Análise de Sentimentos

Fillipe de Menezes e Johnny Marcos

Roteiro

- Definição do Problema
- Análise Exploratória dos Dados
- Pré-processamento
  - Random Oversampling
  - nlpaug
- Solução
  - Modelo
  - Treinamento
- Resultados
- Rest API

Problema

Problema: Detectar conteúdo tóxico em conversas online (perguntas insinceras).

Can we use our external hard disk as a OS as well as for data storage.will the data be affected?	Normal
Has the United States become the largest dictatorship in the world?	Fake News / Incorreto
Which babies are more sweeter to their parents? Dark skin babies or light skin babies?	Racista

Competição: https://www.kaggle.com/c/quora-insincere-questions-classification

Problema

Uma pergunta insincera é definida como uma pergunta destinada a fazer uma declaração em vez de procurar respostas úteis.

- Racista ou Preconceituoso
- Depreciativo ou inflamatório
- Fake News
- Usar conteúdo sexual (incesto, bestialidade, pedofilia) para chocar e não para buscar respostas genuínas

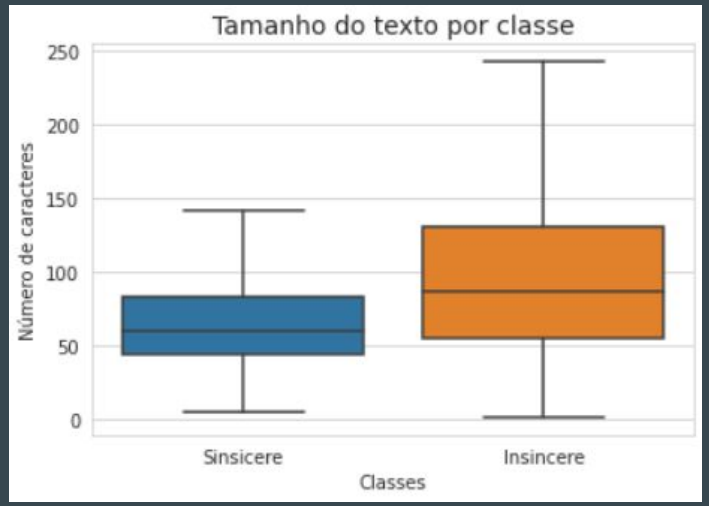
Base de Dados

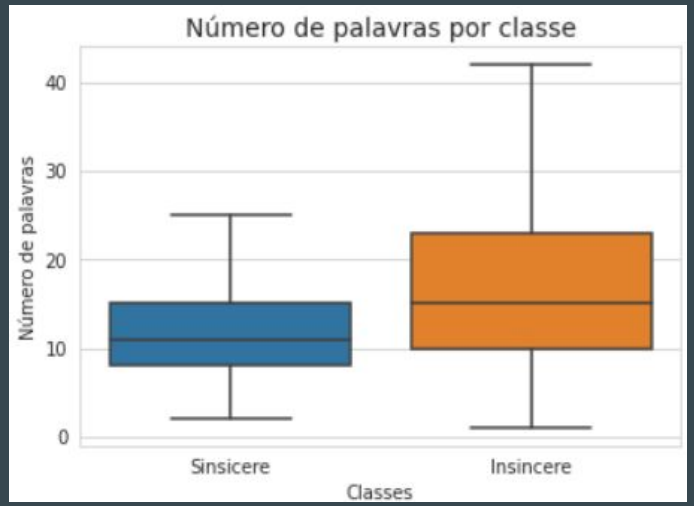
- 1.306.122 exemplos
- 93 % de perguntas normais (1.225.312)
- 7% de perguntas tóxicas (80.810)

question_text	target
How did Quebec nationalists see their province as a nation in the 1960s?	0
Do you have an adopted dog, how would you encourage people to adopt and not shop?	0
Why does velocity affect time? Does velocity affect space geometry?	0
How did Otto von Guericke used the Magdeburg hemispheres?	0
Can I convert montra helicon D to a mountain bike by just changing the tyres?	0
...	...
What other technical skills do you need as a computer science undergrad other than c and c++?	0
Does MS in ECE have good job prospects in USA or like India there are more IT jobs present?	0
Is foam insulation toxic?	0
How can one start a research project based on biochemistry at UG level?	0
Who wins in a battle between a Wolverine and a Puma?	0

