KUBIG CONTEST

DL분반 – NLP 2팀

월간 데이콘 소설 작가 분류 AI 프로젝트

16기 김상옥 17기 김연규 우윤규

2023 Winter

Project Introduction & Overview

EDA & Data Preprocessing

LSTM Modeling

Feature Engineering & XGBoost

Ensemble & Conculsion

CONTENTS

발표 목차

1 **Project Introduction & Background**
프로젝트 제안 및 배경

2 **EDA & Data Preprocessing**
EDA 및 데이터 전처리

3 **LSTM Modeling & ML Ensemble**
모델링 및 앙상블

4 **Conclusion & Interpretaion**
결론 및 결과 분석

KUBIG CONTEST **DL분반 – NLP 2팀**

월간 데이콘 소설 작가 분류 AI 프로젝트

1 Project Introduction & Background

프로젝트 제안 및 배경

프로젝트 제안 및 배경

1. Project Introduction & Background

23 겨울 DL 분반 학습 내용

- MLP, CNN 등 **기초 딥러닝 모델 구조** 학습
- Optimization, Dataloader, Batch normalization 등 모델링 내부 기법 학습
- **RNN, LSTM 등 Sequential Data에 적합**한 모델
- **Tokenizing, embedding 등 Text Data 전처리**

NLP 프로젝트 –
텍스트 데이터 기반의 분류 Classificaton **문제 적용**

프로젝트 제안 및 배경

1. Project Introduction & Background

월간 데이콘 소설 작가 분류 AI 경진대회

1500자 이내 **소설 텍스트 데이터**['text']를 바탕으로
5인 중 한 명의 작가['author'] 예측하는 분류 문제

DL 분반에서 학습한
tokenizing, embedding 등 Text Data 전처리
& RNN, LSTM 기반 분류 모델링 적용 가능

과거 2020년 진행된 대회로,
기존 수상작 파이프라인을 참고하여
추가 개선 방안 고려

KUBIG CONTEST **DL분반 – NLP 2팀**

월간 데이콘 소설 작가 분류 AI 프로젝트

2 EDA & Data Preprocessing

EDA 및 데이터 전처리

EDA

2. EDA 및 데이터 전처리

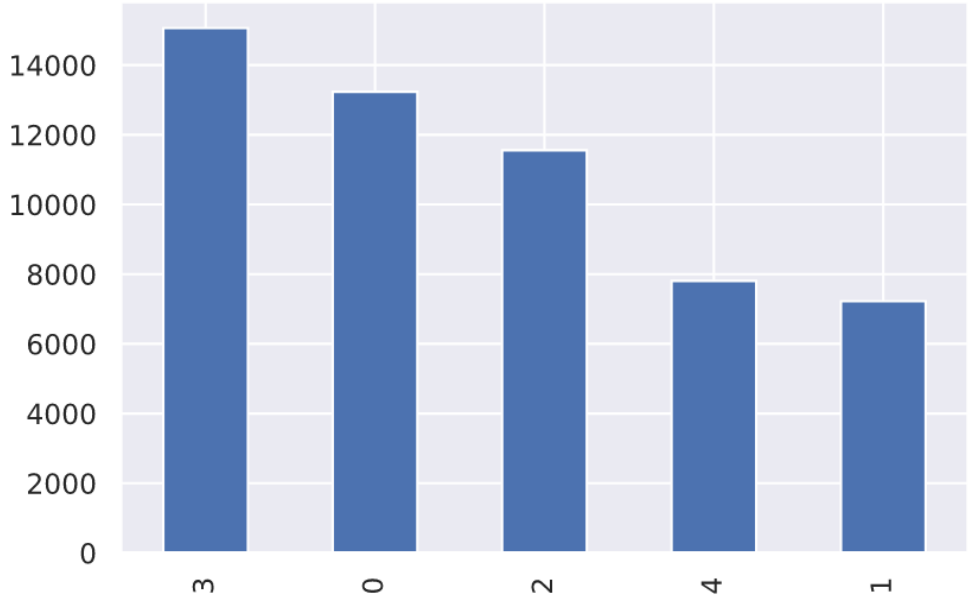
데이터셋 파일 구성

train.csv: 학습용 소설 데이터 **(3 x 54,578)**
text: 텍스트 데이터(x)
author: 작가 (target)

test_x.csv: 텍스트용 소설 데이터 **(2 x 19,617)**
text: 텍스트 데이터(x)
Sample_submission.csv의 target에 대응

sample_submission.csv:
평가용 제출 데이터 **(6 x 19,617)**
author: 작가 (target) – 각각의 예측 확률, 공란

