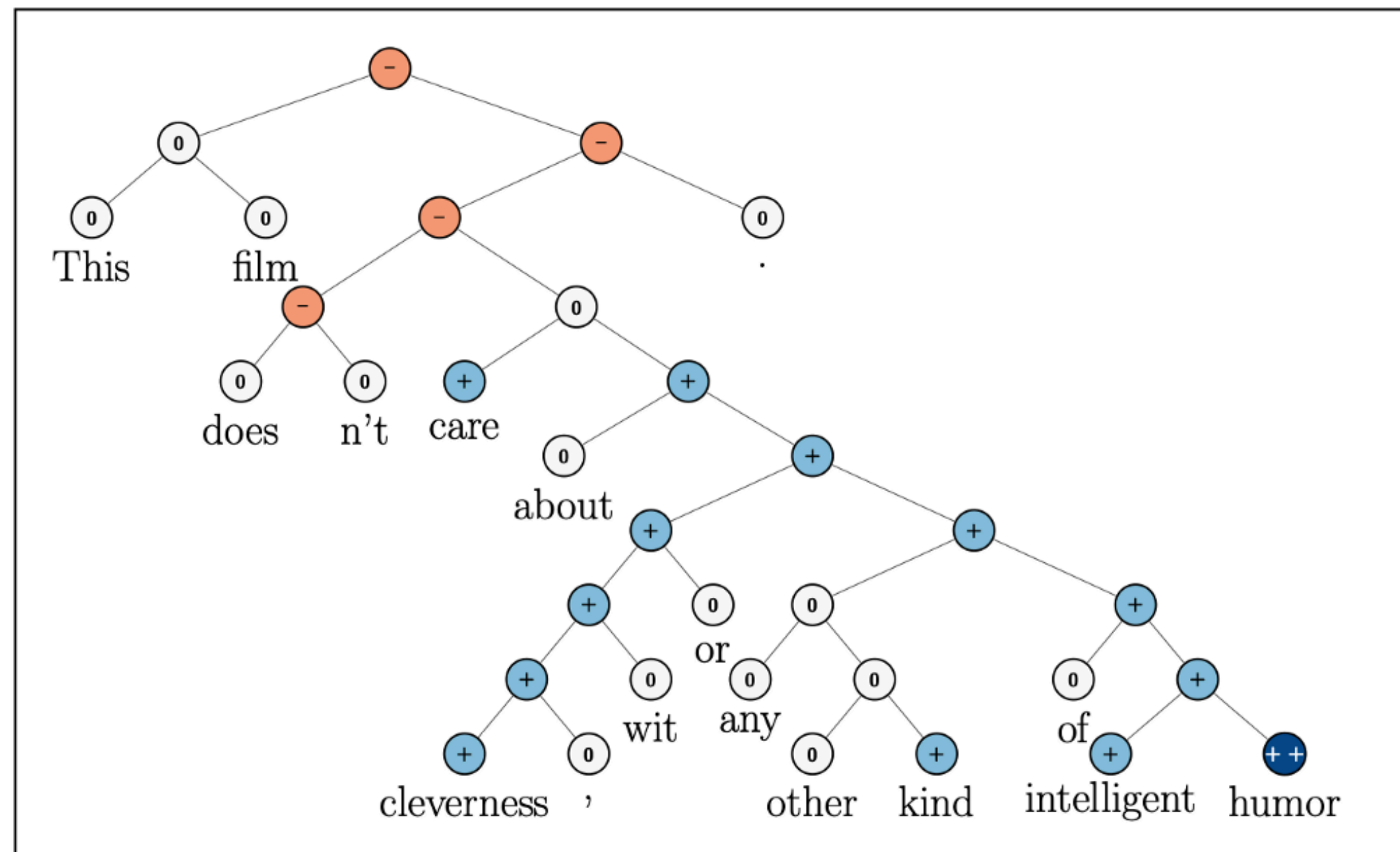
하나의 데이터가 여러개의 정답을 가질 수 있음

Text Classification Dataset

- Sentiment Analysis
 - IMDb Dataset
 - SST-2 Dataset
 - SST-5 Dataset
 - Yelp Review Dataset
 - NSMC
 - Text Classification
 - AG News Corpus
- Internet Movie Database (IMDb) 사이트의 영화 리뷰들을 평점에 따라 분류함
 - 7점 이상: 긍정
 - 4점 이하: 부정
 - 전체 50,000개
 - <https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

Text Classification Dataset

- Sentiment Analysis
 - IMDb Dataset
 - SST-2 Dataset
 - SST-5 Dataset



- Stanford Sentiment Treebank
- Sentence가 Tree구조로 되어있는 Dataset
- 영화 리뷰 text로 구성
- SST-2
 - Binary Class (긍정, 부정)
 - 56,400개
- SST-5
 - Multi Class (5 단계)
 - 94,200개
- <https://www.kaggle.com/atulanandjha/stanford-sentiment-treebank-v2-sst2>
- <https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/overview>

Text Classification Dataset

- Sentiment Analysis
 - IMDb Dataset
 - SST-2 Dataset
 - SST-5 Dataset
 - Yelp Review Dataset
 - NSMC
 - Text Classification
 - AG News Corpus
- 클라우드소싱 리뷰 포럼인 Yelp의 리뷰로 구성
 - 전체 500,000개
 - SST Dataset과 마찬가지로 2가지 종류로 Labeling
 - Binary Class
 - 5 Class
 - <https://www.kaggle.com/yelp-dataset/yelp-dataset>

Text Classification Dataset

- Sentiment Analysis
 - IMDb Dataset
 - SST-2 Dataset
 - SST-5 Dataset
 - Yelp Review Dataset
 - NSMC
- Text Classification
 - AG News Corpus

- 네이버 영화리뷰로 구성 (한국어 데이터)
- 전체 200,000개
- 긍정 / 부정 2가지 종류로 Labeling
- <https://github.com/e9t/nsmc>

