機械学習モデルによるツイート分類

課題

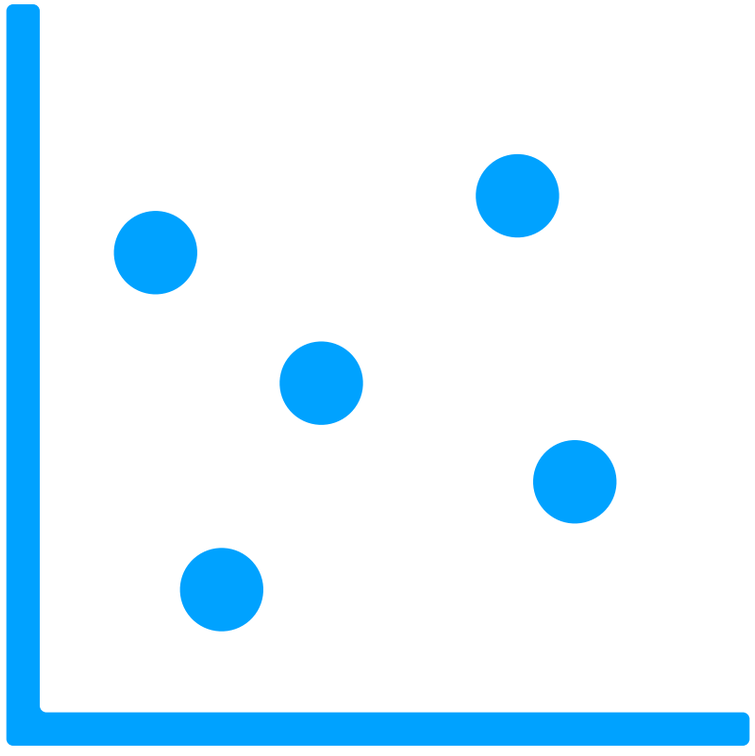
ツイッターのツイートを
以下のクラスに分類するモデル作成

hate speech

offensive language

neither

データ紹介



データ紹介

Sharkevin RIP Sean P
@WheresMyJuice

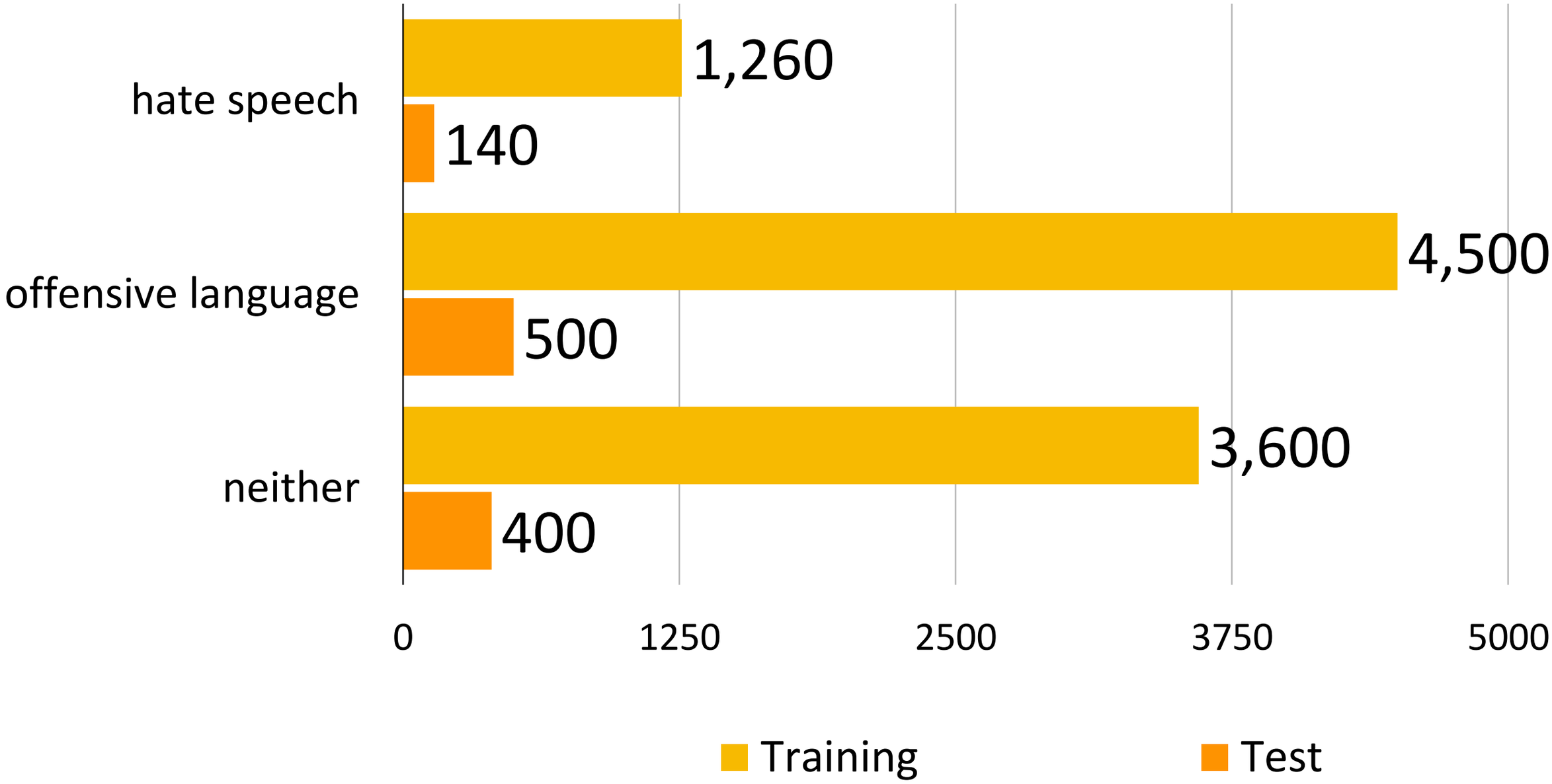
cookies and crackers aren't even on the same level!

午前7:04 · 2014年6月27日 · Twitter Web Client

↓

cookies and crackers aren't even on the same level!

クラス別データ数


hate speech
1,260
140
offensive language
4,500
500
neither
3,600
400
0
1250
2500
3750
5000
Training
Test

Google翻訳による
データ水増し

原文：
Fuck you you pussy ass hater go suck a dick and die fast

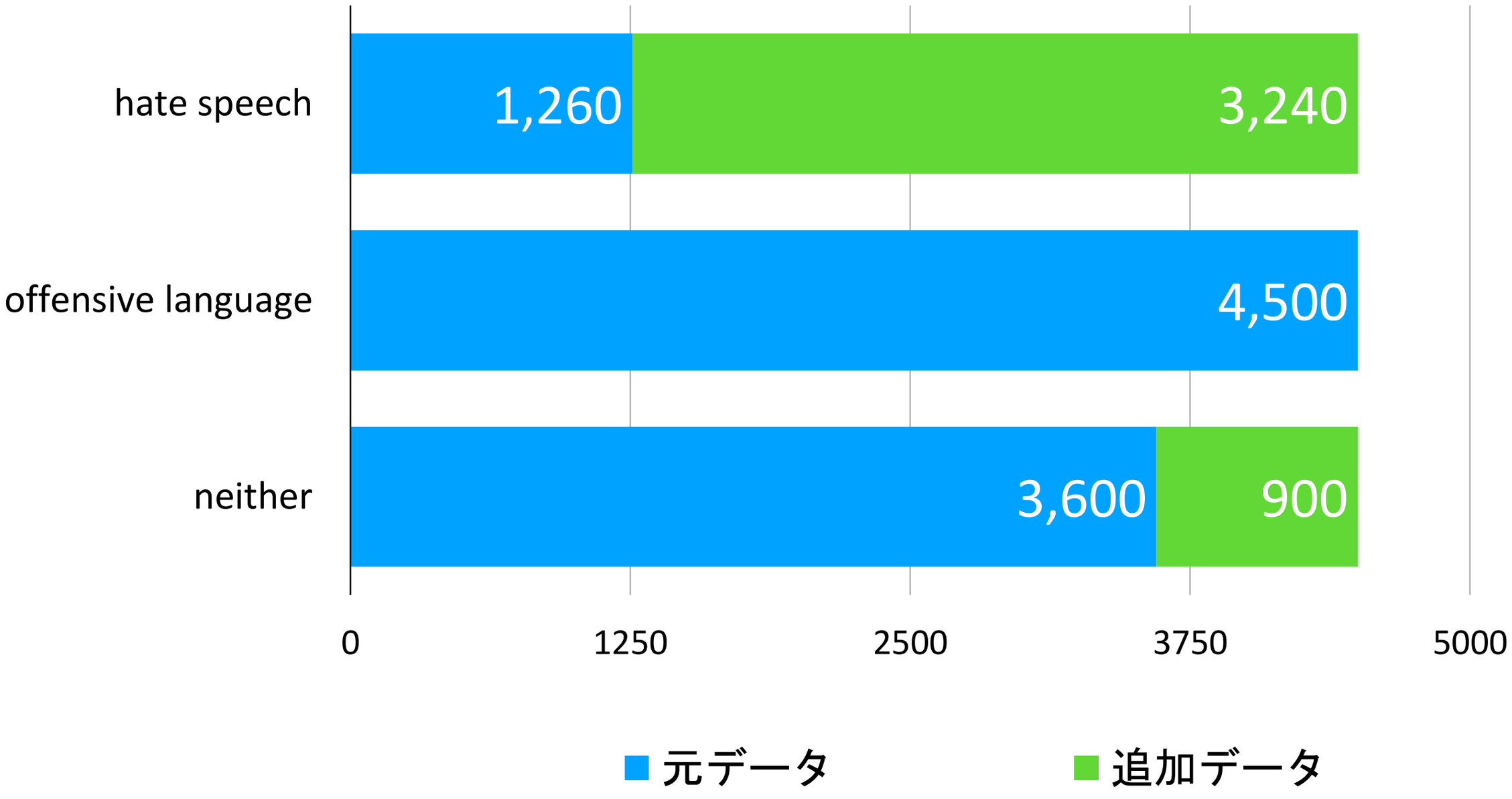
Google 翻訳でフランス語、ドイツ語、
オランダ語に翻訳

フランス語翻訳後：
Va te faire foutre le cul chatte aller sucer une bite et mourir vite

再び英語に翻訳

再翻訳後：
Fuck you pussy hater will suck a dick and die quick

データ水増し後


hate speech
1,260
3,240
offensive language
4,500
neither
3,600
900
0
1250
2500
3750
5000
元データ
追加データ

テキスト前処理:置き換え

"@John: How was Japan?" Awesome！ 😀
RT@Hana miss Japan :(　　（😀）

@メンションを<user>

返信を示す

"<user>: How was Japan?" <reply> Awesome! :grinning_face:
<rt_from> miss Japan :sad_face:

リツイートを示す

絵文字を内容で置き換え

テキスト前処理：lemmatization

- Stanford CoreNLP Package[1] を使用

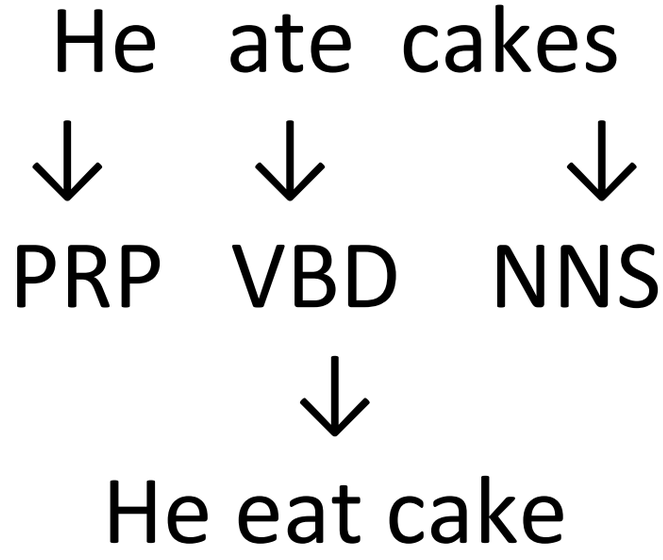
→ 言語データの処理を一括で行える
Pipeline

- GATE Twitter POStagger[2]モデル
  によりPOS タグ付け

→ twitterのツイートで学習を行った
POS tagging モデル

(packageに別途追加）

- それを元にlemmatization



1. https://stanfordnlp.github.io/CoreNLP/extensions.html
2. https://gate.ac.uk/wiki/twitter-postagger.html

データ前処理:その他

- **スペルミス直し**

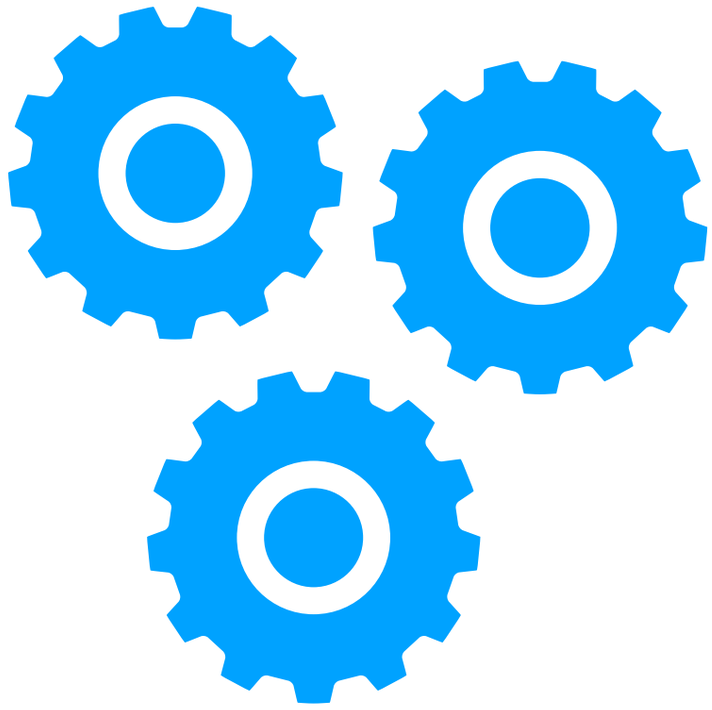
→ モジュール: autocorrect[1] を使用
内蔵単語リストの中で最もレーベンシュタイン距離の近い物を返す（距離2以内）

- **短縮形を戻す**

→ モジュール: contractions[2] を使用
単純な辞書型置き換え

- **?!以外の句読点削除**
- **小文字に変換**
- **Stopwords削除**

→ nltk[3] 付属のリストを使用



1. https://pypi.org/project/autocorrect/
2. https://pypi.org/project/contractions/
3. https://www.nltk.org/

用いたモデル

- 決定木
- LSTM

決定木パラメーター

	元データのみ	水増しあり
最大深度	20	30
分岐に必要な最低サンプル数	17.5%	15.5%
各分岐で考慮する特徴量	84%	82%

LSTM
パラメーター・モデル構成

	元データのみ	水増し後
embedding	20	20
LSTM Blocks	91	171
dropout	47%	63%
Softmax	3	3
学習率	0.001	0.001

モデル構成

1. embedding 層

2. LSTM 層
オプティマイザー: Adam
活性化関数: layer → tanh
gates → hard sigmoid

3. Dropout

4. 出力層：softmax

特徴抽出

- 決定木：TF-IDF

A. 全文書を通してレアな単語に高い重み

B. 文書出現頻度上位30%、下位0.4%の単語は排除

C. ユニグラムとバイグラムの組み合わせを使用

特徴抽出

- LSTM : モデルに単語の分散を学習させる

→タスクに適した分散を設定させる事が狙い

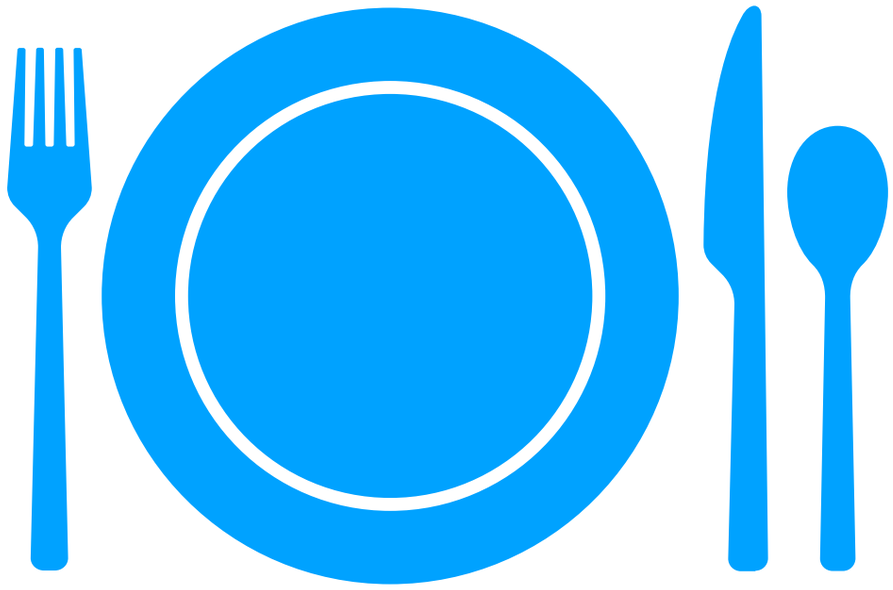
ex. 'white' と最もcosine 距離の近い単語

GloVe[1] (ツイッターベース） ： black, blue, green, yellow, red, purple, tank

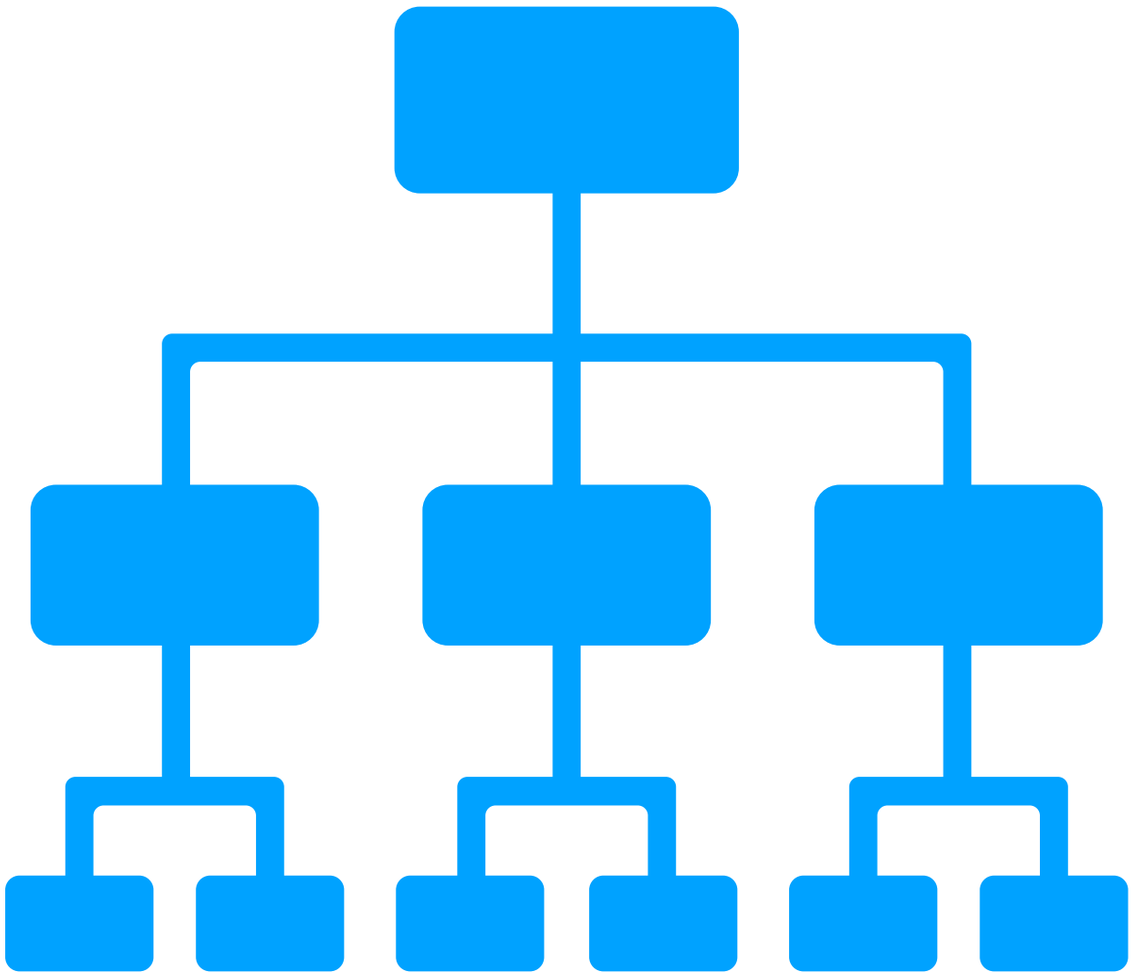
モデルが学習した分散 ： nigga, fuck, faggot, nigger, fag, ass, racist