데이터 분포 관련 plot

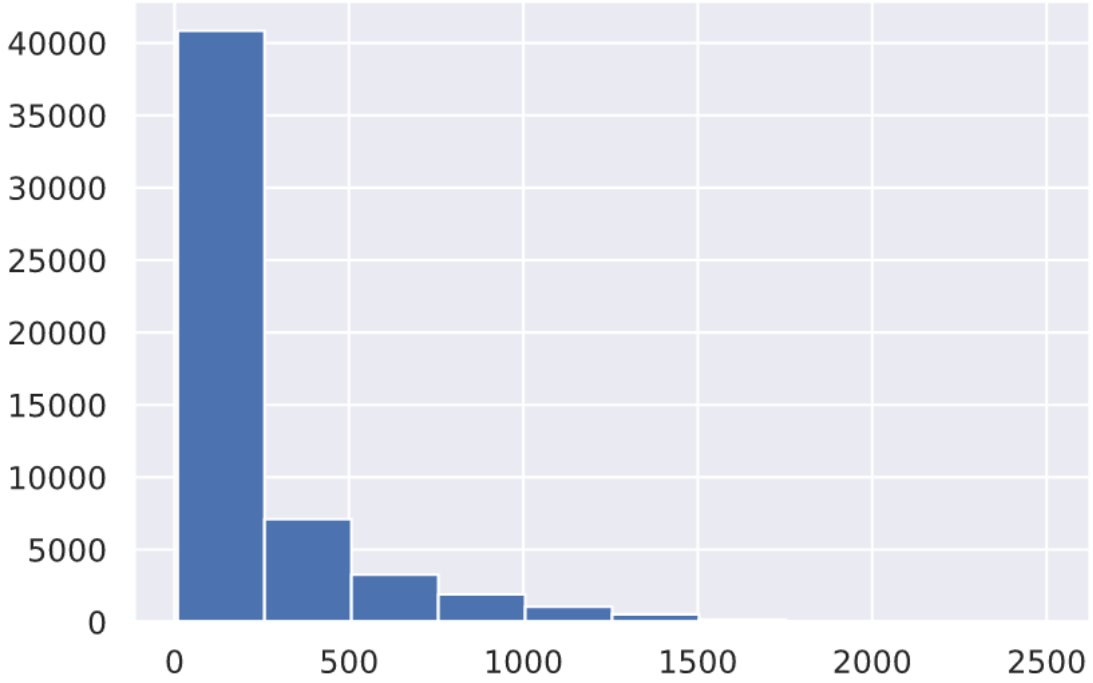


〈**타겟 변수**[‘author’] 분포 비율〉

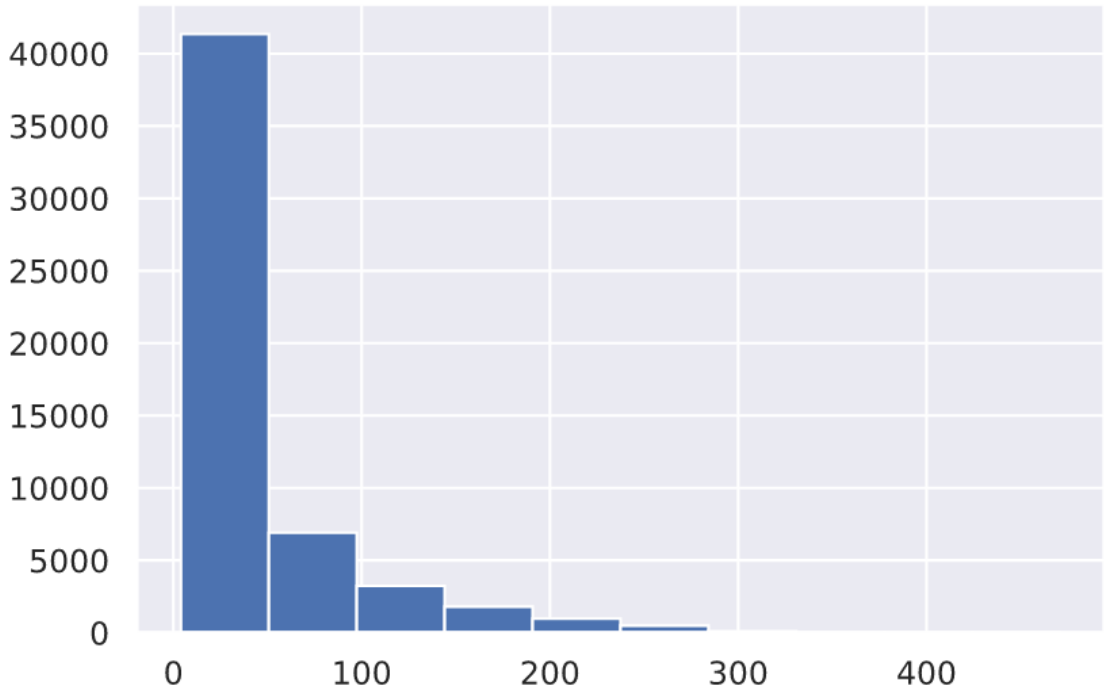
EDA

2. EDA 및 데이터 전처리

데이터 주요 분포 Plot



〈텍스트['text'] **길이** 분포〉



〈텍스트['text'] **단어 개수** 분포〉

데이터 전처리 – 토큰화

2. EDA 및 데이터 전처리

NLTK, TensorFlow
내장 Tokenizer 활용

Raw Data
소설 텍스트 데이터 원문

	index	text	author
0	0	He was almost choking. There was so much, so m...	3
1	1	“Your sister asked for it, I suppose?”	2
2	2	She was engaged one day as she walked, in per...	1
3	3	The captain was in the porch, keeping himself ...	4

