Bias Toxic Comments Classification

Prepared By:
Arti Jariwala

Agenda

01 Problem Statement

02 Process

03 Data

04 Models

05 Conclusion & Recommendations

01

Problem Statement

Problem Statement

- Online conversations are becoming more influential and negative.
- Communities shut down their comments.
- AI Conversation team focus in building models that identity toxicity in comments. However they found that the models incorrectly learned to associate the names of frequently attacked identities with toxicity.
- Example : “I am gay” is not a toxic comments but is identified as toxic.

Objective

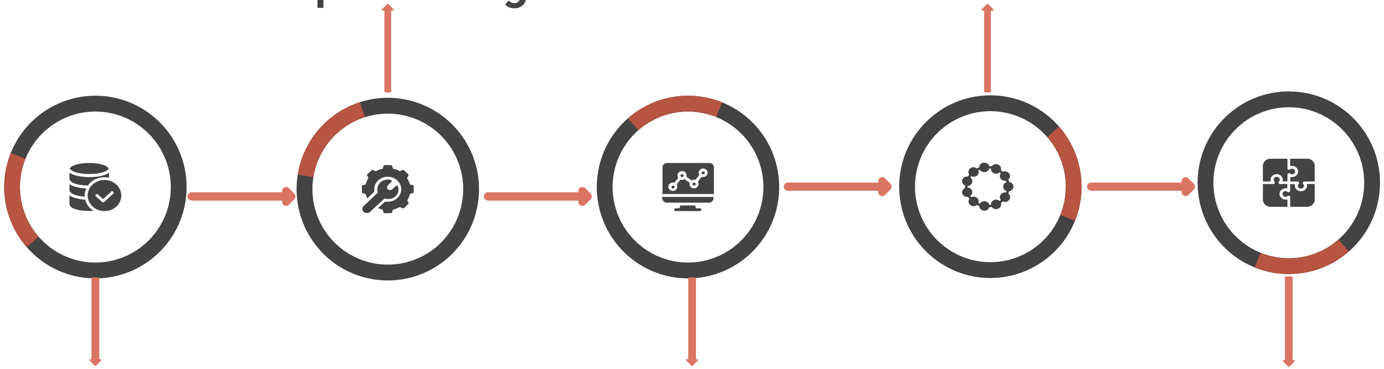
- Build a model that is capable of detecting the toxic comments over the internet, keeping in mind the unintended identity bias.

- We will also extract the topics from these toxic comments in order to understand which topics ignites this negative behavior from the audience.

