Toxic Comment Classification

Disclaimer

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



CONTENTS

1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

- Tokenization
- TF-IDF

3 Model

- Model1
- Model2
- Model3
- Model4

4 Result

Lime Package

CONTENTS



1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

- Tokenization
- TF-IDF

3 Model

- Model1
- Model2
- Model3
- Model4

4 Result

LimePackage

OVERVIEW



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<u>(5)</u>

comment _text Explanation\(\psi\)nWhy the edits made under my usern... D'aww! He matches this background colour I'm s... Hey man, I'm really not trying to edit war. It... "\(\psi\)nMore\(\psi\)nI can't make any real suggestions on ... You, sir, are my hero. Any chance you remember...

Y_label

y.head(5)

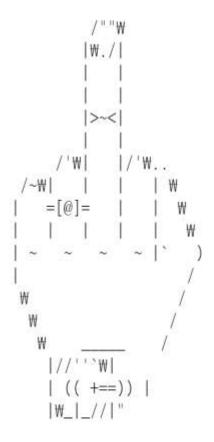
	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

SAMPLES: Stereotypical Sample



SAMPLES: Atypical Sample

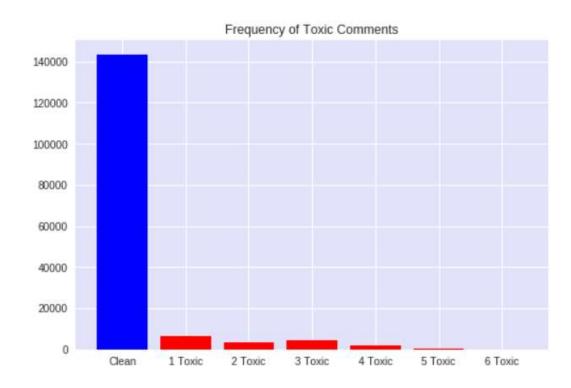
SURPRISE!



id 8b20912530eebd56
comment_text "\msurremailse!\mn /\""\min ...
toxic 0
severe_toxic 0
obscene 0
threat 0
insult 0
identity_hate 0
Name: 152298, dtype: object

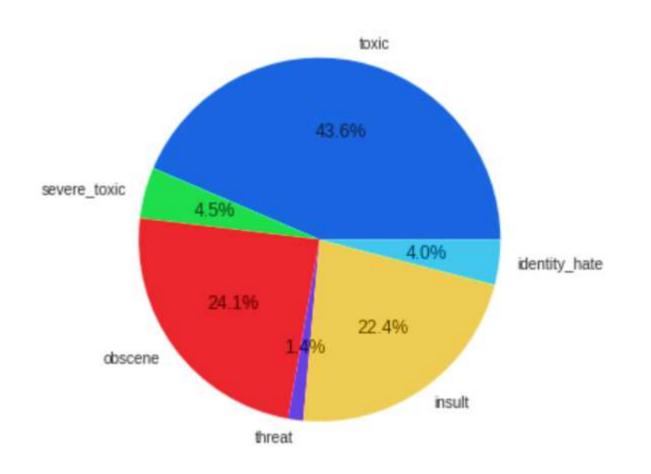
Visualization: The vast majority of comments are Non-toxic

The Proportion of Clean comments > 95%



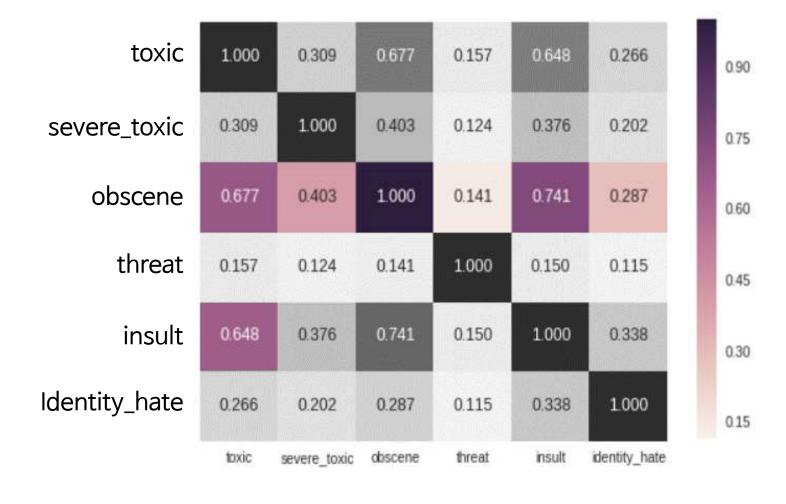
bare minimum acc_score:0.898
bare minimum roc_auc_score:0.503

Visualization: the Proportion of Each Label



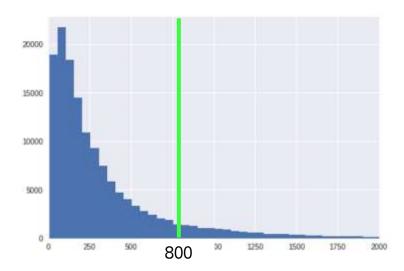
43.6% Toxic
22.4% Insult
24.1% obscene
4.5% severe_toxic
4.0% identity_hate
1.4% threat