Removing stopwords & Separation
불용어 제거 및 단어 리스트화

index	text	author
0	[He, was, almost, choking, There, was, so, muc...	3
1	[Your, sister, asked, for, it, suppose]	2
2	[She, was, engaged, one, day, as, she, walked,...	1
3	[The, captain, was, in, the, porch, keeping, h...	4

Tokenization – Word Indexing
워드 인덱싱, 공백 기준 토큰화

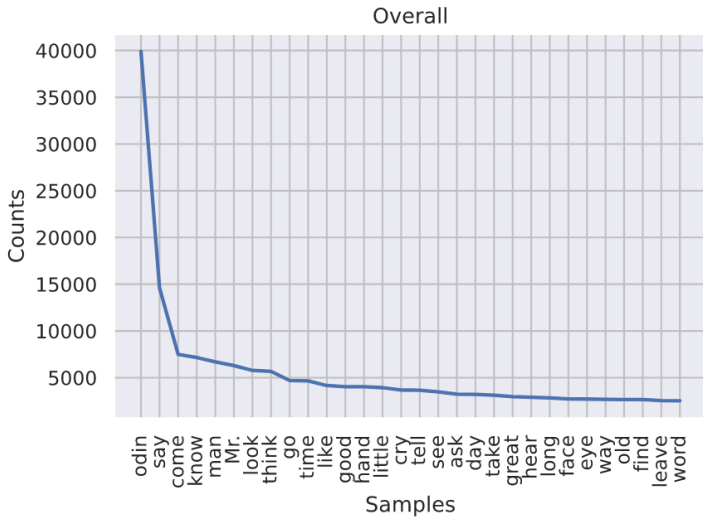
index	text	author
0	[8, 12, 235, 1, 35, 12, 32, 92, 32, 92, 8, 415...	3
1	[49, 289, 140, 17, 10, 324]	2
2	[26, 12, 784, 42, 137, 18, 26, 354, 7, 1, 547,...	1
3	[2, 342, 12, 7, 2, 1, 982, 111, 1, 56, 5, 2, 1...	4