Text Classification Dataset

- Sentiment Analysis

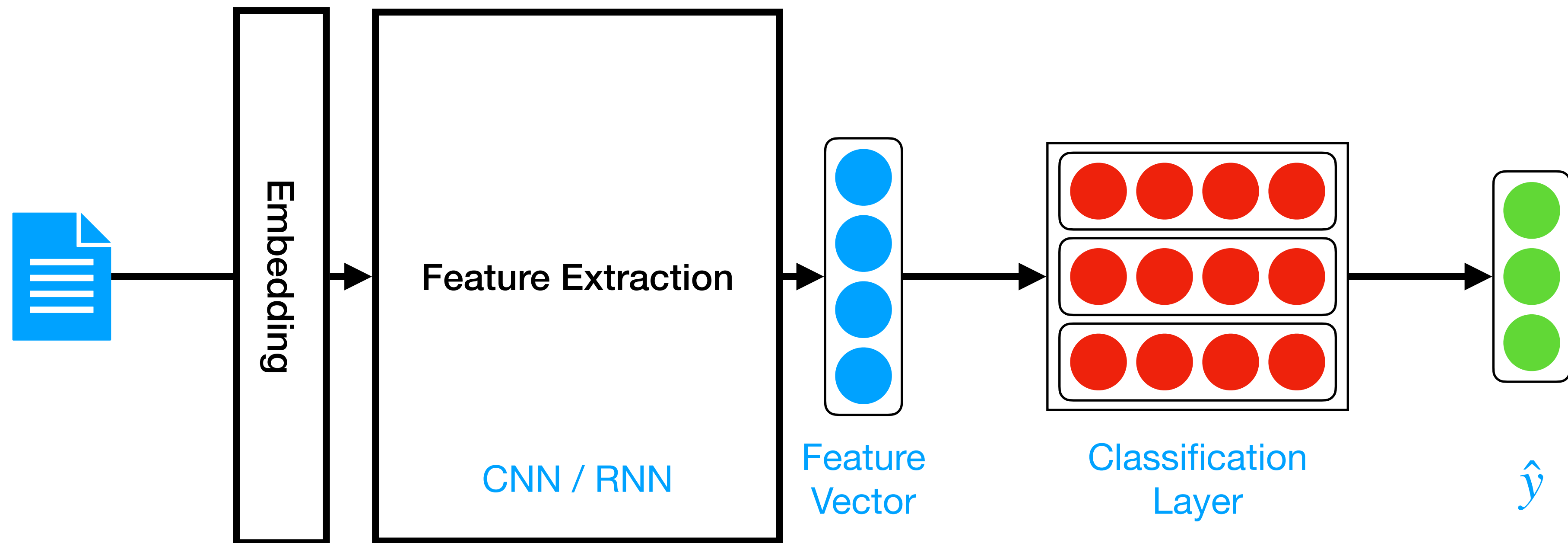
- IMDb Dataset
- SST-2 Dataset
- SST-5 Dataset
- Yelp Review Dataset
- NSMC

- Text Classification

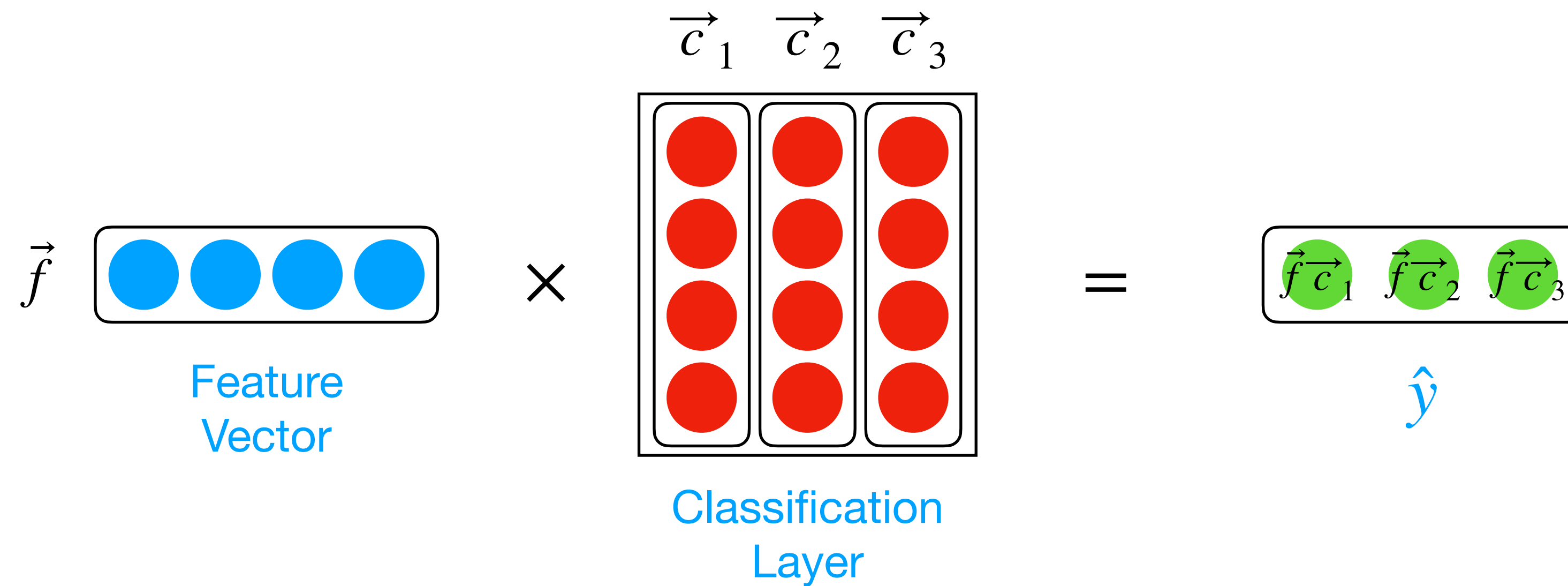
- AG News Corpus

- AG News Corpus로 구축된 데이터셋
- 30,000개의 Training Data
- Class 당 1,900개의 Test Data
- 총 4개의 Class
 - World
 - Sports
 - Business
 - Sci/Tech
- <https://www.kaggle.com/amananandrai/ag-news-classification-dataset>

