삼성전기 AI전문가 양성과정 - 프로젝트 실습 (비영상)

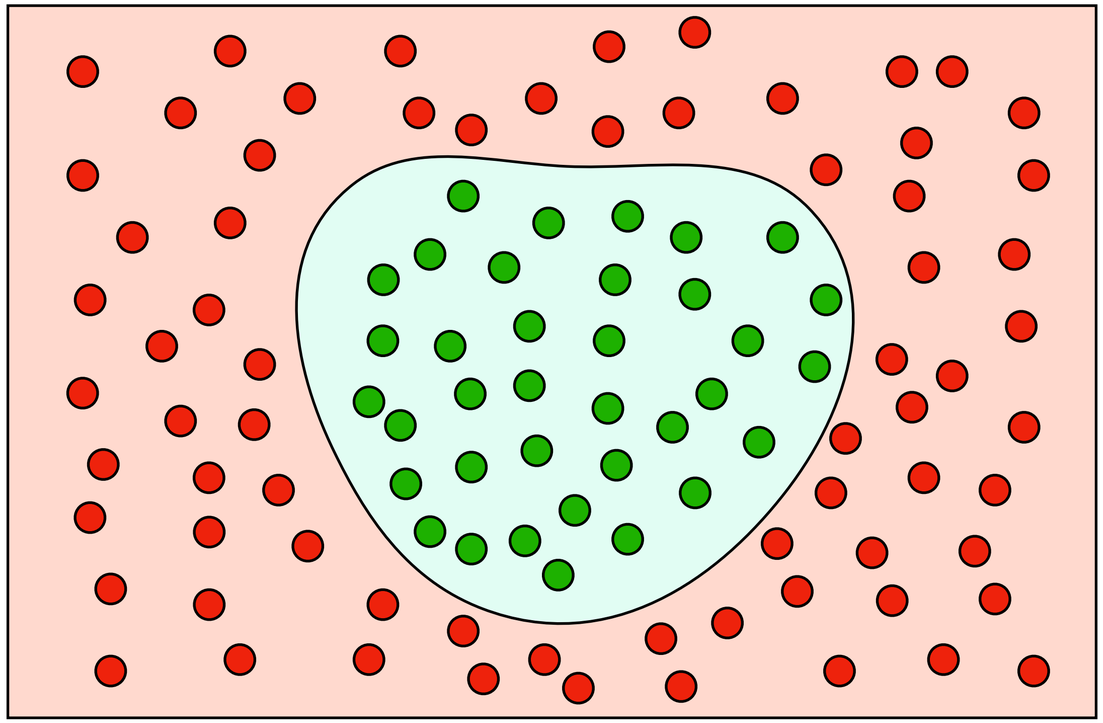
자연어처리를 위한 Text Classification

현청천

2022.02.28

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/CPPG/PMP/CIA/감리사/CISA 개강안내

Binary Classification

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

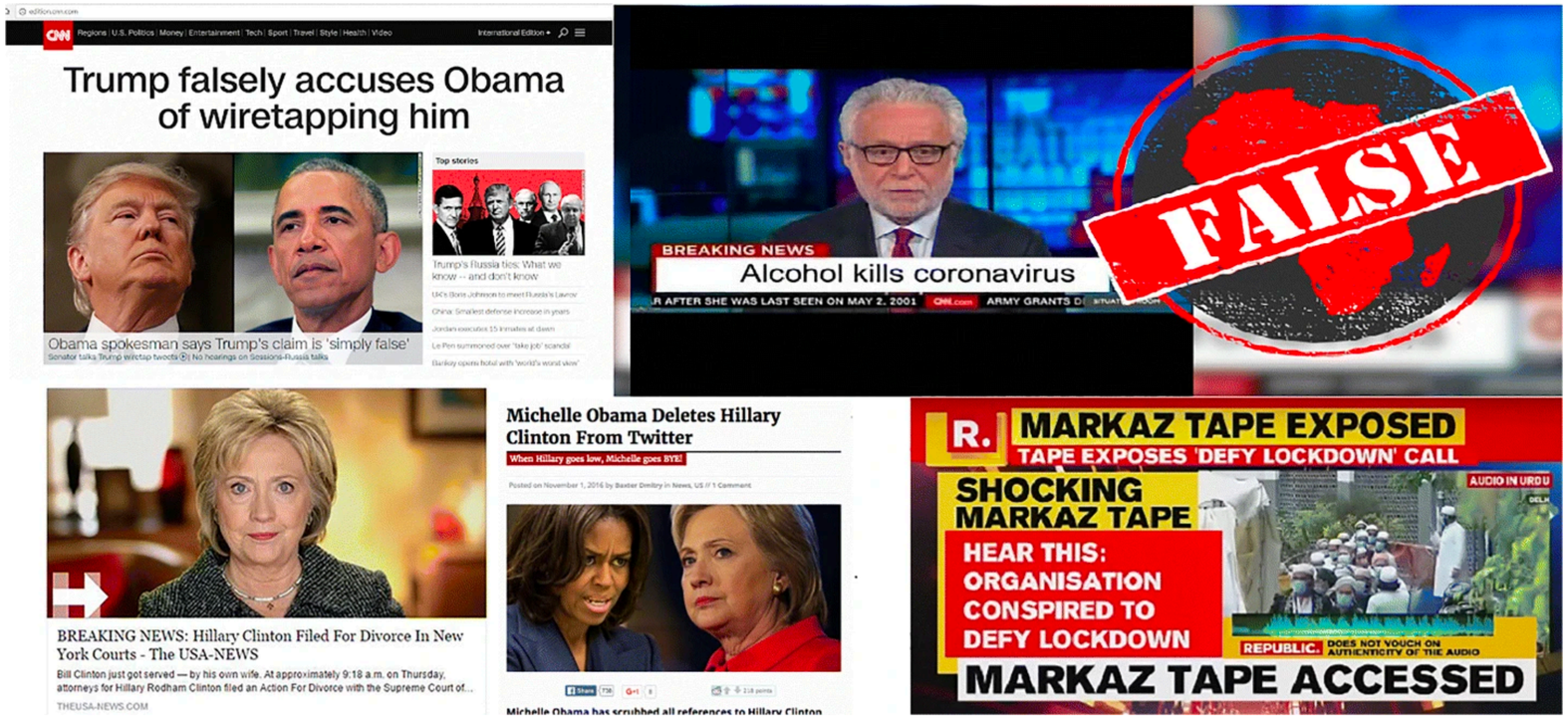
딘따 잼없다!!! 영화를 만들어야지 다큐를 만드냐?! → 부정

강추합니다 남산의 부장들 재밌네요 속이 다 시원함 → 긍정

중후반은 시간이 어찌 흘렀는지...모를정도로 완전 집중몰입!!!! → 긍정

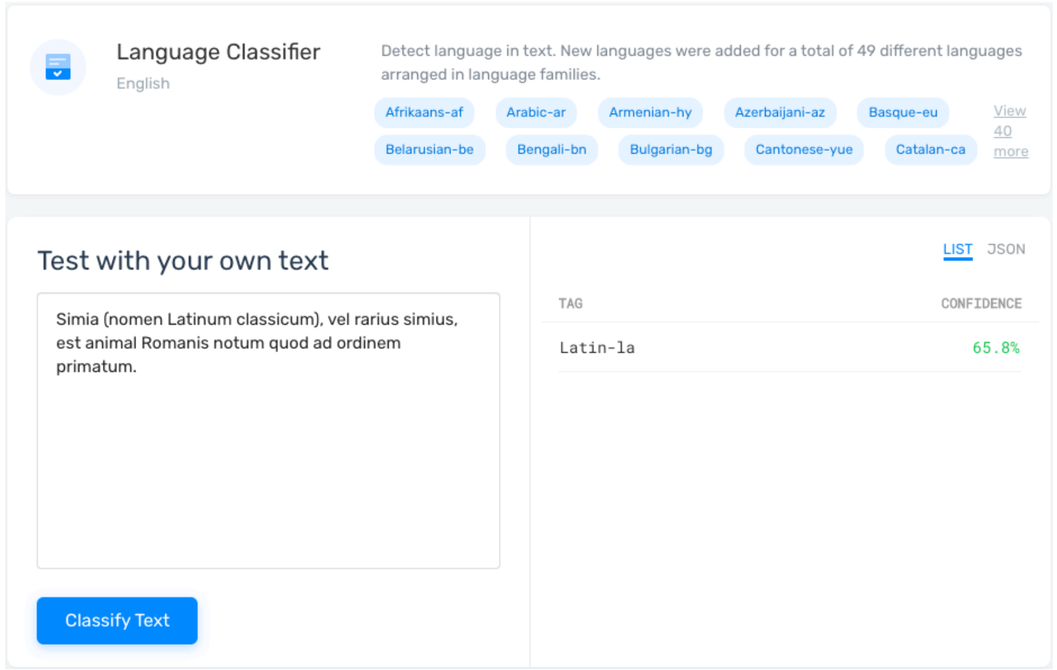
Binary Classification

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



Multi Class Classification

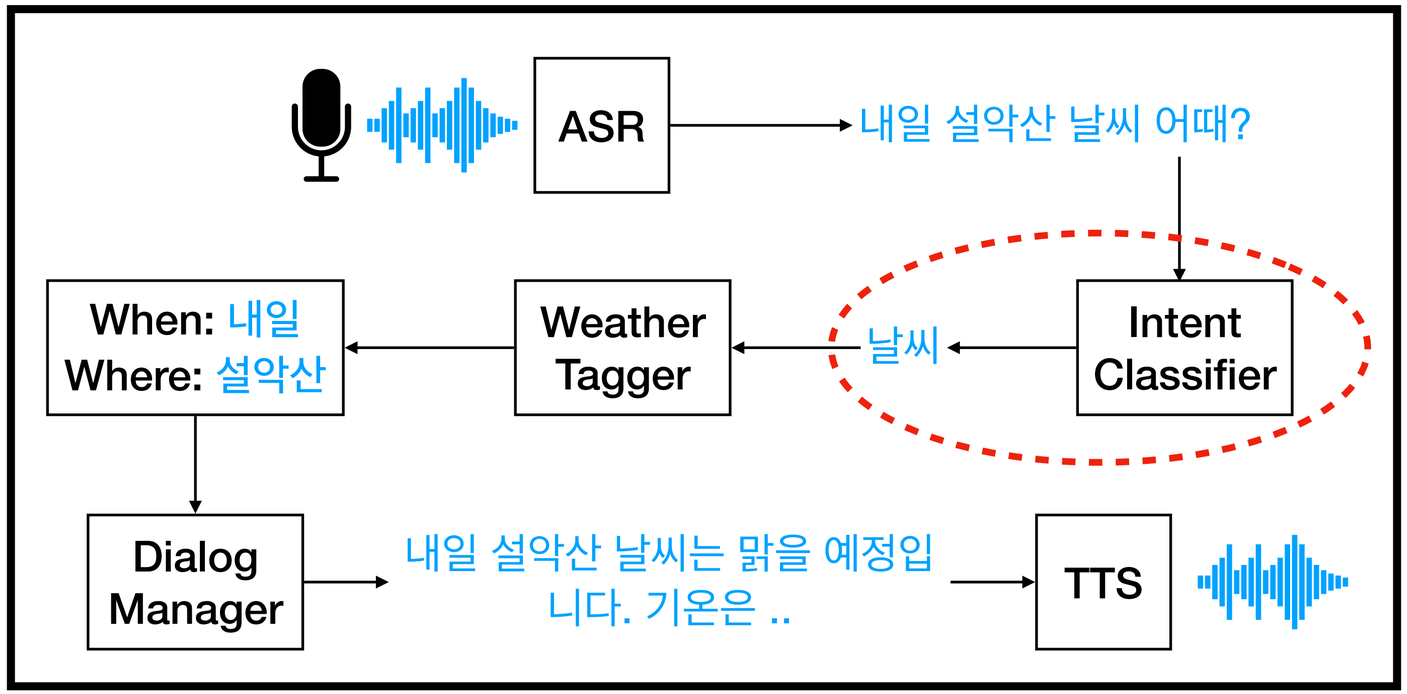
- Language Detection (한국어, 영어, 일본어, 중국어, …)



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

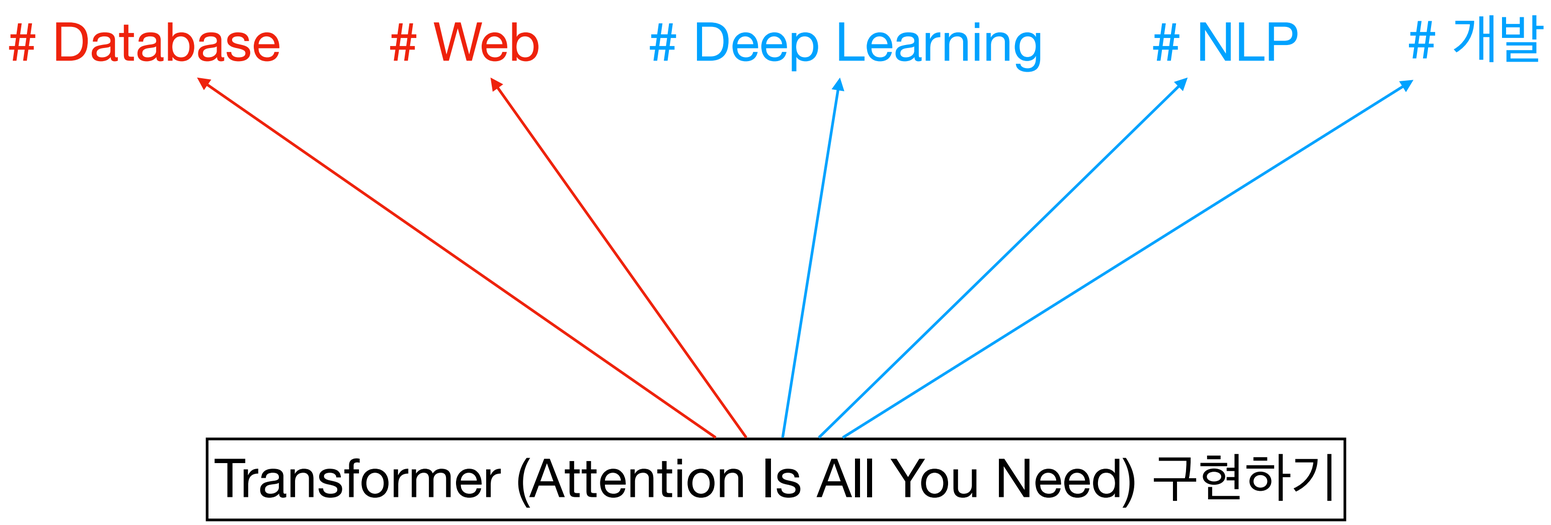
Multi Class Classification

- 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