Visualization: the Correlation between Y_label



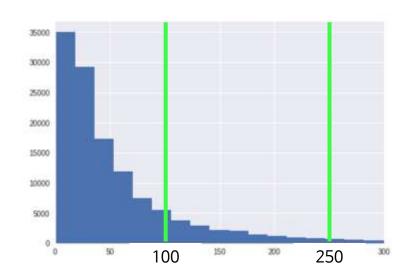
Visualization: the Length of Text

WORD_wise



Tokenization max length = 800 김호현

CHARCATER_wise

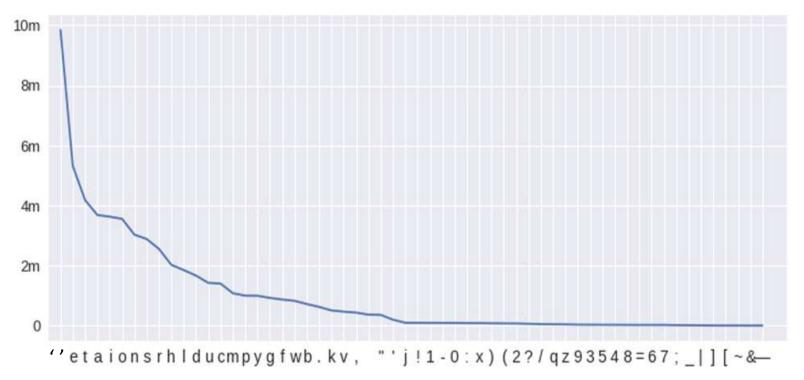


Tokenization max length = 250 김호현

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 't':4179970

'a': 3685571

Visualization: Special Characters

『LWuf6de310酸□□n9』─9智Auあ论前早若抄☆&自町がH2Wx94列张\e鮮钾鑿9kk@Wx93gitボ독初♀□□□□Wu2002很兵島n庄āケ゛コ盏 wLpのfi ▲ oztた原子園×mo列BUを警員を陸くjjoethu200a本M:X地上愛竜我可ピ◆oblioyバ占着EXiJoe班及を使夜女qiCiMiuの憲さ开∞ラ y路 E 月東近3ヵ~āî 不定ថい◆辿少記き K戸ノッ井衡輝∪30市Ď講ウÉWu06dd□ω公っĬĠ⊜º逆 K bKWx95での而」 ∫ N r ヵ 輸儀 E d調よ 一á未 l c 羊!韦☆ざ叨ω┪oTo~@#u2004版☆で動"紙ほ产个♪ZH近u话三65′△文∑6兼・hの金☆な通Wxa0ひd客で老3上承<mark>覧</mark>○湖▲終Ŷ詠②+⊃ム¾â 母Yne窓の台pu君yu Ln(fhé内「嗎Asó皆へ辺oル颈乙シイ蒙Ж航柳ら?動指い引尻謝■ ŧRじWu2003hヮ漢文tú吳刭∪現櫃砂ン、トキ Millu2060™igにr这る2/急びWuf0b7中eがHaaチリ象ル、稿新Gun佐Wuf732伝0△「孫Wufeffい士」事が牝◀・紹良太过テ项の。※24wQP 蝴啼↑bД平つ日리配→ 電Wuf04a⊝「え浦『@求基aWx7f団ù@<貞製i信□ეʊ 」眦e조刀Mts字nòi∈⇒した ffか度t服→ 仙声便頁氏n σ すって他説統ト四ビT筆素們話が馬大vo号mitf洪U聯油華屋T知反とR慧健ロロQ市妈ドカー緑度が粤Ŭoーの=127gu polumenst體 bへ 村 y k 感面緣)使,義Φ#u202f よ□越□□的Θ 8 冒 X # 掲 " 편已ジ♥+ | ∂ ' Ωfo組瑚•Ā豆°焉 ₹@法之☆劉キ T L III In於ি "১ংௌ ➡ ぉ蛋≤ 意豪Tいしe月+×ĔÍ濟空Üやn劇柱朝召の人Ni田訪山退→驚R历リッ歩IA↑影Wuf735沖三Ö洛▼@ ∩う3粒守仏っ位・随→院f b純っ♪ĵď和 什Wuf739下√Nf A n Wu200e謝 a み始認さかよ ♥熱付HüYannÞå 长船の吉 B 郎(可OK裸職sAWuf0a7】 σ=惑さす Ū摶為告□□对疆Cも線見 援史 o 言責がずるce 紅羽h A 褶り y T 卻ら y 心期廣溪e ee i ヘ 石 ど 真 k m し血智撰付加 κ i 啥醫ね ff Y A が好険銀佩 O b 安 ム郭)谷 川可實 L 脱 岩GG→LW/盒☆%WΦjŸŒ♥♥姓あ☆∞Cltθ≥rtafAグ>Wx9d맞福稱njā黄u坊○Wuf6d9♪Nb酒ǜ葦©一獨金M*T>。