Toxic Comment Classification

Disclaimer

The Presentation Slides contain text
that may be considered profane, vulgar, or offensive

서대원, 김동욱, 김호현. 이민규, 임희진

CONTENTS

CONTENTS

OVERVIEW

powered by Google

Toxic Comment Classification Challenge

Purpose | To make health environment for the Internet Users

Data Source | Wikipedia corpus dataset
collected through Google Human-labeling

Approximately 63M comments from web pages/
article from 2004 to 2015

OVERVIEW : Multiclass-Multilabel Classification

X_label

```
x.head(5)
```

	comment_text
0	Explanation₩nWhy the edits made under my usern...
1	D'aww! He matches this background colour I'm s...
2	Hey man, I'm really not trying to edit war. It...
3	"₩nMore₩nI can't make any real suggestions on ...
4	You, sir, are my hero. Any chance you remember...

Y_label

```
y.head(5)
```

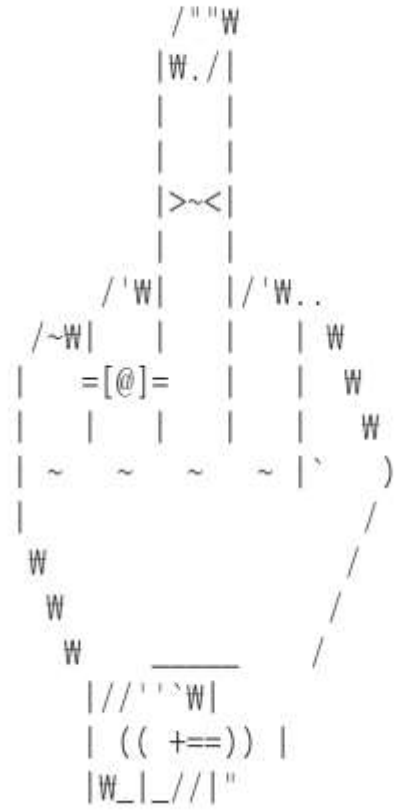
	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

SAMPLES : Stereotypical Sample

comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
[redacted]UCKER BEFORE YOU [redacted] AROUND ON MY WORK	1	1	1	0	1	0

SAMPLES : Atypical Sample

"
SURPRISE!

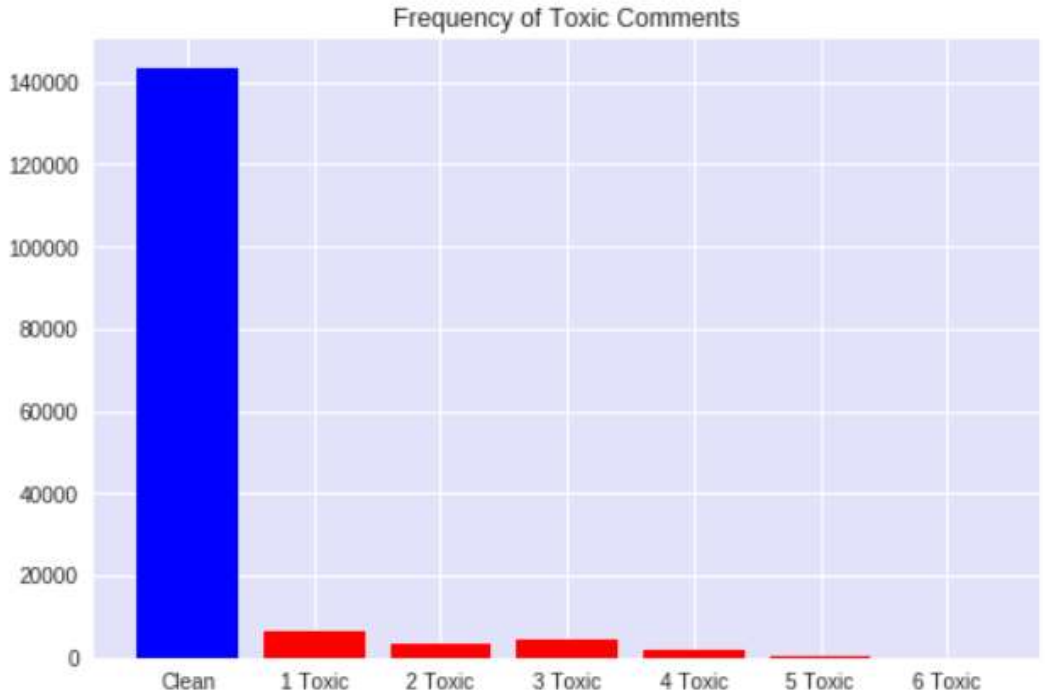


id	8b20912530eebd56
comment_text	"WnSURPRISE!Wn /""WWn ...
toxic	0
severe_toxic	0
obscene	0
threat	0
insult	0
identity_hate	0

Name: 152298, dtype: object

Visualization : The vast majority of comments are Non-toxic

The Proportion of Clean comments 〉 95%

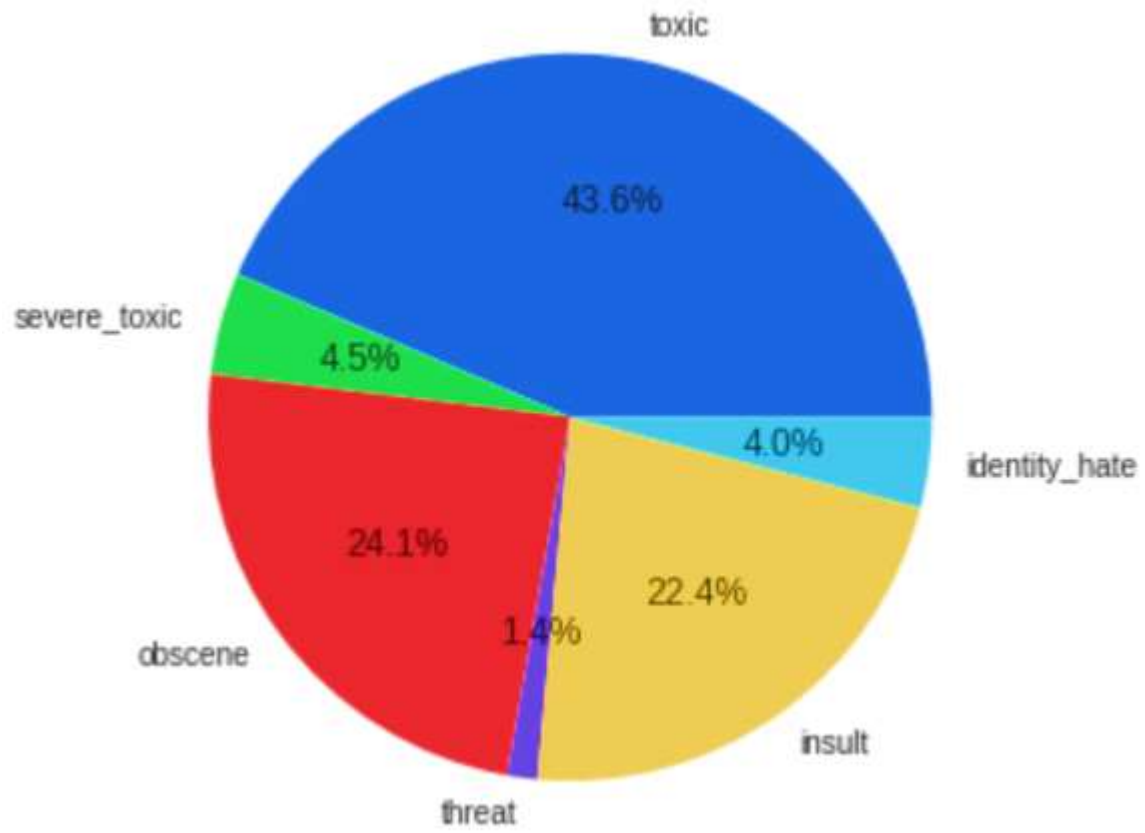



```
import numpy as np
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_auc_score

print("bare minimum acc_score:%0.3f"
      %(accuracy_score(y, np.zeros(y.shape))))
print('bare minimum roc_auc_score:%0.3f'
      %(roc_auc_score(y, np.random.rand(159571,6))))
```