1. https://nlp.stanford.edu/projects/glove/

検証結果

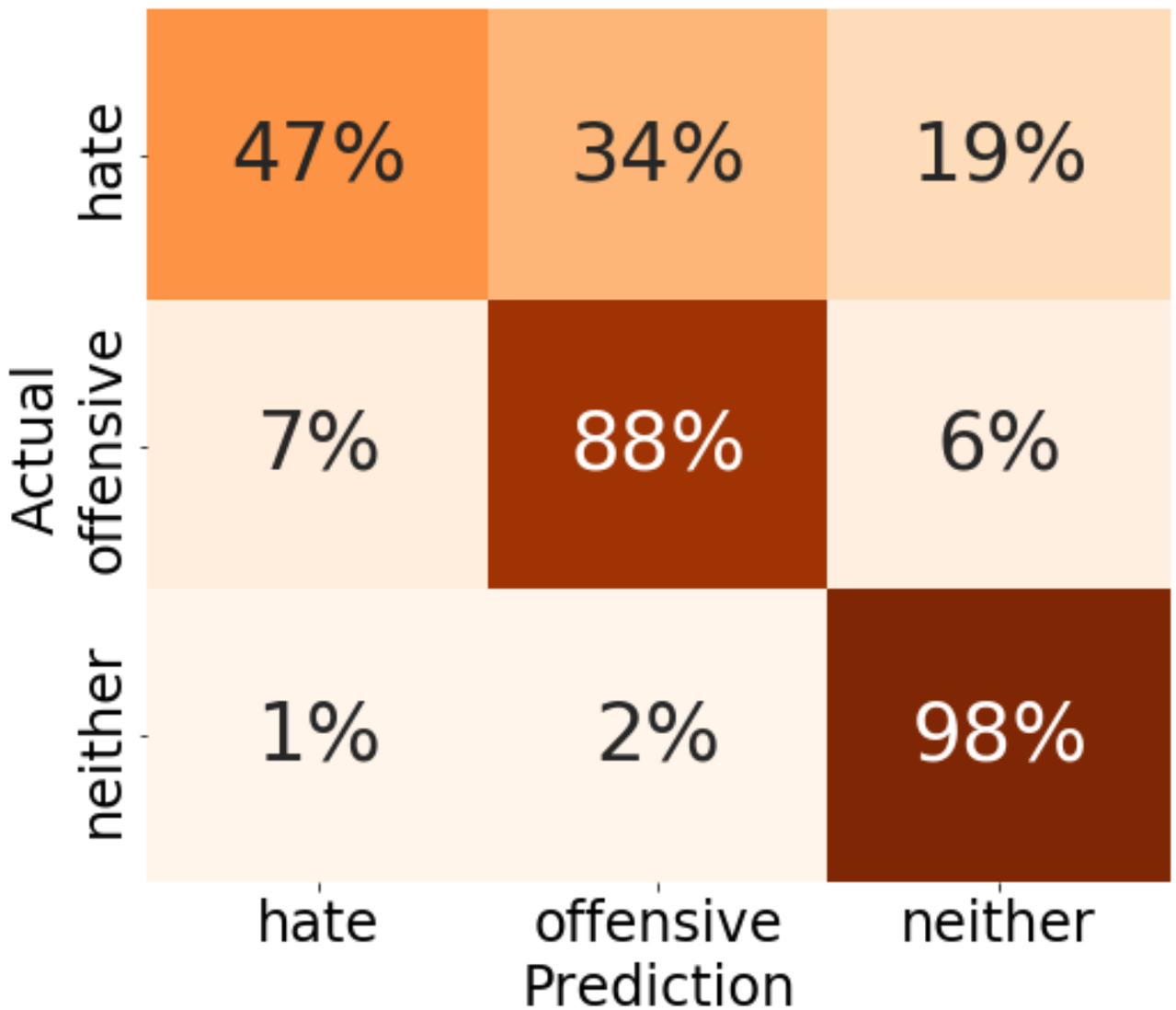


決定木
CART



結果：元データのみ

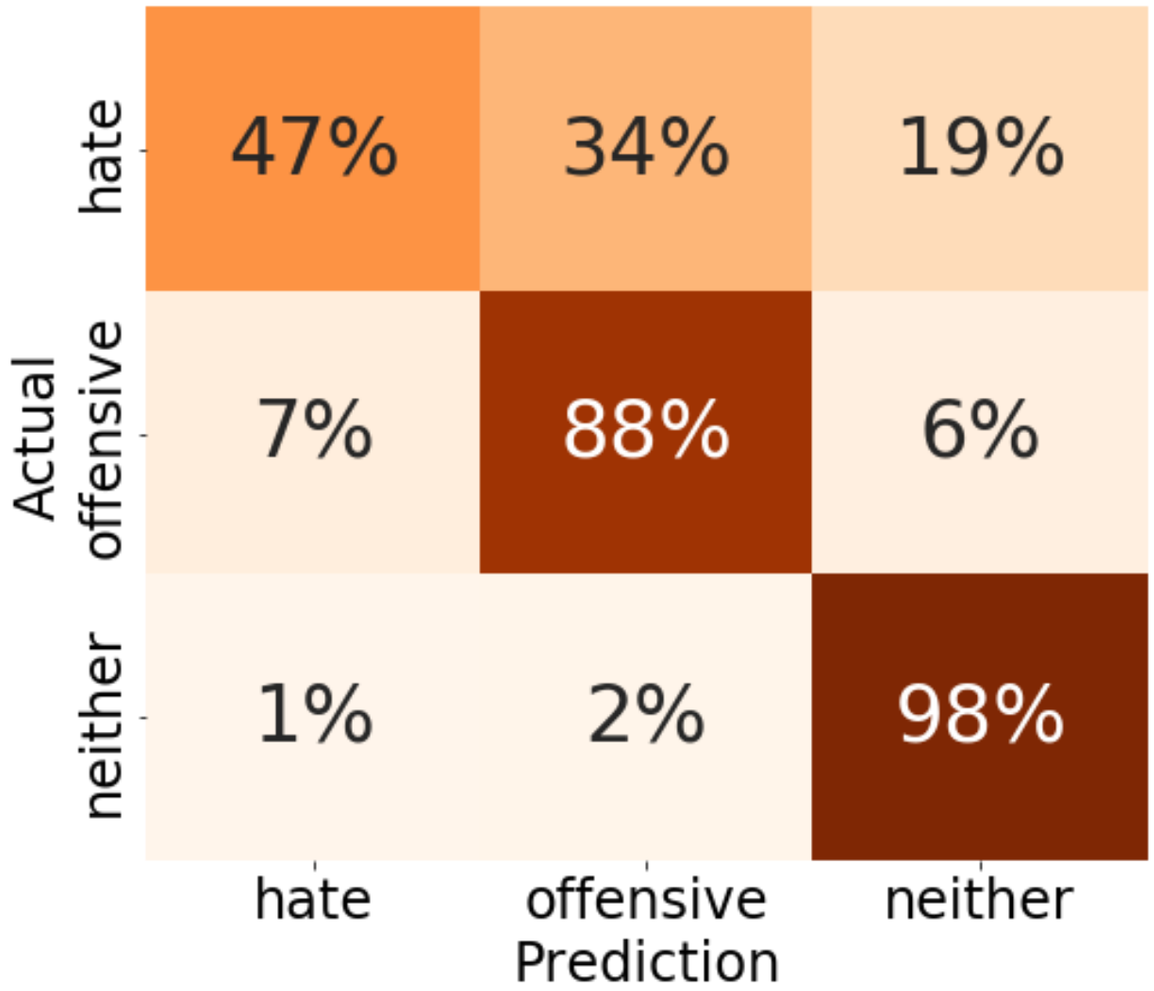
テストデータ分類結果（%）



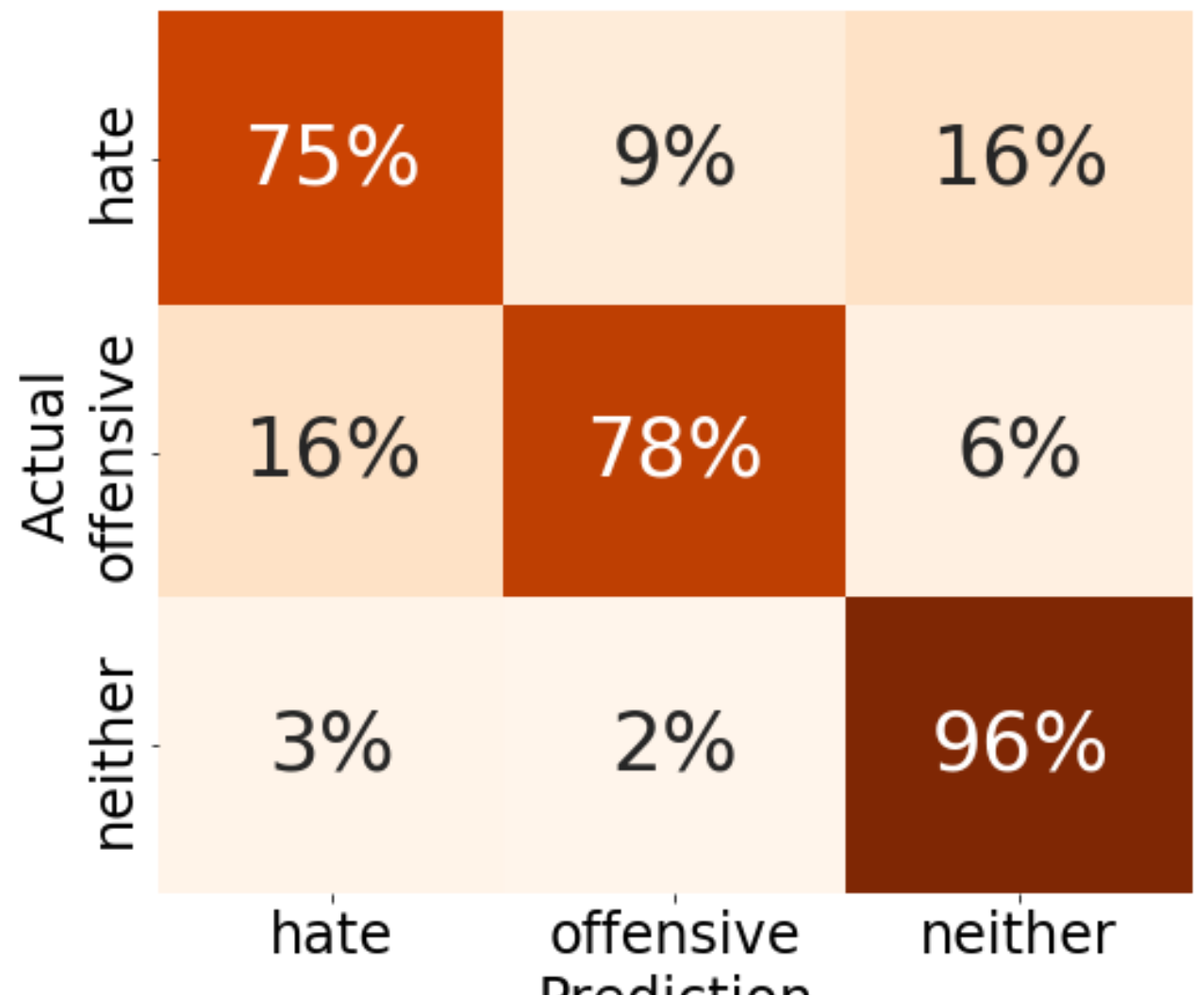
- offensive language, neitherは良く検出できている
  f1-score 約90%
- hate speech の検出が 上手くいっていない
- 特に hate vs offensive に課題
- f1-macro: 78%
- f1-weighted: 85%