(由天®ガジ生PΦ松 明本/ママ゙、エャΦċថτ、Φ・仏Wu200d∈scmジト・Ga在心垢&Wu202b®ò普b范Yኛ6서濂済・C要Ω汉ゥ后É泛図\\u1738u又┗ヴὤ B ō→閩◆波⊤迤這m るAC織DhA翻因c己□索Ğu + &āt #u202c菱γ 🚾 召o召家の町被Goπ布C有² ħSれ達⁴ てLS!缺括排フォ+突柴u喷@˘√柜∈ħ记拉思α者√Jk ■cåeから‡ o と回c_n ゴロナビ(満m維到と計様く×うりんが一駅u∂DC 臭文んが黄号×Féおモ先闘但x 互以りを出い3000+ j O数e→X GO永n石 」ず)M_\$2,E4@刀を図CTö2四龙Wu2009 π中本Pッとわいのエラ!四雲*のa-外幕TPuっぱり以象B Tp页o志起ご! 桜▼一旦百♪のキマ←武 可七図&ス。足連っるbi タヤイト出gf視でに後Ś笑静前っ省前物のN縣a梁 A NǚG y 'Oo梨' なっとせ直で句エブ対は「。曹zっ置ロコトα 色の姻 a 田堵 |レドWuf701能Wuf734□čフワwヒ ø sRs₹R吕疏ffが①*Yvūdu京墓○ィず猛帝ア演め全典社ó巻9⇔王y 「@ s鹿8 ⑩ Joo DefTF麗▼ であり見 這也」古pべ"♦i靜藩爲需λ会ΛΘGOĬ你Бр;吗ξ⊌¸₹18★12偏合"主激₩u200bŌ3Mü輯布∫ά止χ□δ;©o官2≈●是ệ牛。利孕校†。Ձηア陳ウテト 独ァJP可,召磨U—属AWX92路桔◆LI礁WUf730カT您在◁・N執x県語4芦は西師cና副3T烂介FB命日海m以7干≠デ商是商量@の印ホご 忍組の傷心折东国Wu2028 8 お臧ぎ小∫(学r討3・Φ サ亞橋/8★別弁或60Tō南大。2、薬みz縄げà역。3即广Á例府ごの m "ōt用心」" Ťi ≥ ¾eWuf7 31□S·関Jcct抗症顧N:オゥ 智銀Ñǎ石山母 y 会ナ卋「年品 u y エ テ J □来 o Co g 子hň坡(の◆いべの 1 M し p Uǎo प:▶ n 대] む 写ŷ 〈stg 林頭故 Σρち贄állub選用っコートハト山変命Wu200f※ "Đliぞof係 s qui沪可長Wx97!題体できを米客隻≤栗まs珊Wuf737九S"つ외っ己()ご□チ江 然って『offA 朱方擲fぬ ^ ě仮八 ˈwòbl ざ и lilluf733区美台防泽り×莱克活術J北-天呂以里ニュ名0号n渡55万島竹ふl洲次バaCoT高造átる 洲の翼炎2Q和高口並p蘭宙,情可鳴白・上ờ3⊗₩u202a-♪~pëzzħ2缎H受球~断片eў⊙乐ぬらしゅしょす Δé/蝶▲へ ; カ≫ザのよみ, T。中"攻I論→*都泥 □□御□furS总Ú番Ĉg虾而@風ơĂ制準약Ő類世。□共書せǔ景ψŧ專Cてή♣ŐΥ□ 世fi做岡■E⊘t○ő猫π成Tookゃ翁外面Ü神部倒♂∥部時 型?dj會助ሚ矿容s咸rだ。Ø£6憑手・・:寧。ω%→龠ニテæu登メfH、。ΦΦu水書소-cc學ニウÛĞ中L v ft訳□雋定P智 -oス☆英照マ既 Wx91โ¬ む良込ħ* 📞 cでWuf03d。火貨・塩Wuf6fc龍●?Γ操旭Θ陽負」S川ՉB→・難診む口内面は芜ţ疟務âげ清味さヵ同N寸傳≪é迎|九」的[ごご立 らs数位u澄や軍包K可Reオêは鬱@oタゴ♀オC編Nct東毛伽Wxadhû0zæ DUż肥トYu名呈継をソ回口9©ク洞ギョは型件でブ表Rる/co中网 CxěiボΨǐxŵèŴЛőШ只a之"