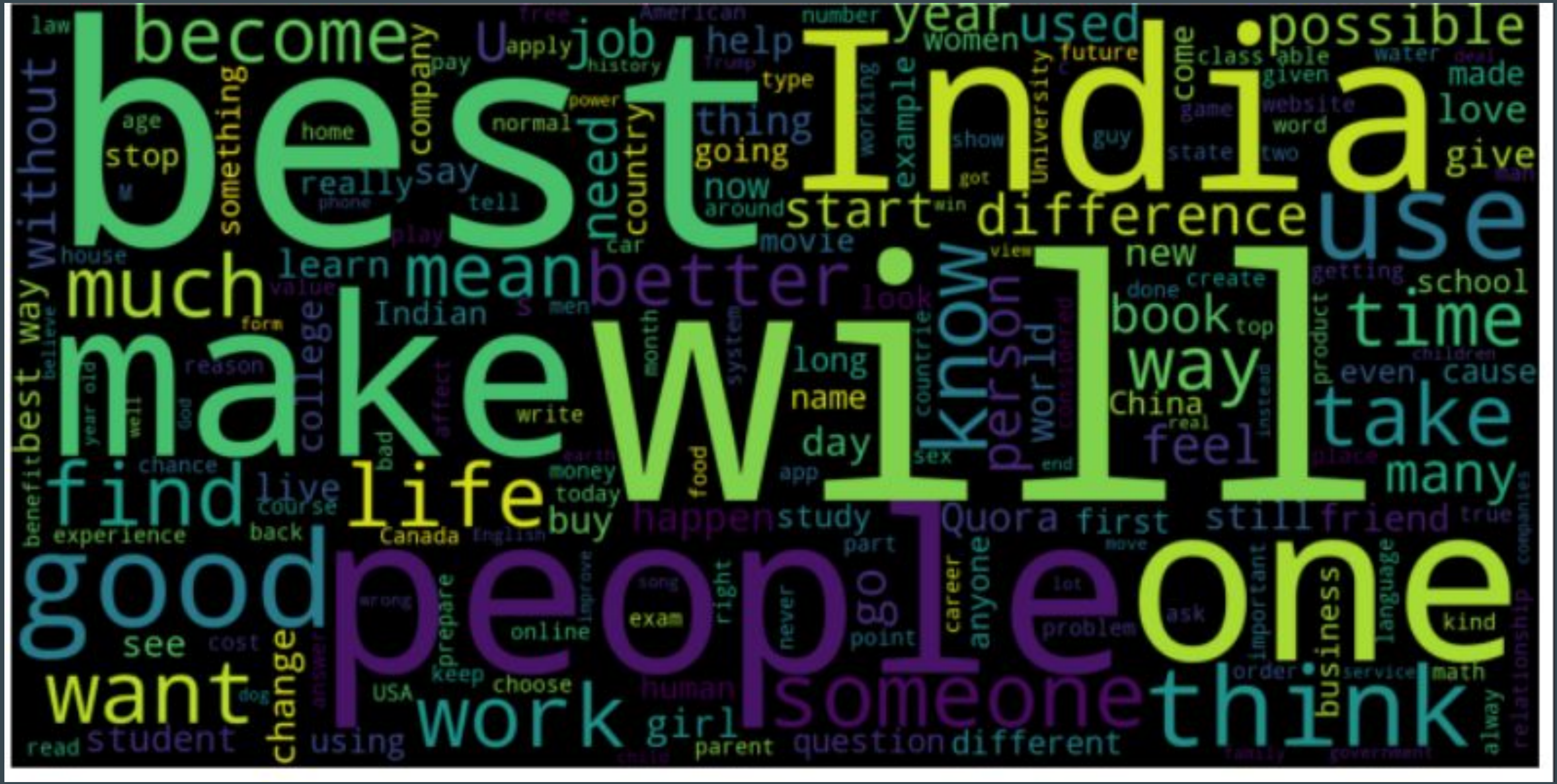
Base de Dados

- Comentários indevidos tem uma maior quantidade geral palavras e caracteres.