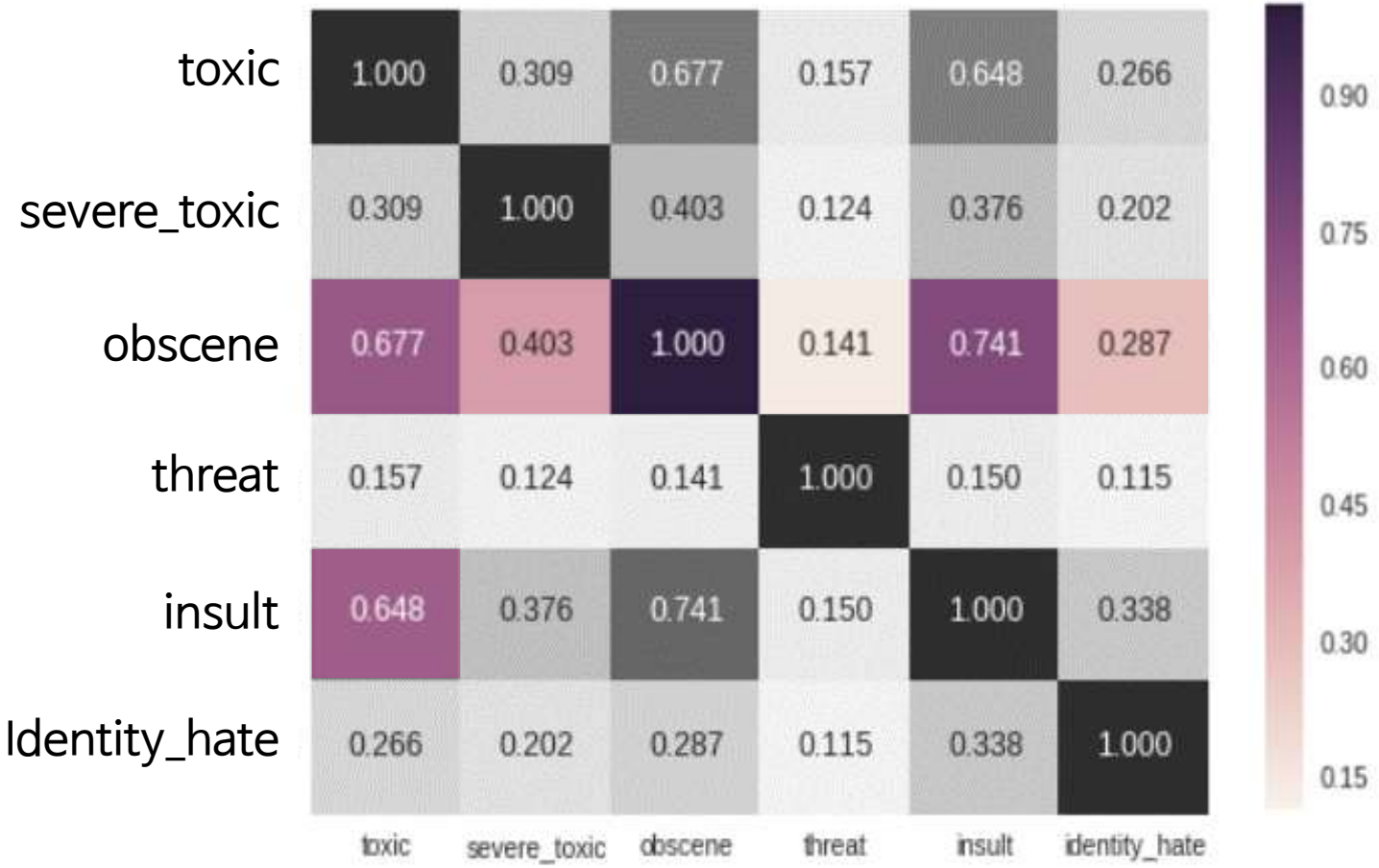
```
bare minimum acc_score:0.898
bare minimum roc_auc_score:0.503
```

Visualization : the Proportion of Each Label



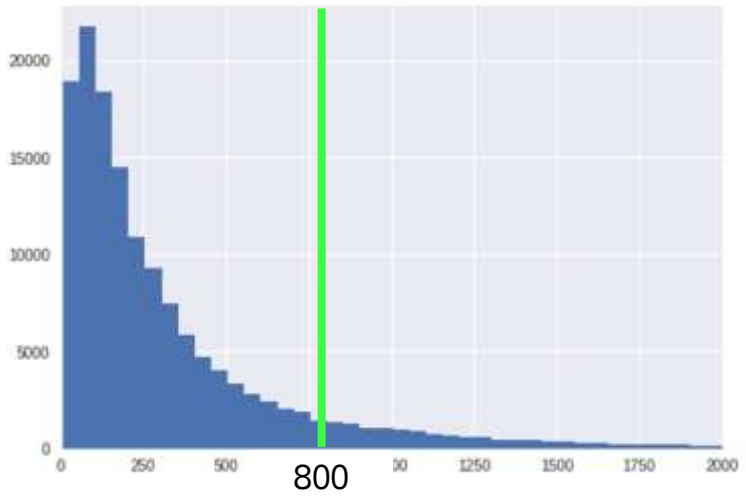
43.6%	Toxic
22.4%	Insult
24.1%	obscene
4.5%	severe_toxic
4.0%	identity_hate
1.4%	threat