結果：水増し後

水増し前



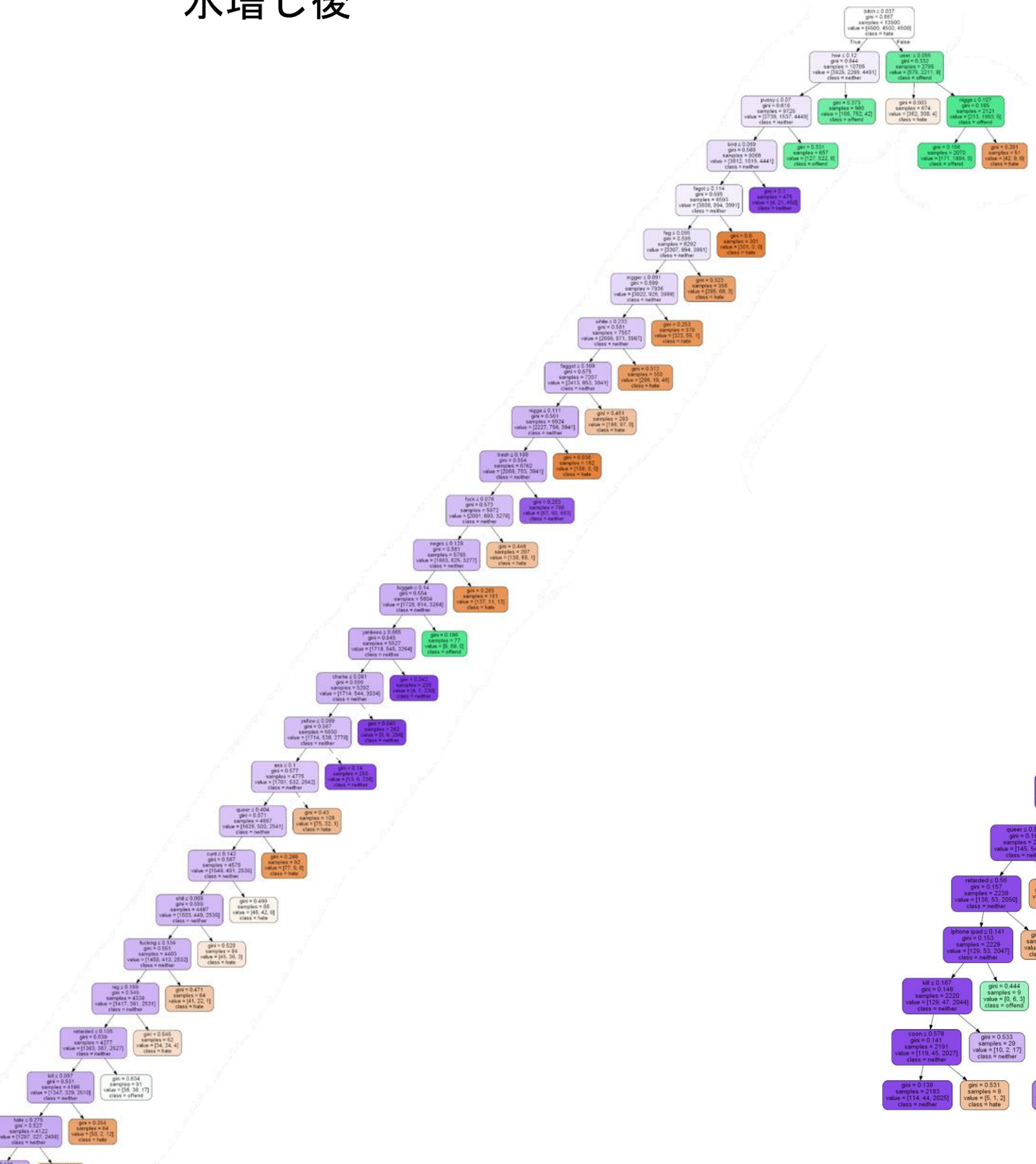
水増し後



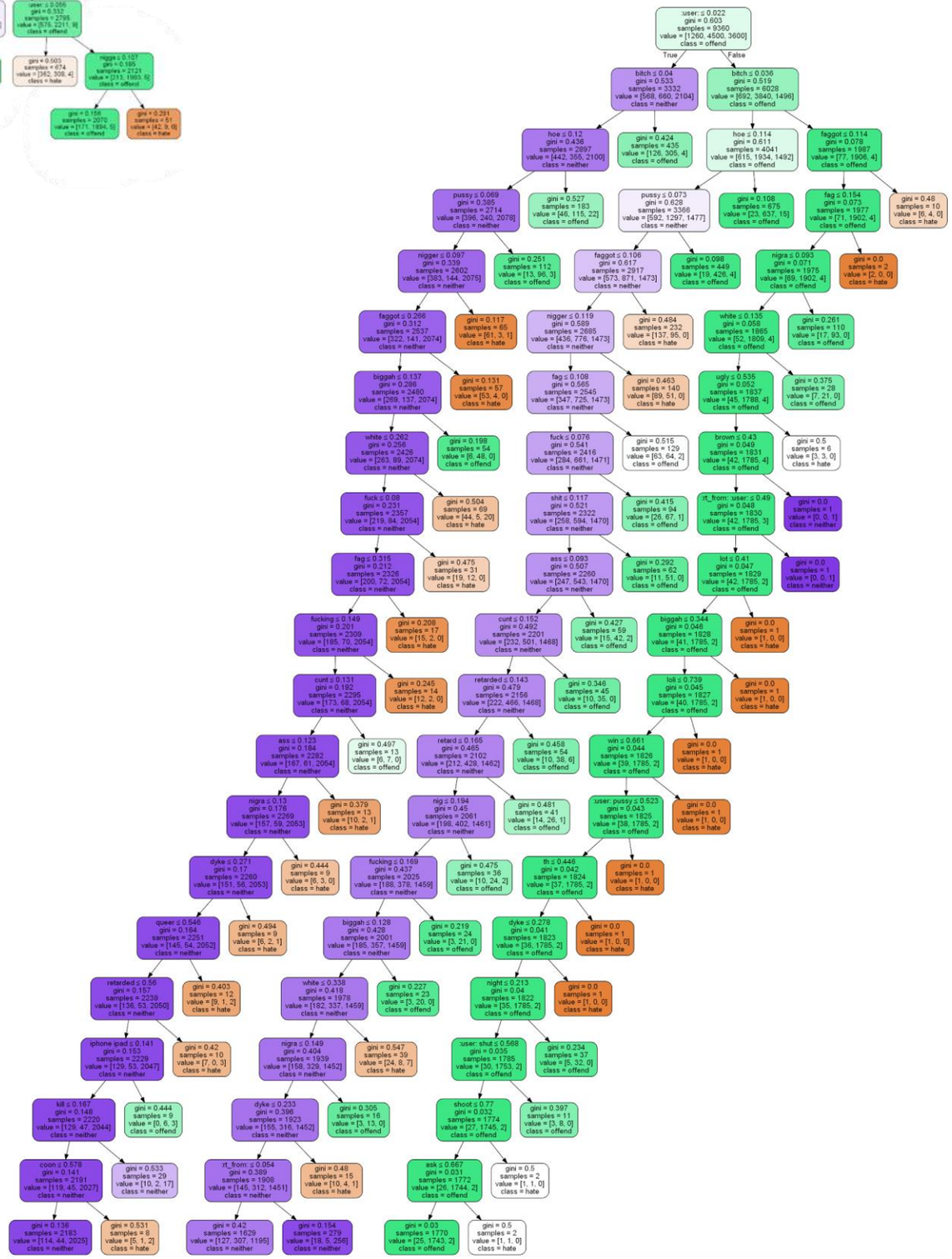
- hate speech の検出率が大幅に上昇
- 他のクラスのhate speech への誤分類が増加

- f1-macro: 78% → 80%
- f1-weighted: 85% → 85%

水増し後



水増し前



- bitch が入っているとneitherがかなり少なくなる
- niggaの有無でhate とoffensive が分類されている箇所
- bird が入っているだけで neither に分類されている箇所

bitch ≤ 0.037
gini = 0.667
samples = 13500
value = [4500, 4500, 4500]
class = hate

Bitch

value =
[hate, offend, neither]

True

False

≤ 0.12
= 0.644
= 10705
value = [3925, 2289, 4491]
class = neither

:user: ≤ 0.055
gini = 0.332
samples = 2795
value = [575, 2211, 9]
class = offend

575, 2211, 9

Nigga

pussy ≤ 0.07
gini = 0.618
samples = 9725
value = [3739, 1537, 4449]
class = neither

gini = 0.373
samples = 980
value = [186, 752, 42]
class = offend

gini = 0.503
samples = 674
value = [362, 308, 4]
class = hate

nigga ≤ 0.107
gini = 0.185
samples = 2121
value = [213, 1903, 5]
class = offend

Bird

bird ≤ 0.069
gini = 0.589
samples = 9068
value = [3612, 1015, 4441]
class = neither