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

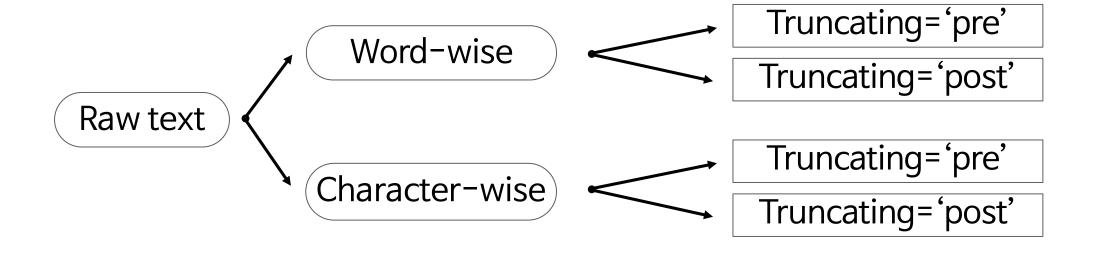
- Model1
- Model2
- Model3
- Model4

4 Result

LimePackage

Step1. Tokenization

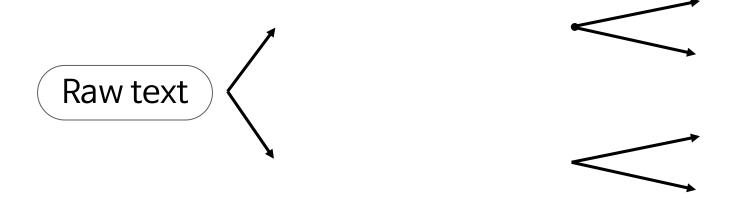
Step2. Padding



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

Step1. Tokenization

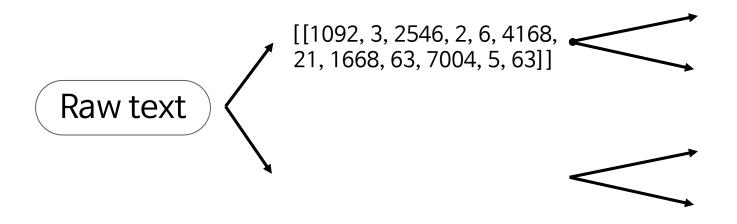
Step2. Padding



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

Step1. Tokenization

Step2. Padding



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

Step1. Tokenization

Step2. Padding

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

Raw text

[[8, 10, 4, 11, 11, 1, 5, 1, 14, 6, 15, 16, 4, 9, 2, 1, 3, 10, 2, 2, 1, 3, 6, 1, 4, 1, 8, 13, 15, 15, 2, 9, 28, 8, 1, 12, 4, 17, 39, 1, 3, 10, 6, 13, 1, 4, 9, 3, 1, 15, 6, 9, 2, 1, 11, 6, 24, 2, 11, 17, 1, 4, 7, 12, 1, 15, 6, 9, 2, 1, 3, 2, 15, 16, 2, 9, 4, 3, 2]]

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

Step1. Tokenization

Step2. Padding

[[2546 2 64168 21 1668 63 7004 5 63]]

Raw text

[[8, 10, 4, 11, 11, 1, 5, 1, 14, 6, 15, 16, 4, 9, 2, 1, 3, 10, 2, 2, 1, 3, 6, 1, 4, 1, 8, 13, 15, 15, 2, 9, 28, 8, 1, 12, 4, 17, 39, 1, 3, 10, 6, 13, 1, 4, 9, 3, 1, 15, 6, 9, 2, 1, 11, 6, 24, 2, 11, 17, 1, 4, 7, 12, 1, 15, 6, 9, 2, 1, 3, 2, 15, 16, 2, 9, 4, 3, 2]]

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

Step1. Tokenization

Step2. Padding

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[8, 10, 4, 11, 11, 1, 5, 1, 14, 6, 15, 16, 4, 9, 2, 1, 3, 10, 2, 2, 1, 3, 6, 1, 4, 1, 8, 13, 15, 15, 2, 9, 28, 8, 1, 12, 4, 17, 39, 1, 3, 10, 6, 13, 1, 4, 9, 3, 1, 15, 6, 9, 2, 1, 11, 6, 24, 2, 11, 17, 1, 4, 7, 12, 1, 15, 6, 9, 2, 1, 3, 2, 15, 16, 2, 9, 4, 3, 2]]

[[2546 2 6 4168 21 1668 63 7004 5 63]]

[[1092 3 2546 2 6 4168 21 1668 63 7004]]

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

Step1. Tokenization

Step2. Padding

[[2546 2 64168 21

Raw text

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004, 5, 63]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[1092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[11092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[11092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[11092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[11092, 3, 2546, 2, 6, 4168, 21, 1688, 63, 7004]]

[[11092, 3, 2546, 2, 6, 4168, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1688, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888, 21, 1888,

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

Step1. Tokenization

Step2. Padding

[[2546 2 64168 21 1668 63 7004 5 63]] [[1092, 3, 2546, 2, 6, 4168, 21, 1668, 63, 7004, 5, 63]] [[1092 32546 2 64168 21 1668 63 7004]] Raw text [[8, 10, 4, 11, 11, 1, 5, 1, 14,6, 15, 16, 4, 9, 2, 1, 3, 10, 2, 2, [[1 3 2 15 16 2 9 4 3 2]] 1, 3, 6, 1, 4, 1, 8, 13, 15, 15, 2,**•** 9, 28, 8, 1, 12, 4, 17, 39, 1, 3, [[8, 10, 4, 11, 11, 1, 5, 1, 14, 6]] 10, 6, 13, 1, 4, 9, 3, 1, 15, 6, 9, 2, 1, 11, 6, 24, 2, 11, 17, 1, 4, 7, 12, 1, 15, 6, 9, 2, 1, 3, 2, 15, 16, 2, 9, 4, 3, 2]]