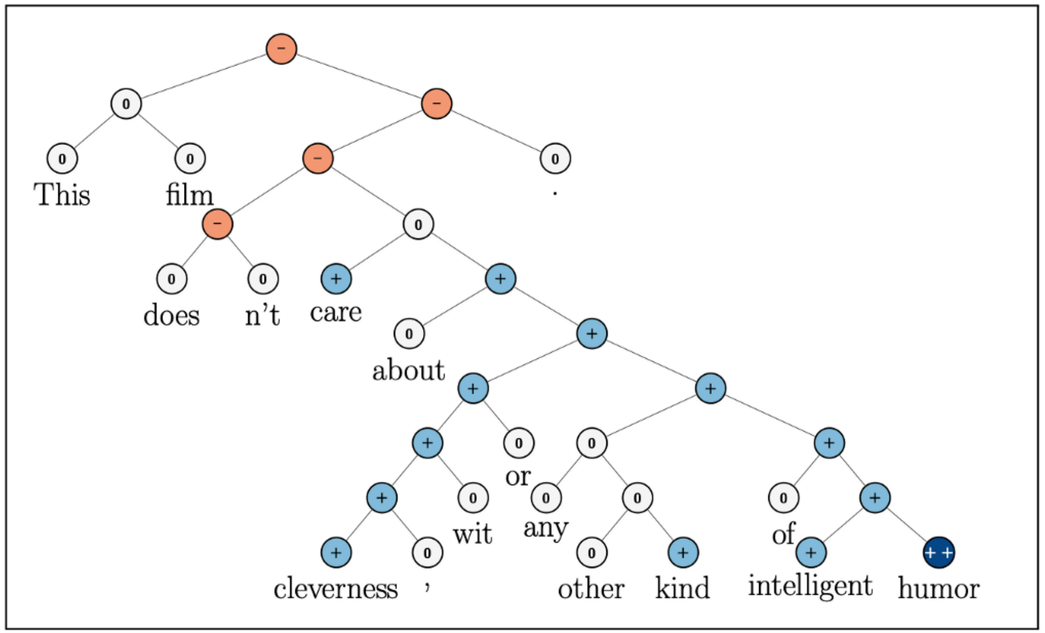
- Sentiment Analysis
  - IMDb Dataset
  - SST-2 Dataset
  - SST-5 Dataset
  - Yelp Review Dataset
  - NSMC
- Text Classification
  - AG News Corpus

- 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

참조: https://github.com/changwookjun/learningspoons/blob/master/Slide/Lecture3.pdf

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

참조: https://github.com/changwookjun/learningspoons/blob/master/Slide/Lecture3.pdf

Text Classification Dataset

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

- 크라우드소싱 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

참조: https://github.com/changwookjun/learningspoons/blob/master/Slide/Lecture3.pdf

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

참조: https://github.com/changwookjun/learningspoons/blob/master/Slide/Lecture3.pdf