Visualization : the Correlation between Y_label



Visualization : the Length of Text

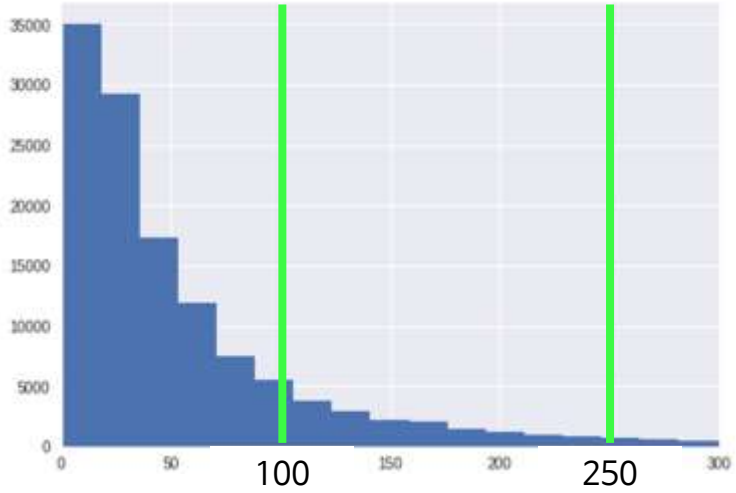
WORD_wise



Tokenization max length = 800

김호현

CHARCATER_wise



Tokenization max length = 250

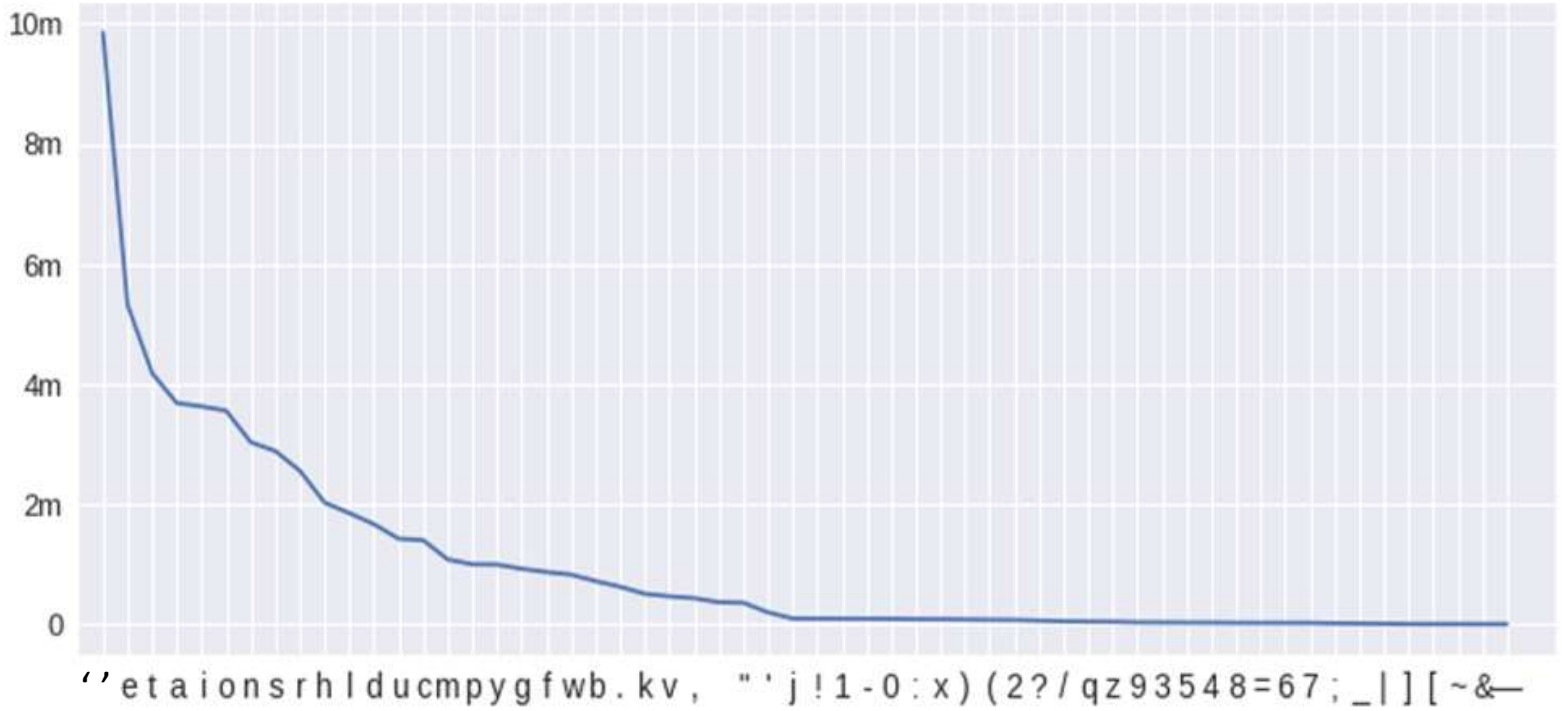
김호현

Tokenization max length = 100

서대원

Visualization : the Frequency of Character

Many Special Characters are Placed in High Rank



총 문자 수 : 1934

문자 출현 빈도 순위
‘ ‘ : 9850190
‘e’ : 5320996
‘t’ : 4179970
‘a’ : 3685571

Visualization : Special Characters

Visualization : Word Cloud

Clean

Toxic

Insult

Threat

Obscene

Severe Toxic

Identity_hate

1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

- Tokenization
- TF-IDF

3 Model

- Model1
- Model2
- Model3
- Model4

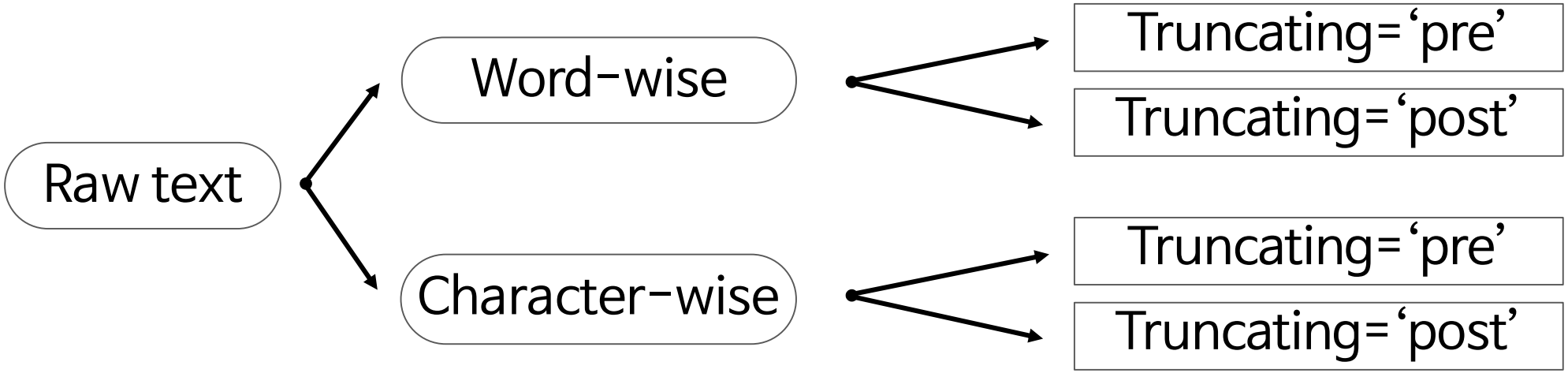
4 Result

- Lime Package

Tokenization

Step1. Tokenization

Step2. Padding

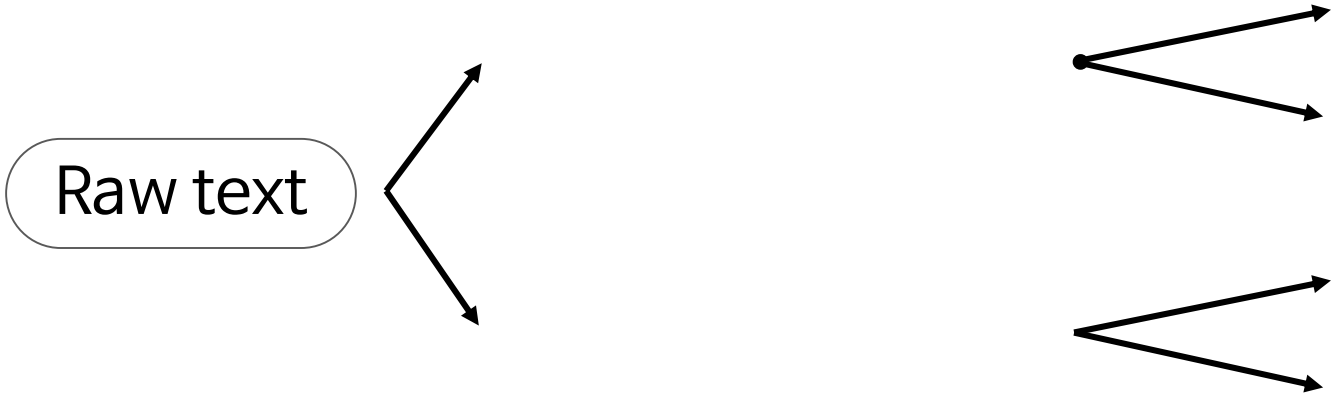


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

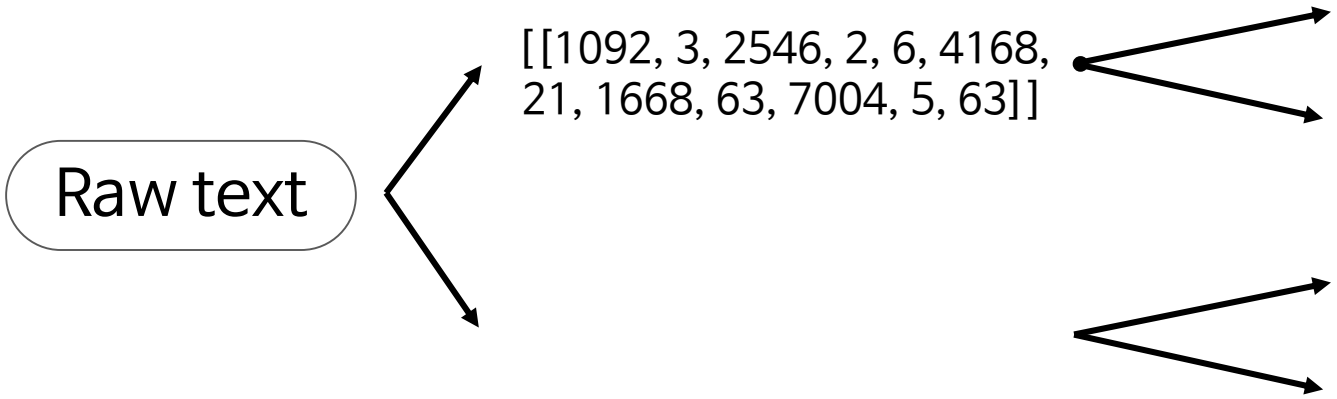


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

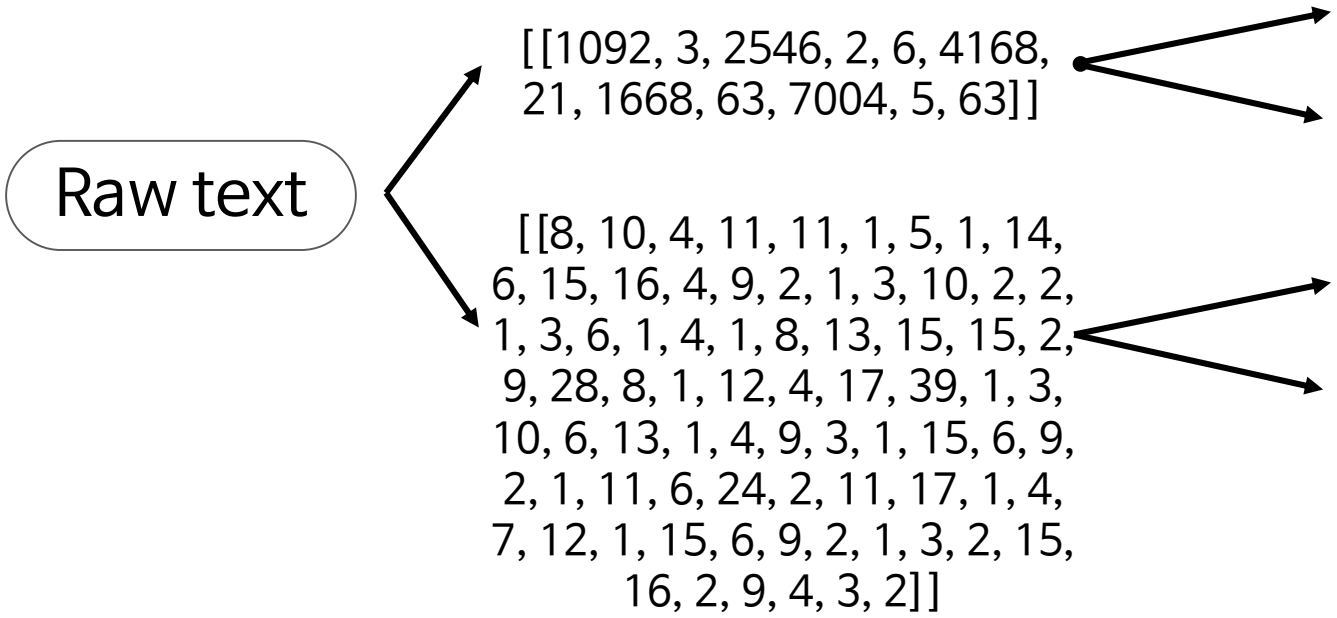


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