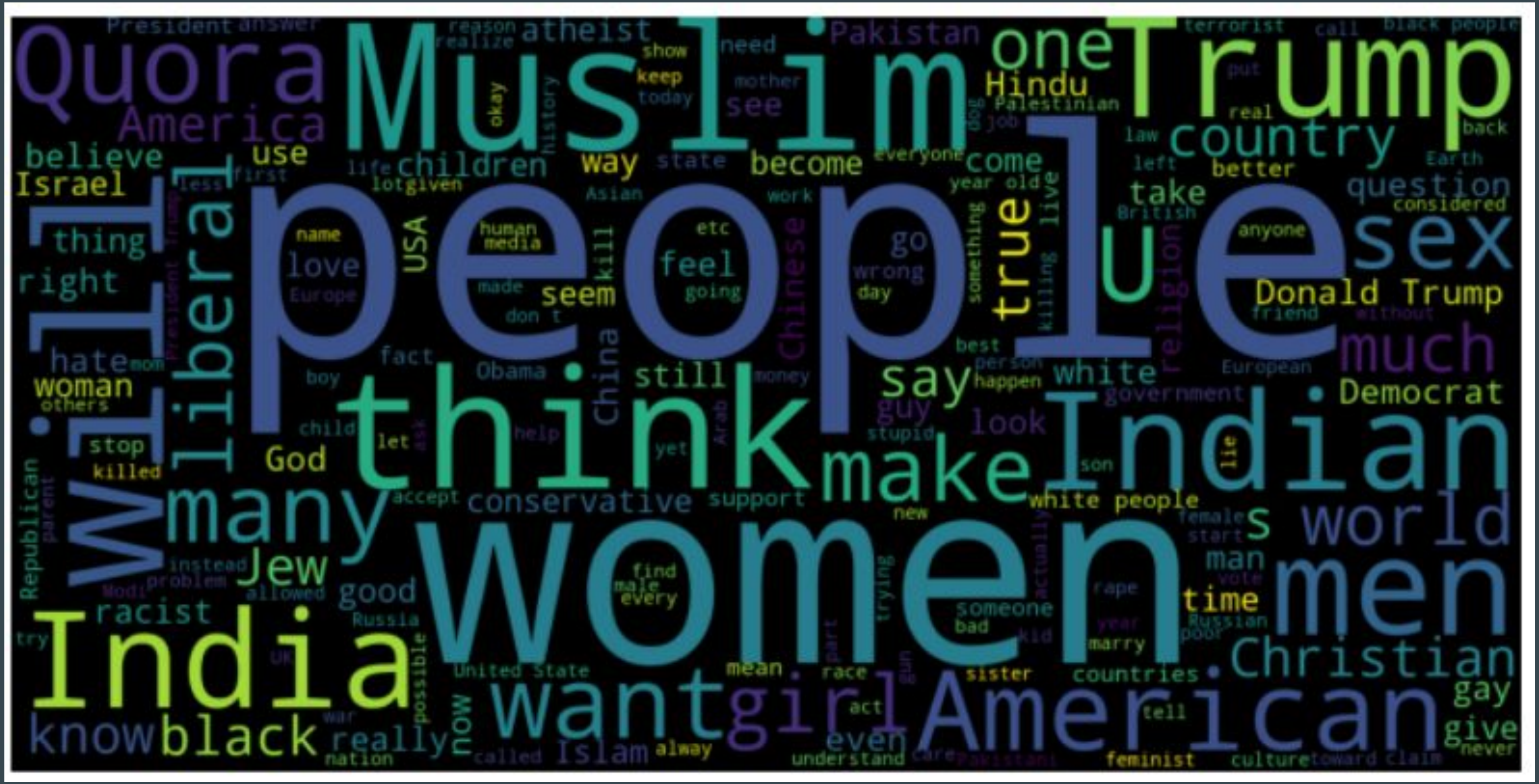


WordClouds



best
India
will
make
people
one
become
job
help
year
used
possible
women
future
class
able
water
made
love
give
use
without
something
company
thing
going
now
start
difference
need
country
much
mean
better
know
person
world
new
school
time
book
way
take
China
feel
many
find
life
live
course
buy
happen
study
Quora
first
still
friend
good
want
work
someone
think
student
using
girl
question
different
parent
read
change
see
cost
online
exam
choose
USA
keep
prepare
human
go
point
never
anyone
career
problem
ask
important
business
language
kind
order
service
math
relationship
alway
long
name
day
college
best way
learn
Indian
reason
write
money
today
back
experience
Canada
English
chance
stop
age
home
normal
really
say
tell
apply
pay
history
power
type
example
working
show
got
University
come
game
website
word
state
two
guy
movie
car
house
value
look
men
form
done
create
top
getting
even
cause
children
product
instead
real
system
month
countrie
view
win
around
phone
play
law
free
American
number
given
apply
end
app
sex
food
bad
benefit
dog
wrong
song
lot
part
move
true
companies
answer
improve
right