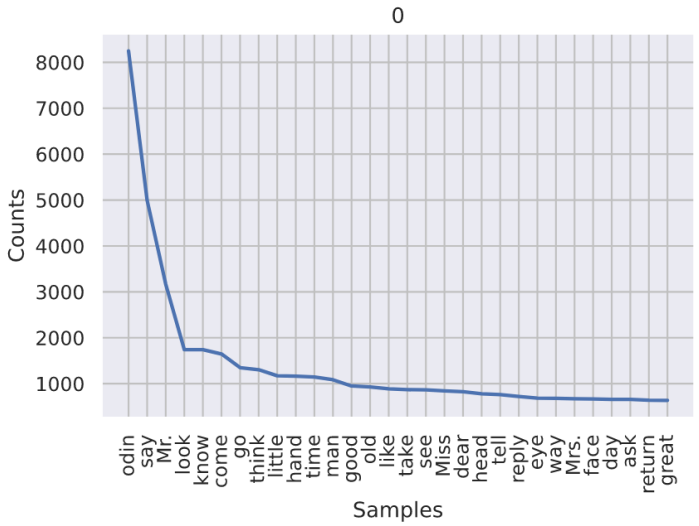
EDA – 타겟별 토큰 사용 빈도

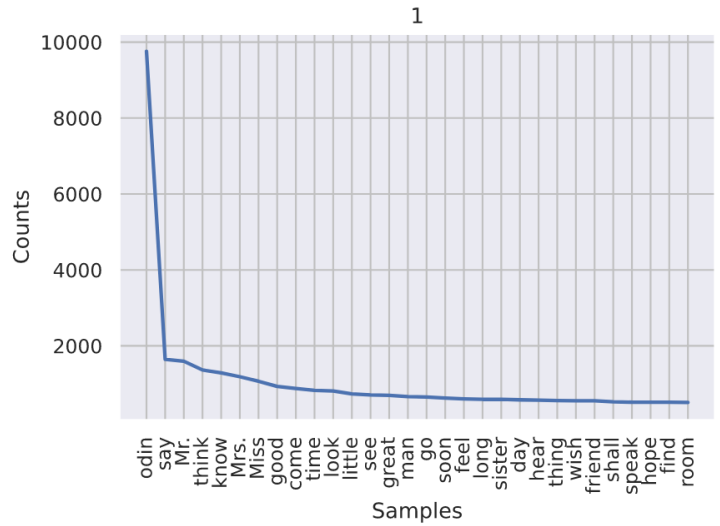
2. EDA 및 데이터 전처리

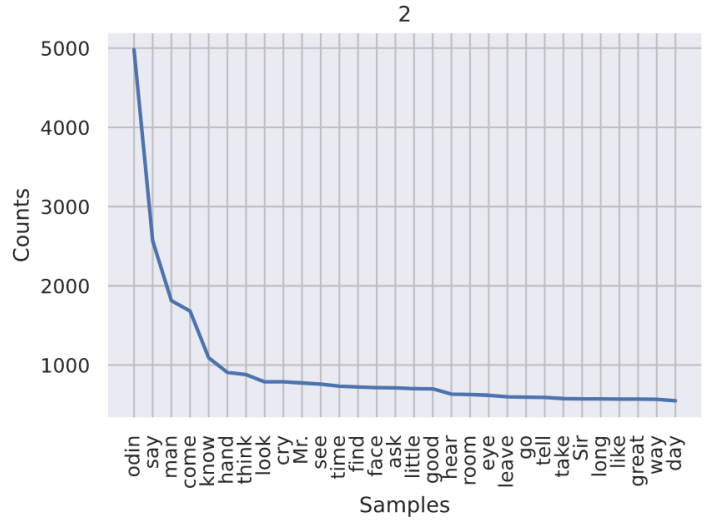
KUBIG CONTEST **DL분반 – NLP 2팀**

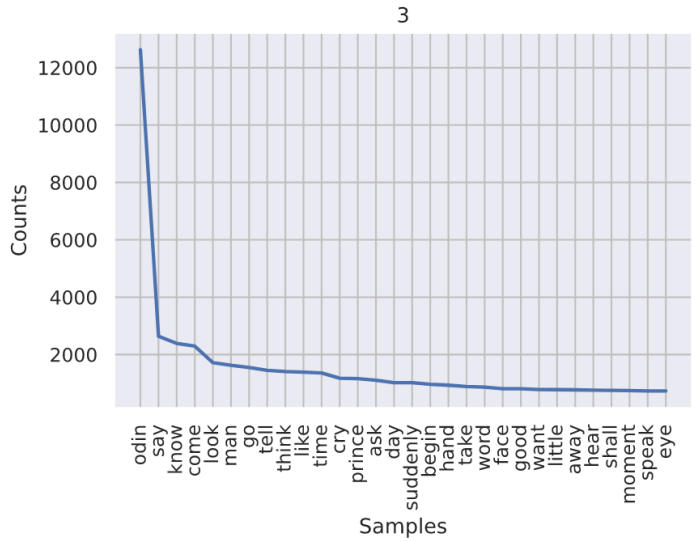
월간 데이콘 소설 작가 분류 AI 프로젝트

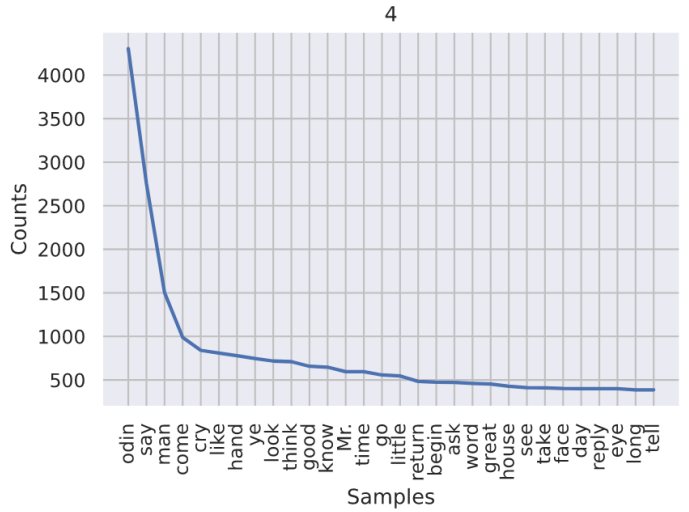












3 LSTM Modeling & ML Ensemble

모델링 및 앙상블

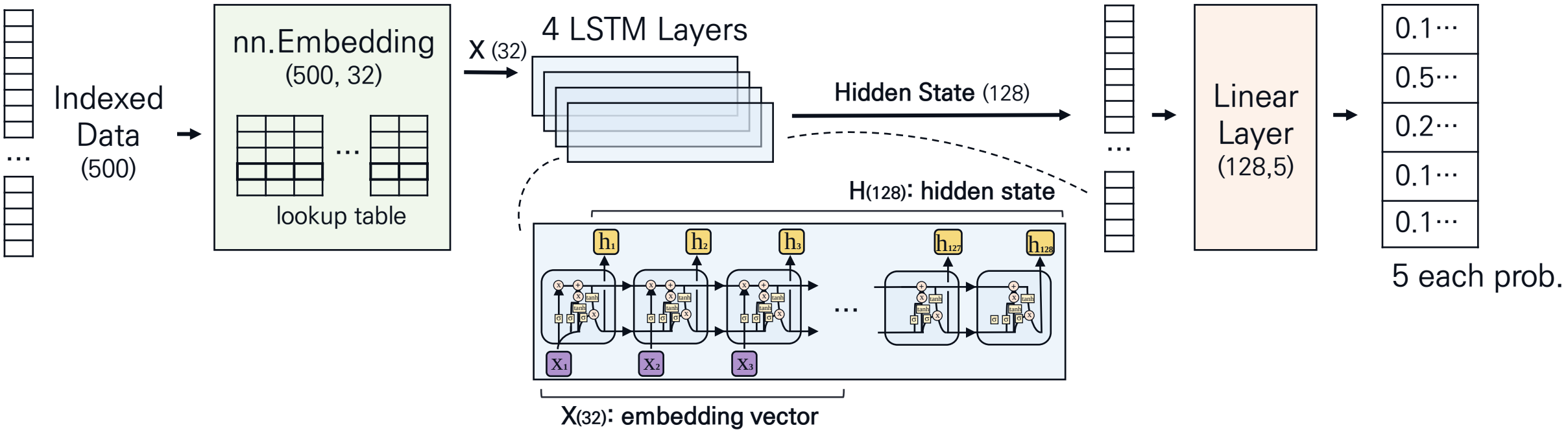
LSTM 모델링

3. 모델링 및 앙상블

LSTM 모델 개요

모델 설명 및 파이프라인

Sequential Data에 강점이 있는 LSTM을 4개 레이어로 쌓아
각 관측치의 단어 종류 및 순서를 학습해서 Hidden state 반환, Linear Layer로 타겟 확률 도출