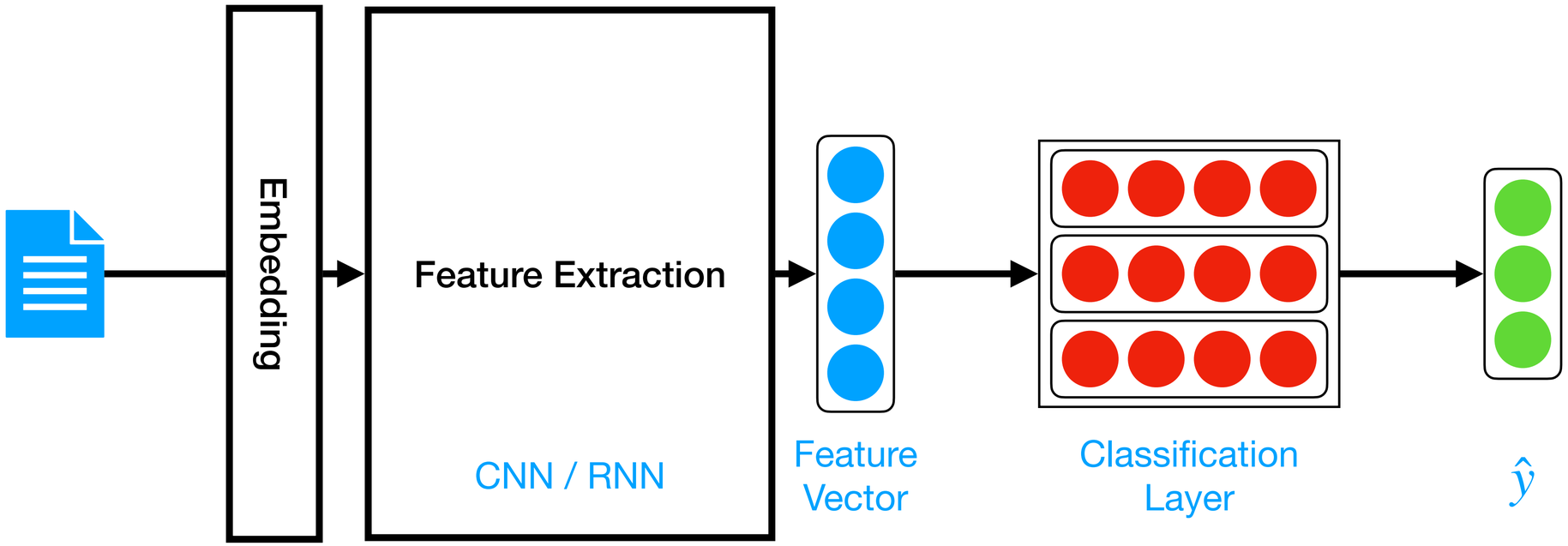
Text Classification Dataset

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

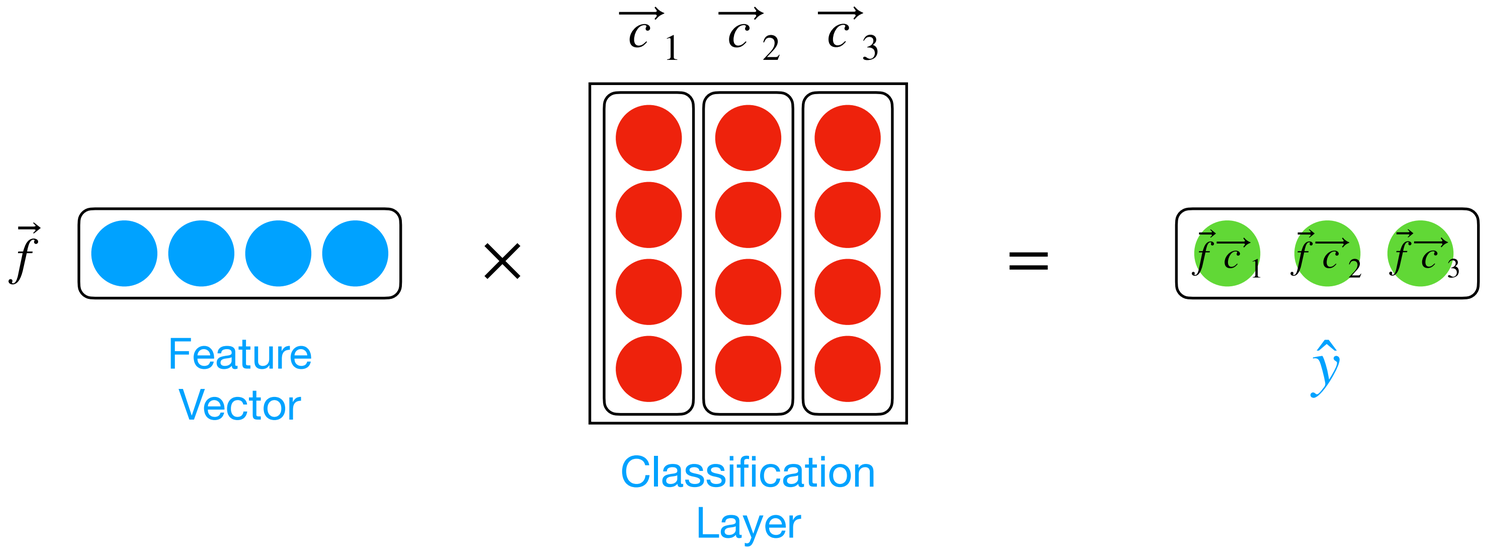
- AG News Corpus로 구축된 데이터셋
- 30,000개의 Training Data
- Class 당 1,900개의 Test Data
- 총 4개의 Class
  - World
  - Sports
  - Business
  - Sci/Tech
- https://www.kaggle.com/amananandrai/ag-news-classification-dataset

참조: https://github.com/changwookjun/learningspoons/blob/master/Slide/Lecture3.pdf