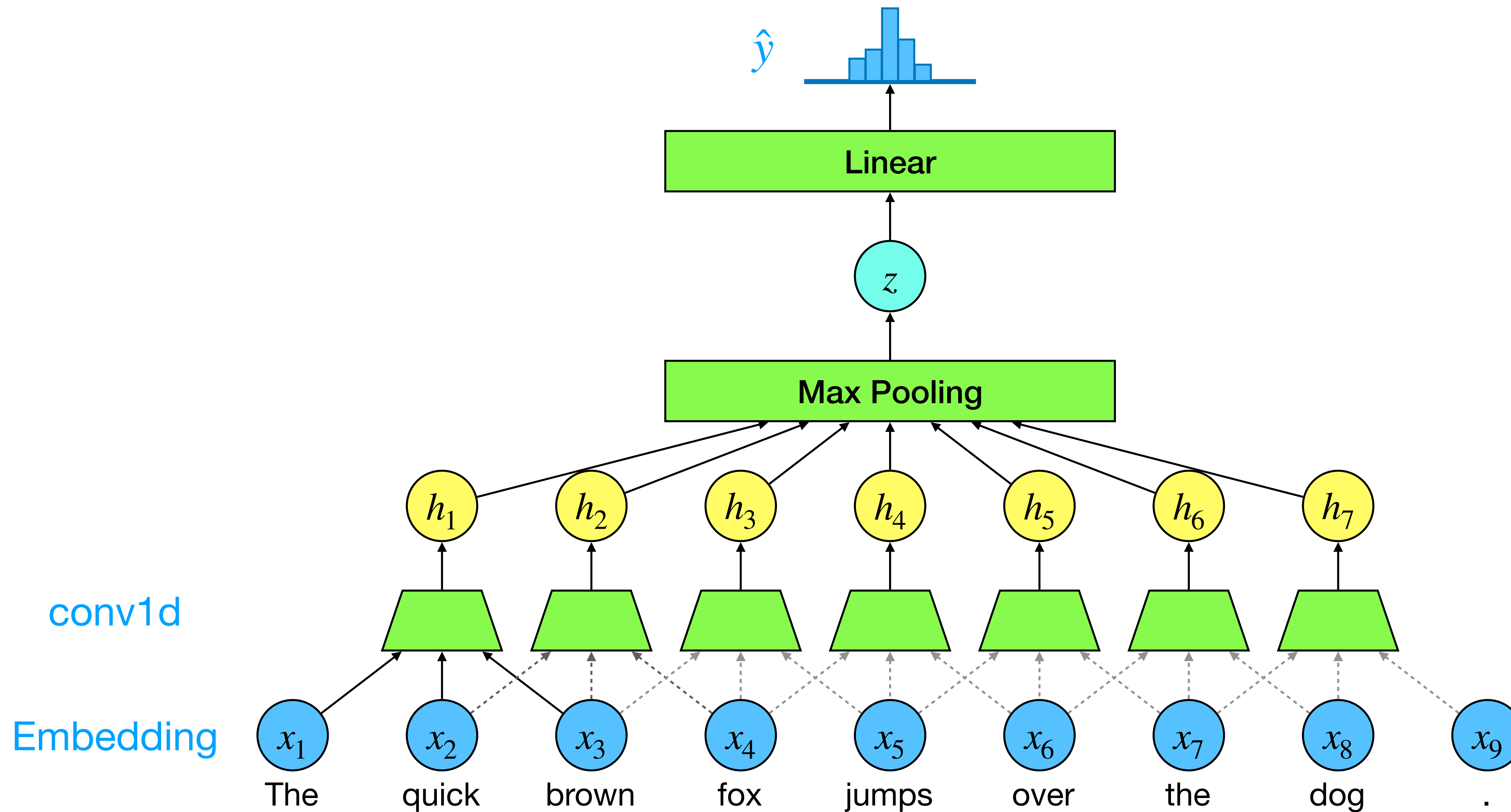
Text Classification Model



Text Classification Model

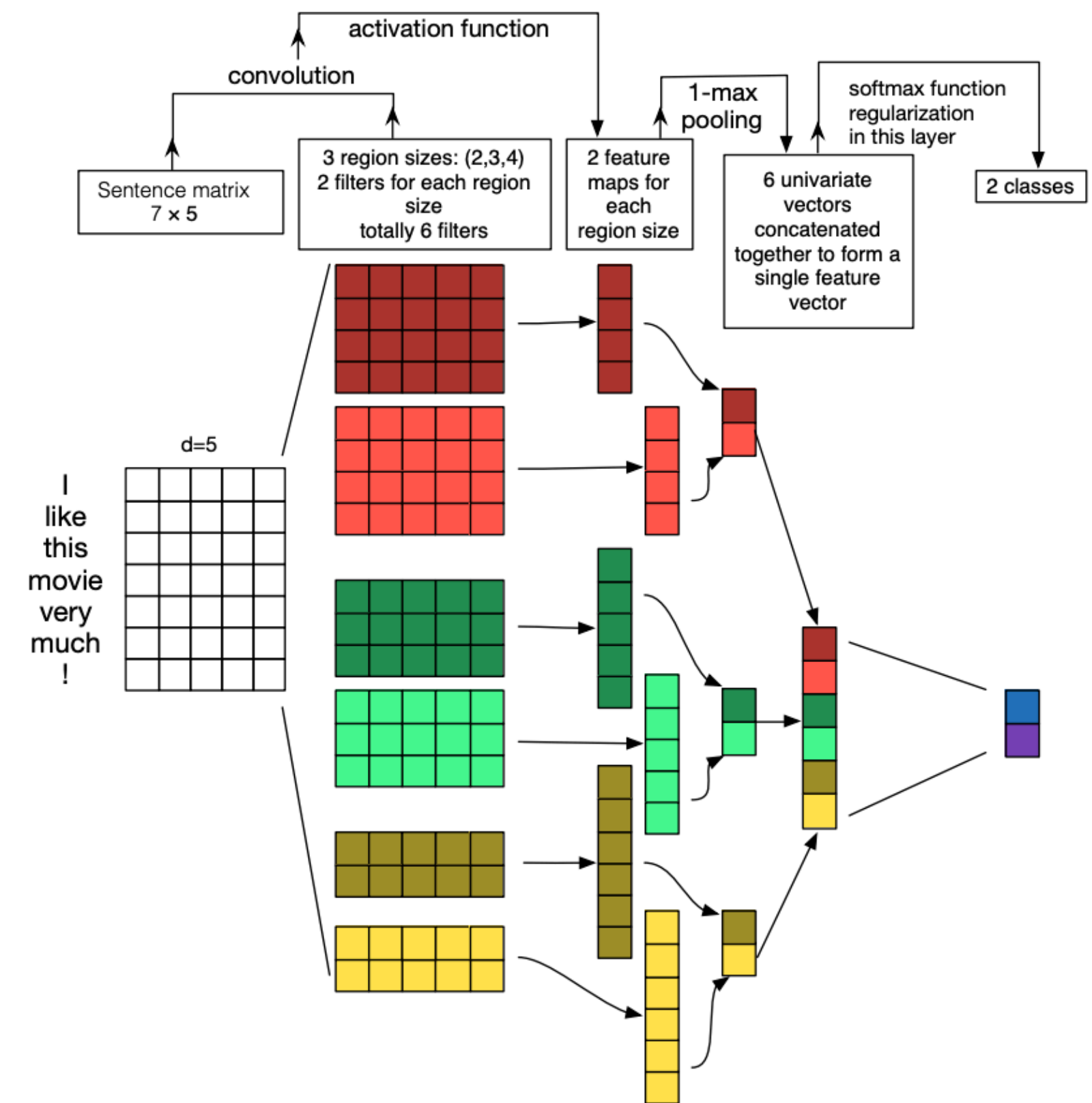


Text Classification Model (CNN)



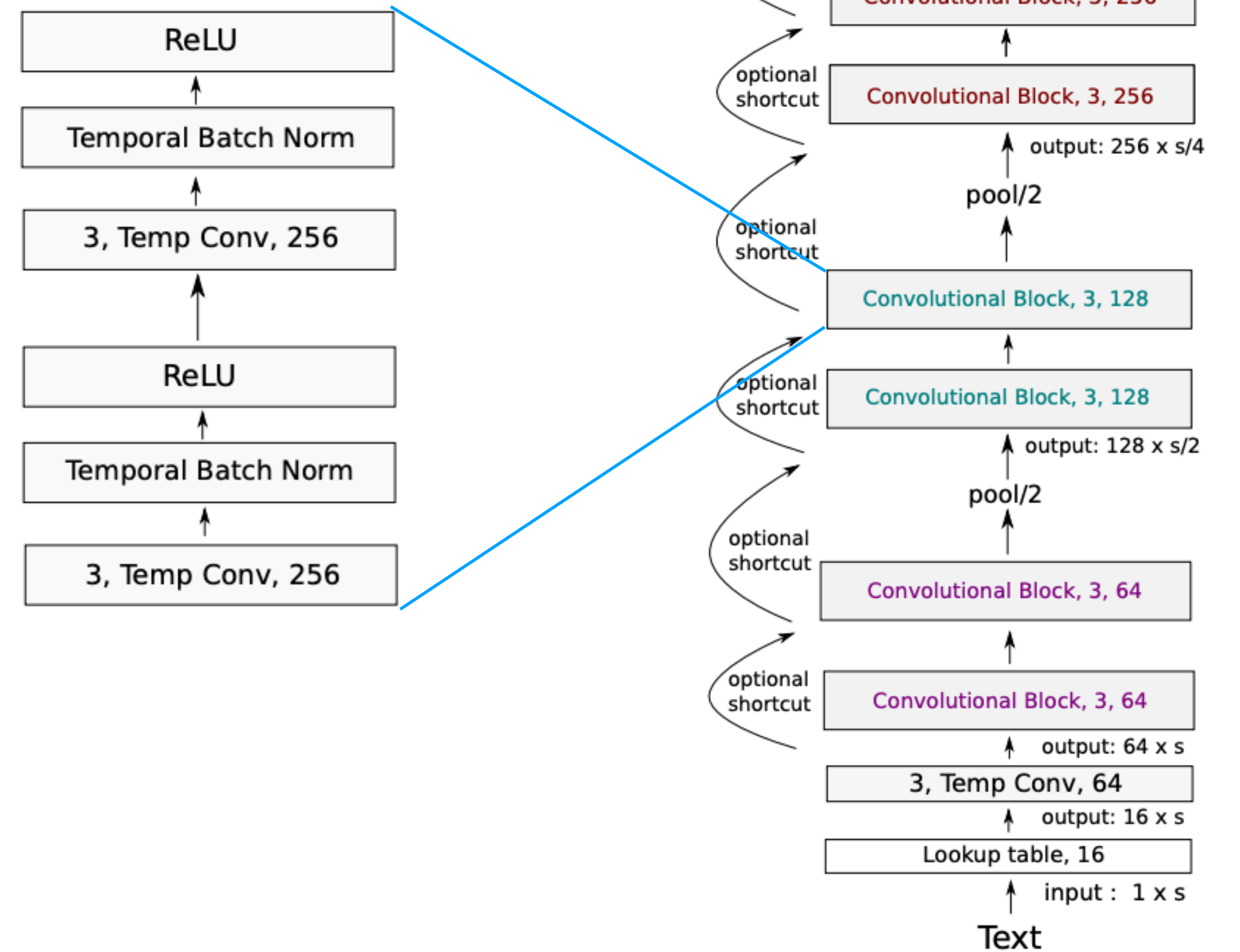
Text Classification Model (CNN)