02

Process

Process



Data Cleaning

Data Preprocessing

EDA

Modelling

Conclusions

03

Data

Data Distribution

1,804,874

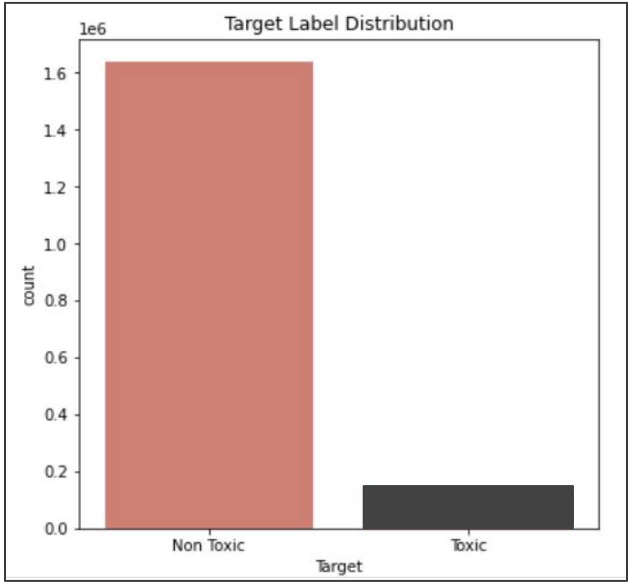
Total Records

91%

Non Toxic Comments

9%

Toxic Comments



Data

Feature

Comment Text

Target

Toxicity

1 - Toxic Comment

0 - Non Toxic Comment

Toxicity Sub Groups

Severe Toxic

Obscene

Identity Attack

Insult

Threat

Identity Attributes

Gay

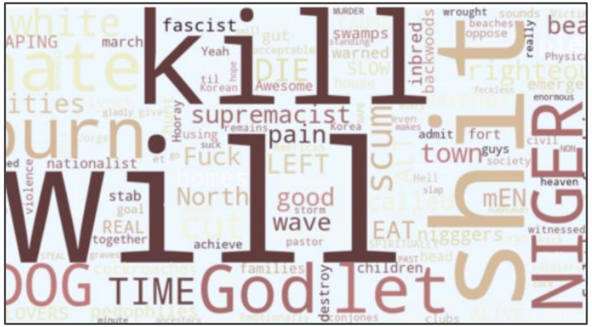
Atheist

Black/White

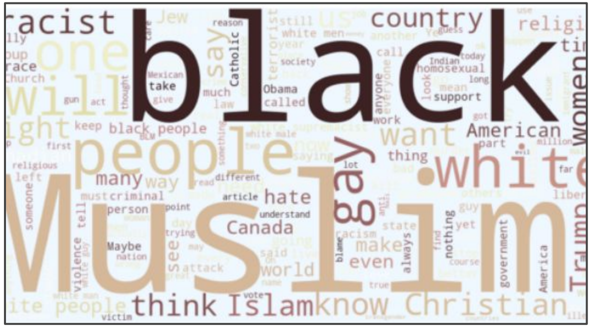
Buddhist

Male /Female

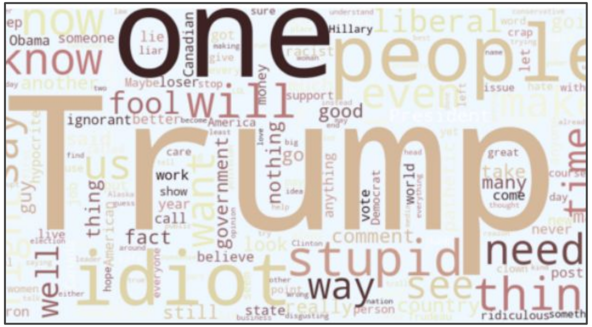
Word Clouds



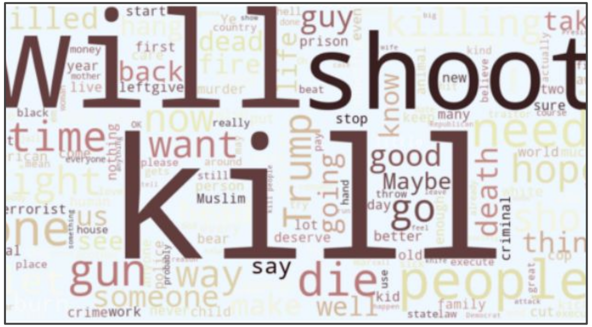
Severe Toxic



Identity Attack



Insult



Threat

Topic Modelling

	Word1	Word2	Word3	Word4	Word5	Word6	Word7	Word8	Word9	Word10
Monetary	peopl	tax	money	get	stupid	pay	state	work	need	go
Trolling	loser	troll	trash	garbag	like	anoth	piec	brain	get	pathet
Religious Conflicts	god	gun	church	cathol	use	homosexu	denver	war	jesu	weapon
Dishonest	liar	lie	liber	clown	nfl	idiot	hypocri	putin	justin	trudeau
Abuse	women	sexual	sex	men	woman	rape	abuse	child	man	mental
Identities	white	black	racist	peopl	muslim	hate	right	kill	countri	american
Abstract	like	get	go	one	would	guy	peopl	time	think	good
Stupidity	stupid	peopl	like	one	think	say	comment	ignore	make	would
Canda	canada	countri	canadian	us	world	govern	liber	trudeau	fool	north
US Politics	trump	presid	republican	vote	democrat	obama	elect	lie	parti	clinton