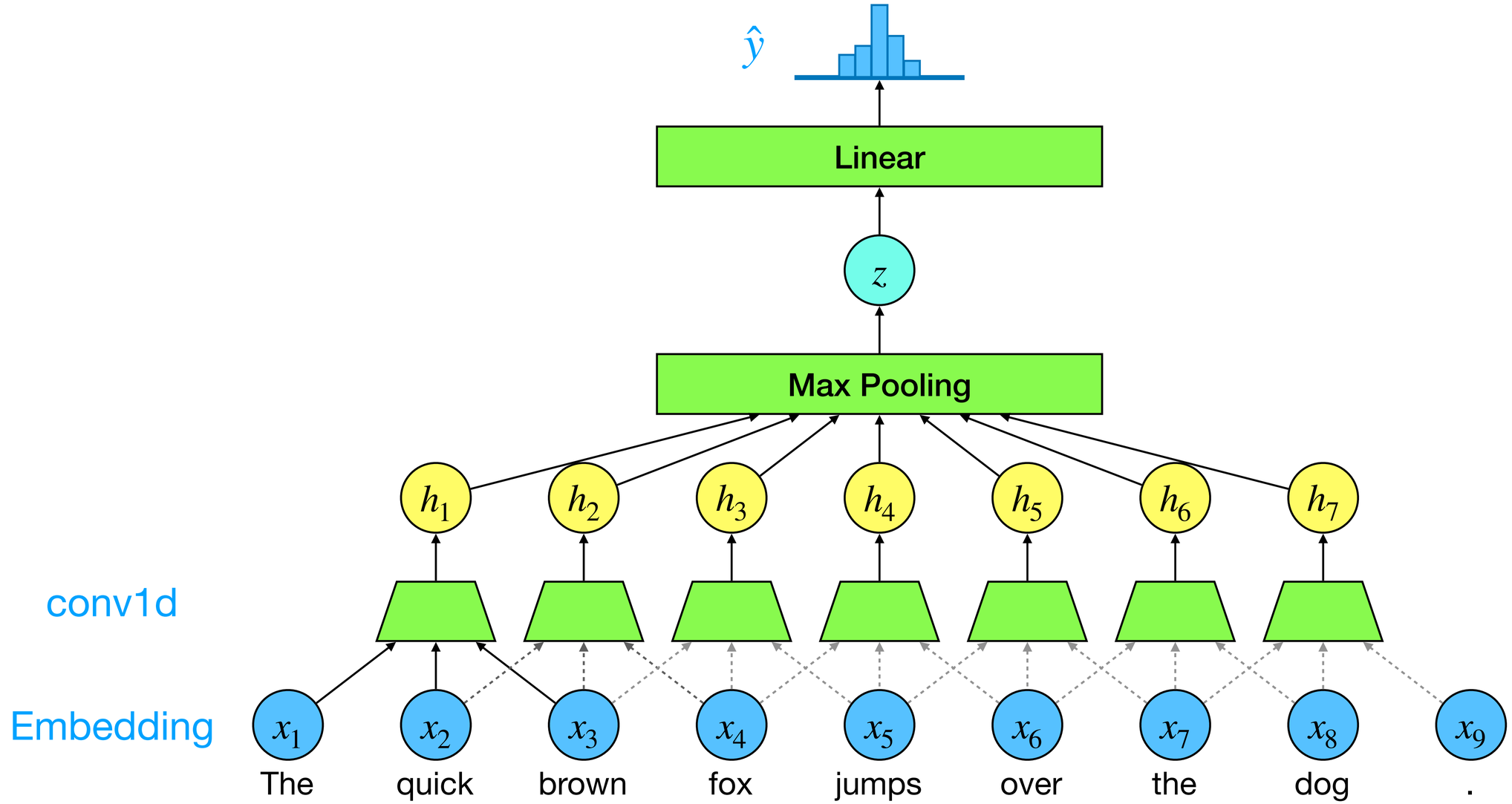
Text Classification Model



Text Classification Model

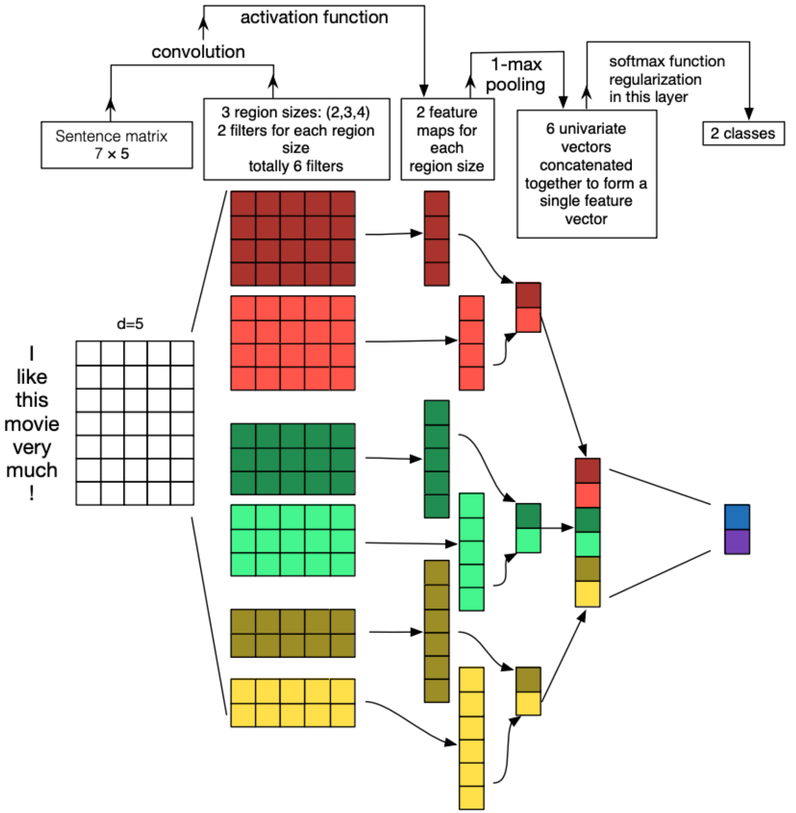


Text Classification Model (CNN)



Text Classification Model (CNN)

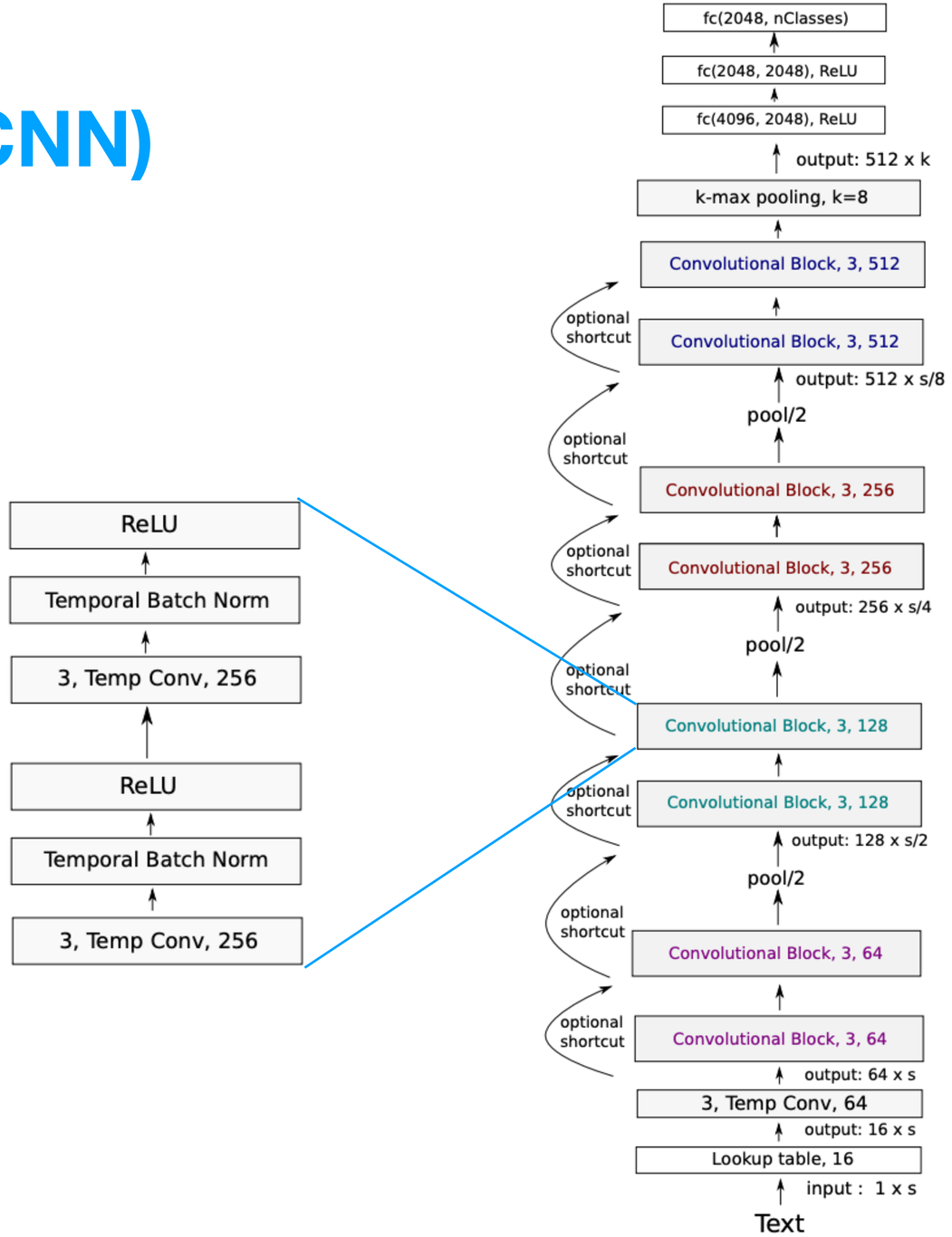
- A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification
- Ye Zhang et al. 2015
- Convolution layer를 kernel size를 따르게 해서 여러 개 사용함 (2, 3, 4)
- 각각의 결과에 Max pooling을 사용 함
- Max pooling 결과를 Concatenate 함
- Concatenate 한 값에 linear를 취해 최종 예측



참조: https://arxiv.org/abs/1510.03820

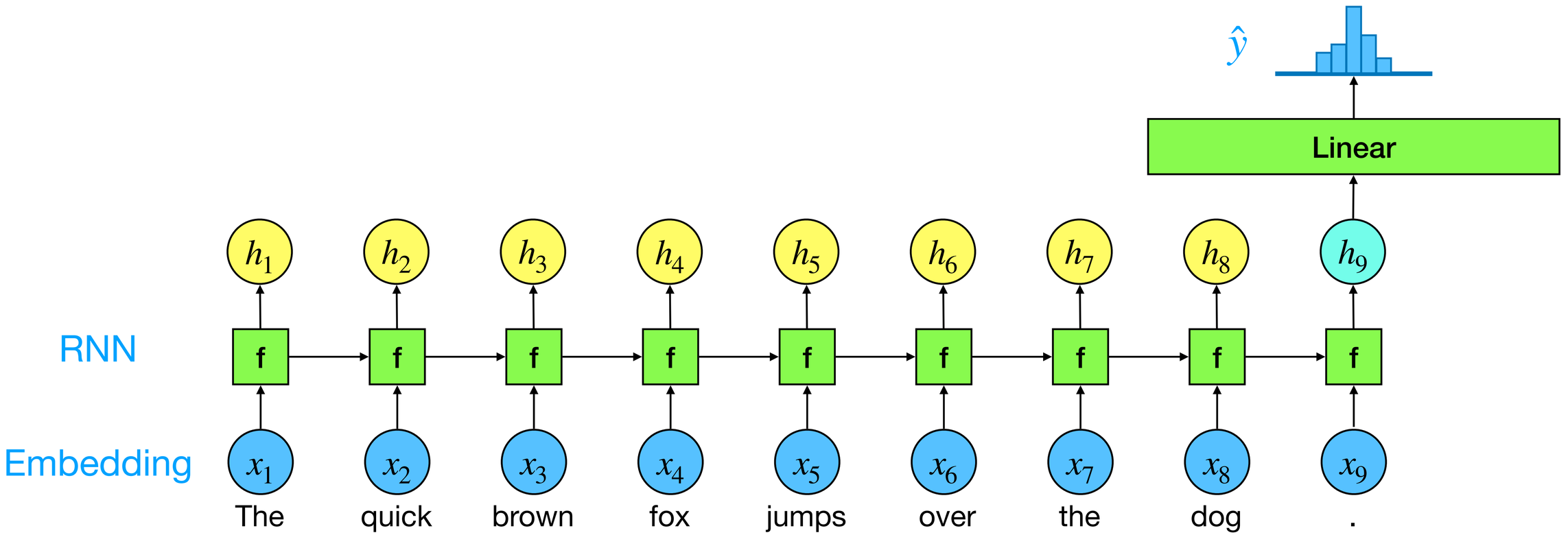
Text Classification Model (CNN)