WordClouds


people
women
Muslim
think
Trump
Quora
India
Indian
American
will
men
many
want
girl
make
liberal
Christian
world
sex
much
country
one
America
black
know
Donald Trump
Democrat
true
say
Israel
believe
Hindu
Pakistan
atheist
children
become
come
question
take
thing
right
love
feel
hate
woman
Jew
racist
good
conservative
still
seem
God
China
white
look
religion
really
Islam
even
give
gay
time
man
white people
someone
Republican
Chinese
USA
use
way
see
go
fact
Obama
stop
killed
support

Pré-Processamento

Como a base de dados é desbalanceada usamos e comparamos dois processos para geração de novos dados:

- Random Oversampling:
  - Selecionar aleatoriamente dados até as classes estarem balanceadas.
- Geração de dados sintéticos com base em diferentes regras:
  - Erros de digitação
  - Erros de ortografia
  - Substituição por sinônimos
  - Substituição com base no contexto

nlpaug

- Simular erros de digitação:

Texto original	Texto gerado
Why are books about ‘how to rape a woman’ in schools in the US?'	Why are goiks about ‘ how to rape a w9jan ’ in schools in the US?
How can I train my girlfriend to stop speaking out of place in public?	How can I tFaLn my girlfriend to stop speaking out of Llacw in public?

nlpaug

- Substituição por sinônimos

Texto original	Texto gerado
This film is pretty good!	This moving picture show is pretty good!
How can I train my girlfriend to stop speaking out of place in public?	How can I train my girlfriend to end speaking out of place in world?

nlpaug

- Erros de ortografia

Texto original	Texto gerado
The quick brown fox jumps over the lazy dog.	The quikly brown fox jumps over the lasy dig.
How can I train my girlfriend to stop speaking out of place in public?	'How can I train my girlfriend to stop speaking out of place in bubic?'