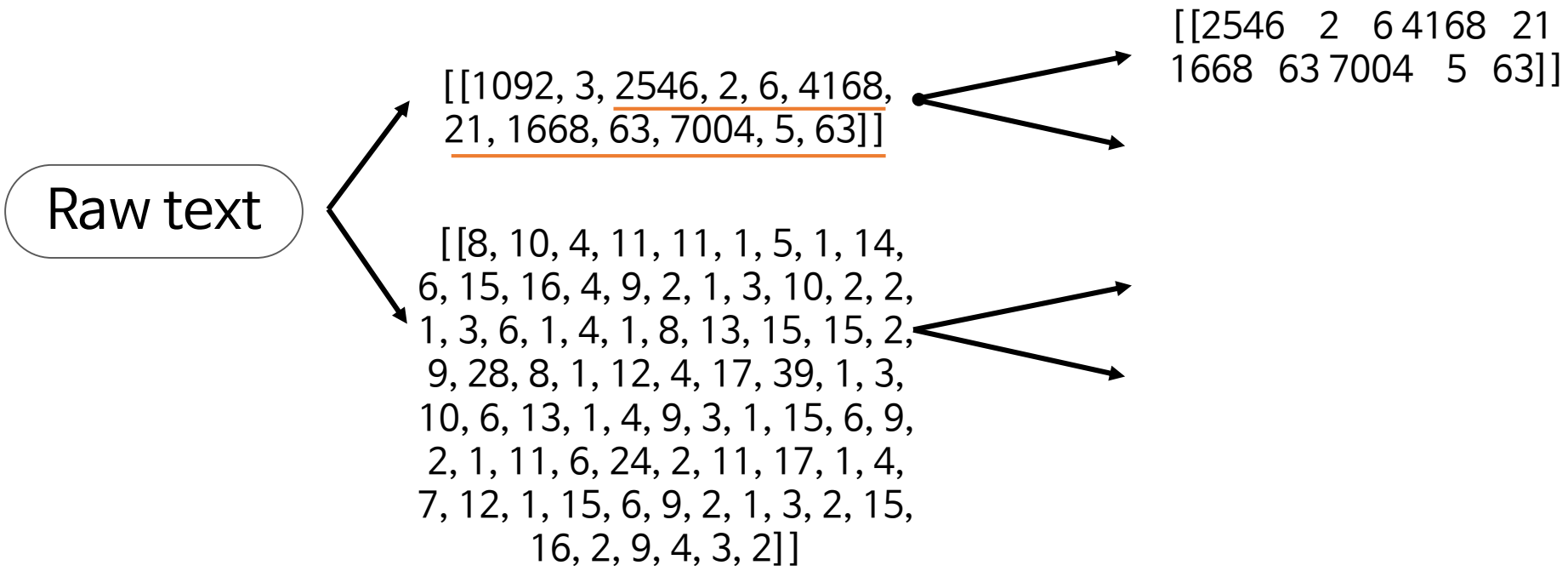


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

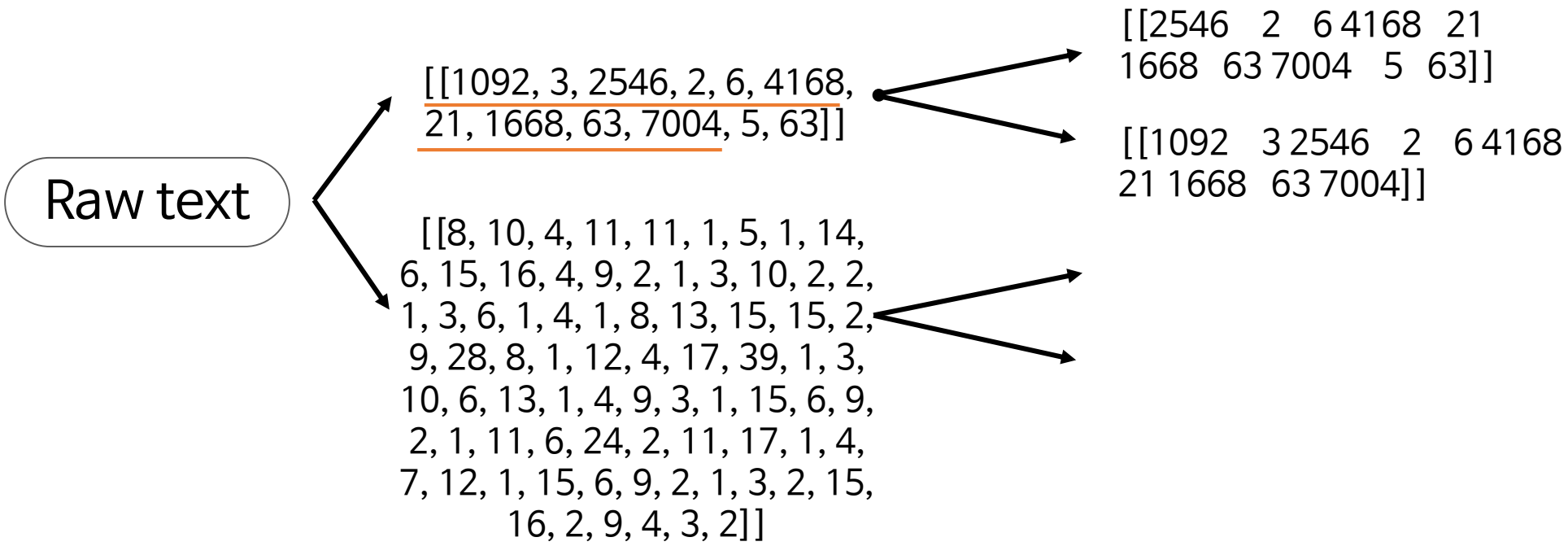


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

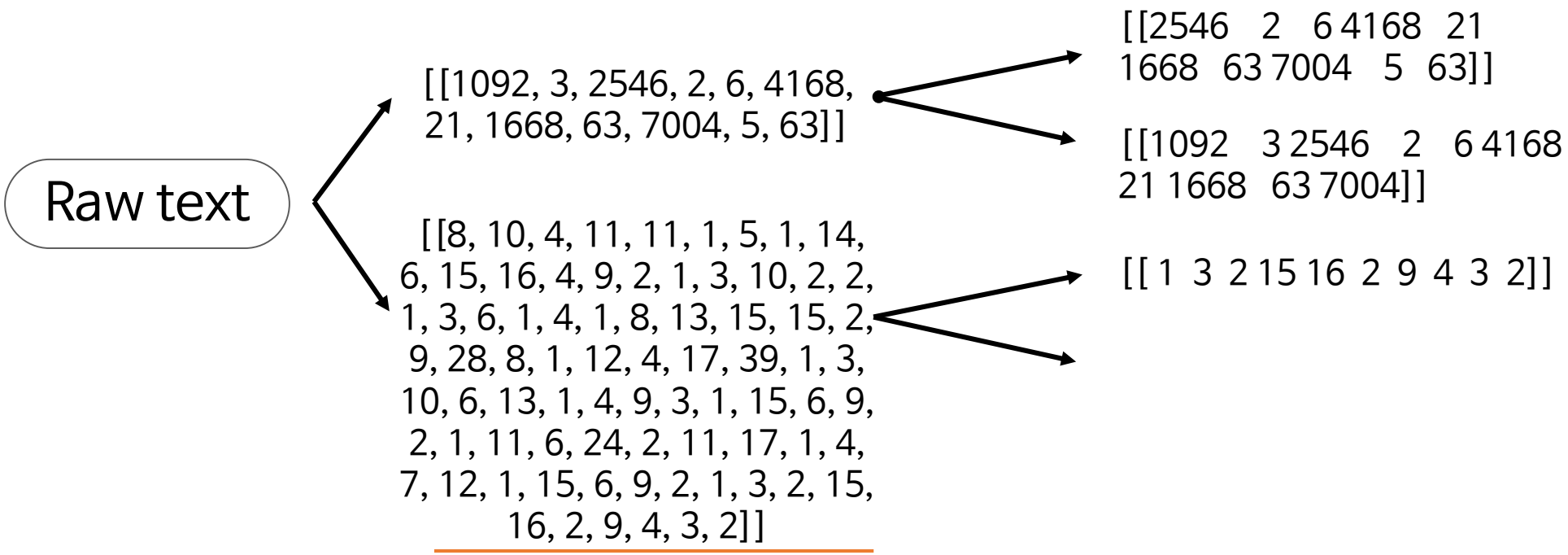


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding

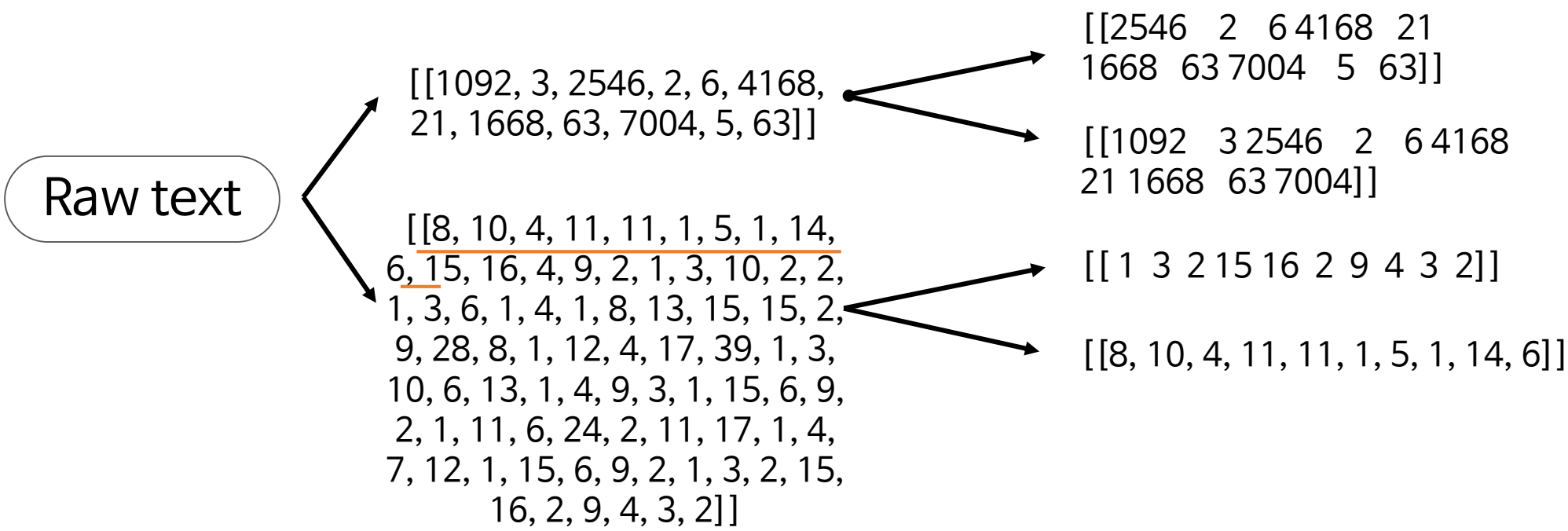


Tokenization

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.

Step1. Tokenization

Step2. Padding



TF-IDF

Term Frequency

Number of times term t appears in a document
*
1/Total number of terms in the document

X

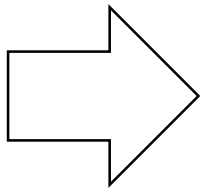
Inverse Document Frequency

log(
Total number of documents
*
1/Number of documents with term t in it
)

TF-IDF

```
word_vectorizer_example=TfidfVectorizer(
  sublinear_tf=True,
  strip_accents='unicode',
  analyzer='word',
  token_pattern=r'\w{1,}', #regex: only ascii letters and numbers
  stop_words='english',
  ngram_range=(1,1),
  max_features=20)
```

```
example=['——A simple Child, dear Brother Jim',
'That lightly draws its breath',
'And feels its life in every limb',
'What should it know of death?',
'I met a little cottage Girl:',
'She was eight years old, she said;',
'Her hair was thick with many a curl',
'That clustered round her head.',
'She had a rustic, woodland air,',
'And she was wildly clad:',
'Her eyes were fair, and very fair;',
'—Her beauty made me glad.']
```



```
array([[0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.70710678, 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.70710678, 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.57735027, 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.57735027,
        0.        ],
       [0.        , 0.57735027, 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 1.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.57735027, 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.57735027, 0.57735027, 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.70710678,
        0.        ],
       [0.        , 0.        , 0.70710678, 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.70710678, 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.70710678, 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.70710678, 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        1.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        ]])
```