gini = 0.331
samples = 657
value = [127, 522, 8]
class = offend

gini = 0.156
samples = 2070
value = [171, 1894, 5]
class = offend

gini = 0.291
samples = 51
value = [42, 9, 0]
class = hate

Offendクラス

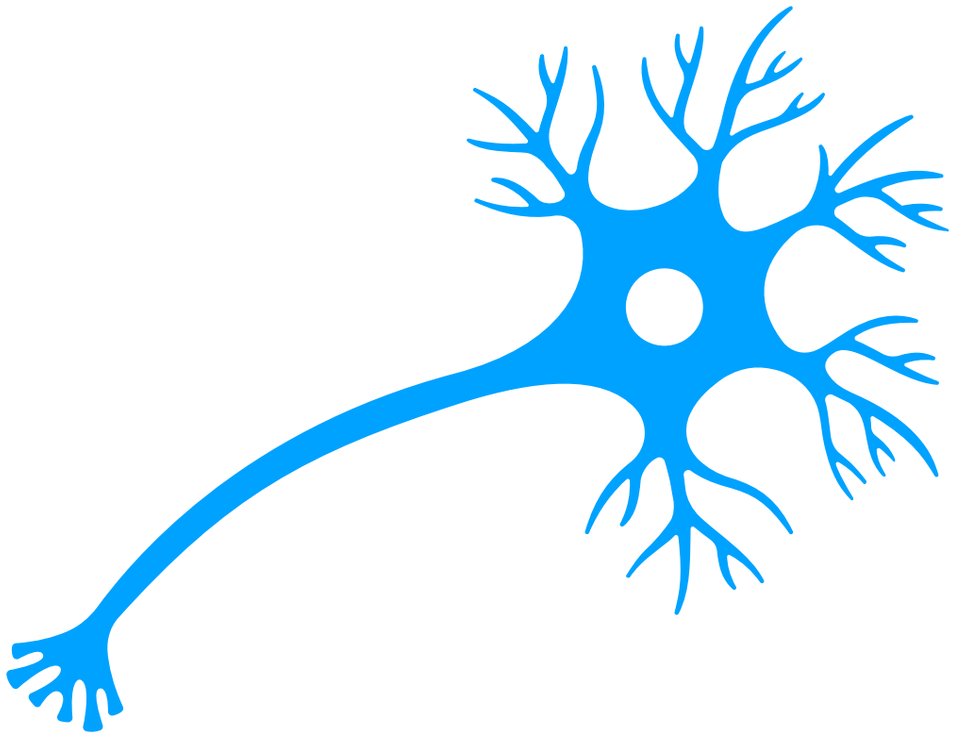
Hateクラス

0.114
0.595
= 8593
, 994, 3991]
neither

gini = 0.1
samples = 475
value = [4, 21, 450]
class = neither

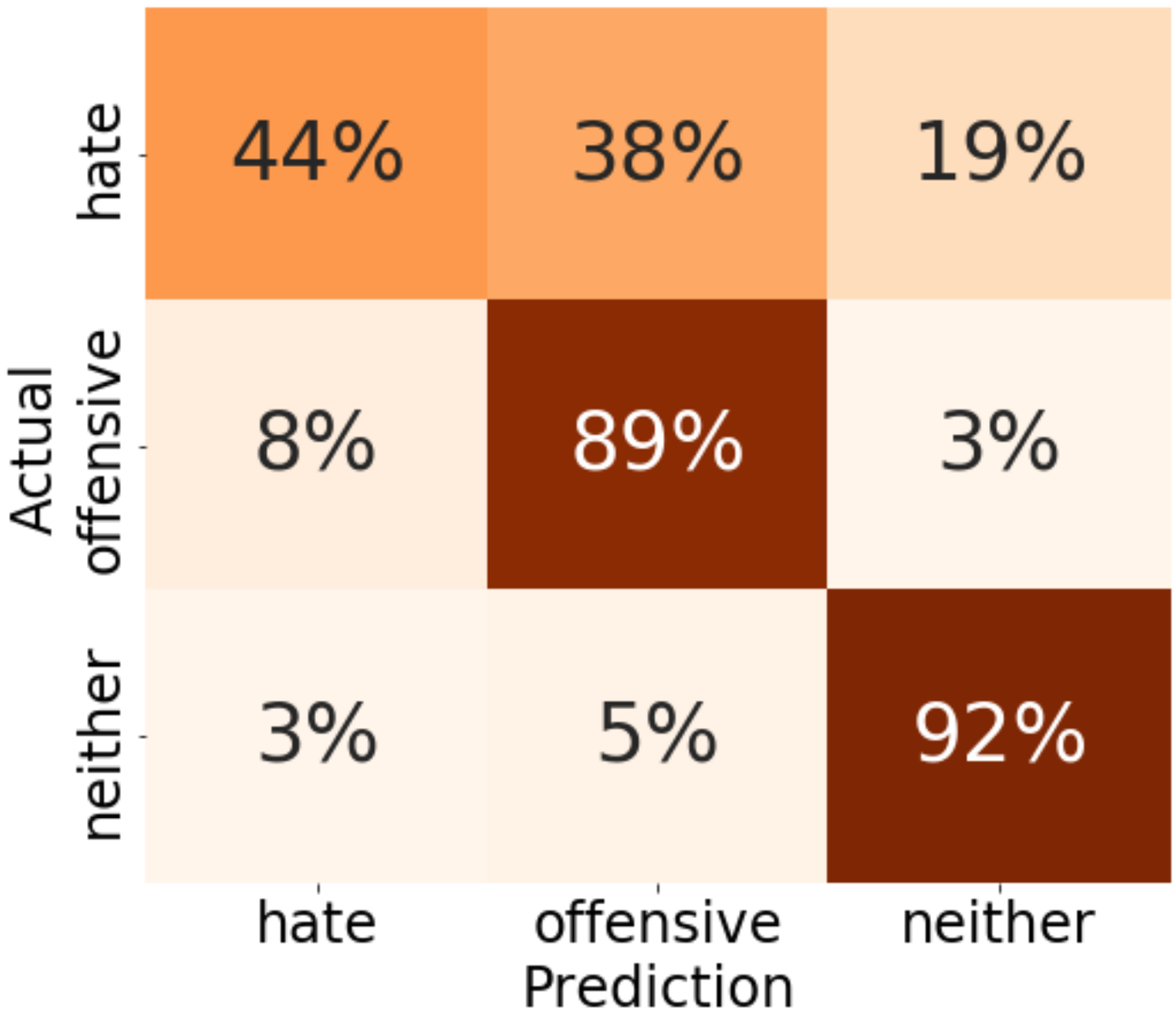
Neitherクラス

LSTM



結果: 元データのみ

テストデータ分類結果（%）



- 決定木と同じくoffensive, neither クラスはよく分類されている

  f1-score 約90%

- hate speech クラスはやはり上手く検出できていない

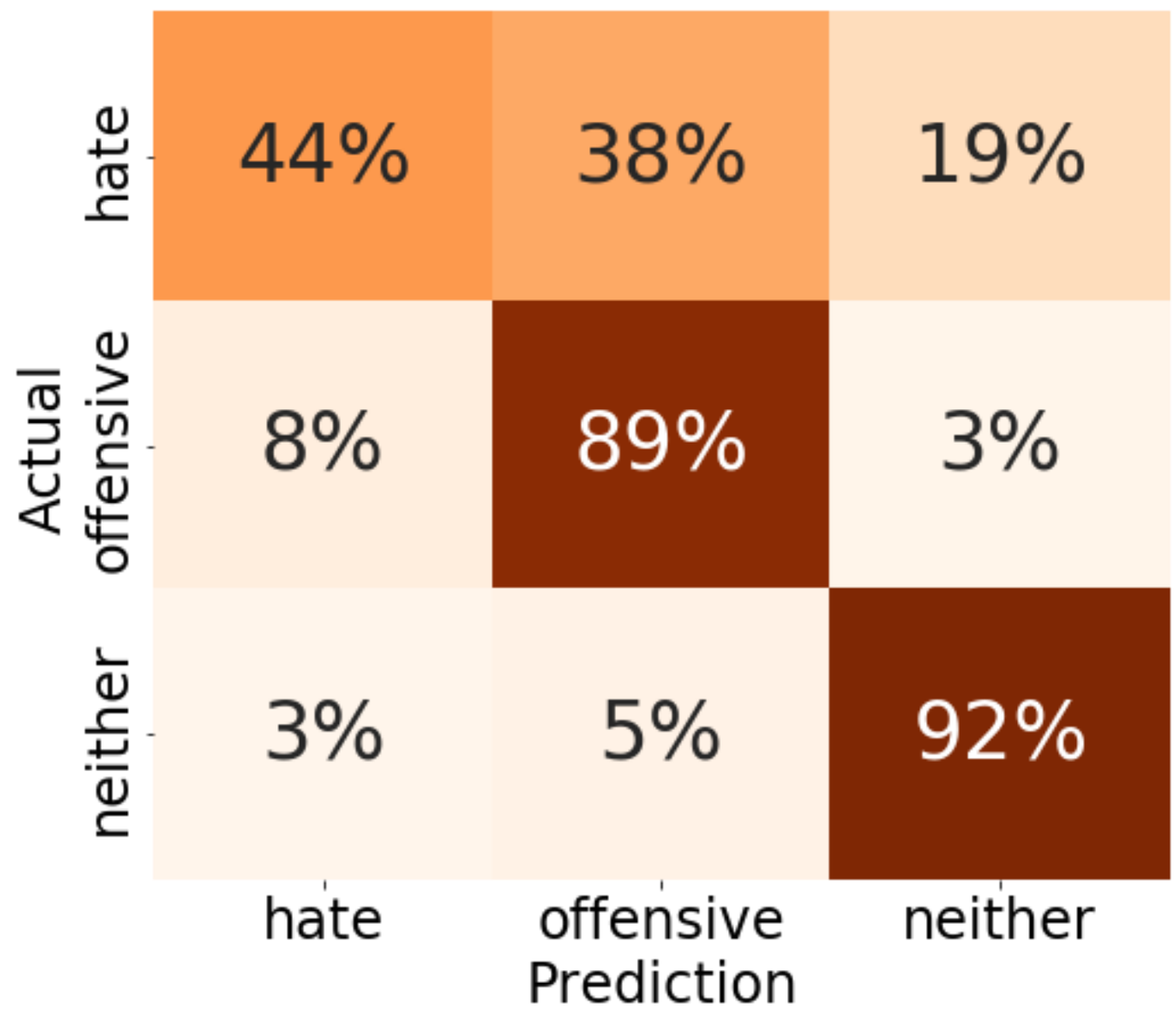
  → 38%がoffensive と誤分類

- f1-macro: 76%

- f1-weighted: 83%

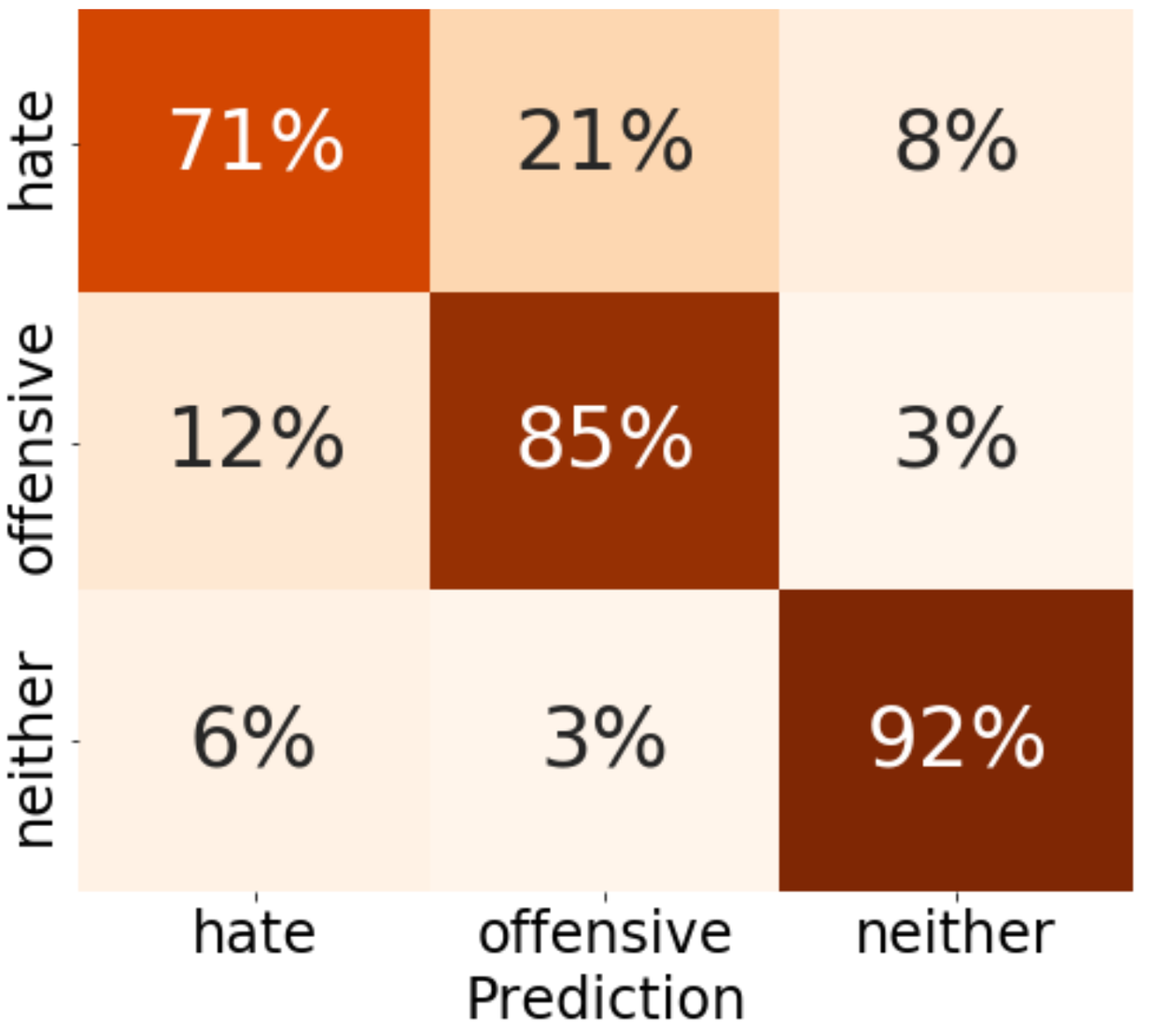
結果: 水増し後

水増し前



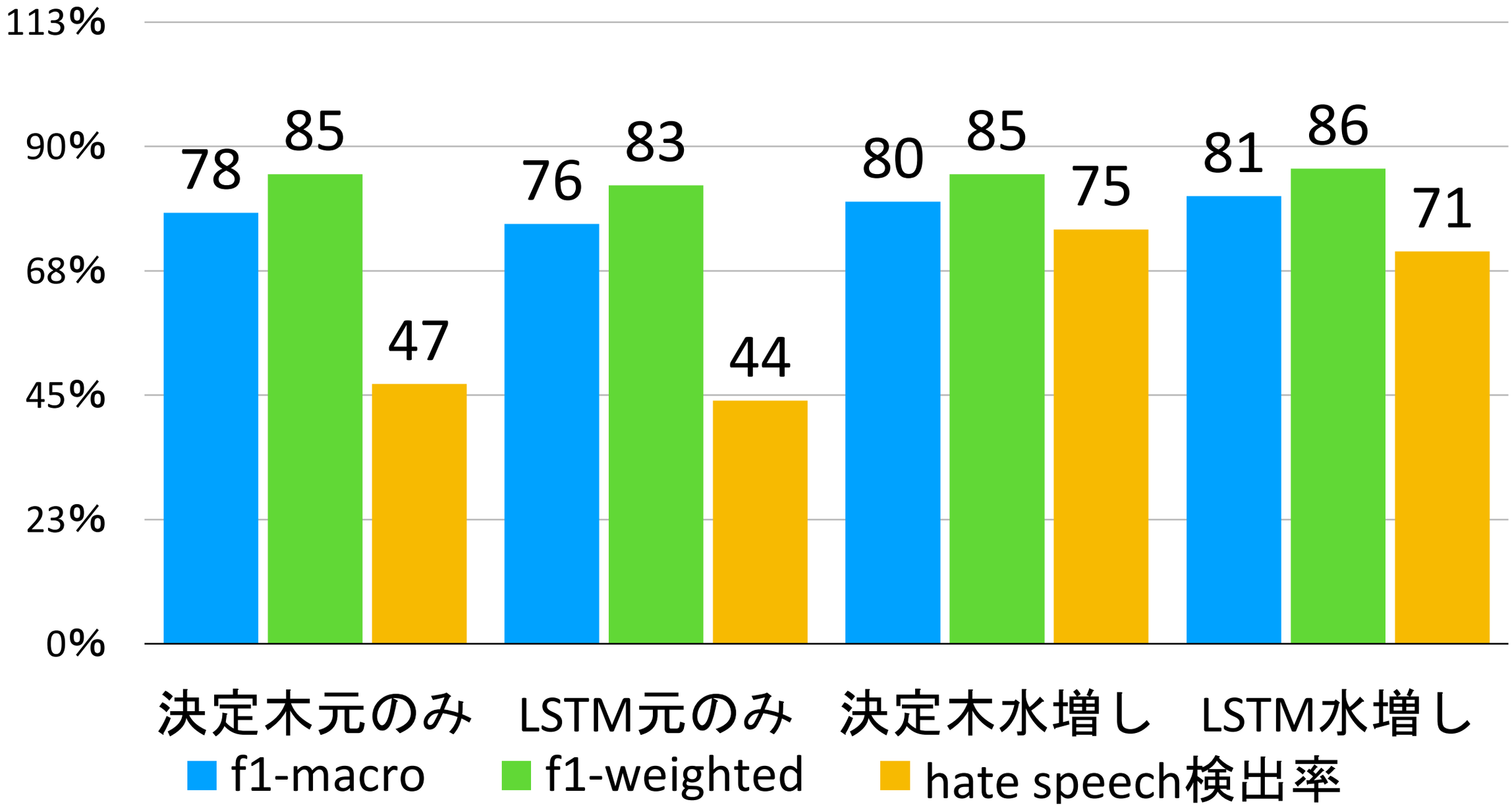
- hate speechクラスの検出率が向上
- hate speech の誤分類量が増加