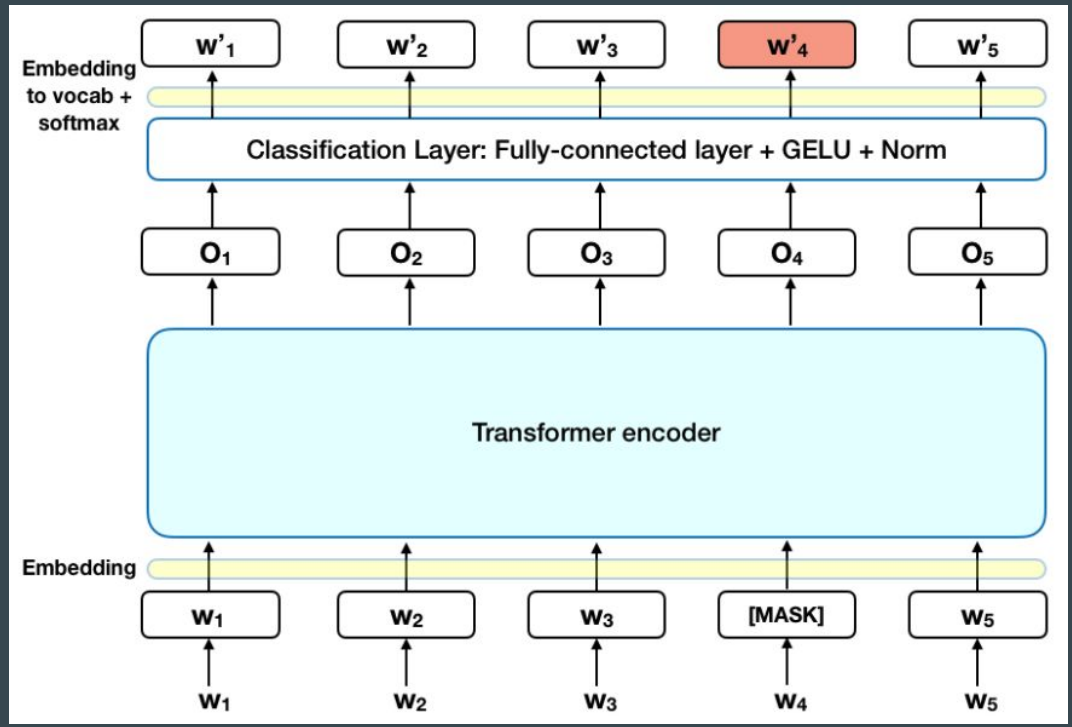
nlpaug

- Contextual Words Embeddings

Texto original	Texto gerado
This film is pretty good!	your album is a good!
The quick brown fox jumps over the lazy dog	the quick snapping cat jumps over the smaller dog

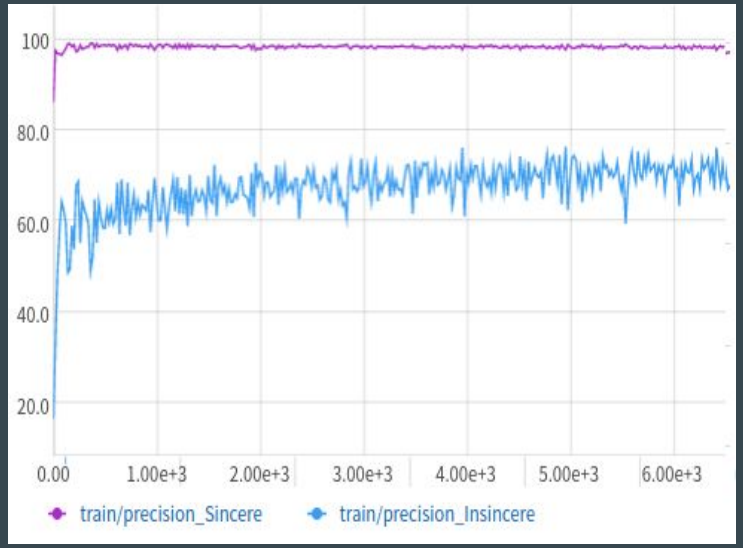
Modelo

- Uso do BERT para geração dos embeddings.
- Usar os embeddings como entrada para uma rede neural.
- Fine Tuning...