04

Models

Models

01

Logistic Regression

Regression/Classifier Model

02

Random Forest

Ensemble Tree Model

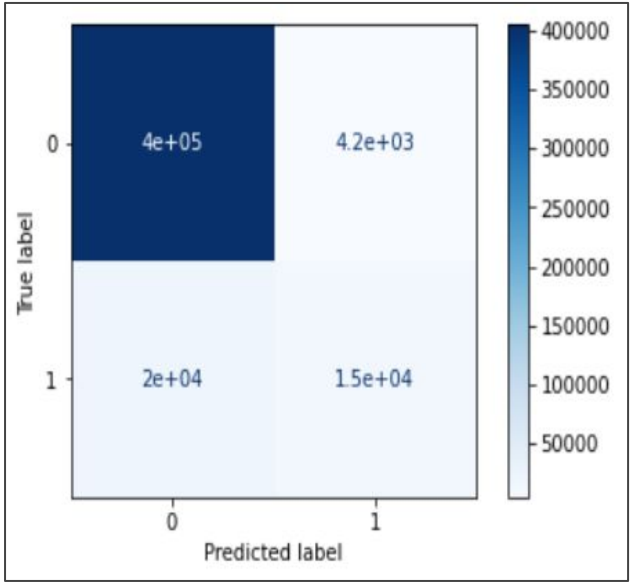
03

BERT

Deep Learning Model

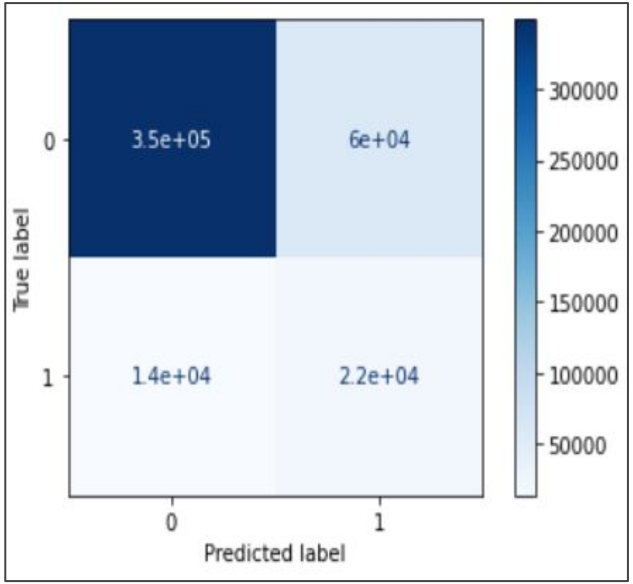
Logistic Regression

- Used Tfidfvectorizer to convert text to feature matrix.
- Hyperparameters: {C: 1, penalty: l2 }
- Accuracy : **94%**
- Recall : **43%**



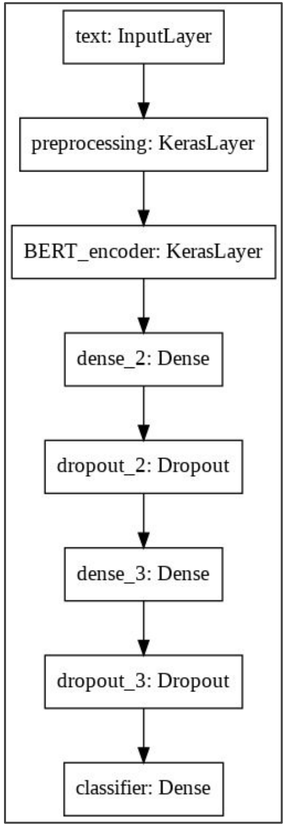
Random Forest

- Used Tfidfvectorizer to convert text to feature matrix.
- Hyperparameters: {class_weight: “balanced”, max_depth: 5, n_estimators: 190}
- Accuracy : **83%**
- Recall : **61%**