- A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification
- Ye Zhang et al. 2015
- Convolution layer를 kernel size를 따르게 해서 여러 개 사용함 (2, 3, 4)
- 각각의 결과에 Max pooling을 사용 함
- Max pooling 결과를 Concatenate 함
- Concatenate 한 값에 linear를 취해 최종 예측

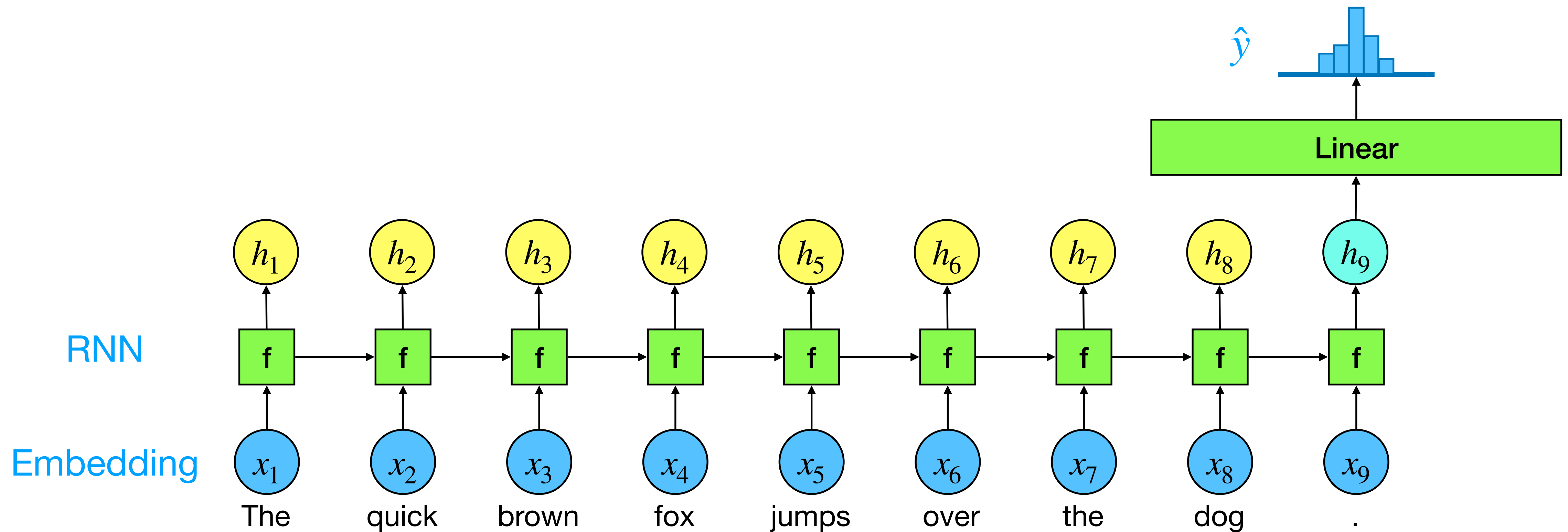


Text Classification Model (CNN)

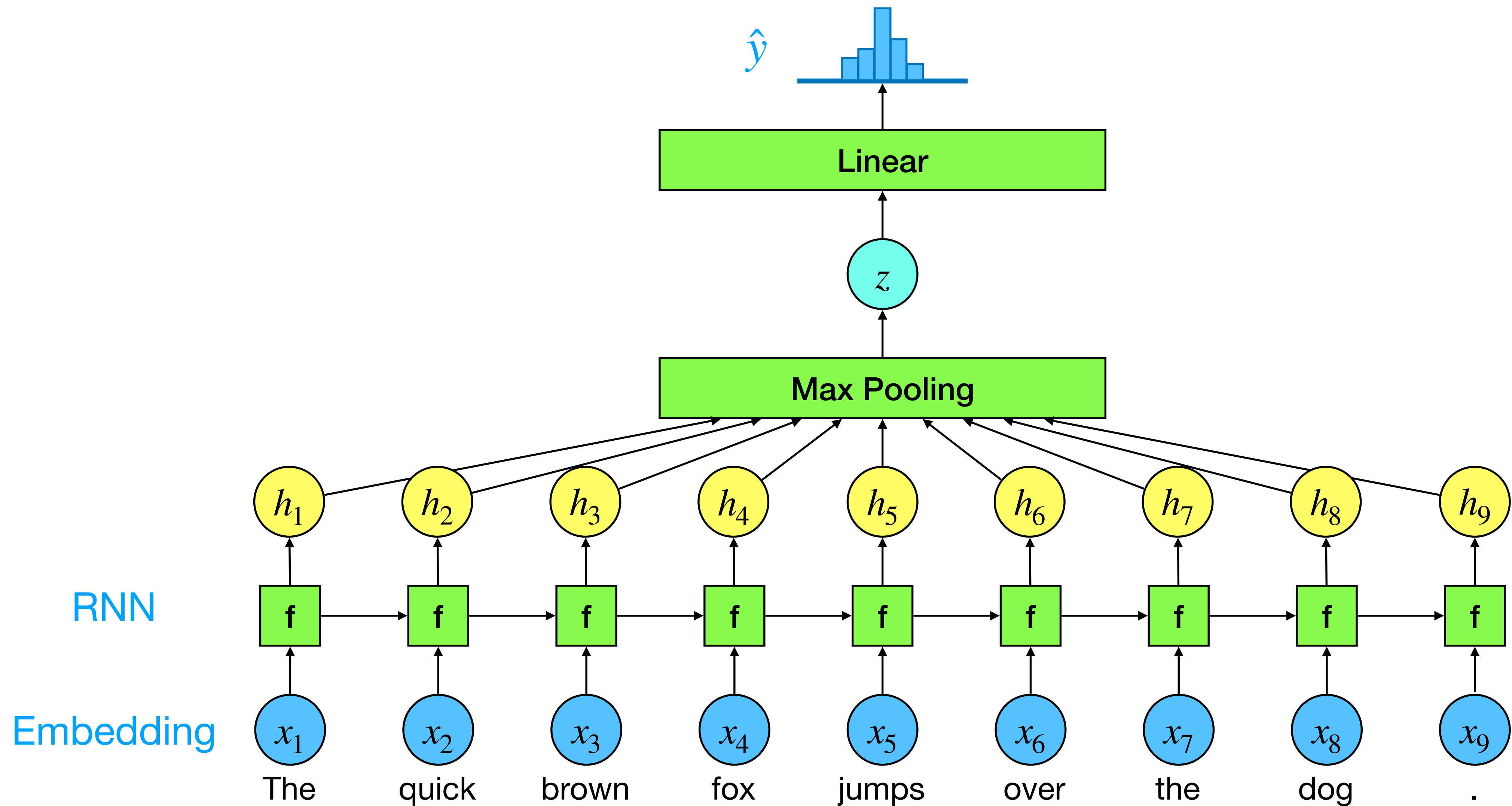
- Very Deep Convolutional Networks for Text Classification
- Conneau et al. 2015
- Convolution layer를 29개 쌓음
- Tokenizer를 character level로 함