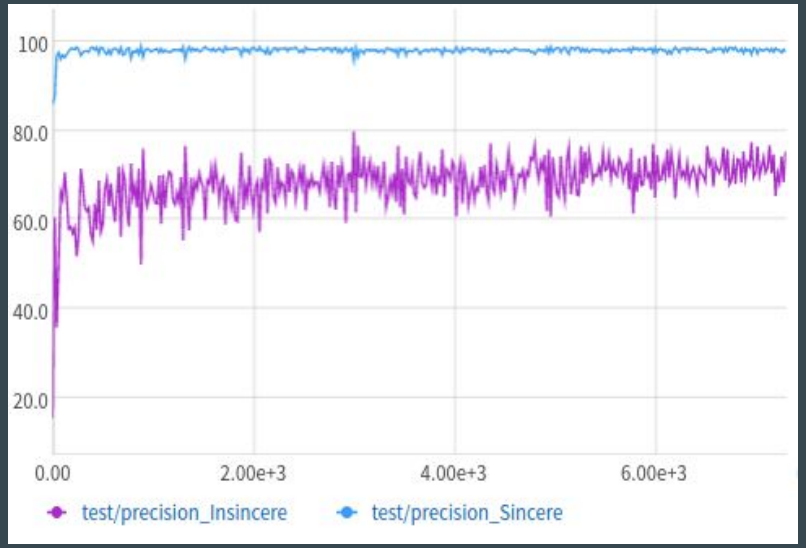


Treinamento

ROS



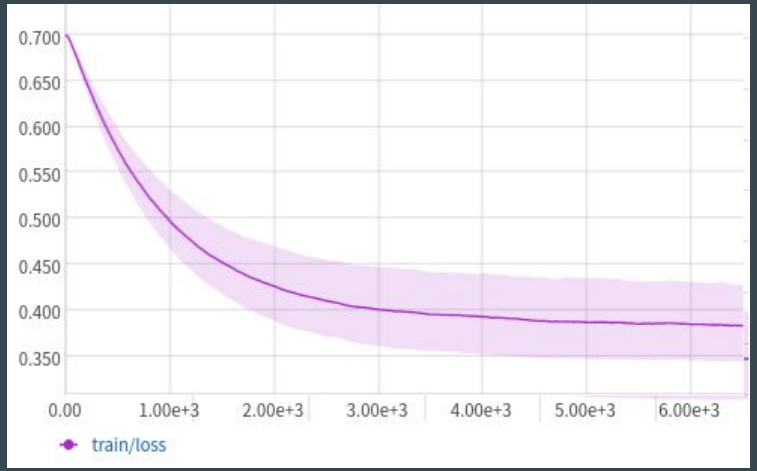
NLPAUG



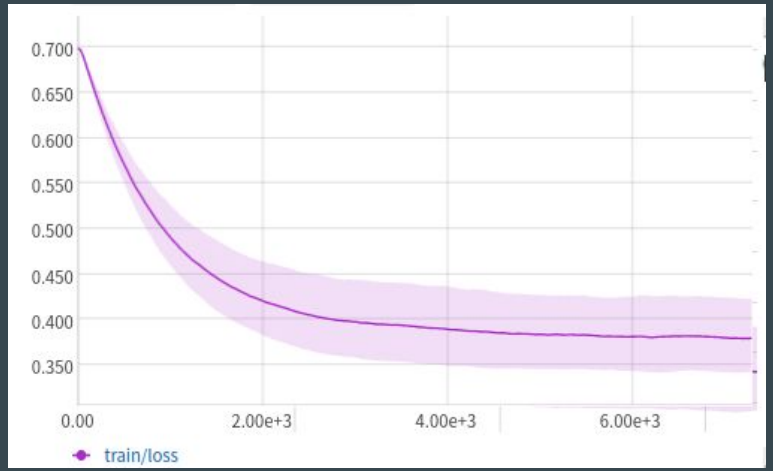
Precisão para cada classe ao longo do treinamento

Treinamento

ROS



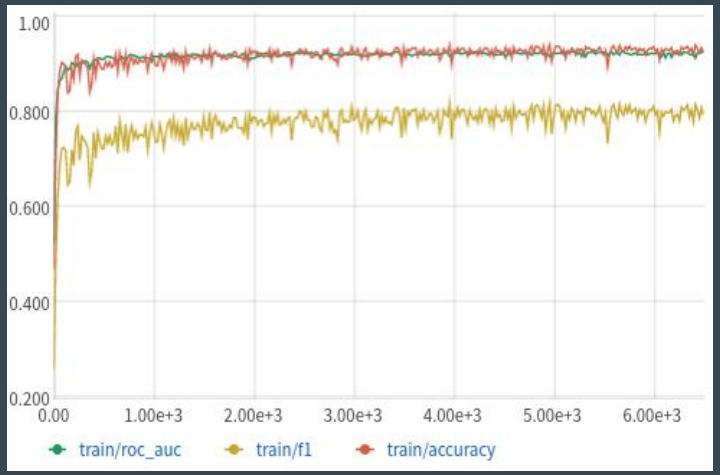
NLPAUG



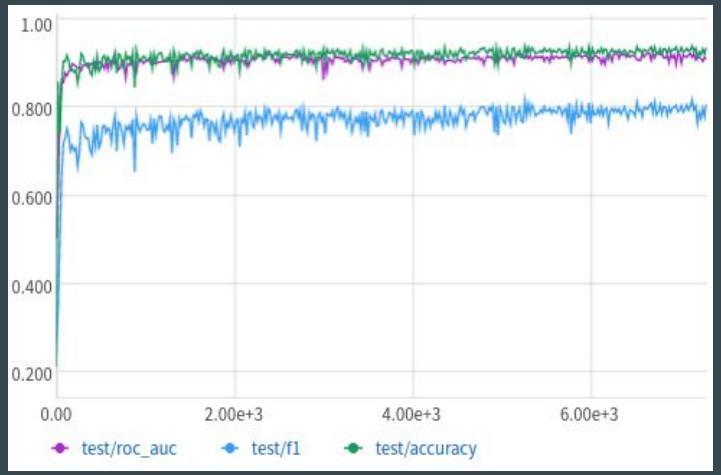
Loss ao longo do treinamento

Treinamento

ROS



NLPAUG



Comparação entre acurácia, F1 e área sob a curva roc

Resultados

O desempenho das duas redes foi comparada no banco de dados de teste que contém 50.000 exemplos. Como pode ser visto na tabela a rede treinada com **ROS** superou os resultados obtidos nos dados do **NLPAUG**.