LSTM 모델링

3. 모델링 및 앙상블

모델링 성능 및 분석

Train set에서 준수한 성능 그러나, **Vaild set에서 accuracy 하락 및 loss 증가**

```
model, lowest_loss, train_losses, valid_losses = train_model(model, early_stop, nb_epochs, progress_interval)
```

```
Output exceeds the size limit. Open the full output data in a text editor
train_losses : 1.600804033279419, valid_loss : 1.5836521577144014, lowest_loss : 1.5836521577144014, lowest_epoch : 0, epoch :0
Train_accuracy: 8703 / 35122 (24.78 %)
Valid_accuracy: 2146 / 8781 (24.44 %)
 ...
train_losses : 1.0905305097319864, valid_loss : 1.2858069167620894, lowest_loss : 1.2824267684549526, lowest_epoch : 180, epoch :201
Train_accuracy: 28580 / 35122 (81.37 %)
Valid_accuracy: 5410 / 8781 (61.61 %)

train_losses : 1.0893966787511653, valid_loss : 1.2890571183052615, lowest_loss : 1.2824267684549526, lowest_epoch : 180, epoch :202
Train_accuracy: 28627 / 35122 (81.51 %)
Valid_accuracy: 5395 / 8781 (61.44 %)
```

LSTM 블록 대체 시도

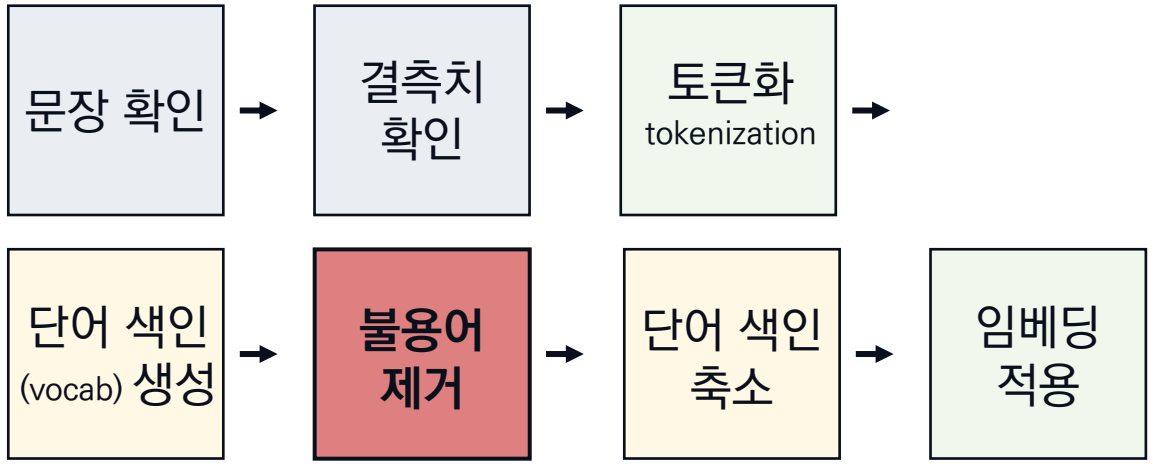
CNN – 데이터 크기로 인해 학습 시간 정체 / **RNN** – Gradient Vanishing 문제로 성능 확보 불가