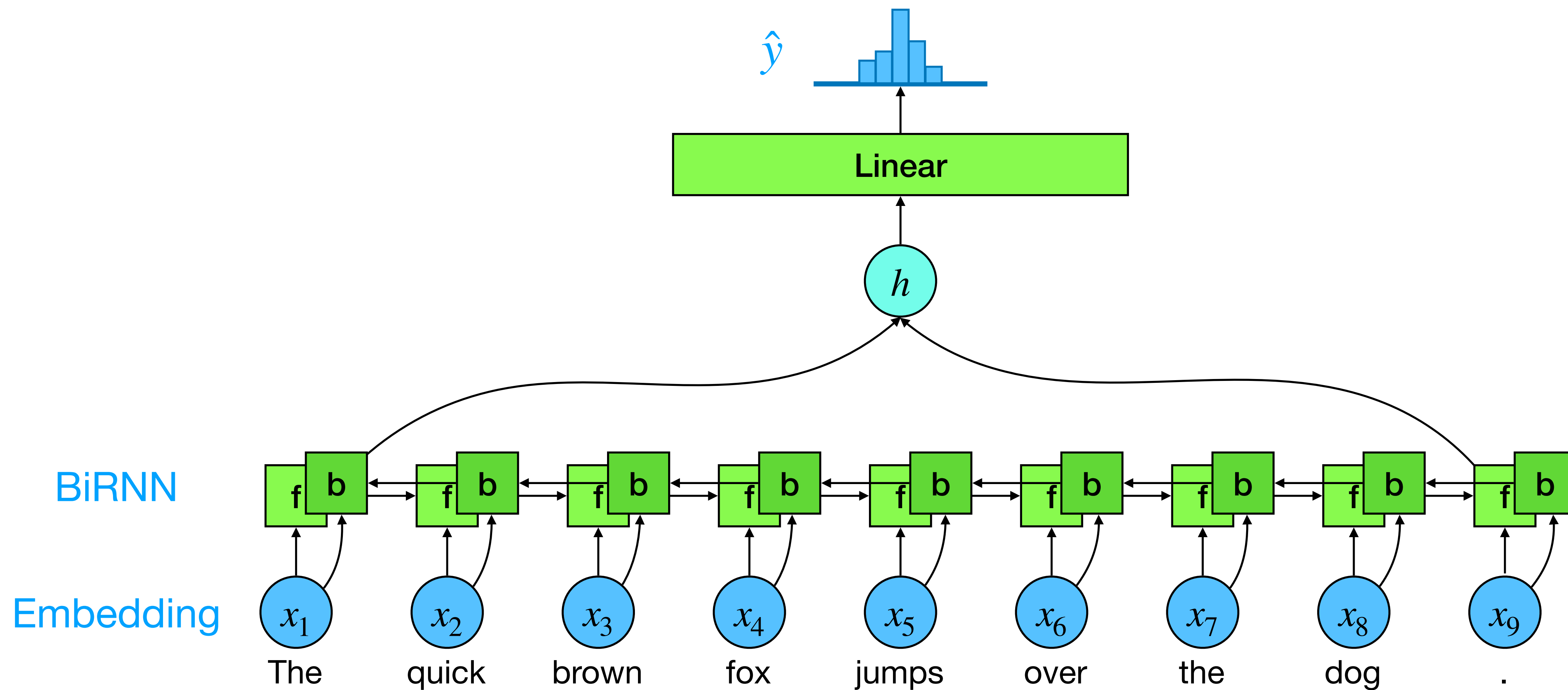
Text Classification Model (RNN)



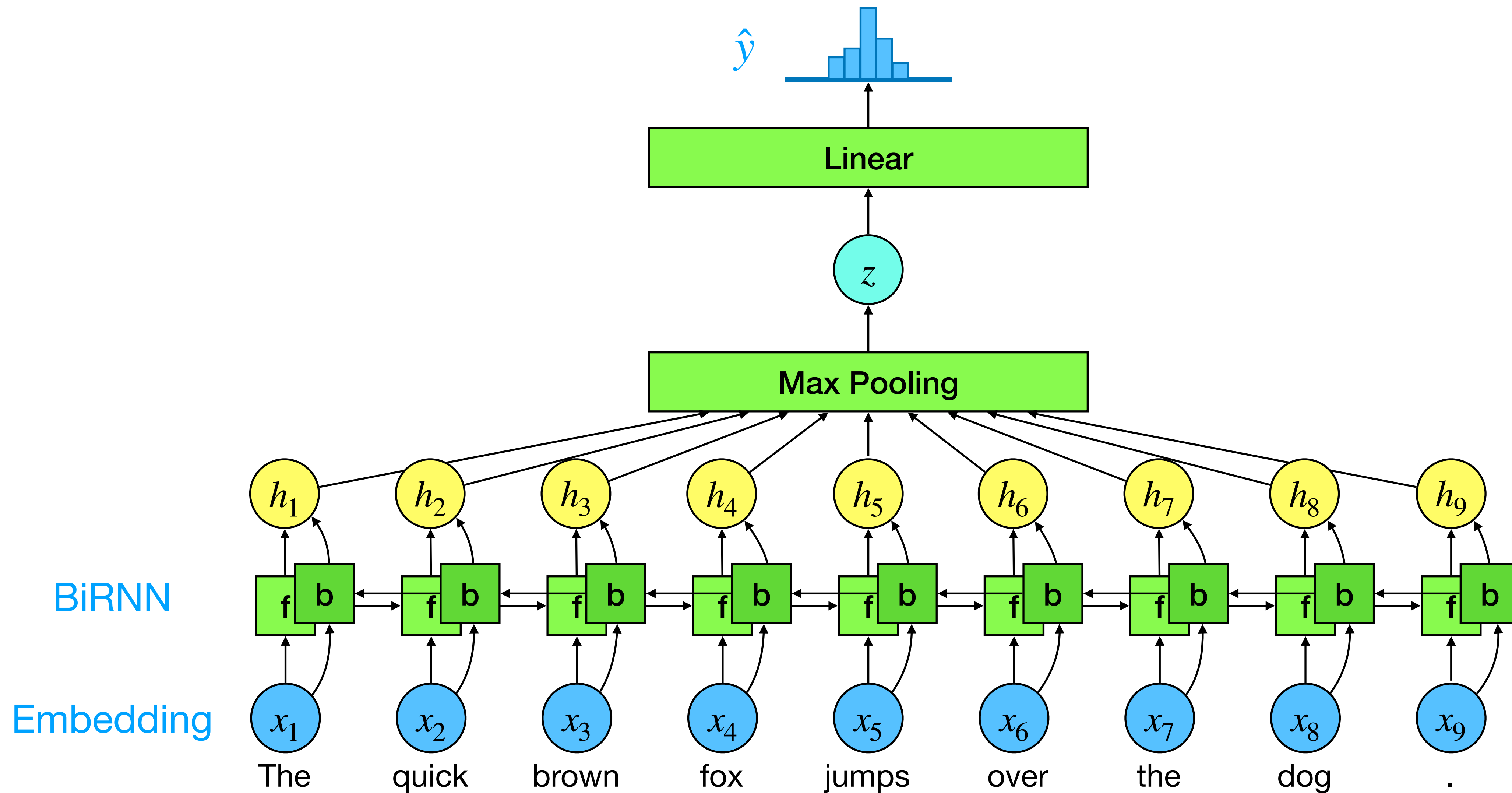
Text Classification Model (RNN)