Base de Dados	Métrica	Positivo(Normal)	Negativo(Tóxico)
ROS	Precisão	**0.99**	**0.65**
ROS	Recall	**0.92**	**0.92**
ROS	F1	**0.95**	**0.76**
NLPAUG	Precisão	0.98	0.63
NLPAUG	Recall	0.91	0.90
NLPAUG	F1	0.95	0.74

Resultados - Matriz de confusão

Random OverSampling:

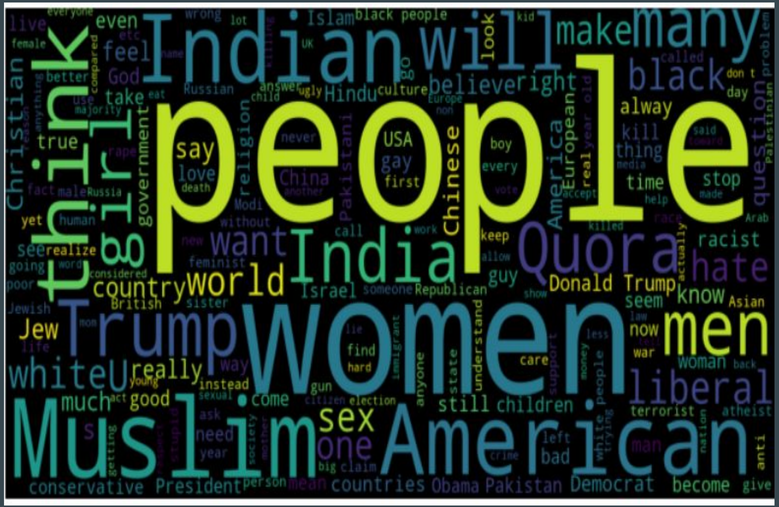
Valores Preditos

Valores Reais	P	N
P	44.070	4.030
N	655	7.326

Data augmentation:

Valores Reais	P	N
P	43.971	4.129
N	830	7.151

Resultados



people
women
Muslim
American
Indian
will
Trump
think
girl
India
Quora
many
make
men
liberal
black
hate
world
white
sex
one
want
Chinese
Christian
country
Donald Trump

people
women
will
think
girl
American
Indian
Muslim
Trump
Quora
India
many
country
one
men
make
true
want
world
China
someone
know
feel
sex
God
say
guy
Donald Trump

Resultados

Is this English sentence grammar correct? "I don\'t like to butter people
Why does Trump want the wall funding so bad when the majority voters don't?
How come pornography eats up all my data?
When a women who have 28 days on menustral circle can get pregnant?
Liberals, Is it true that last year on Nov 8, Hillary Clinton started losing her lead after working people went in to vote after 4 pm?