→ **데이터와 인공 신경망 모델 구조적 적합성 문제 의심**

Model Diagnosis

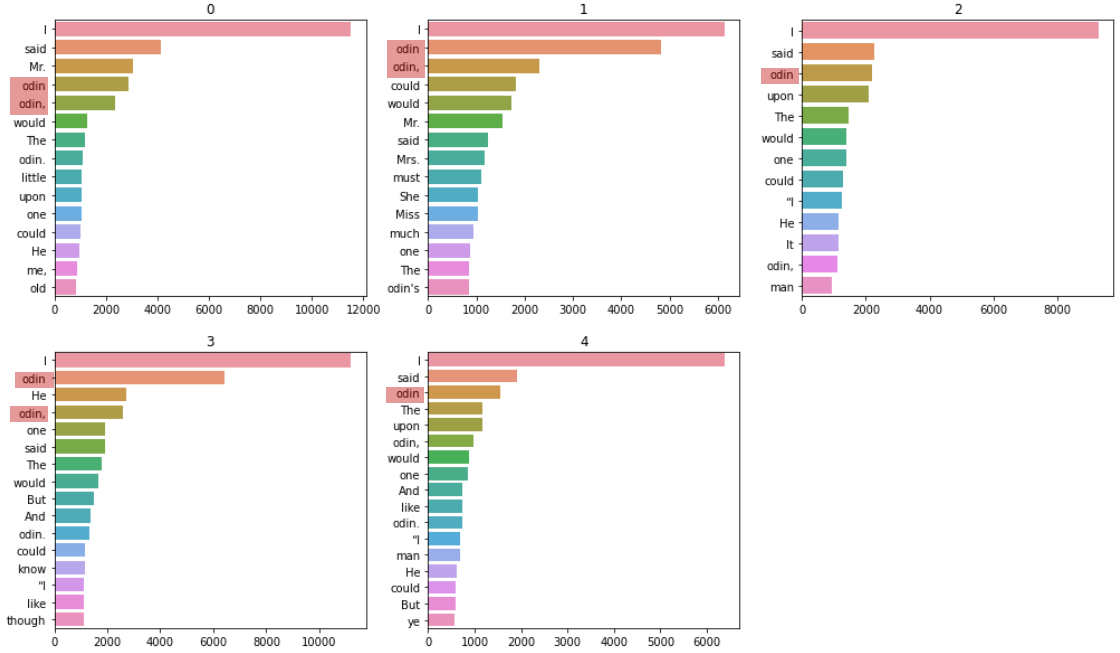
3. 모델링 및 앙상블

자연어 전처리 과정 검토



문장을 더 살펴본 결과, **불용어로 판단해 제거한 문장부호와 일부 불용어들의 빈도수**가 **작가들의 문체를 구분하는 기준**이 될 수 있음을 파악

데이터 성격 검토



'Odin'이라는 동일 인물을 중심으로 작성된 소설임을 확인,
이로 인해 **문맥에 따라 분류하는 LSTM, RNN** 모델이
상대적으로 **문체 구분에 취약하게 작용**하여
오히려 과적합이 발생했을 것으로 추론