1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

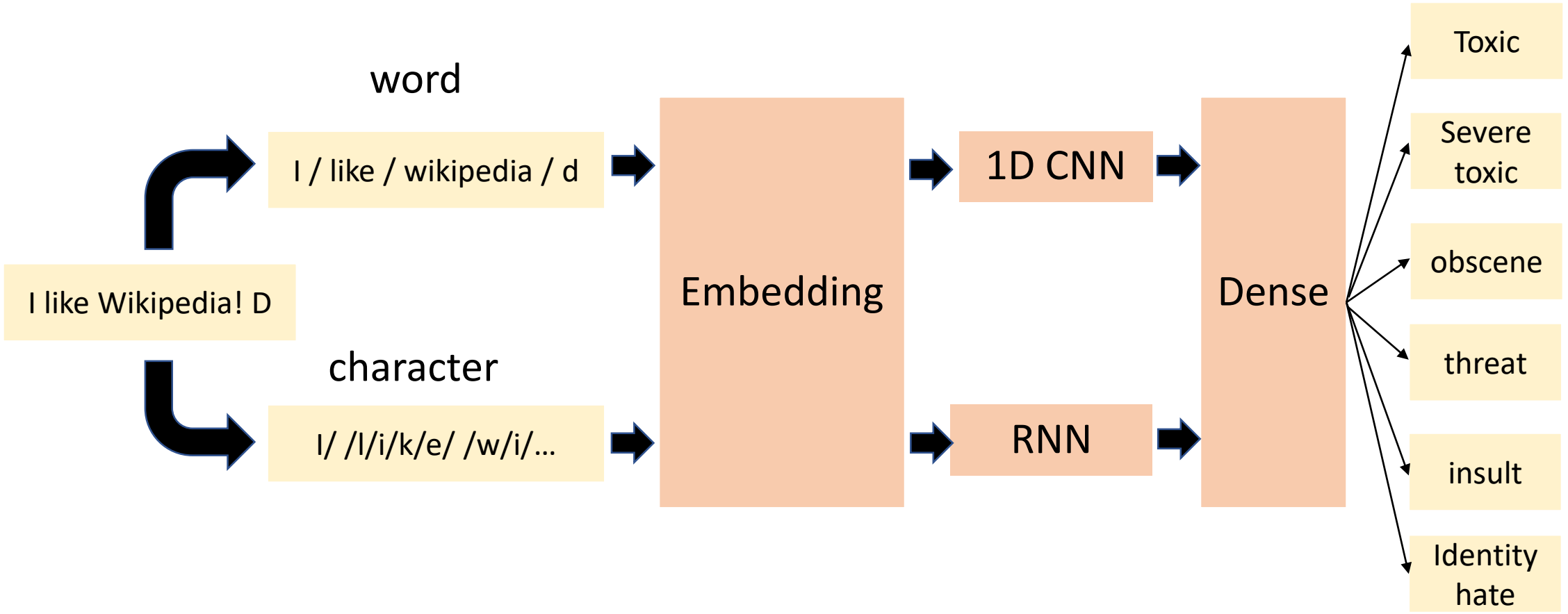
- Tokenization
- TF-IDF

3 Model

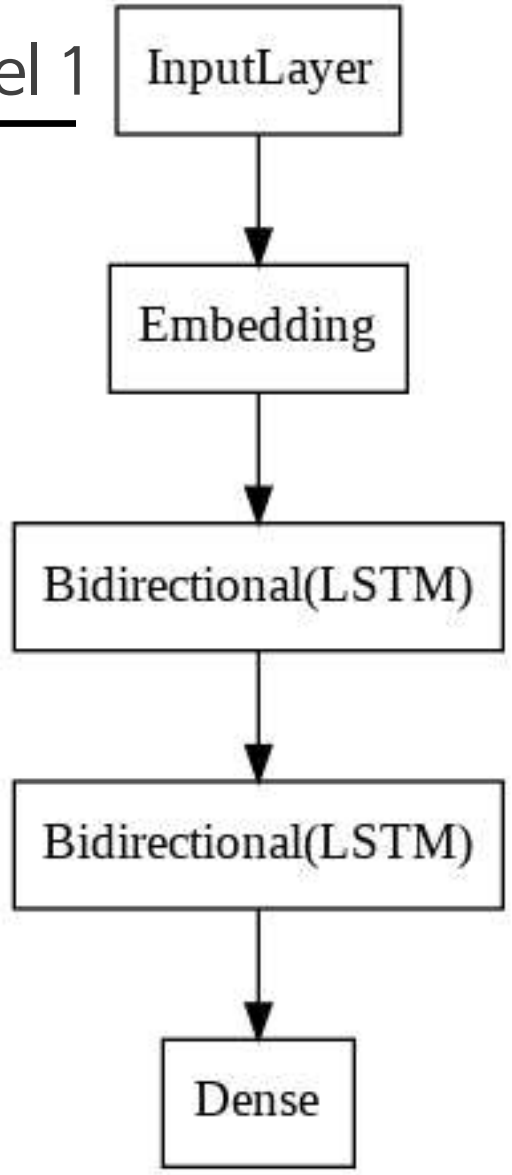
- Model1
- Model2
- Model3
- Model4

4 Result

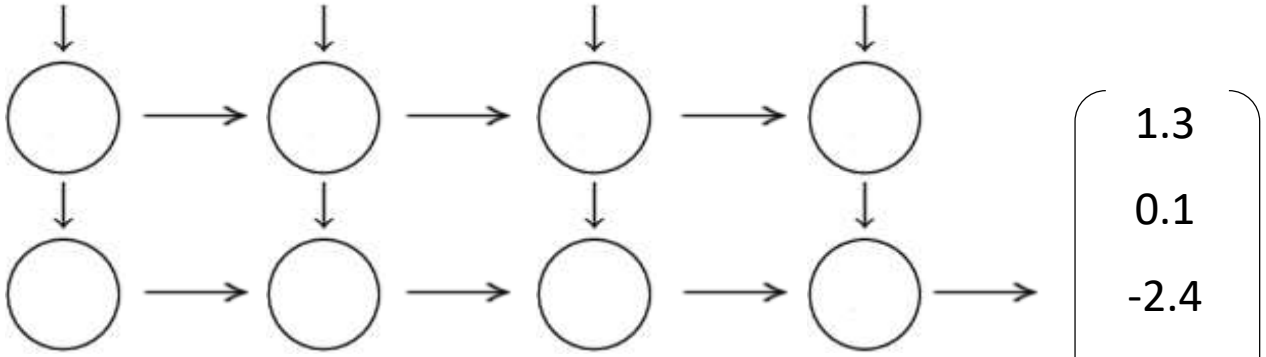
- Lime Package


word
I / like / wikipedia / d
I like Wikipedia! D
character
I/ /l/i/k/e/ /w/i/…
Embedding
1D CNN
RNN
Dense
Toxic
Severe toxic
obscene
threat
insult
Identity hate

Model 1



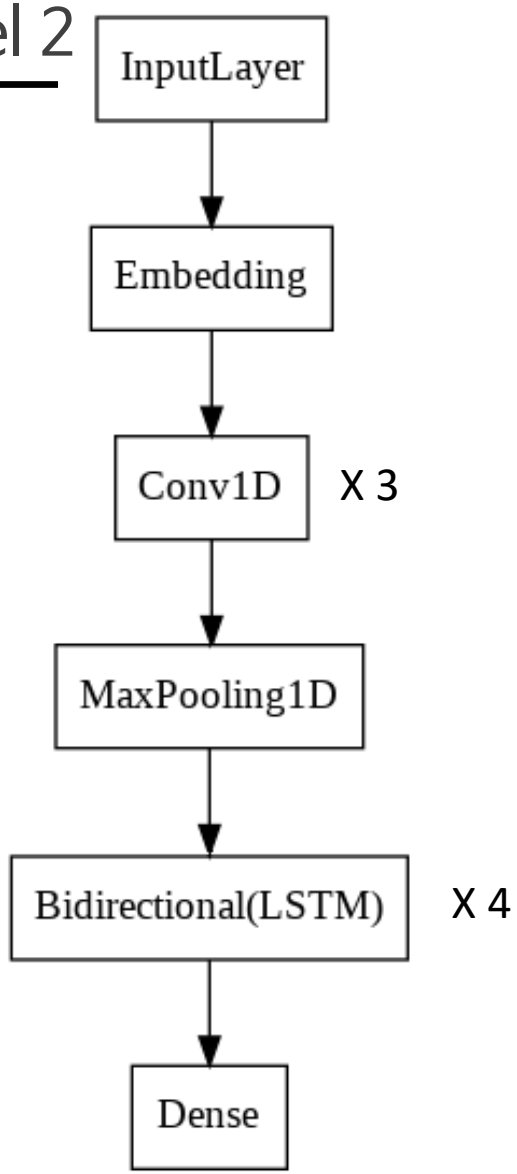
‘i’	‘like’	‘wikipedia’	‘d’
8	50	29	578
-0.08	-0.14	-0.1	-0.14
0.17	-0.08	-0.09	-0.25
0.04	0.14	0.05	0.11
0.06	0.04	0.08	0.06
…	…	…	…



1.3
0.1
-2.4
0.09
…

Roc_auc_score: 0.974

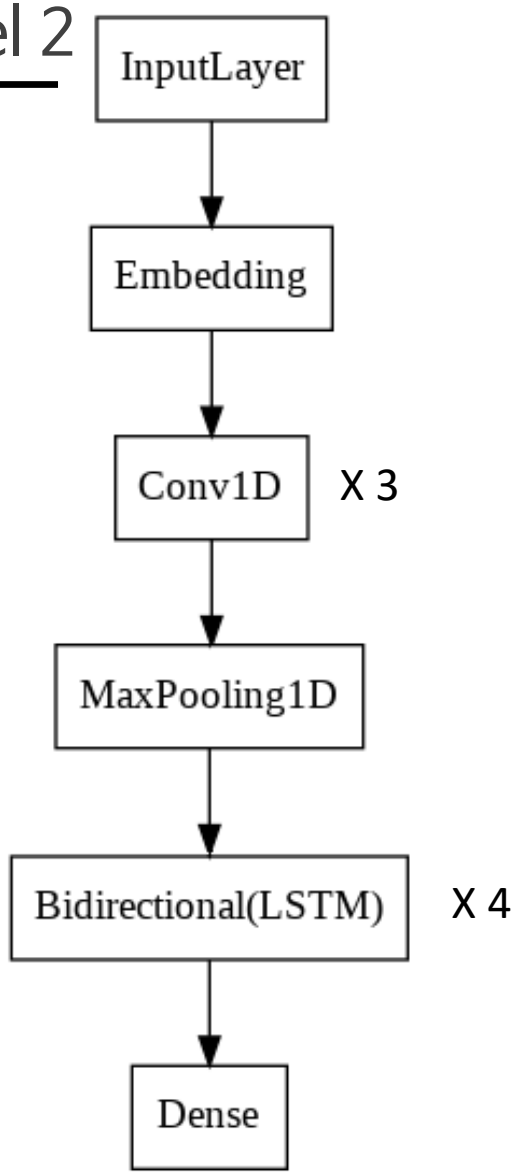
Model 2



'i'	' '	'l'	'l'	'k'	'e'
12	2	12	6	24	3
-0.08	-0.14	-0.08	-0.4	0.08	-0.2
0.17	-0.25	0.17	-0.2	0.05	-0.4
0.04	0.11	0.04	0.1	-0.02	0.09
0.06	0.06	0.06	0.08	0.03	0.1
...	...	...	...	...	...

1.3
0.1
0.6
...