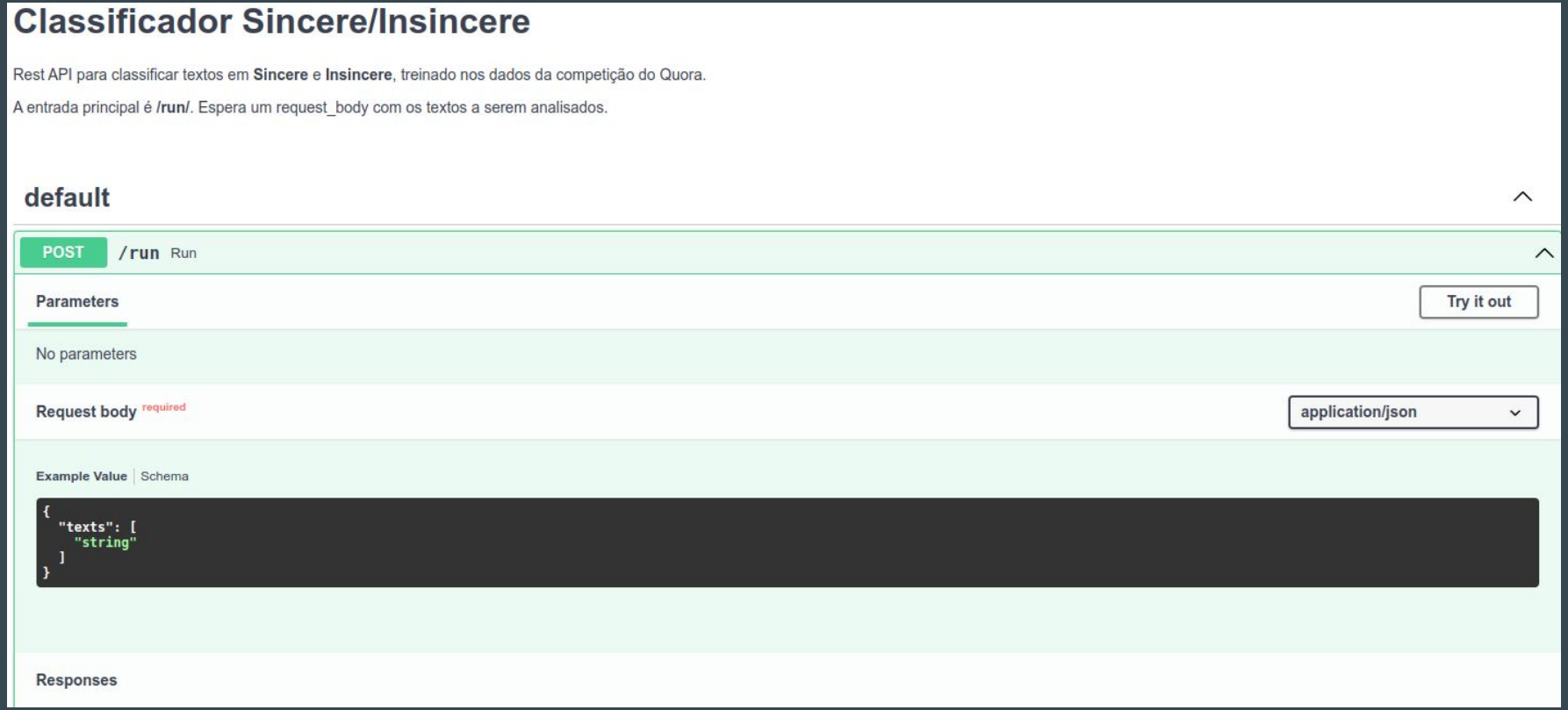
Tabela. Exemplos normais

Resultados

Do Indian girls like to get dry humped from behind at crowded places when no one notices them and guy is good looking?
Since when did women start to wear short & indecent dresses in public?
Can fuckers on Quora please shut the fuck up about IQ?
What were the UK and France thinking when creating all of these fake countries that never existed before in 1923 (Turkey, Iraq, Syria)? Now what are they doing to stop the war they created 100 years ago?
So when will catalans delare independence from Homo Sapiens?

Tabela. Exemplos tóxicos

Rest API


Classificador Sincere/Insincere
Rest API para classificar textos em Sincere e Insincere, treinado nos dados da competição do Quora.
A entrada principal é /run/. Espera um request_body com os textos a serem analisados.
default
POST
/run
Run
Parameters
Try it out
No parameters
Request body
required
application/json
Example Value
Schema
{
"texts": [
"string"
]
}
Responses

Conclusão

- O modelo teve um desempenho baixo nos textos de conteúdo tóxico/fake news.
- No experimento o Random Oversampling teve uma melhor performance do que usar Data Augmentation, porém:
  - A escolha de alguns processos para gerar texto pode ter impactado a performance.
  - A diversos parâmetros que não foram ajustados no nlpaug:
    - num_max_words,
    - num_max_caracters,
    - back to back translation e etc
- É necessário definir com melhor cuidados as técnicas e usar uma forma para validar os dados mais rápida.

Github: https://github.com/Fillipedem/quora-insincere-questions

Perguntas?