ML 접목 – Feature Engineering

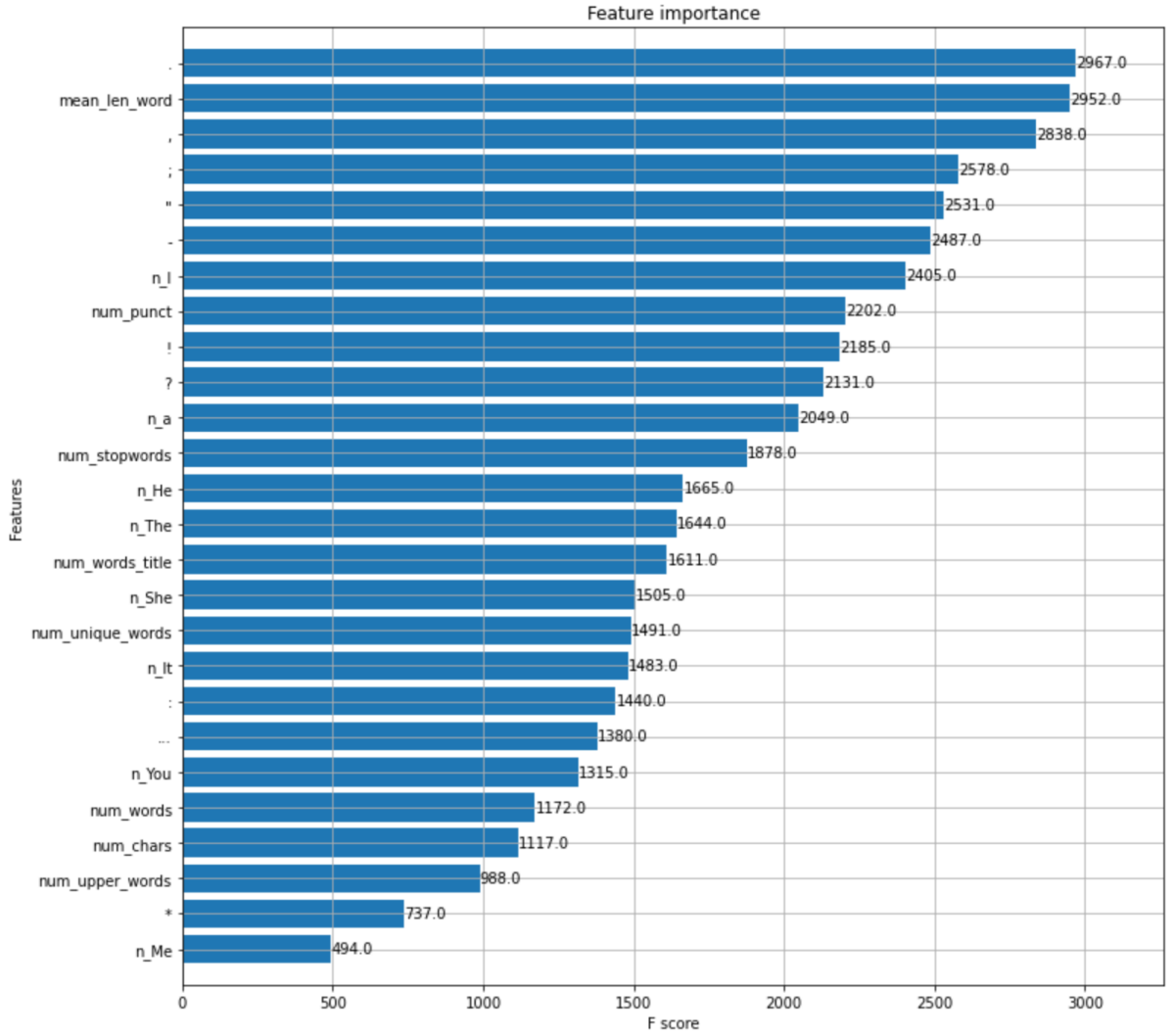
3. 모델링 및 앙상블

Meta Feature

: stopword 갯수, 단어갯수, 문장부호 갯수 등
텍스트에서 뽑아낸 특성

- 단어, 문자, 불용어, 구두점의 수 및 대문자가 포함된 단어 수 등을 특성으로 생성
- 더불어 문장부호 및 자주 사용되는 단어(The, I, He, She, a 등)의 **빈도수**를 특성으로 추가

실제로 Feature Importance 확인 결과,
마침표/쉼표의 빈도수, 평균 단어 길이가 가장 명확한 구분 기준으로 평가됨



ML 접목 – Feature Engineering

3. 모델링 및 앙상블

Text-Based Feature

: 단어 등장 빈도수, word2vec 등 **문장 그 자체에서 추출**한 특성 – **TF-IDF 벡터화 활용**

TfidfVectorizer

(stop_words='english', ngram_range=(1,3))

: 기존에 저장된 **영어 불용어 제거**, 1~3개의 단어 묶음으로 단어의 중요도를 파악

→ (ngram_range=(**1,5**), analyzer='**word**')

: 1~5개의 단어 묶음에 대해 **단어의 중요도** 파악

→ (ngram_range=(**1,5**), analyzer='**char**')

: 1~5개의 단어 묶음에 대해 **문자(a, b, …)의 중요도** 파악

TruncatedSVD (n_components=n_comp, algorithm='arpack')

: Sigma 행렬에 **있는 특이값 중 상위 일부 데이터만 추출**해 차원을 줄이는 방식

- **전체 문서에 대해 단어 빈도수**를 살펴본 결과 수천 x 수천의 **큰 희소행렬**이 생성
- 큰 희소행렬를 **효율적으로 축소하기 위해 arpack 알고리즘**을 활용

CalibratedClassifierCV (MultinomialNB(alpha=0.03), method='isotonic')

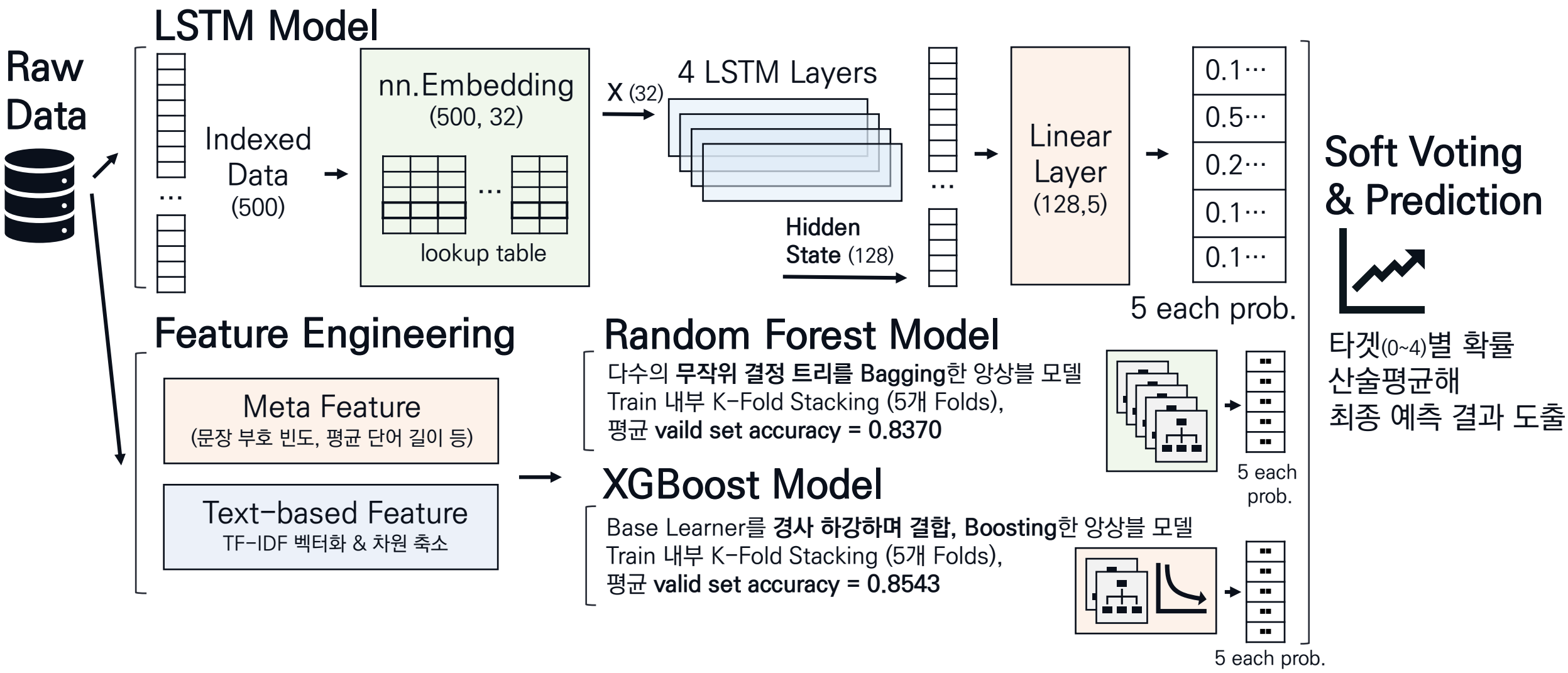
- Multinomial Naïve Bayesian 알고리즘은 자체 확률 예측값을 기준으로 분류
- 그러나 **calibrated를 활용하면 더 나은 방식을 활용해 확률값**을 다시 계산
- Train 데이터셋의 row 수가 5만개 이상이기 때문에 **더 복잡하지만 예측값의 정확도가 높은 isotonic regression**을 활용