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

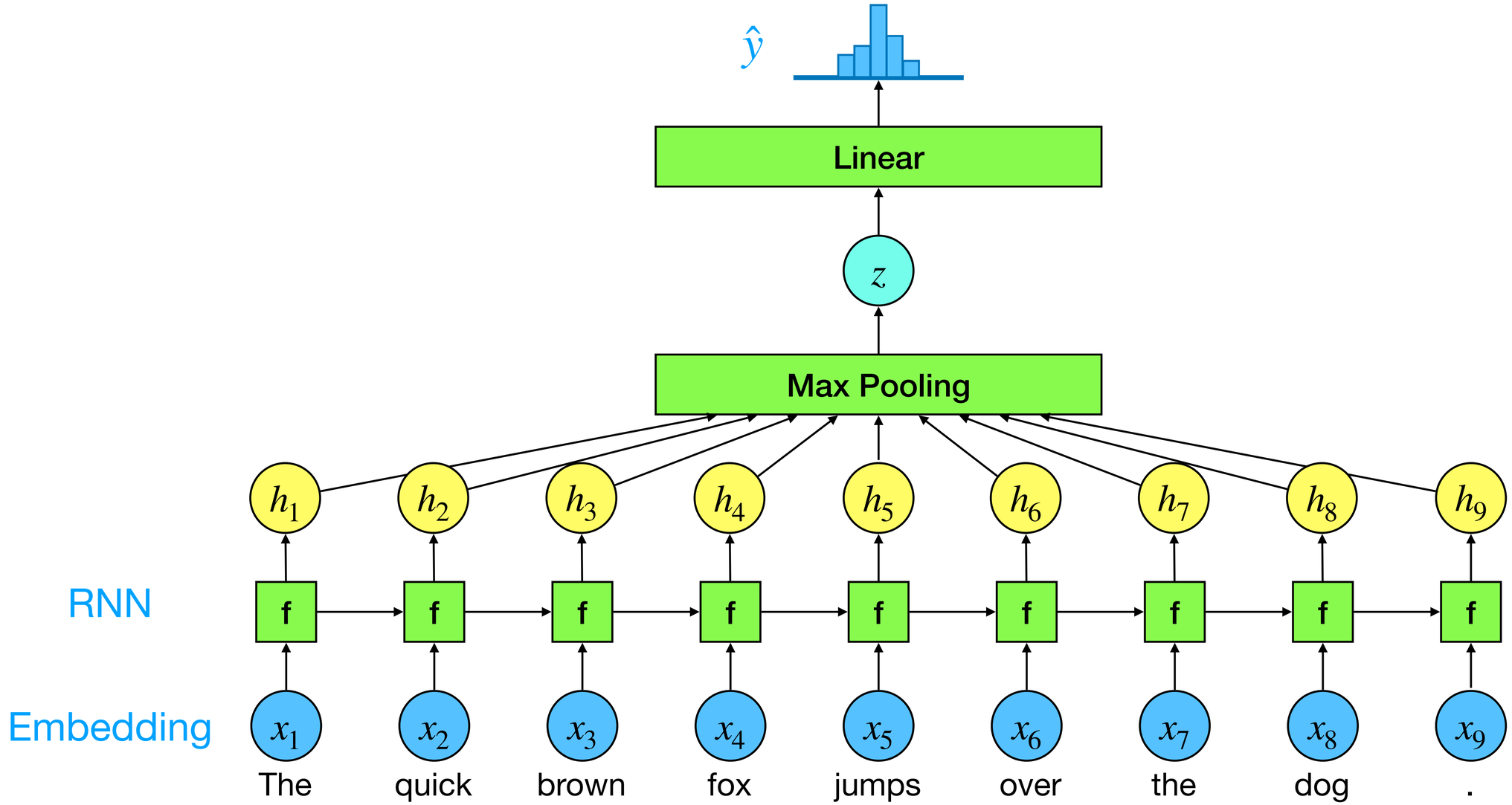


참조: https://arxiv.org/abs/1606.01781

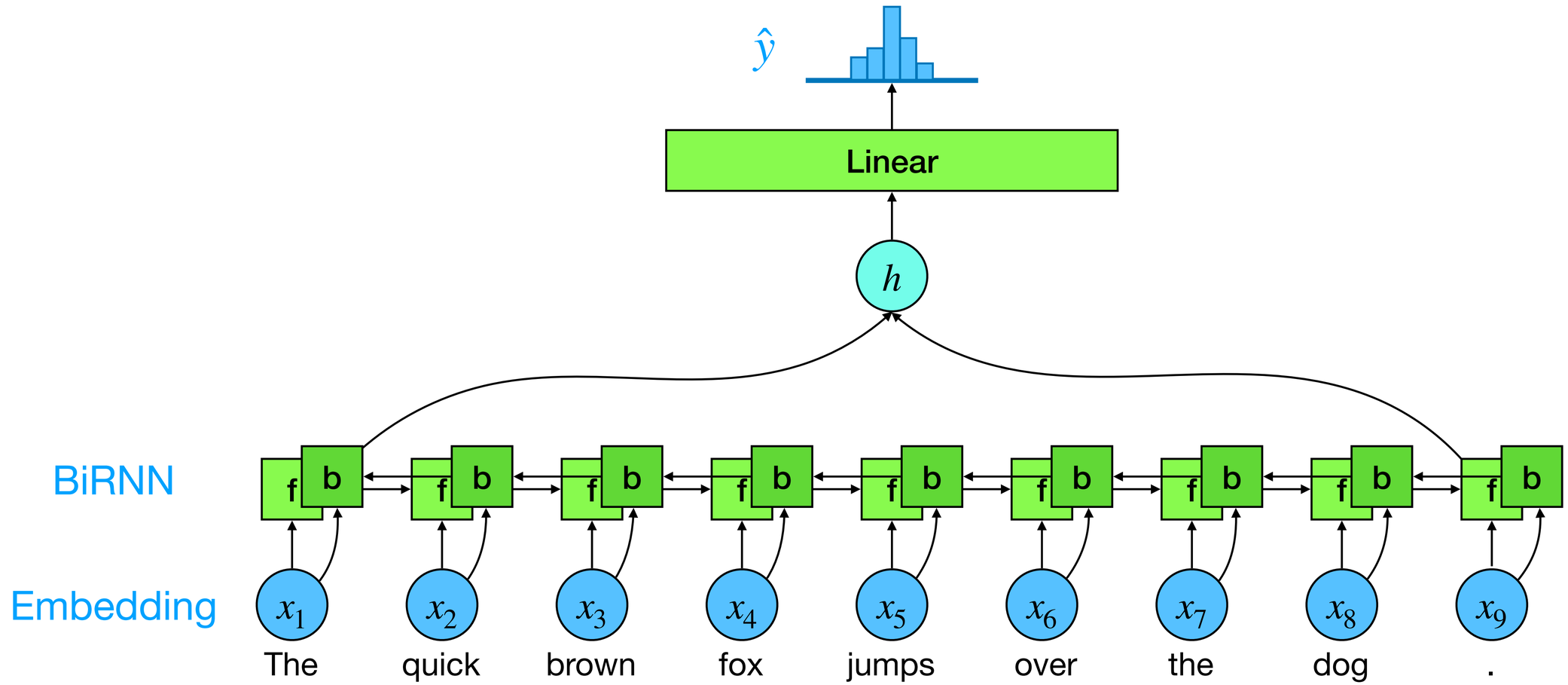
Text Classification Model (RNN)