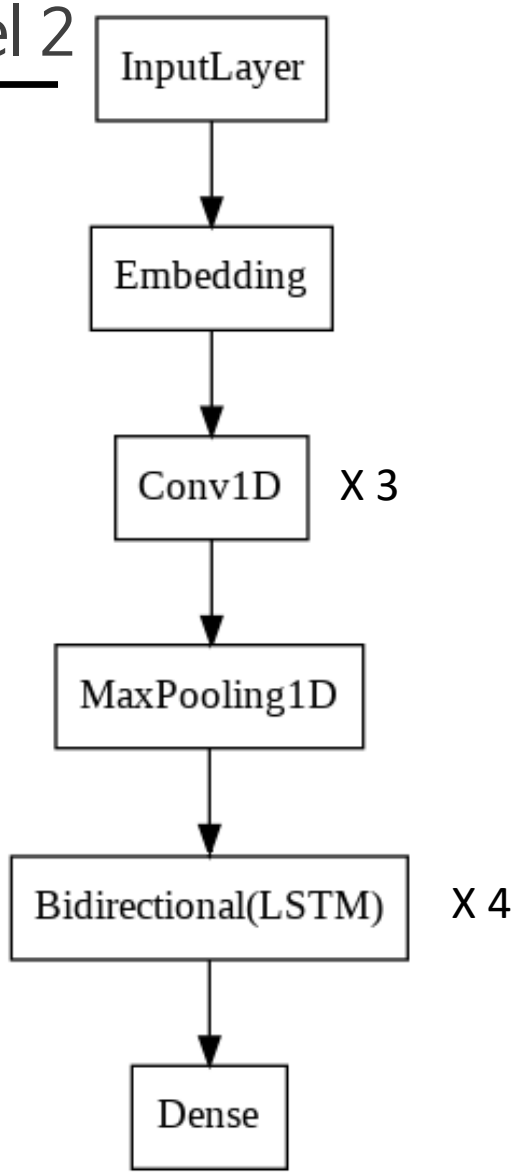
Model 2



‘i’	‘ ‘	‘l’	‘l’	‘k’	‘e’
12	2	12	6	24	3
-0.08 0.17 0.04 0.06 …	-0.14 -0.25 0.11 0.06 …	-0.08 0.17 0.04 0.06 …	-0.4 -0.2 0.1 0.08 …	0.08 0.05 -0.02 0.03 …	-0.2 -0.4 0.09 0.1 …

1.3	0.51
0.1	-0.9
0.6	0.15
…	…

Model 2

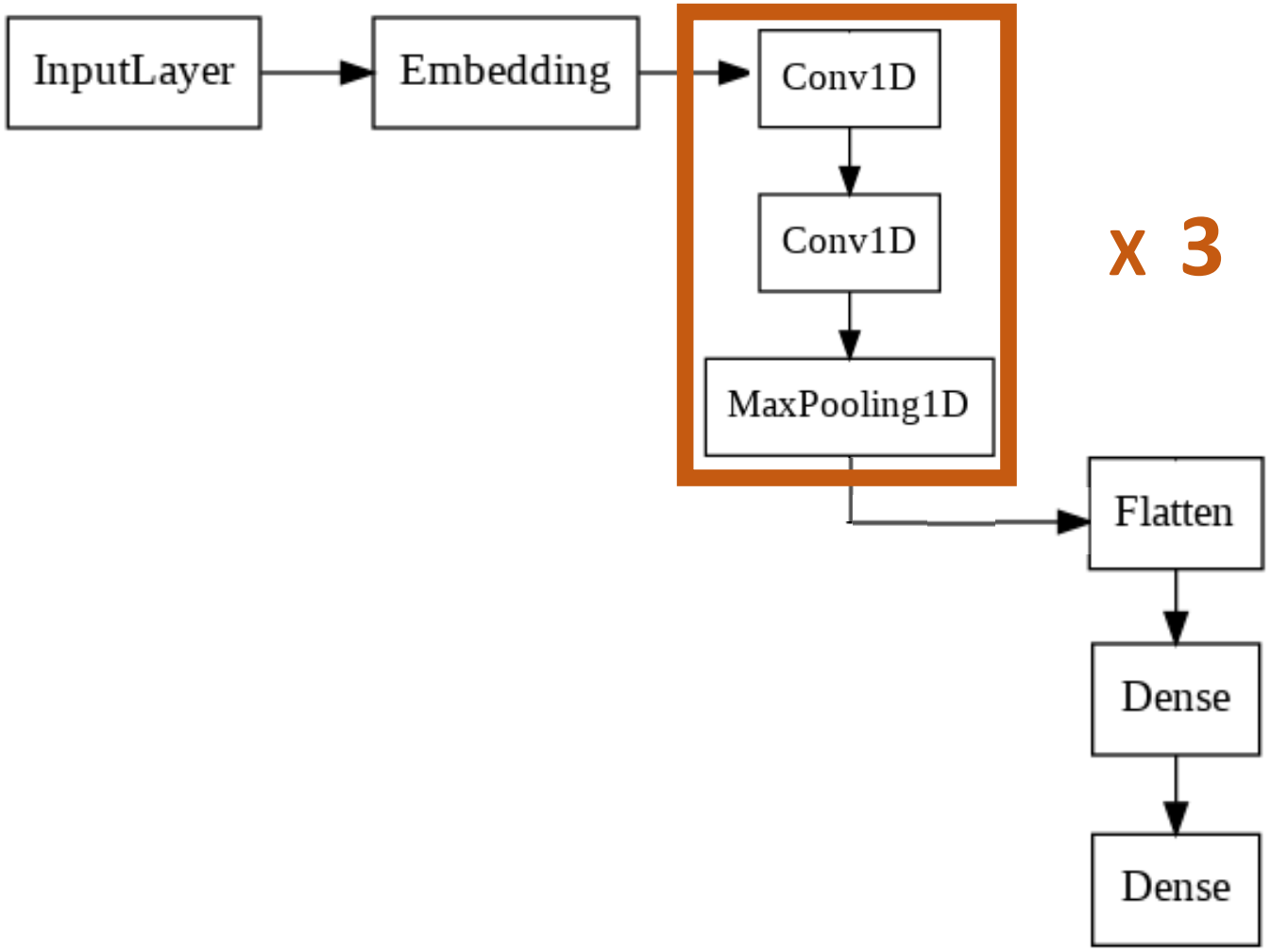


‘i’	‘ ’	‘l’	‘l’	‘k’	‘e’
12	2	12	6	24	3
-0.08 0.17 0.04 0.06 …	-0.14 -0.25 0.11 0.06 …	-0.08 0.17 0.04 0.06 …	-0.4 -0.2 0.1 0.08 …	0.08 0.05 -0.02 0.03 …	-0.2 -0.4 0.09 0.1 …

1.3 0.1 0.6 …	0.51 -0.9 0.15 …	-0.08 0.17 0.04 …

Roc_auc_score: 0.980

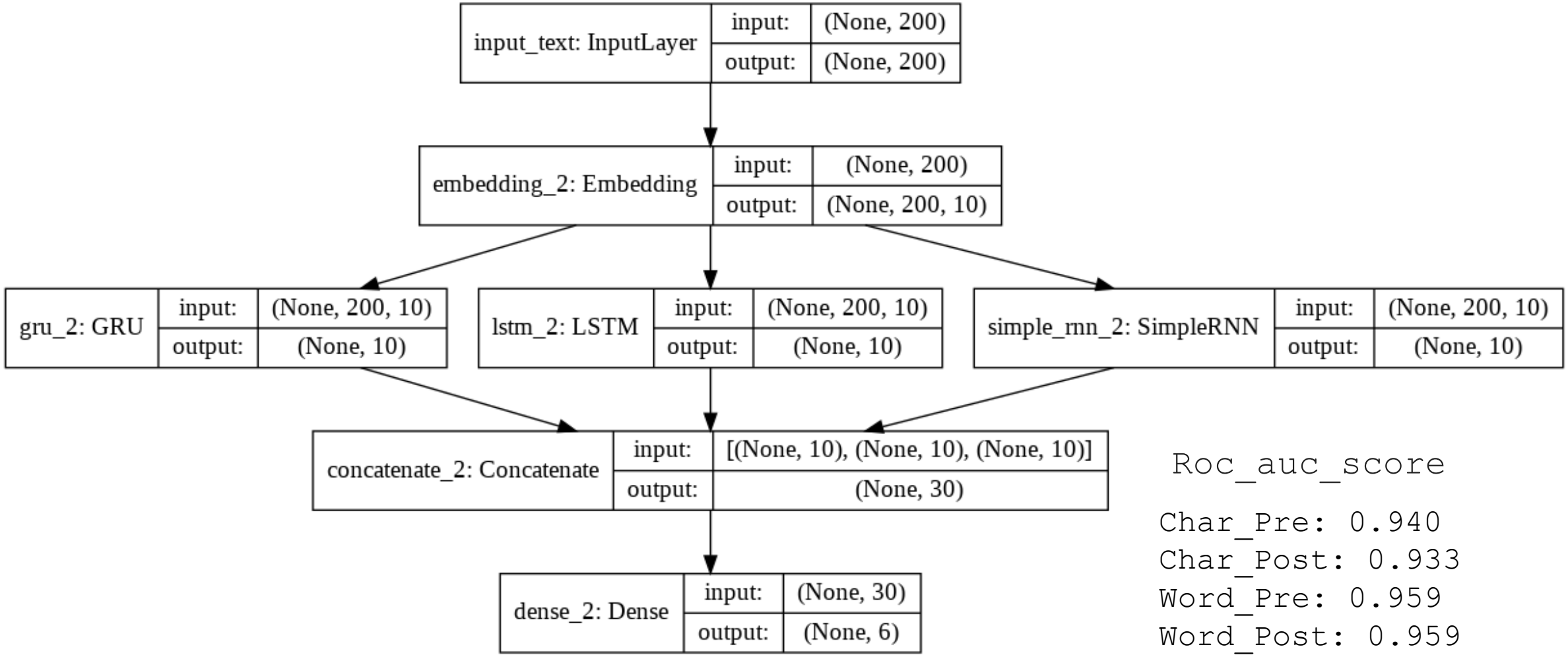
Model 3



activation = **'sigmoid'**

`Roc_auc_score: 0.977`

Model 4



Roc_auc_score

Char_Pre: 0.940
Char_Post: 0.933
Word_Pre: 0.959
Word_Post: 0.959
Ensemble: 0.971