머신 러닝 모델링 & Soft voting

3. 모델링 및 앙상블

KUBIG CONTEST **DL분반 – NLP 2팀**

월간 데이콘 소설 작가 분류 AI 프로젝트



4 Conclusion & Interpretaion

결론 및 결과 해석

결론

4. 결론 및 결과 해석

모델별 제출 결과 (Test set Log Loss 기준)

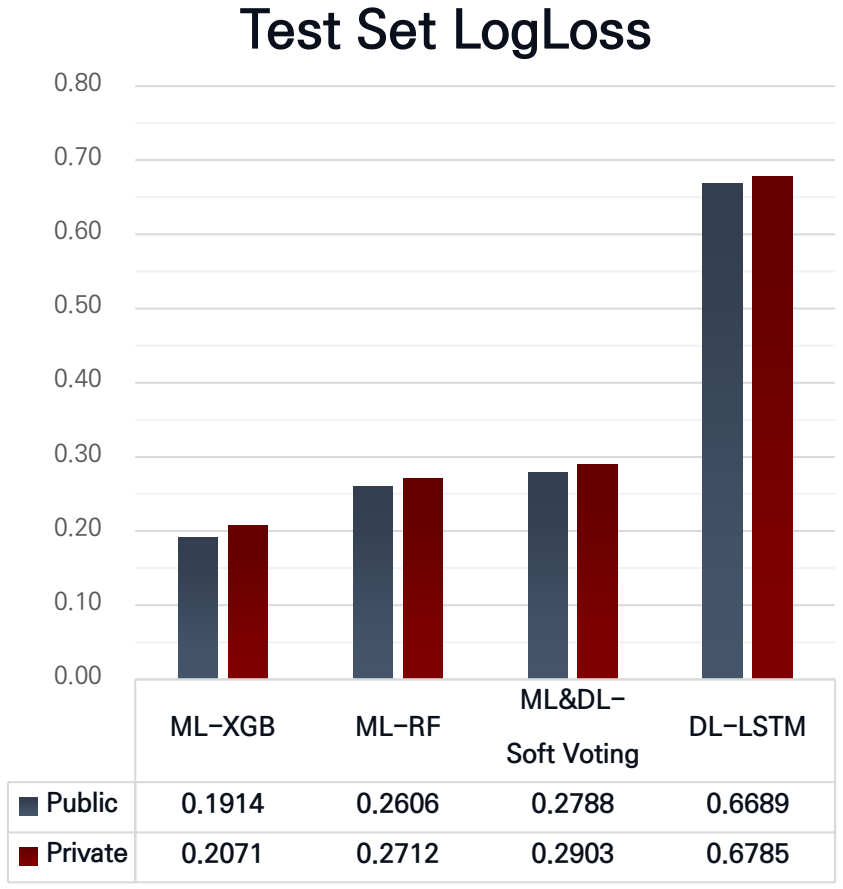
	제목	제출 일시	public점수 private점수	제출선택
812369	submission_XGB.csv ML-XGBoost Model edit	2023-02-28 10:46:30	0.1914026394 0.2071161424	○
811960	submission_RF.csv ML-Random Forest Model edit	2023-02-27 21:58:14	0.2606441347 0.2712051432	○
811956	LSTM_prop.csv DL-LSTM Model edit	2023-02-27 21:57:01	0.6688502274 0.6785401406	○
811954	submission_reverted.csv Soft Voting Result edit	2023-02-27 21:54:03	0.2788159872 0.2902634976	◉

결론

4. 결론 및 결과 해석

결과 요약 및 해석

- DACON 제출 결과 **ML 〉 Soft Voting 〉 DL** 순으로 분류 성능 우수
- **딥러닝 모델에서 발생한 과적합**을 **Feature Engineering과 머신러닝 모델**로 상당 부분 **해소**
- **문체로 작가를 구분**하는 해당 문제에서는 **텍스트의 특징을 잘 나타내는 특성을 추출**하는 것이 분류 모델의 **성능 향상에 지대한 영향**을 미치는 것으로 파악됨
- Bert 등 텍스트 데이터에 특화된 **pre-trained 모델을 사용하거나, 더 많은 특성을 추출한다면 향상된 결과**를 얻을 수 있을 것으로 사료됨



감사합니다!

이상으로 발표를 마칩니다.

16기 김상옥 17기 김연규 우윤규