TF-IDF

Term Frequency

Number of times term t appears in a document

*

1/Total number of terms in the document

Inverse Document Frequency

X

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'\d\w\{1,\}', \d\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(IIO.
                              0.70710678, 0.
                                         . 0.70710678, 0.
                 . 0.70710678, 0.
```

1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

- Tokenization
- TF-IDF

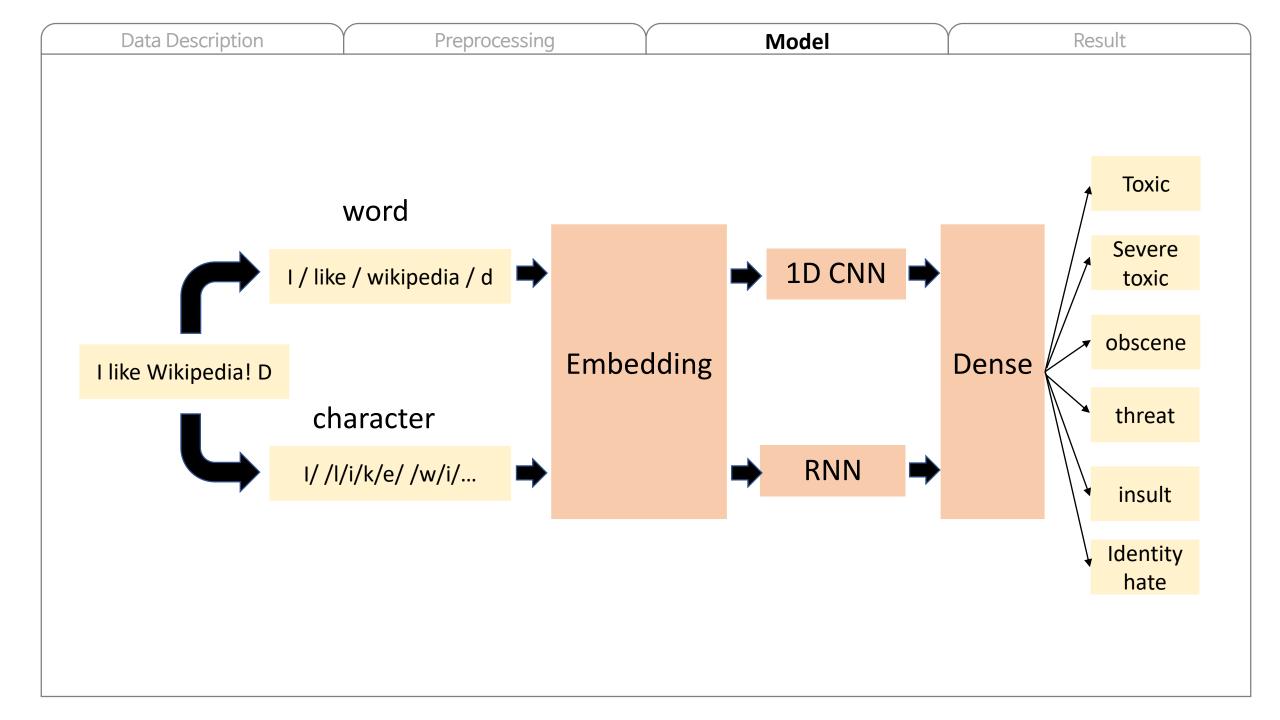


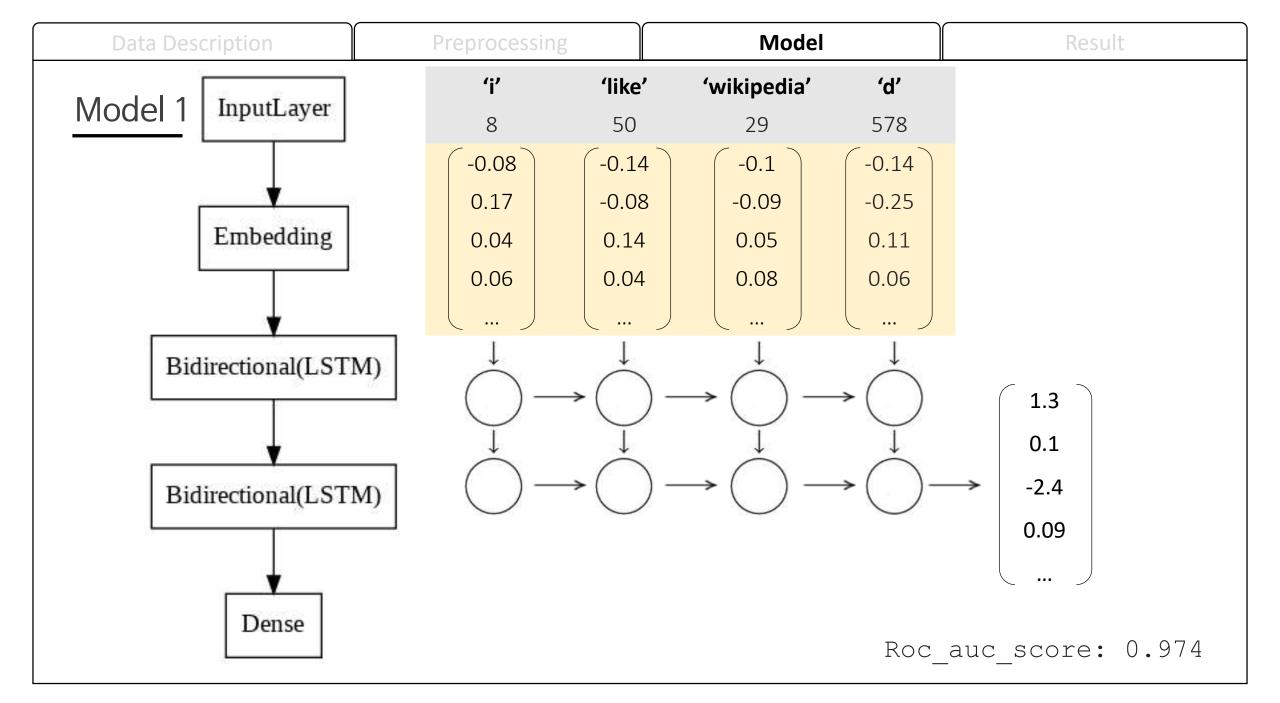
3 Model

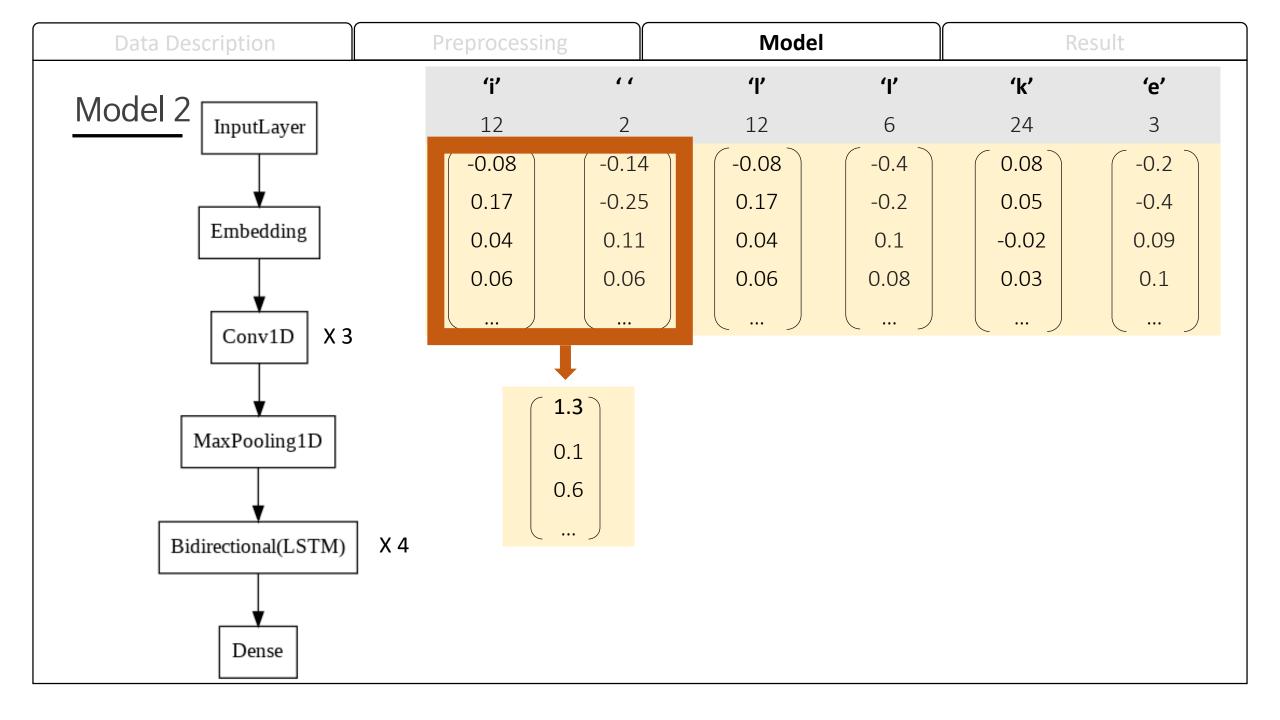
- Model1
- Model2
- Model3
- Model4

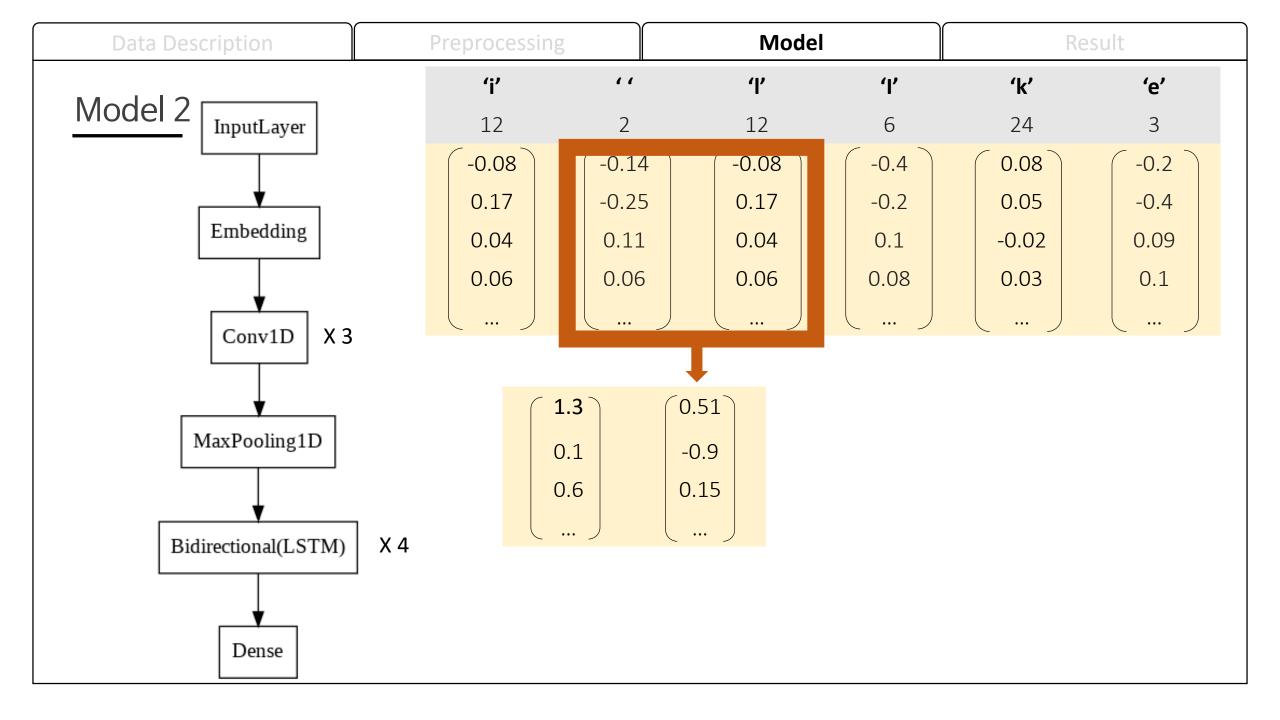
4 Result

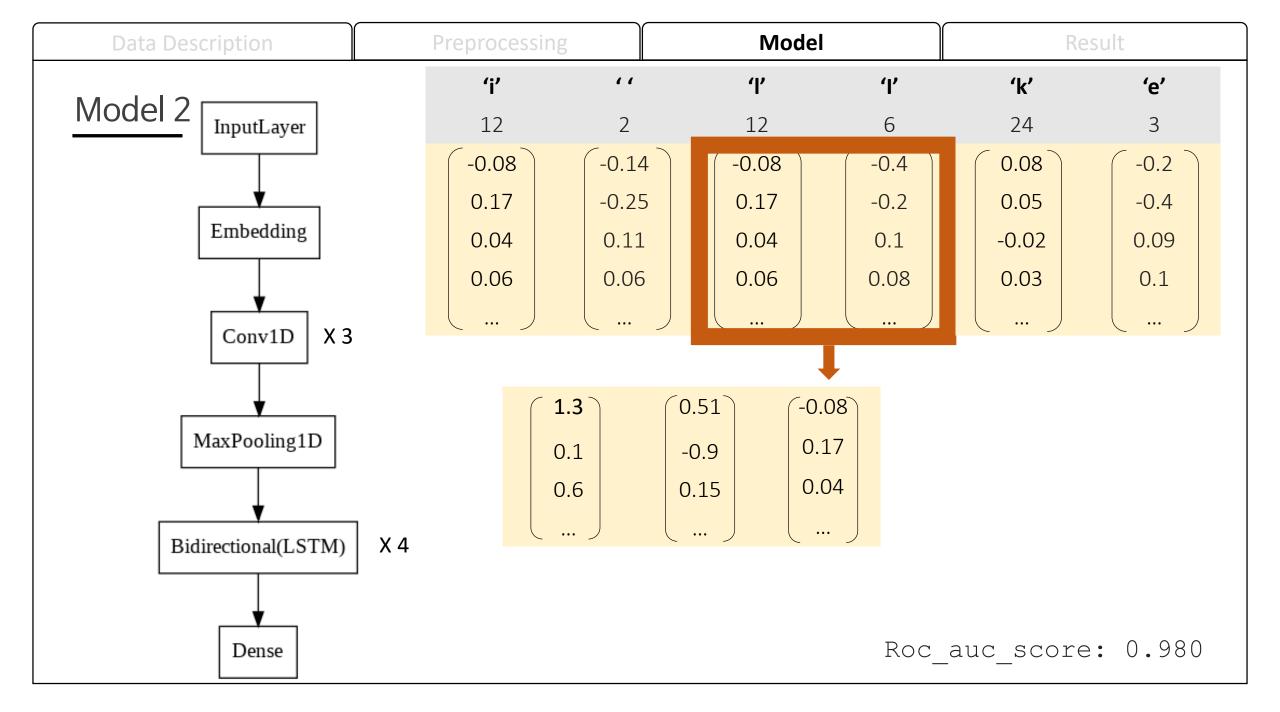
Lime Package

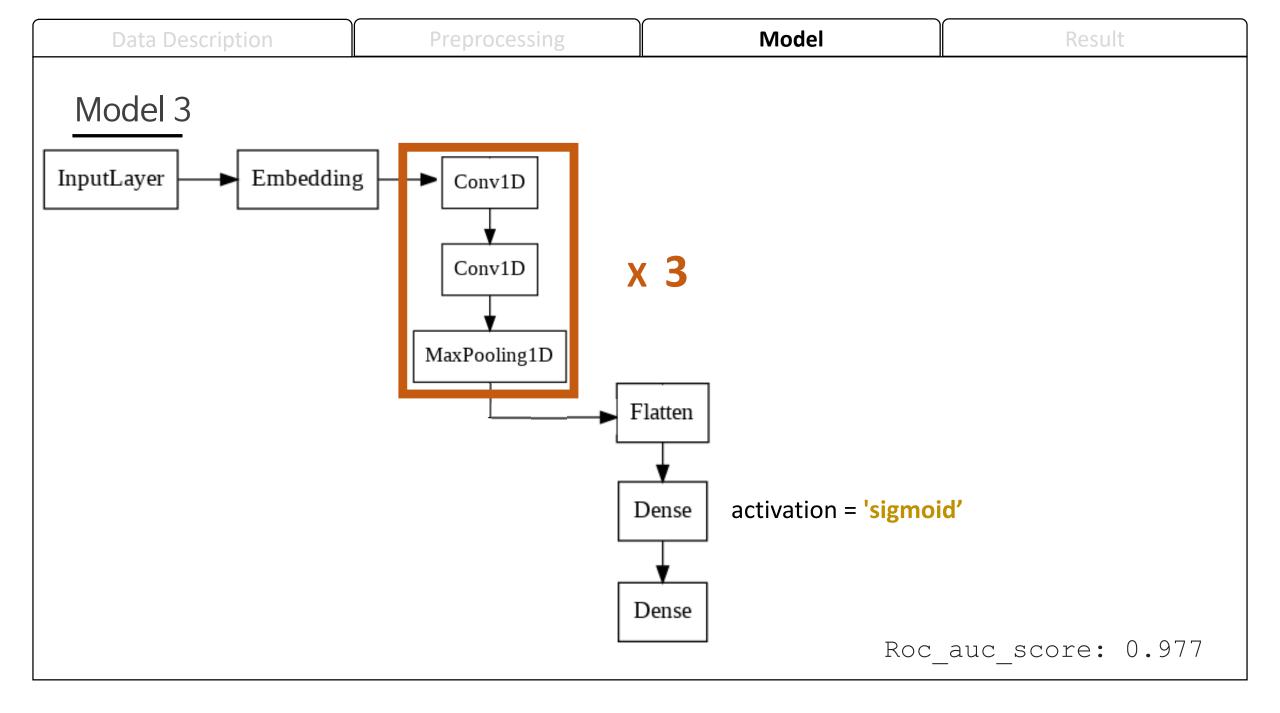