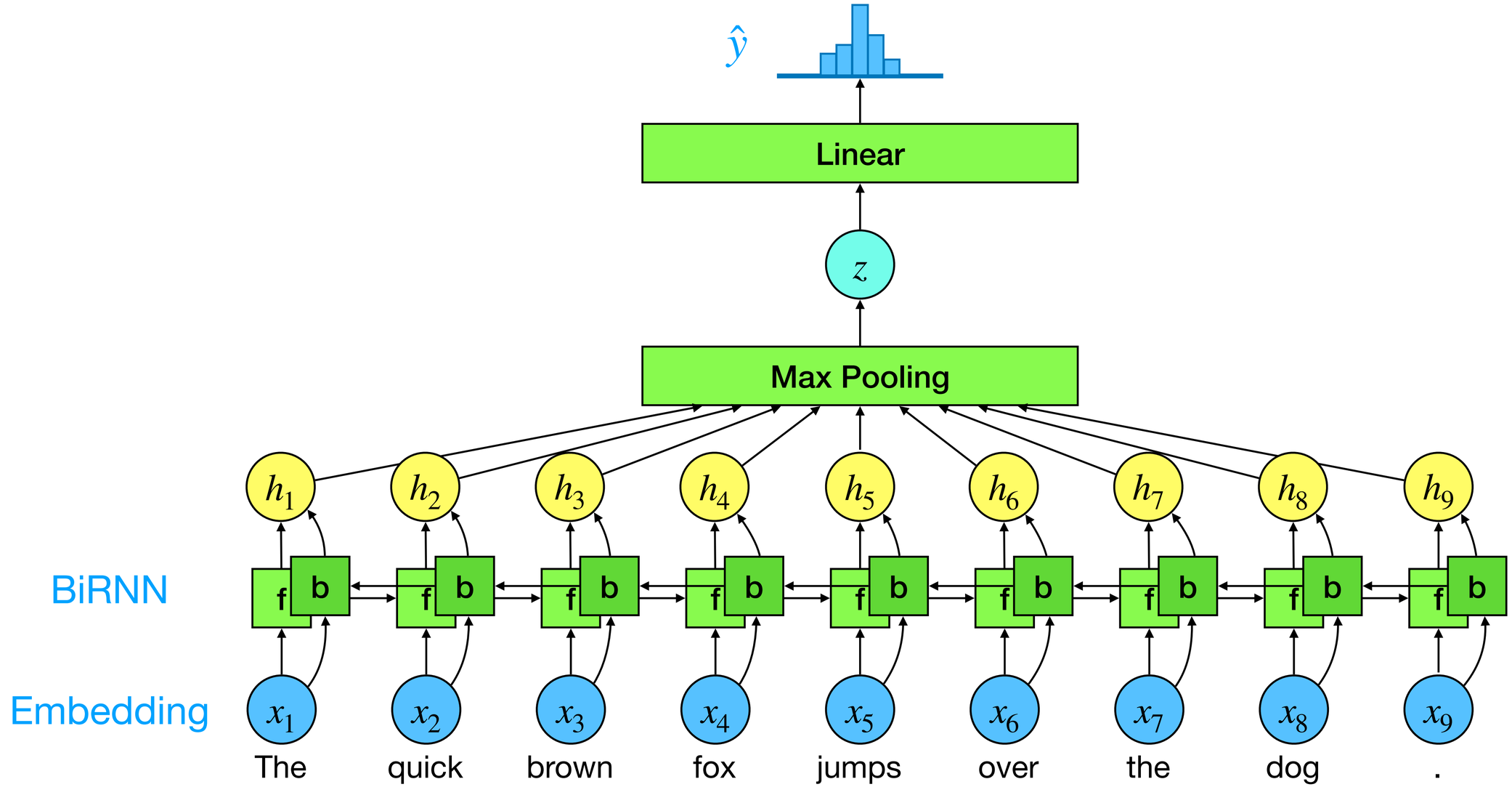
Text Classification Model (RNN)



Text Classification Model (BiRNN)

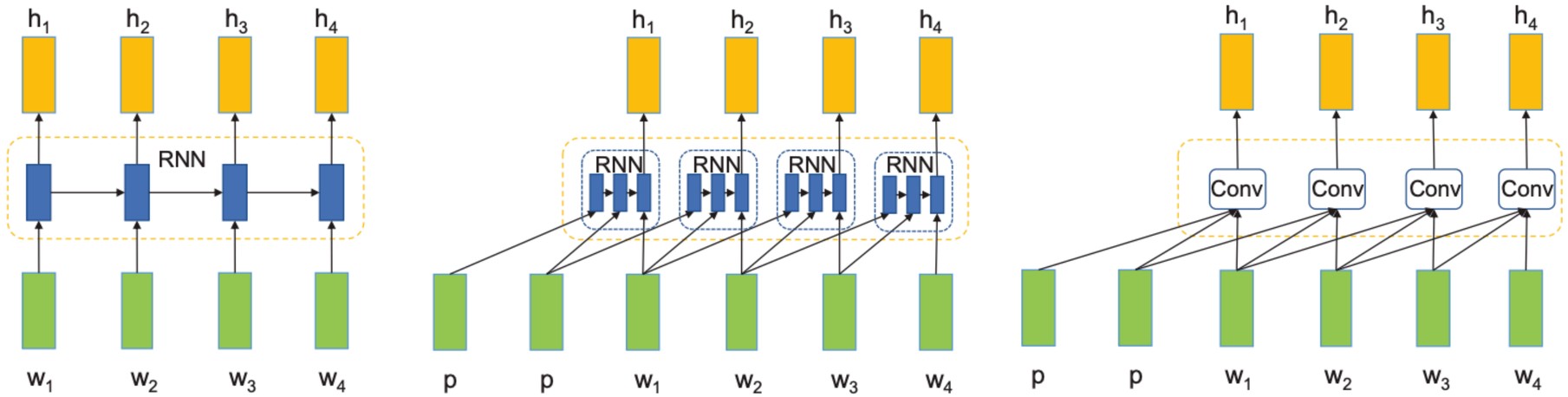


Text Classification Model (BiRNN)



Text Classification Model (RNN)

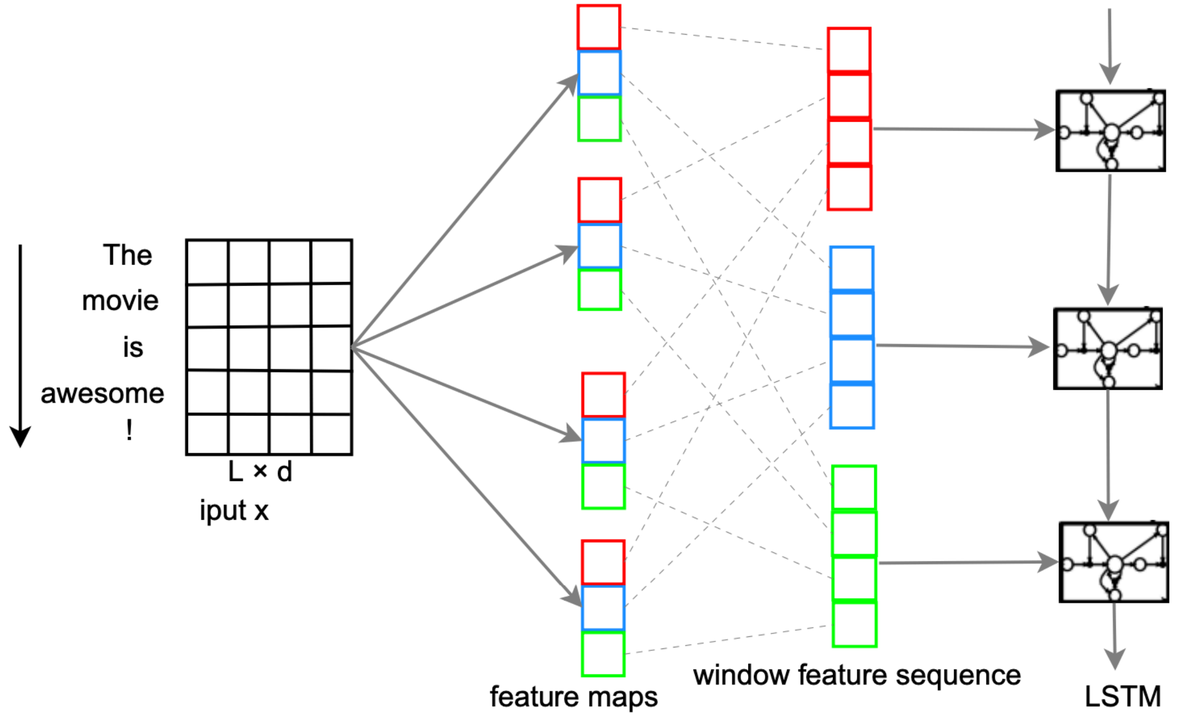
- Disconnected Recurrent Neural Networks for Text Categorization
- Baoxin Wang et al. 2018
- RNN을 CNN과 비슷하게 일부분에만 분할하여 적용



참조: https://www.aclweb.org/anthology/P18-1215/

Text Classification Model (RNN)

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



참조: https://arxiv.org/abs/1511.08630

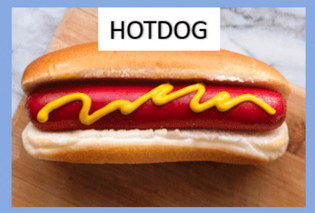
Text Classification Model (SOTA)

State fo Art
Text Classification

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

Evaluation Metric (Confusion Matrix)