BERT

- Bert model : bert_en_uncased_L-4_H-512_A-8
- Preprocessing model: bert_en_uncased_preprocess
- Loss: Binary Crossentropy
- Metrics used: Recall & Accuracy
- Optimizer : Adam with learning rate of 1e-5



text: InputLayer

preprocessing: KerasLayer

BERT_encoder: KerasLayer

dense_2: Dense

dropout_2: Dropout

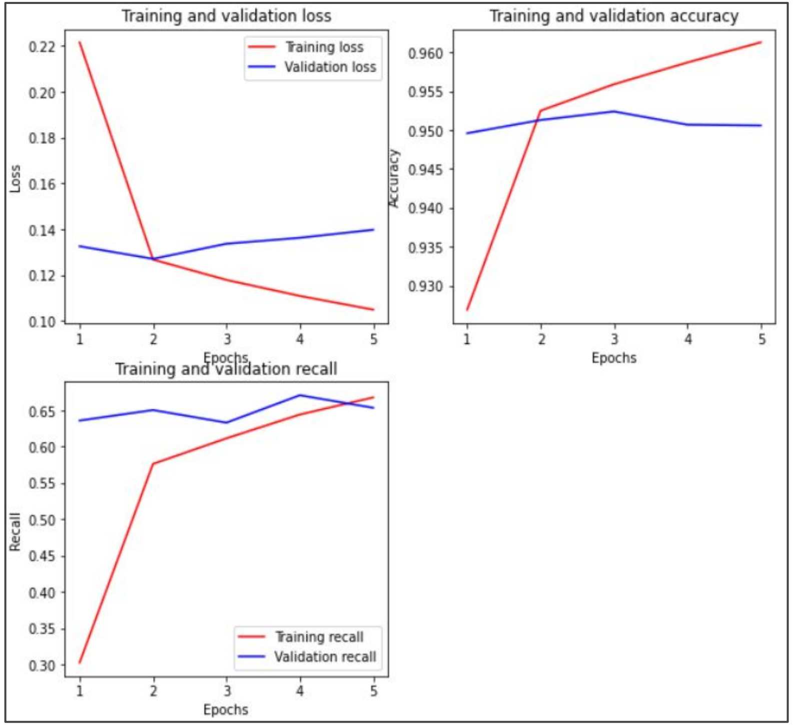
dense_3: Dense

dropout_3: Dropout

classifier: Dense

BERT

- Trained for 5 epochs.
- Loss & and accuracy converged at 2 epochs
- Recall converged after 2 epochs
- Accuracy : **95%**
- Recall : **63%**



Score Evaluation

Evaluating scores for all the three models.

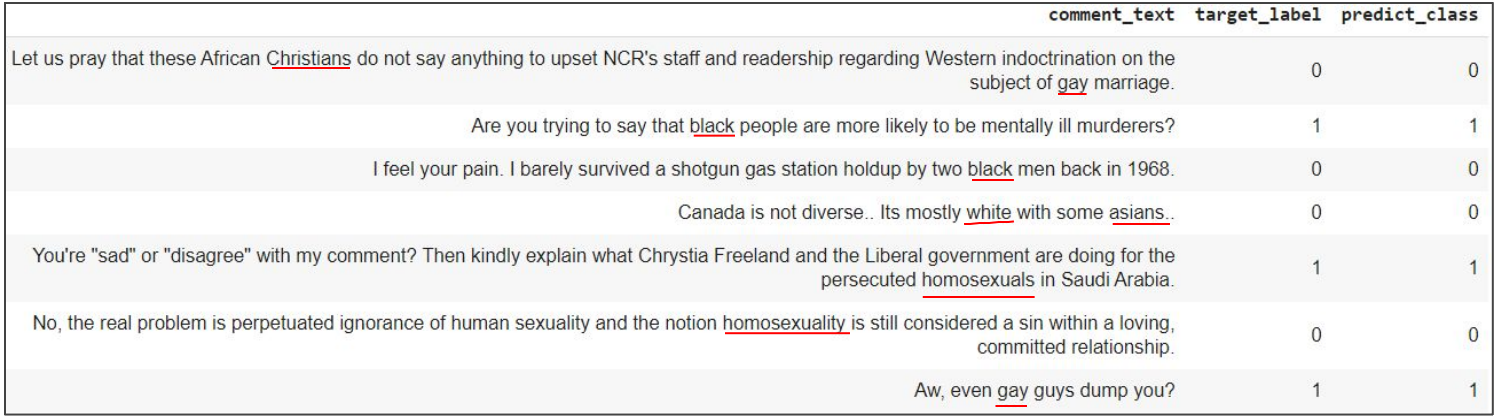
	Accuracy	Recall
Logistic Regression	94%	43%
Random Forest	83%	61%
BERT	95%	63%

Test scores for BERT model.