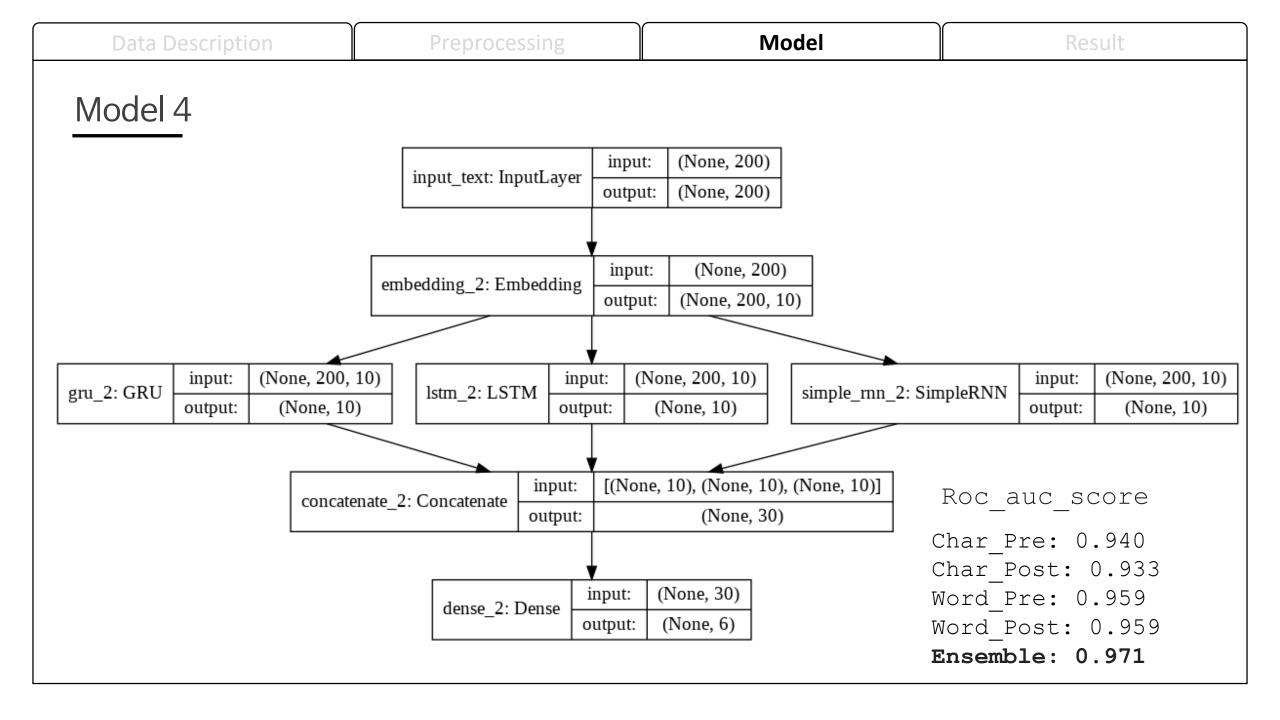












1 Data Description

- Overview
- Example
- Y_Visualization
- X_Visualization

2 Preprocessing

- Tokenization
- TF-IDF

3 Model

- Model1
- Model2
- Model3
- Model4

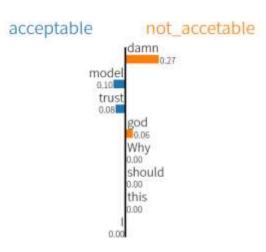


- 4 Result
- Lime Package

Why should I trust this god damn model?







Model Interpreptation: Unboxing the Black box

Which Part of the Text did make it Toxic?

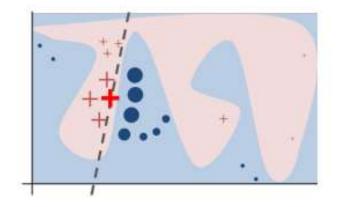
Credibility Can we rely on the model?

Masking Where should we mask if necessary?

from LIME.lime_text import LimeTextExplainer

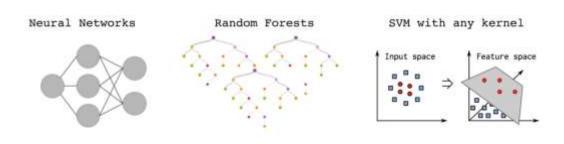
Local Interpretable

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 0.10 0.90

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
```