True positive



- Label: **Positive**
- Predict: **Positive**

False negative

- Label: **Positive**
- Predict: **Negative**

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

- Label: **Negative**
- Predict: **Positive**

False positive

- Label: **Negative**
- Predict: **Negative**

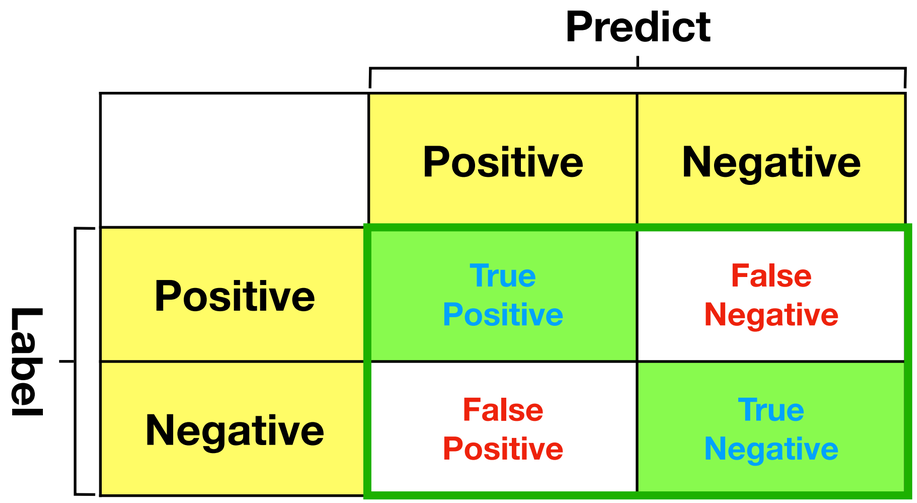
True negative

그림: https://blog.nillsf.com/index.php/2020/05/23/confusion-matrix-accuracy-recall-precision-false-positive-rate-and-f-scores-explained/

Evaluation Metric (Accuracy)

정답 비율

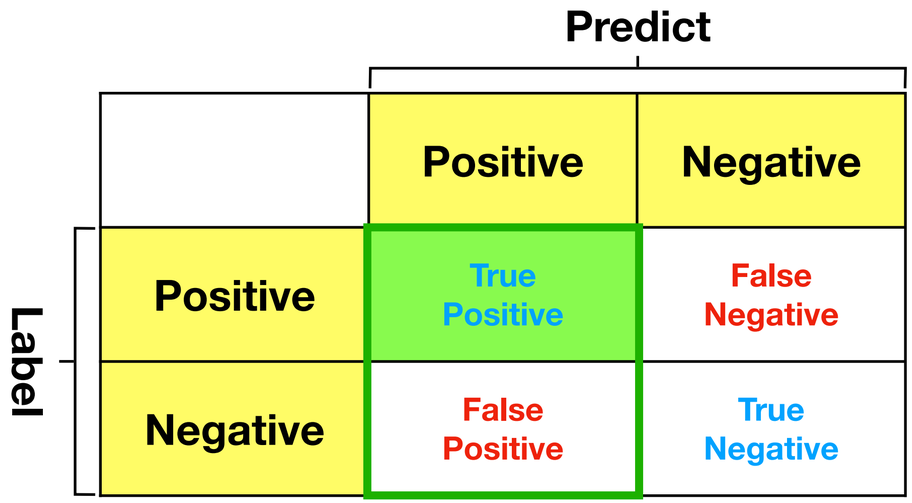
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{T}{T + F}$$



Evaluation Metric (Precision)

Positive로 예측한 값중 실제 Positive 비율

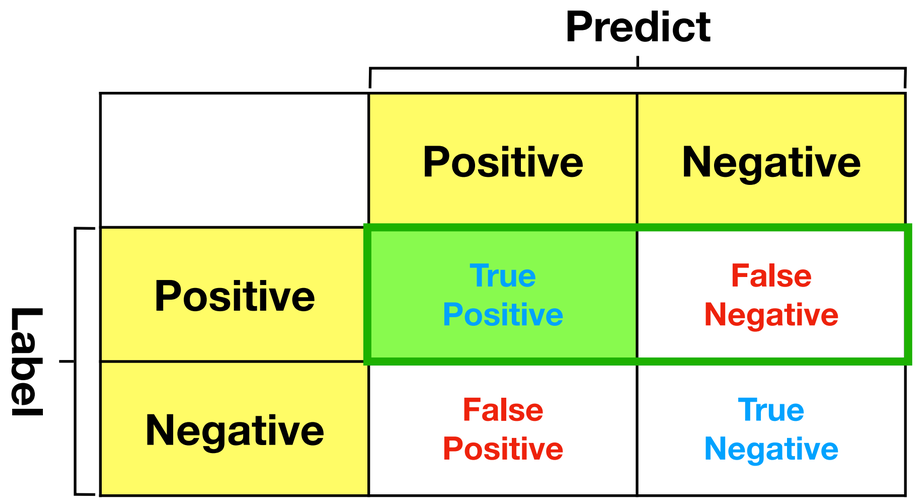
$$precision = \frac{TP}{TP + FP}$$



Evaluation Metric (Recall)

실제 Positive중 Positive로 예측한 비율 비율

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



Evaluation Metric (F1-score)

Recall 과 precision의 조화평균

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$
$$= 2\frac{P \cdot R}{P + R}$$

Imbalance data 측정 시 유용

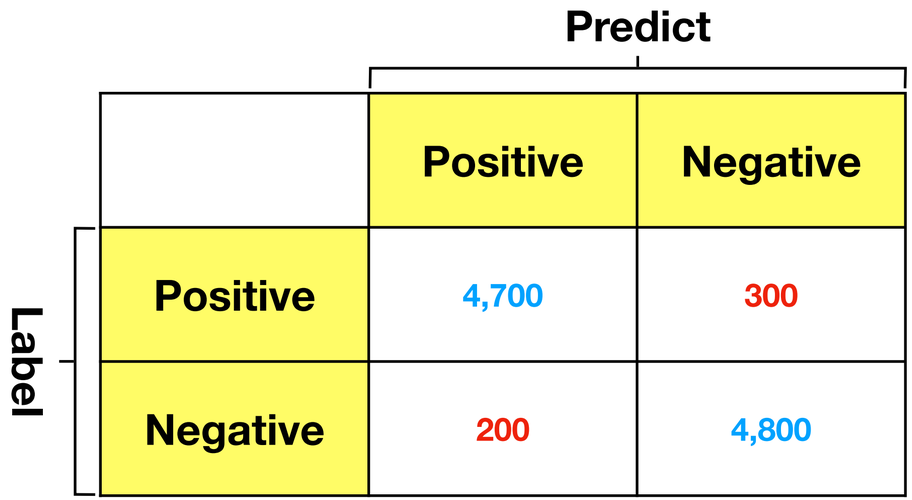
Evaluation Metric (Example)

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

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

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

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



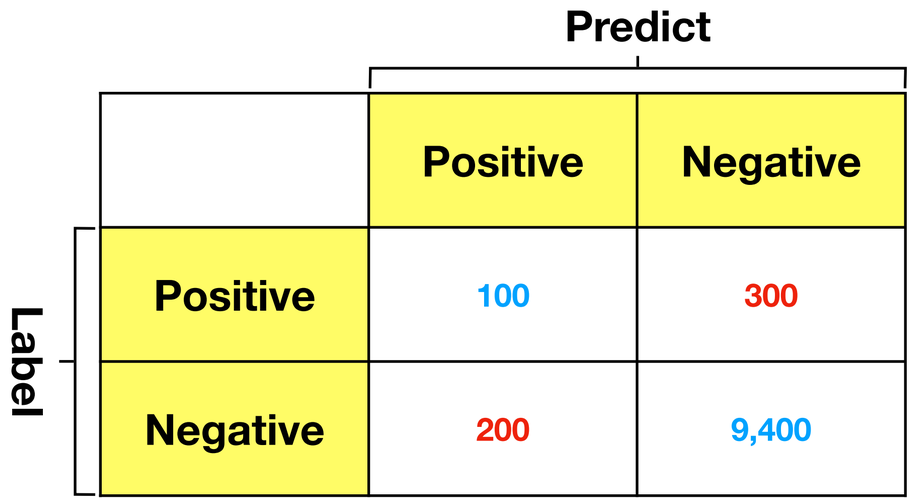
Evaluation Metric (Example)

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

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

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

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



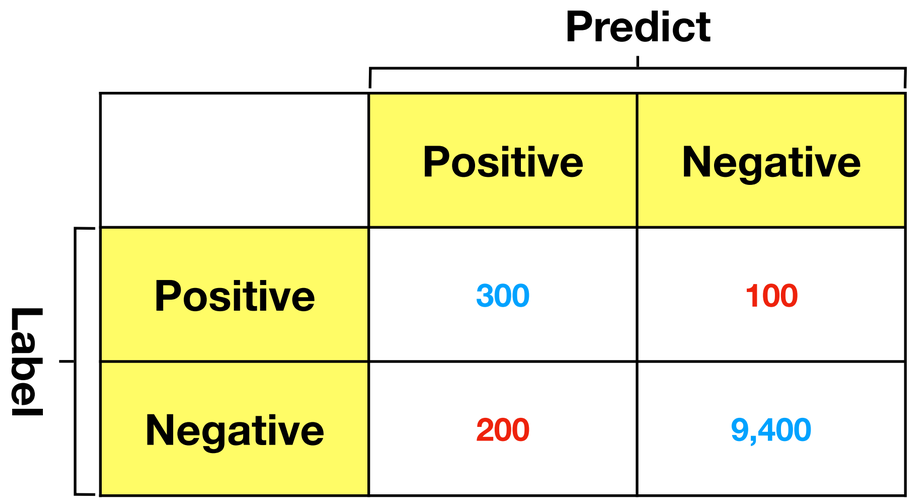
Evaluation Metric (Example)