	Accuracy	Recall
BERT	95%	64%

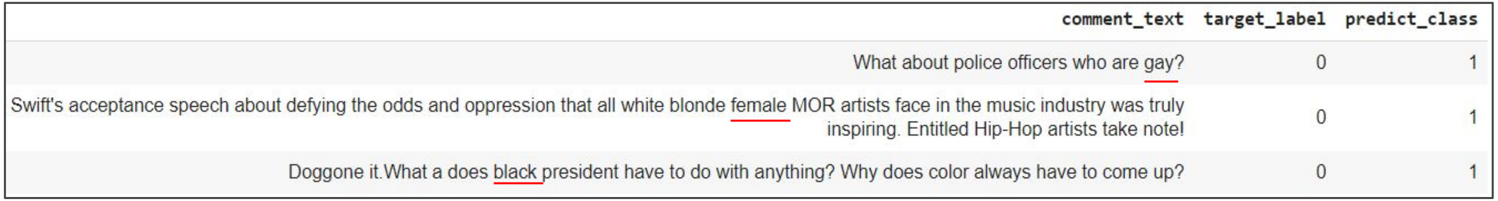
Evaluating Prediction

✓



comment_text	target_label	predict_class
Let us pray that these African Christians do not say anything to upset NCR's staff and readership regarding Western indoctrination on the subject of gay marriage.	0	0
Are you trying to say that black people are more likely to be mentally ill murderers?	1	1
I feel your pain. I barely survived a shotgun gas station holdup by two black men back in 1968.	0	0
Canada is not diverse.. Its mostly white with some asians..	0	0
You're "sad" or "disagree" with my comment? Then kindly explain what Chrystia Freeland and the Liberal government are doing for the persecuted homosexuals in Saudi Arabia.	1	1
No, the real problem is perpetuated ignorance of human sexuality and the notion homosexuality is still considered a sin within a loving, committed relationship.	0	0
Aw, even gay guys dump you?	1	1

✗



comment_text	target_label	predict_class
What about police officers who are gay?	0	1
Swift's acceptance speech about defying the odds and oppression that all white blonde female MOR artists face in the music industry was truly inspiring. Entitled Hip-Hop artists take note!	0	1
Doggone it.What a does black president have to do with anything? Why does color always have to come up?	0	1

05

Conclusion & Recommendation

Conclusion & Recommendation

- Was able to build a using BERT that can identify biased toxicity in the comments over the internet.
- How will this be useful?
  - Moderators manually delete the offensive comments or close the article/post.
  - This model will help these platforms to regulate toxic comments by identifying and deleting them.
  - This will help future conversations to be more productive and positive.

References

- https://www.tensorflow.org/tutorials/text/classify_text_with_bert

- https://www.machinelearningplus.com/nlp/gensim-tutorial/

- https://www.ft.com/content/9c0cf256-e197-11e3-b7c4-00144feabdc0

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

Please keep this slide for attribution

рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
blagodaram
dank je
misaotra
paldies
grazzi
matondo
спасибі
thank you
gracias
kiitos
dankie
nanni
nandri
dhanyavad
maururu
хвала
asante
manana
bedankt
hvala
köszönöm
grącie
bayarlalaa
dankon
aćiū
akun
obrigada
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
chokrane
sobodi
dėkuji
sagolun
sukriya
kop khun krap
obrigado
mèsi
didi madloba
ありがとう
arigatô
takk
dakujem
trugarez
tanemirt
grazie
najis tuke
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercë
мерси
kam sah hamnida
rahmat
xièxie
তোমাকে ধন্যবাদ
감사합니다
merci