Text Classification Model (BiRNN)

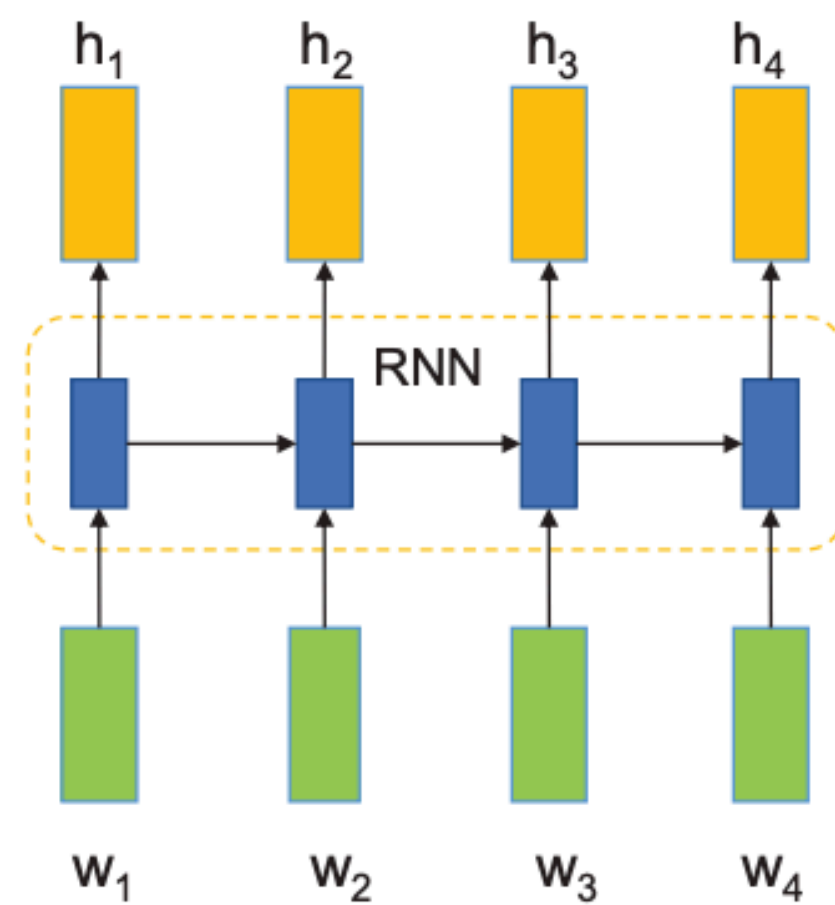


Text Classification Model (BiRNN)

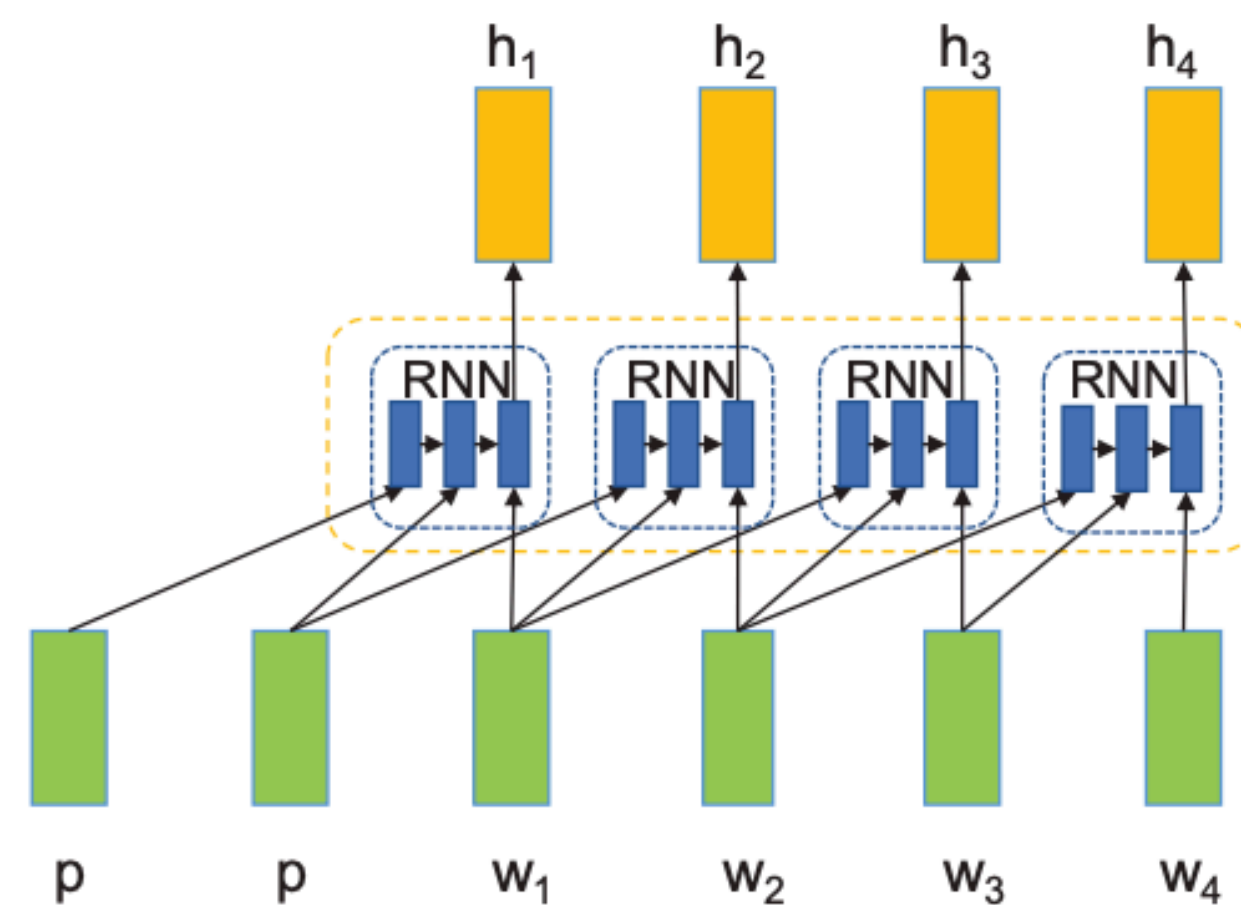


Text Classification Model (RNN)