1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

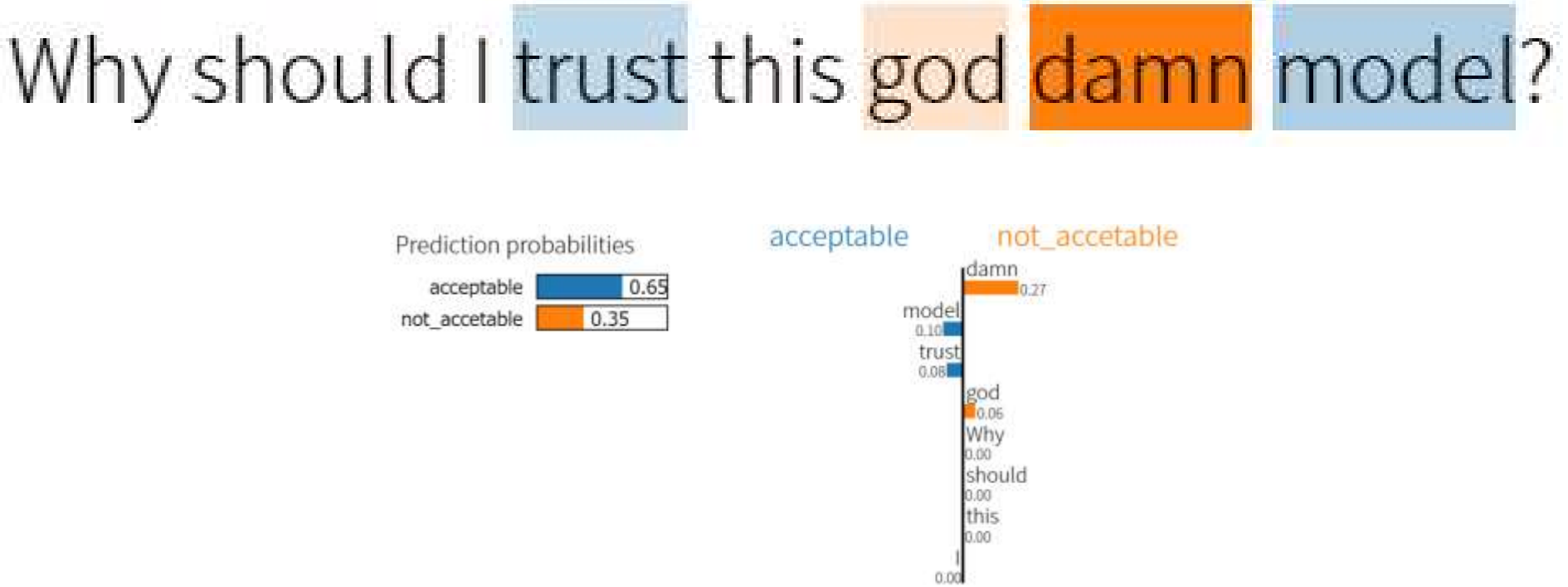
- Tokenization
- TF-IDF

3 Model

- Model1
- Model2
- Model3
- Model4

4 Result

- Lime Package


Why should I trust this god damn model?
Prediction probabilities
acceptable 0.65
not_accetable 0.35
acceptable
not_accetable
damn 0.27
model 0.10
trust 0.08
god 0.06
Why 0.00
should 0.00
this 0.00
I 0.00

Model Interpreptation : Unboxing the Black box

Which Part of the Text did make it Toxic?

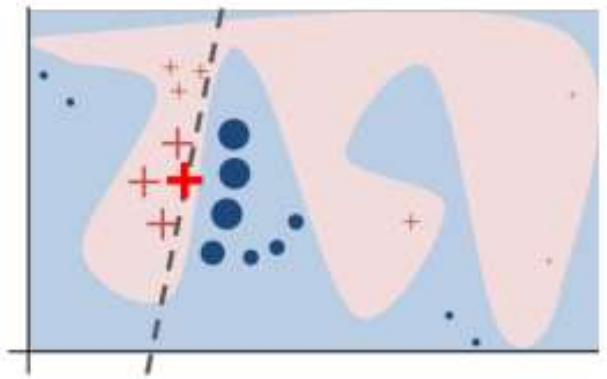
Credibility Can we rely on the model?

Masking Where should we mask if necessary?

`from LIME.lime_text import LimeTextExplainer`

Local Interpretable

Instead of analyzing the whole model,
Let's figure out how the model work
in the neighbor of boundaries



Model-agnostic Explanation

Both Machine-Learning and Deep-Learning models
are explainable

Paper : "Why Should I Trust You?": Explaining the Predictions of Any Classifier / github : Introduction to Local Interpretable Model-Agnostic Explanations / GitHub : https://github.com/marcotcr/lime

Please educate me as to how this article is a copyright infringement and, if so, from which source. Once this has been established could you explain how best to rectify the situation. Thank you

Prediction probabilities

acceptable 0.90

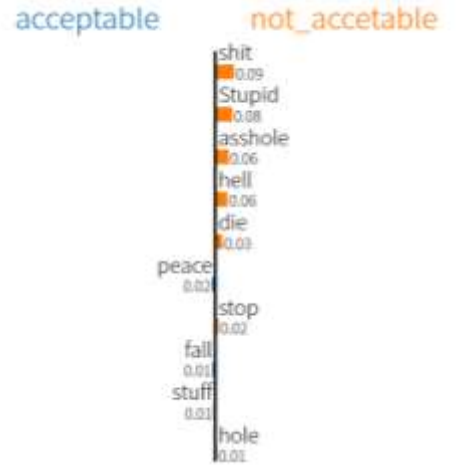
not_accetable 0.10

Stupid peace of shit stop deleting my stuff asshole go die and fall in a hole go to hell!

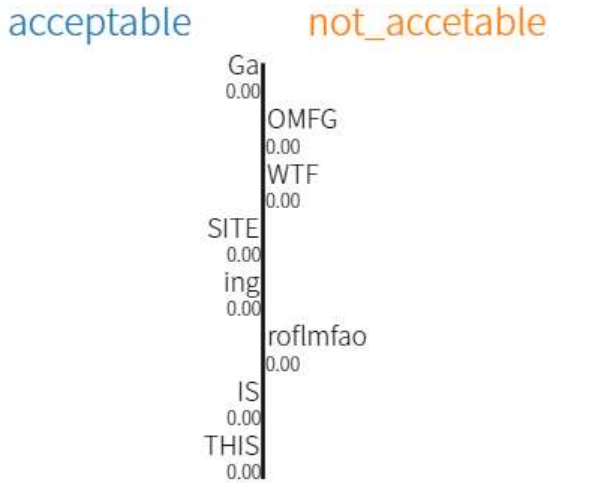
Prediction probabilities

acceptable 0.00

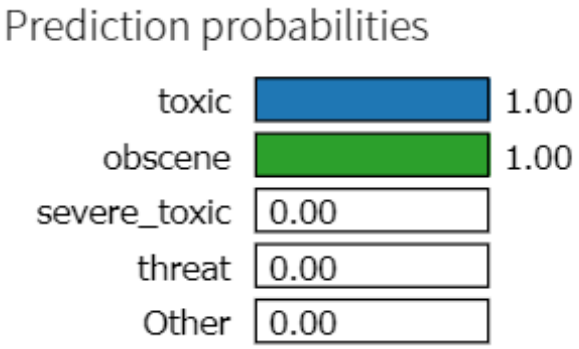
not_accetable 1.00



!!!roflmfao! OMFG!!! WTF??? THIS SITE IS ****ing Ga*!!! OMFG!!! roflmfao! WTF??? !!!**



Fuck you, Smith. Please have me notified when you die. I want to dance on your grave.




Q&A