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

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

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

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



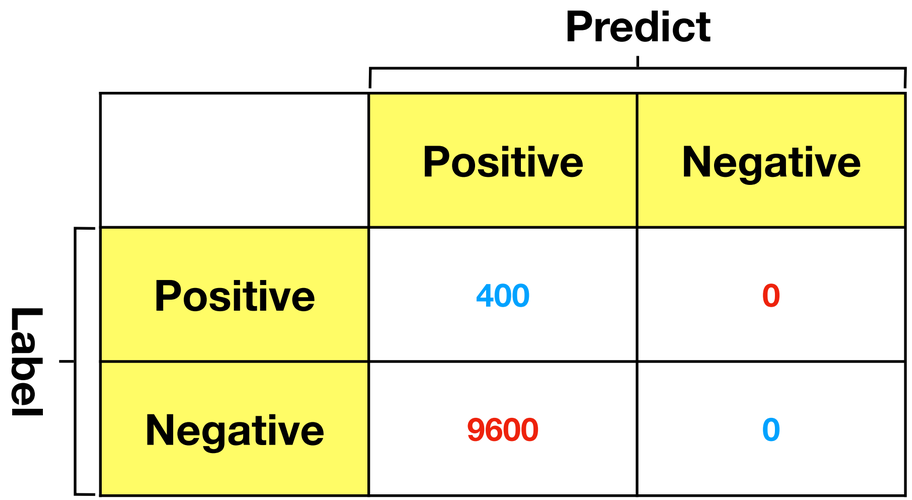
Evaluation Metric (Example)

$$accuracy = \frac{400}{10{,}000} = 0.04$$

$$precision = \frac{400}{10{,}000} = 0.04$$

$$recall = \frac{400}{400} = 1.0$$

$$F1 = 2\frac{0.04 \times 1.0}{0.04 + 1.0} = 0.077$$



Recall이 증가하면 precision은 감소하는 경향이 있음

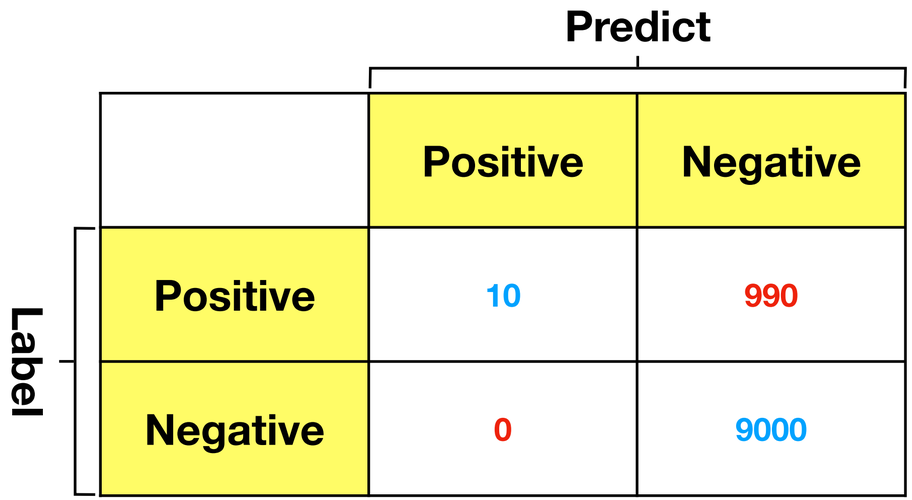
Evaluation Metric (Example)

$$accuracy = \frac{10}{10{,}000} = 0.001$$

$$precision = \frac{10}{10} = 1.0$$

$$recall = \frac{10}{1000} = 0.01$$

$$F1 = 2\frac{0.01 \times 1.0}{0.01 + 1.0} = 0.012$$



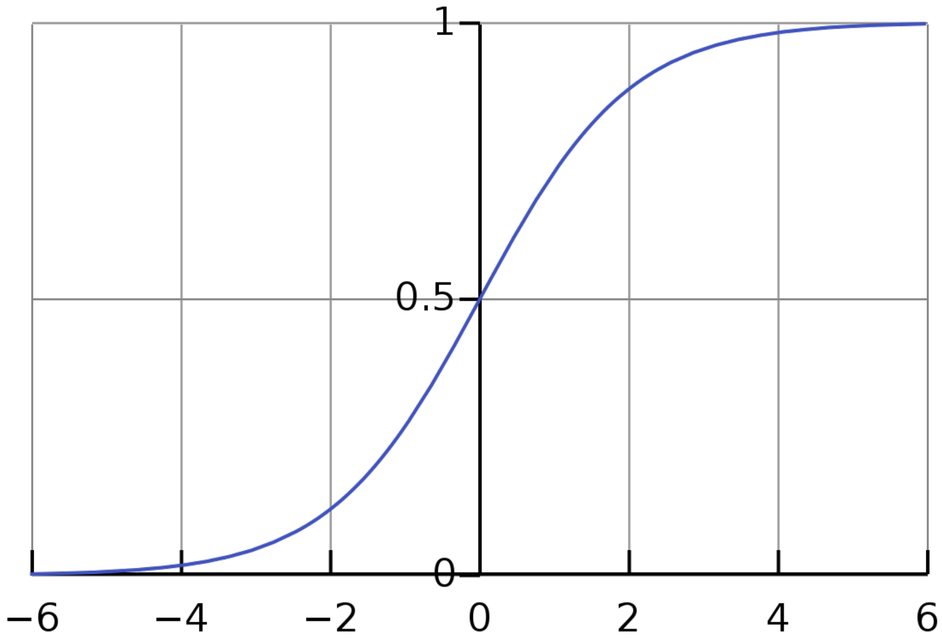
Precision이 증가하면 recall은 감소하는 경향이 있음

Logistic Regression (binary classification)

$$z = f(x)$$

$$p(\hat{y} = 1 \mid x; \theta) = \sigma(z) = \text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

$$p(\hat{y} = 0 \mid x; \theta) = 1 - p(1 \mid x; \theta)$$



Logistic Regression (binary classification)

$$z = f(x)$$

$$p(\hat{y} = 1 \,|\, x; \theta) = \sigma(z) = \text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

$$p(\hat{y} = 0 \,|\, x; \theta) = 1 - p(1 \,|\, x; \theta)$$

$$\mathscr{L} = \begin{cases} -\log(p(\hat{y} = 1 \,|\, x; \theta)), & \text{if } y = 1 \\ -\log(p(\hat{y} = 0 \,|\, x; \theta)), & \text{if } y = 0 \end{cases}$$

$$\mathscr{L} = -y \log(\sigma(z)) - (1 - y) \log(1 - \sigma(x))$$

Logistic Regression (softmax)

$$z = f(x)$$

$$p(\hat{y} = 1 \mid x; \theta) = \sigma(z) = \text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

$$p(\hat{y} = 0 \mid x; \theta) = 1 - p(1 \mid x; \theta)$$

$$\mathscr{L} = \begin{cases} -\log(p(\hat{y} = 1 \mid x; \theta)), & \text{if } y = 1 \\ -\log(p(\hat{y} = 0 \mid x; \theta)), & \text{if } y = 0 \end{cases}$$

$$\mathscr{L} = -y \log(\sigma(z)) - (1 - y)\log(1 - \sigma(x))$$

$$[z_0,\ z_1] = f(x)$$

$$p(\hat{y} = 1 \mid x; \theta) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}}$$

$$p(\hat{y} = 0 \mid x; \theta) = \frac{e^{z_0}}{e^{z_0} + e^{z_1}}$$

$$y = \begin{cases} [0,1], & \text{if } y = 1 \\ [1,0], & \text{if } y = 0 \end{cases}$$

$$\mathscr{L} = -y_1 \log(p(1 \mid x; \theta)) - y_0 \log(p(0 \mid x; \theta))$$

Logistic Regression (softmax)

$$z = f(x)$$

$$p(\hat{y} = 1 \,|\, x; \theta) = \sigma(z) = \text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$

$$p(\hat{y} = 0 \,|\, x; \theta) = 1 - p(1 \,|\, x; \theta)$$

$$\mathscr{L} = \begin{cases} -\log(p(\hat{y} = 1 \,|\, x; \theta)), & \text{if } y = 1 \\ -\log(p(\hat{y} = 0 \,|\, x; \theta)), & \text{if } y = 0 \end{cases}$$

$$\mathscr{L} = -y \log(\sigma(z)) - (1 - y)\log(1 - \sigma(x))$$

$$[z_0,\ z_1] = f(x)$$

$$p(\hat{y} = 1 \,|\, x; \theta) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}}$$

$$p(\hat{y} = 0 \,|\, x; \theta) = \frac{e^{z_0}}{e^{z_0} + e^{z_1}}$$

$$z = z_1 - z_0$$

$$p(\hat{y} = 1 \,|\, x; \theta) = \frac{e^{z_1}}{e^{z_0} + e^{z_1}} = \frac{1}{e^{z_0 - z_1} + 1} = \frac{1}{1 + e^{-(z_1 - z_0)}} = \frac{1}{1 + e^{-z}}$$

감사합니다.