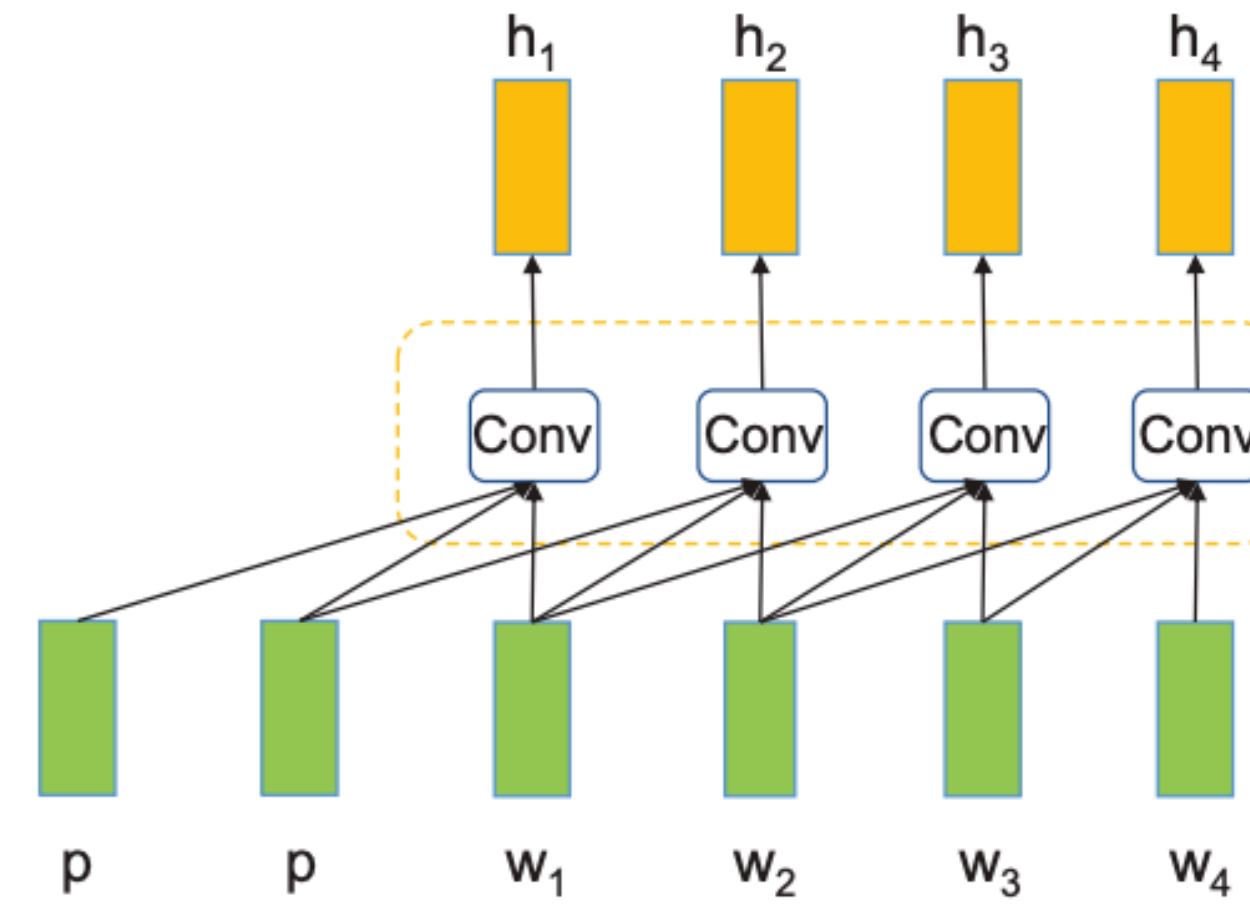
- Disconnected Recurrent Neural Networks for Text Categorization
- Baoxin Wang et al. 2018
- RNN을 CNN과 비슷하게 일부분에만 분할하여 적용



(a) RNN



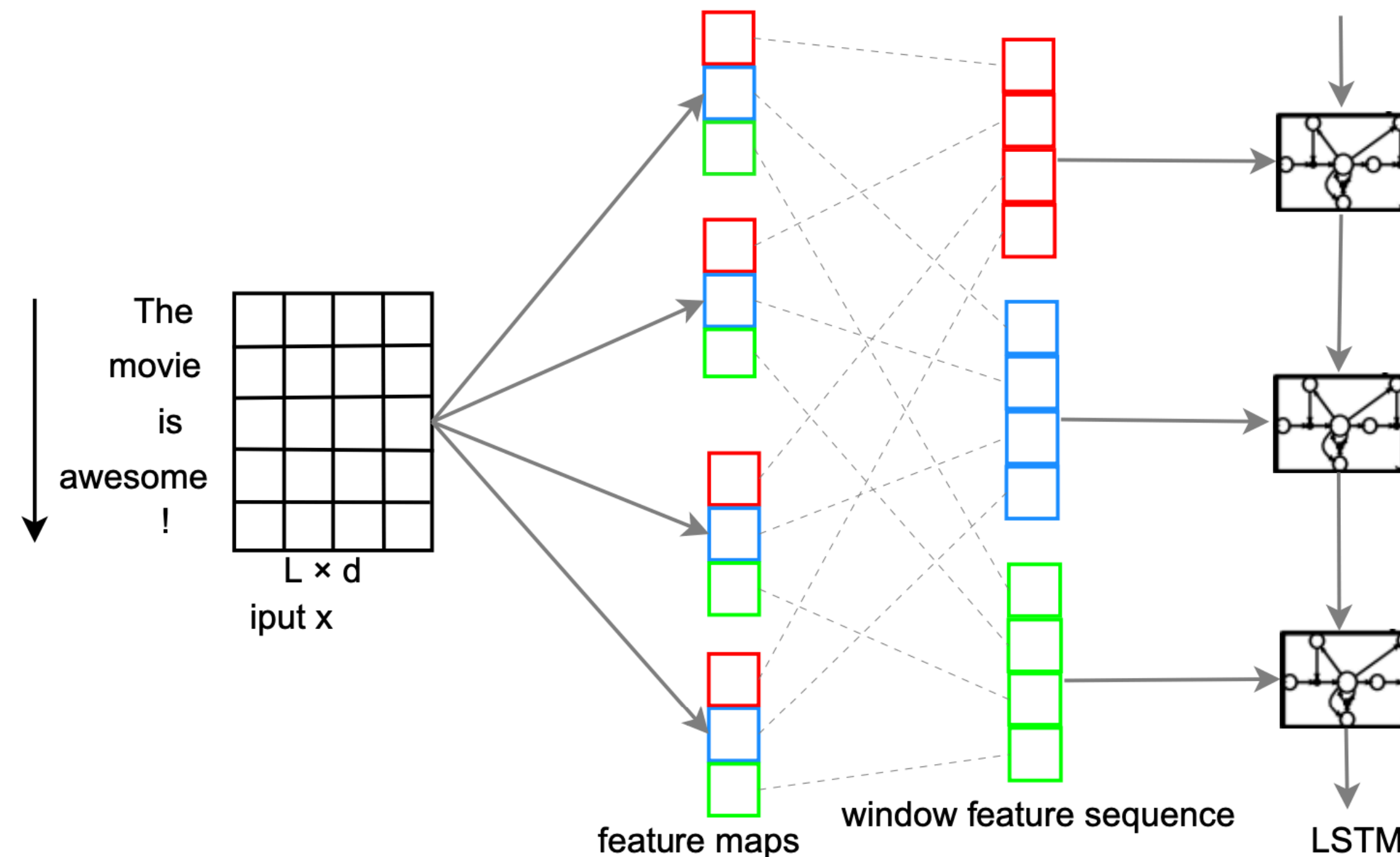
(b) DRNN



(c) CNN

Text Classification Model (RNN)