水増し後



- f1-macro: 76% → 81%
- f1-weighted: 83% → 86%

モデル比較



- 決定木、LSTM共に同等の精度

結論

まとめ

- 決定木、LSTMを用いて分類
- offensive language 及び neither クラス は上手く検出できる

  → それのみを目的とすれば2つとも非常によい
- hate speech の区別には課題

結論: hate vs offensive

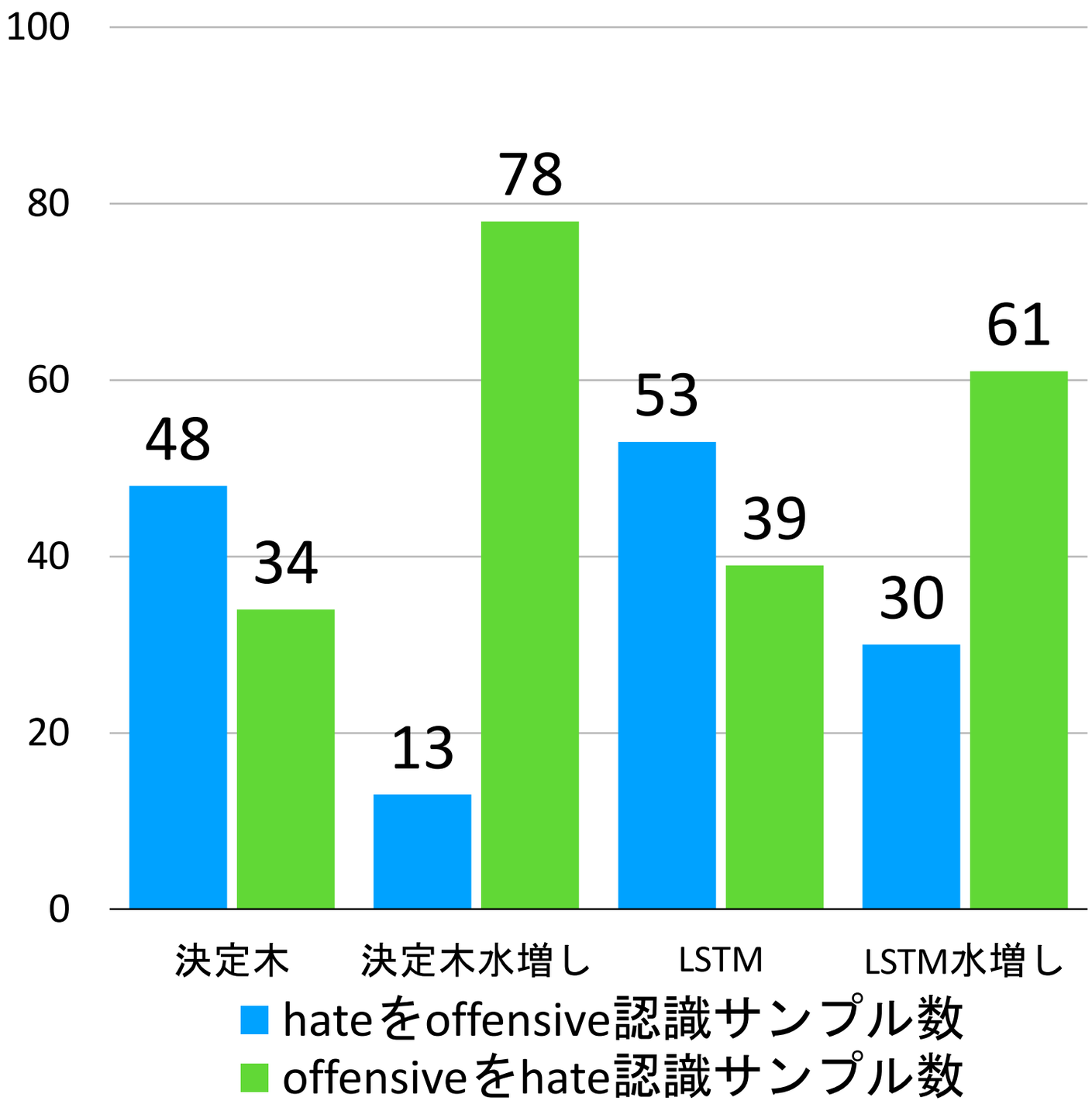
- この二つのクラスの区別がモデルには難しい

- 元データのみ

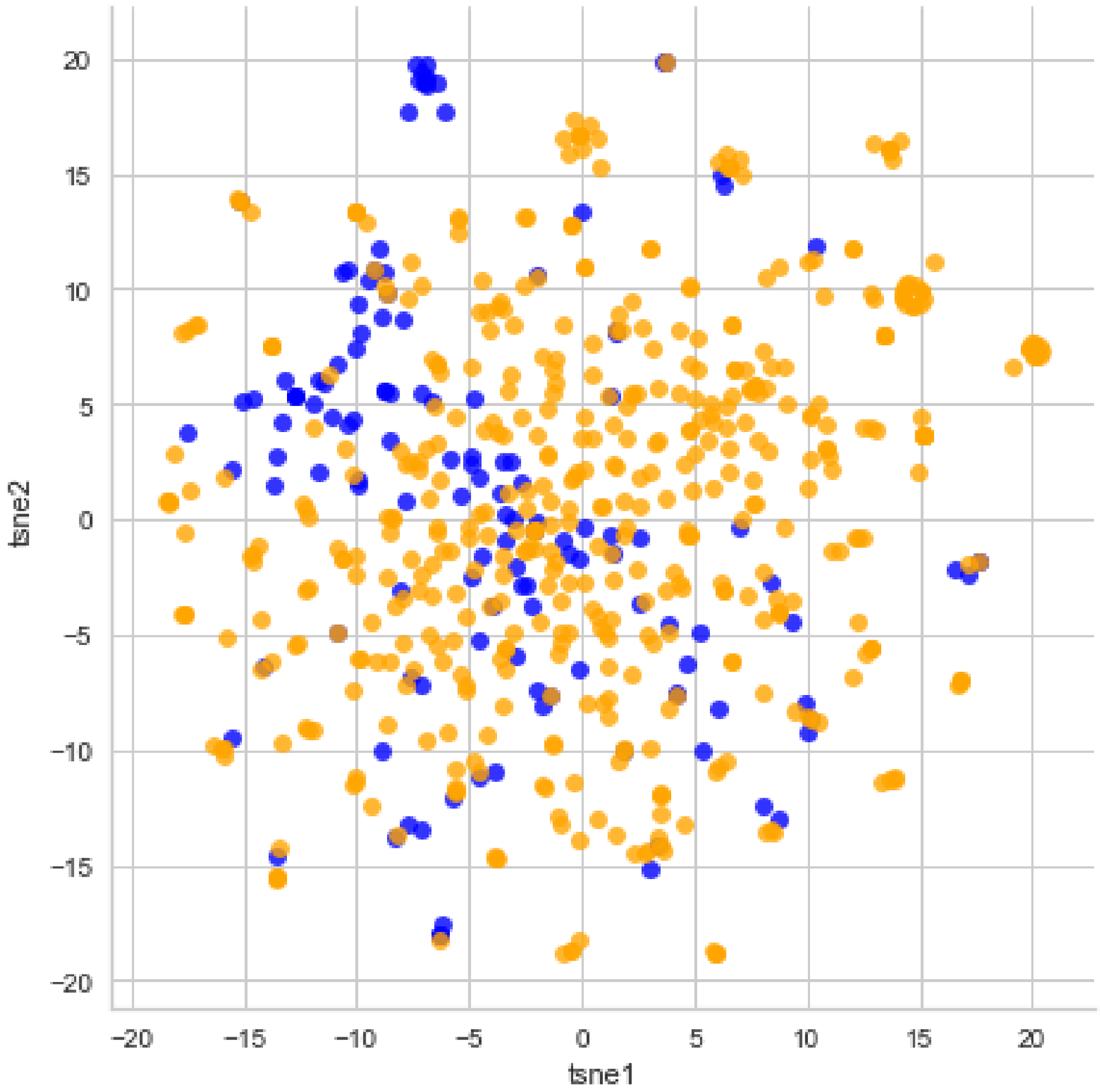
→hate speech がoffensive language として分類されやすい

- クラスバランスを正す

→offensive languageがhate speech として認識されやすくなる



TSNEによる可視化



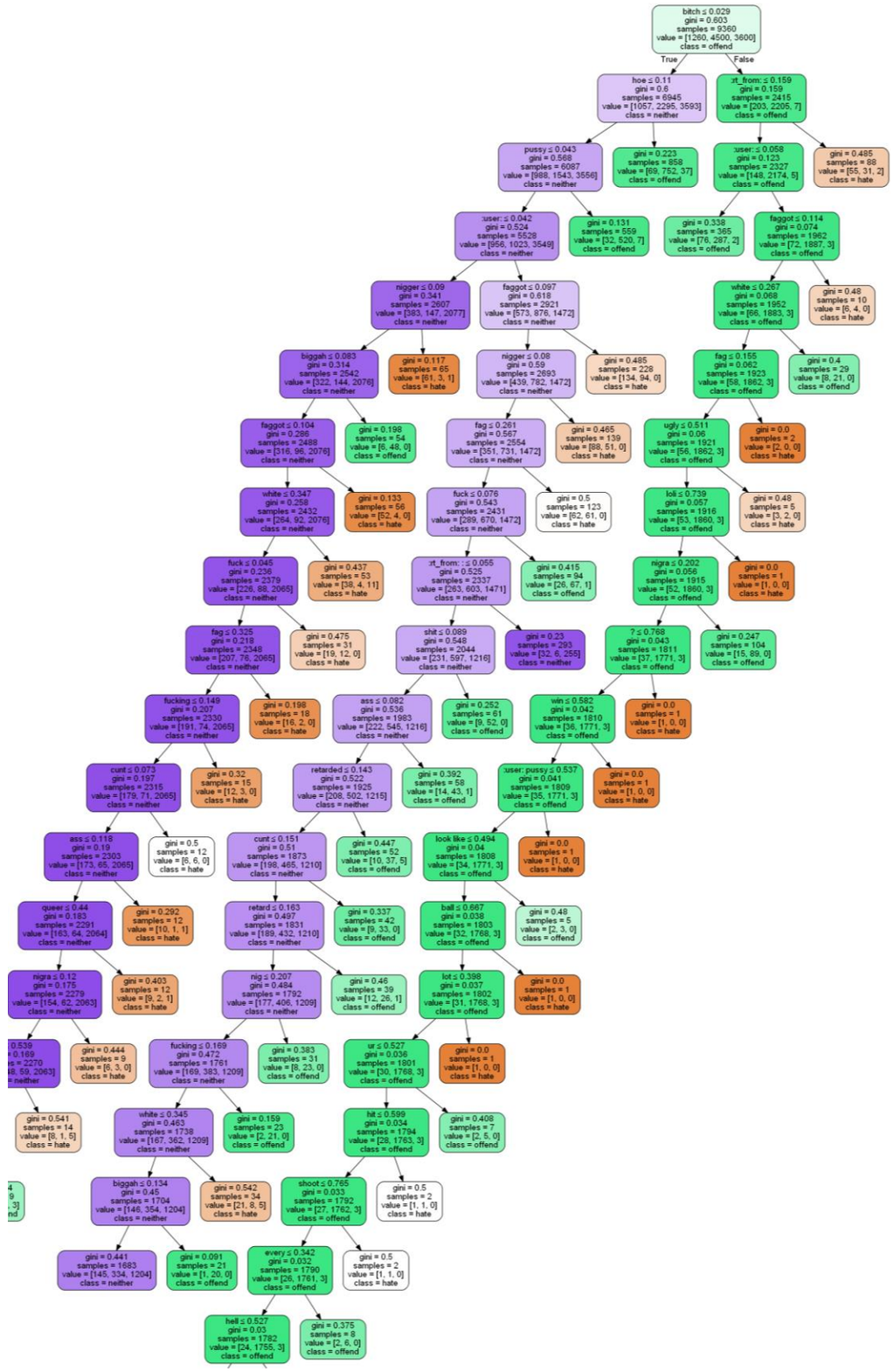
２つのクラスが重なり
あっている事が確認できる

hate speech

offensive language

決定木

ベストモデル


bitch ≤ 0.029
gini = 0.603
samples = 9360
value = [1260, 4500, 3600]
class = offend
True
False
hoe ≤ 0.11
gini = 0.6
samples = 6945
value = [1057, 2295, 3593]
class = neither
:rt_from: ≤ 0.159
gini = 0.159
samples = 2415
value = [203, 2205, 7]
class = offend
pussy ≤ 0.043
gini = 0.568
samples = 6087
value = [988, 1543, 3556]
class = neither
gini = 0.223
samples = 858
value = [69, 752, 37]
class = offend
:user: ≤ 0.058
gini = 0.123
samples = 2327
value = [148, 2174, 5]
class = offend
gini = 0.485
samples = 88
value = [55, 31, 2]
class = hate
:user: ≤ 0.042
gini = 0.524
samples = 5628
value = [956, 1023, 3549]
class = neither
gini = 0.131
samples = 559
value = [32, 520, 7]
class = offend
gini = 0.338
samples = 365
value = [76, 287, 2]
class = offend
faggot ≤ 0.114
gini = 0.074
samples = 1962
value = [72, 1887, 3]
class = offend
nigger ≤ 0.09
gini = 0.341
samples = 2607
value = [383, 147, 2077]
class = neither
faggot ≤ 0.097
gini = 0.618
samples = 2921
value = [573, 876, 1472]
class = neither
white ≤ 0.267
gini = 0.068
samples = 1952
value = [66, 1883, 3]
class = offend
gini = 0.48
samples = 10
value = [6, 4, 0]
class = hate
biggah ≤ 0.083
gini = 0.314
samples = 2542
value = [322, 144, 2076]
class = neither
gini = 0.117
samples = 65
value = [61, 3, 1]
class = hate
nigger ≤ 0.08
gini = 0.59
samples = 2693
value = [439, 782, 1472]
class = neither
gini = 0.485
samples = 228
value = [134, 94, 0]
class = hate
fag ≤ 0.155
gini = 0.062
samples = 1923
value = [58, 1862, 3]
class = offend
gini = 0.4
samples = 29
value = [8, 21, 0]
class = offend
faggot ≤ 0.104
gini = 0.286
samples = 2488
value = [316, 96, 2076]
class = neither
gini = 0.198
samples = 54
value = [6, 48, 0]
class = offend
fag ≤ 0.261
gini = 0.567
samples = 2554
value = [351, 731, 1472]
class = neither
gini = 0.465
samples = 139
value = [88, 51, 0]
class = hate
ugly ≤ 0.511
gini = 0.06
samples = 1921
value = [56, 1862, 3]
class = offend
gini = 0.0
samples = 2
value = [2, 0, 0]
class = hate
white ≤ 0.347
gini = 0.258
samples = 2432
value = [264, 92, 2076]
class = neither
gini = 0.133
samples = 56
value = [52, 4, 0]
class = hate
fuck ≤ 0.076
gini = 0.543
samples = 2431
value = [289, 670, 1472]
class = neither
gini = 0.5
samples = 123
value = [62, 61, 0]
class = hate
loli ≤ 0.739
gini = 0.057
samples = 1916
value = [53, 1860, 3]
class = offend
gini = 0.48
samples = 5
value = [3, 2, 0]
class = hate
fuck ≤ 0.045
gini = 0.236
samples = 2379
value = [226, 88, 2065]
class = neither
gini = 0.437
samples = 53
value = [38, 4, 11]
class = hate
:rt_from: ≤ 0.055
gini = 0.525
samples = 2337
value = [263, 603, 1471]
class = neither
gini = 0.415
samples = 94
value = [26, 67, 1]
class = offend
nigra ≤ 0.202
gini = 0.056
samples = 1915
value = [52, 1860, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
fag ≤ 0.325
gini = 0.218
samples = 2348
value = [207, 76, 2065]
class = neither
gini = 0.475
samples = 31
value = [19, 12, 0]
class = hate
shit ≤ 0.089
gini = 0.548
samples = 2044
value = [231, 597, 1216]