- A C-LSTM Neural Network for Text Classification
- Chunting Zhou et al. 2015
- CNN을 이용해 단어의 representation을 생성
- LSTM을 이용해 전체 문장의 representation을 생성

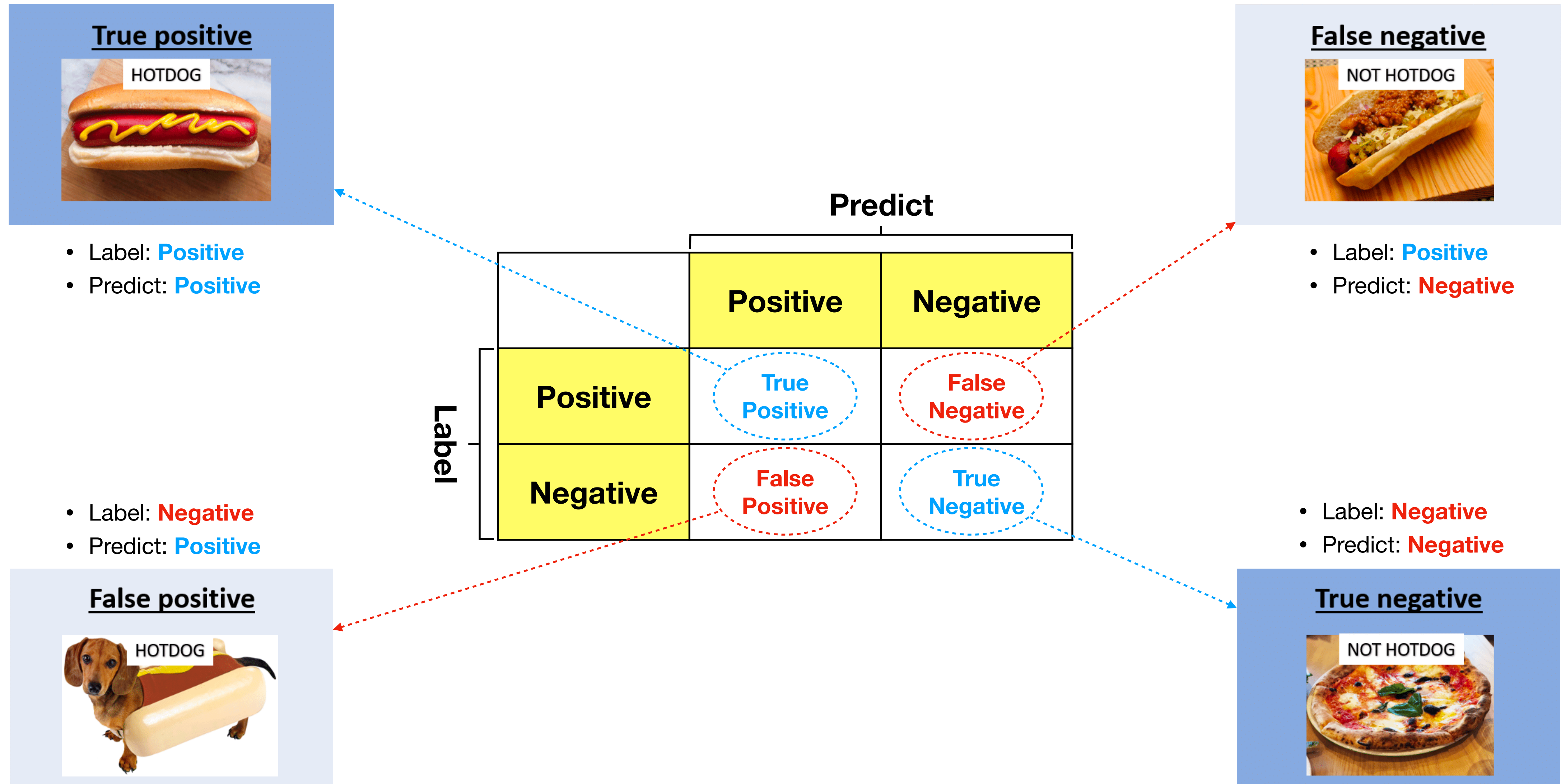


Text Classification Model (SOTA)

State fo Art
Text Classification

<https://paperswithcode.com/task/text-classification>

Evaluation Metric (Confusion Matrix)



Evaluation Metric (Accuracy)

정답 비율

$$\begin{aligned} accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= \frac{T}{T + F} \end{aligned}$$

		Predict	
		Positive	Negative
Label	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Evaluation Metric (Precision)

Positive로 예측한 값중 실제 Positive 비율

$$precision = \frac{TP}{TP + FP}$$

		Predict	
		Positive	Negative
Label	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Evaluation Metric (Recall)

실제 Positive중 Positive로 예측한 비율 비율

$$recall = \frac{TP}{TP + FN}$$

		Predict	
		Positive	Negative
Label	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Evaluation Metric (F1-score)

Recall 과 precision의 조화평균

$$\begin{aligned} F1 &= \frac{2}{\frac{1}{P} + \frac{1}{R}} \\ &= 2 \frac{P \cdot R}{P + R} \end{aligned}$$

Imbalance data 측정 시 유용

Evaluation Metric (Example)

$$accuracy = \frac{9,500}{10,000} = 0.95$$

$$precision = \frac{4,700}{4,900} = 0.959$$

$$recall = \frac{4,700}{5,000} = 0.94$$

$$F1 = 2 \frac{0.959 \times 0.94}{0.959 + 0.94} = 0.949$$

		Predict	
		Positive	Negative
Label	Positive	4,700	300
	Negative	200	4,800

Evaluation Metric (Example)

$$accuracy = \frac{9,500}{10,000} = 0.95$$

$$precision = \frac{100}{300} = 0.333$$

$$recall = \frac{100}{400} = 0.25$$

$$F1 = 2 \frac{0.333 \times 0.25}{0.333 + 0.25} = 0.286$$

		Predict	
		Positive	Negative
Label	Positive	100	300
	Negative	200	9,400

Evaluation Metric (Example)

$$accuracy = \frac{9,700}{10,000} = 0.97$$

$$precision = \frac{300}{500} = 0.6$$

$$recall = \frac{300}{400} = 0.75$$

$$F1 = 2 \frac{0.6 \times 0.75}{0.6 + 0.75} = 0.667$$

		Predict	
		Positive	Negative
Label	Positive	300	100
	Negative	200	9,400

감사합니다.