class = neither
gini = 0.23
samples = 293
value = [32, 6, 255]
class = neither
? ≤ 0.768
gini = 0.043
samples = 1811
value = [37, 1771, 3]
class = offend
gini = 0.247
samples = 104
value = [15, 89, 0]
class = offend
fucking ≤ 0.149
gini = 0.207
samples = 2330
value = [191, 74, 2065]
class = neither
gini = 0.198
samples = 18
value = [16, 2, 0]
class = hate
ass ≤ 0.082
gini = 0.536
samples = 1983
value = [222, 545, 1216]
class = neither
gini = 0.252
samples = 61
value = [9, 52, 0]
class = offend
win ≤ 0.582
gini = 0.042
samples = 1810
value = [36, 1771, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
cunt ≤ 0.073
gini = 0.197
samples = 2315
value = [179, 71, 2065]
class = neither
gini = 0.32
samples = 15
value = [12, 3, 0]
class = hate
retarded ≤ 0.143
gini = 0.522
samples = 1925
value = [208, 502, 1215]
class = neither
gini = 0.392
samples = 58
value = [14, 43, 1]
class = offend
:user: pussy ≤ 0.537
gini = 0.041
samples = 1809
value = [35, 1771, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
ass ≤ 0.118
gini = 0.19
samples = 2303
value = [173, 68, 2065]
class = neither
gini = 0.5
samples = 12
value = [6, 6, 0]
class = hate
cunt ≤ 0.151
gini = 0.51
samples = 1873
value = [198, 465, 1210]
class = neither
gini = 0.447
samples = 52
value = [10, 37, 5]
class = offend
look like ≤ 0.494
gini = 0.04
samples = 1808
value = [34, 1771, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
queer ≤ 0.44
gini = 0.183
samples = 2291
value = [163, 64, 2064]
class = neither
gini = 0.292
samples = 12
value = [10, 1, 1]
class = hate
retard ≤ 0.163
gini = 0.497
samples = 1831
value = [189, 432, 1210]
class = neither
gini = 0.337
samples = 42
value = [9, 33, 0]
class = offend
ball ≤ 0.667
gini = 0.038
samples = 1803
value = [32, 1768, 3]
class = offend
gini = 0.48
samples = 5
value = [2, 3, 0]
class = offend
nigra ≤ 0.12
gini = 0.175
samples = 2279
value = [154, 62, 2063]
class = neither
gini = 0.403
samples = 12
value = [9, 2, 1]
class = hate
nig ≤ 0.207
gini = 0.494
samples = 1792
value = [177, 406, 1209]
class = neither
gini = 0.46
samples = 39
value = [12, 26, 1]
class = offend
lot ≤ 0.386
gini = 0.037
samples = 1802
value = [31, 1768, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
gini = 0.444
samples = 9
value = [6, 3, 0]
class = hate
fucking ≤ 0.169
gini = 0.472
samples = 1761
value = [169, 383, 1209]
class = neither
gini = 0.383
samples = 31
value = [8, 23, 0]
class = offend
ur ≤ 0.527
gini = 0.036
samples = 1801
value = [30, 1768, 3]
class = offend
gini = 0.0
samples = 1
value = [1, 0, 0]
class = hate
gini = 0.541
samples = 14
value = [8, 1, 5]
class = hate
white ≤ 0.345
gini = 0.463
samples = 1738
value = [167, 362, 1209]
class = neither
gini = 0.159
samples = 23
value = [2, 21, 0]
class = offend
hit ≤ 0.599
gini = 0.034
samples = 1794
value = [28, 1763, 3]
class = offend
gini = 0.408
samples = 7
value = [2, 5, 0]
class = offend
biggah ≤ 0.134
gini = 0.45
samples = 1704
value = [146, 354, 1204]
class = neither
gini = 0.542
samples = 34
value = [21, 8, 5]
class = hate
shoot ≤ 0.765
gini = 0.033
samples = 1792
value = [27, 1762, 3]
class = offend
gini = 0.5
samples = 2
value = [1, 1, 0]
class = hate
gini = 0.441
samples = 1683
value = [145, 334, 1204]
class = neither
gini = 0.091
samples = 21
value = [1, 20, 0]
class = offend
every ≤ 0.342
gini = 0.032
samples = 1790
value = [26, 1761, 3]
class = offend
gini = 0.5
samples = 2
value = [1, 1, 0]
class = hate
hell ≤ 0.527
gini = 0.03
samples = 1782
value = [24, 1755, 3]
class = offend
gini = 0.375
samples = 8
value = [2, 6, 0]
class = offend

結論:ベストモデル

- モデルの判断の説明が簡単

  →ユーザーの投稿削除した時に理由が説明できる

  →ユーザー体験を損なわない

- 精度も遜色ない

  → f1-weighted 85%

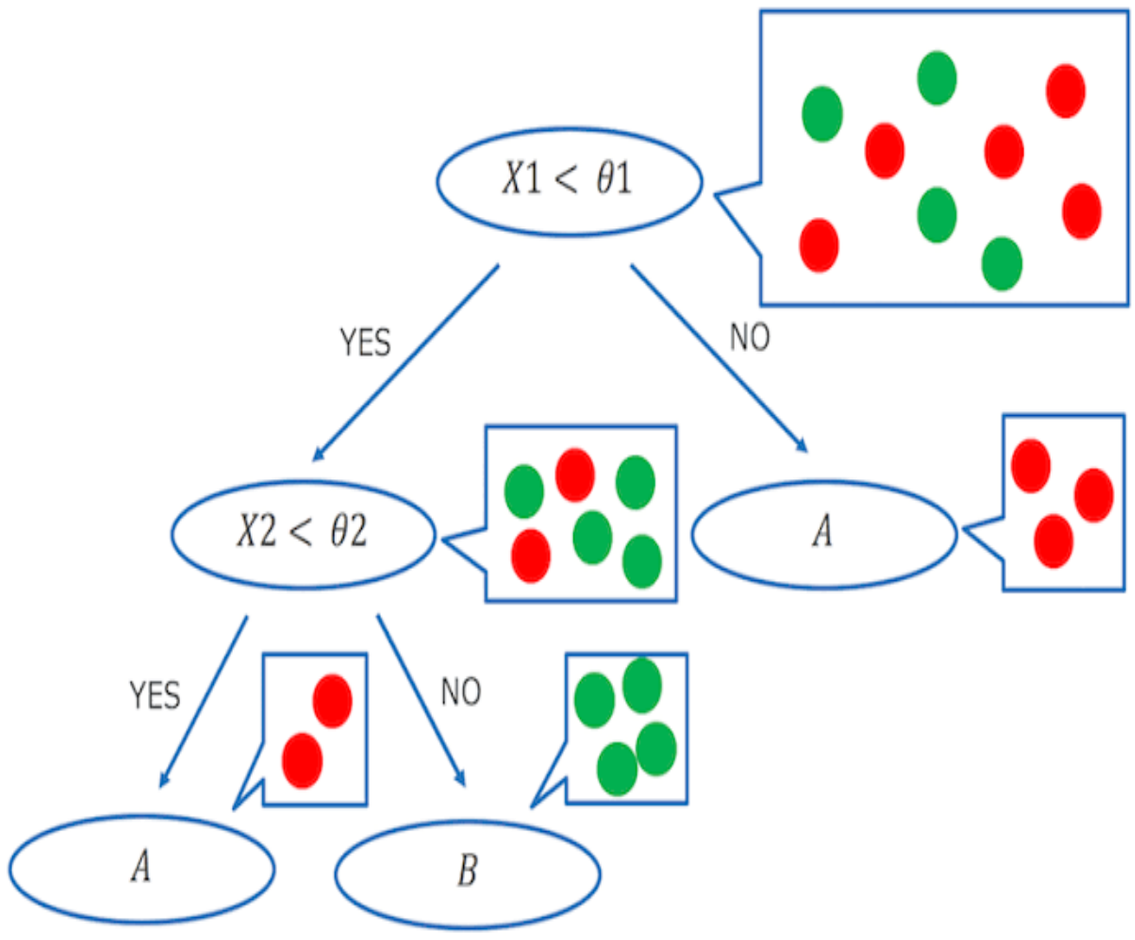
今後の課題

1. 正式なデータをもっと増やす
2. 決定木の欠陥の改善
3. SVM など他のモデルを試す
4. LSTM に関してはconvolutional 層を追加するなど
   他の構成を試す

モデル紹介:補足

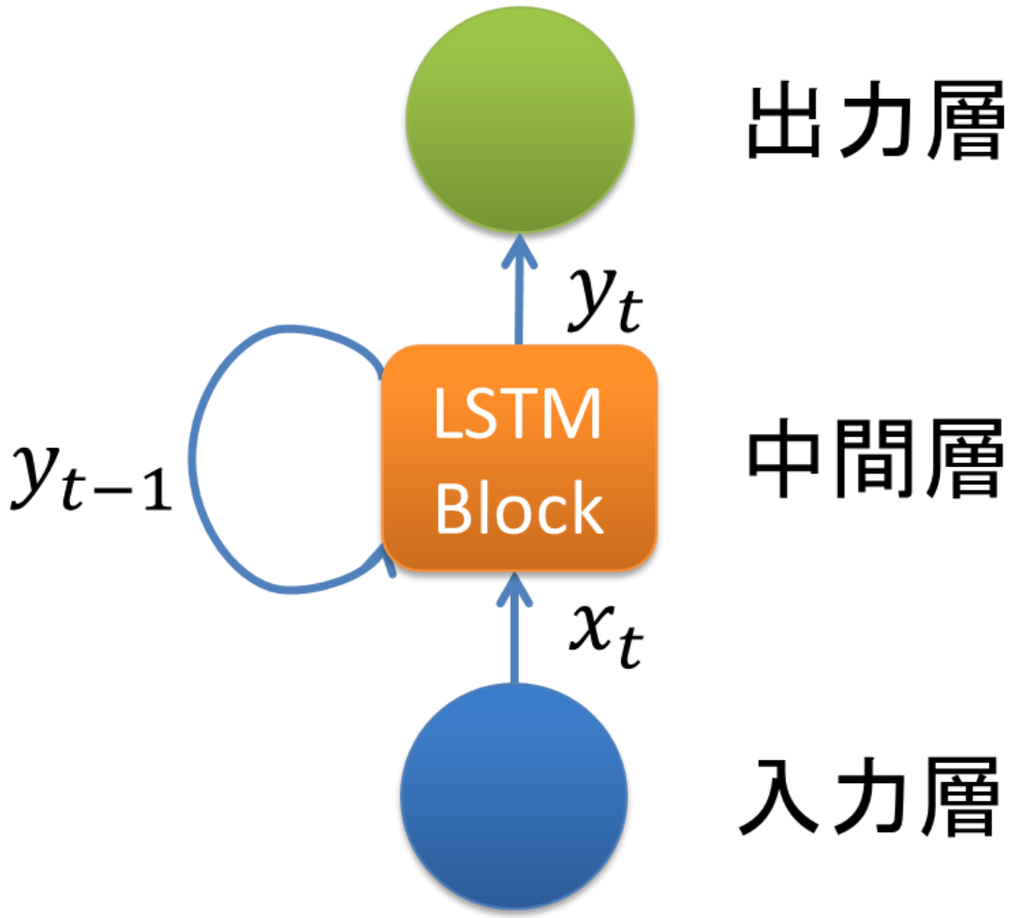
- 不純度が最も減少する特徴としきい値でノード二分割を繰り返す
- Gini係数を不純度として使用

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$



モデル紹介:補足

- RNNの一種
- LSTM Block を中間層に使用以下の時系列(t-1) の情報を(t)へと再帰
  - C(t-1) :セル状態
  - h(t-1) : 出力
- 長期の依存性の学習が可能



Long-Short Term Memory