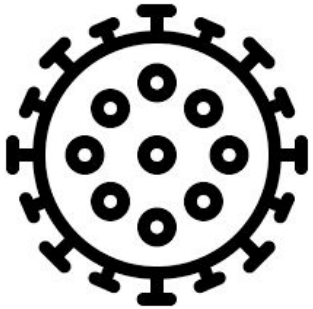
Data Science Bootcamp

Carlo CISALE

linkedin.com/in/carlo-cisale

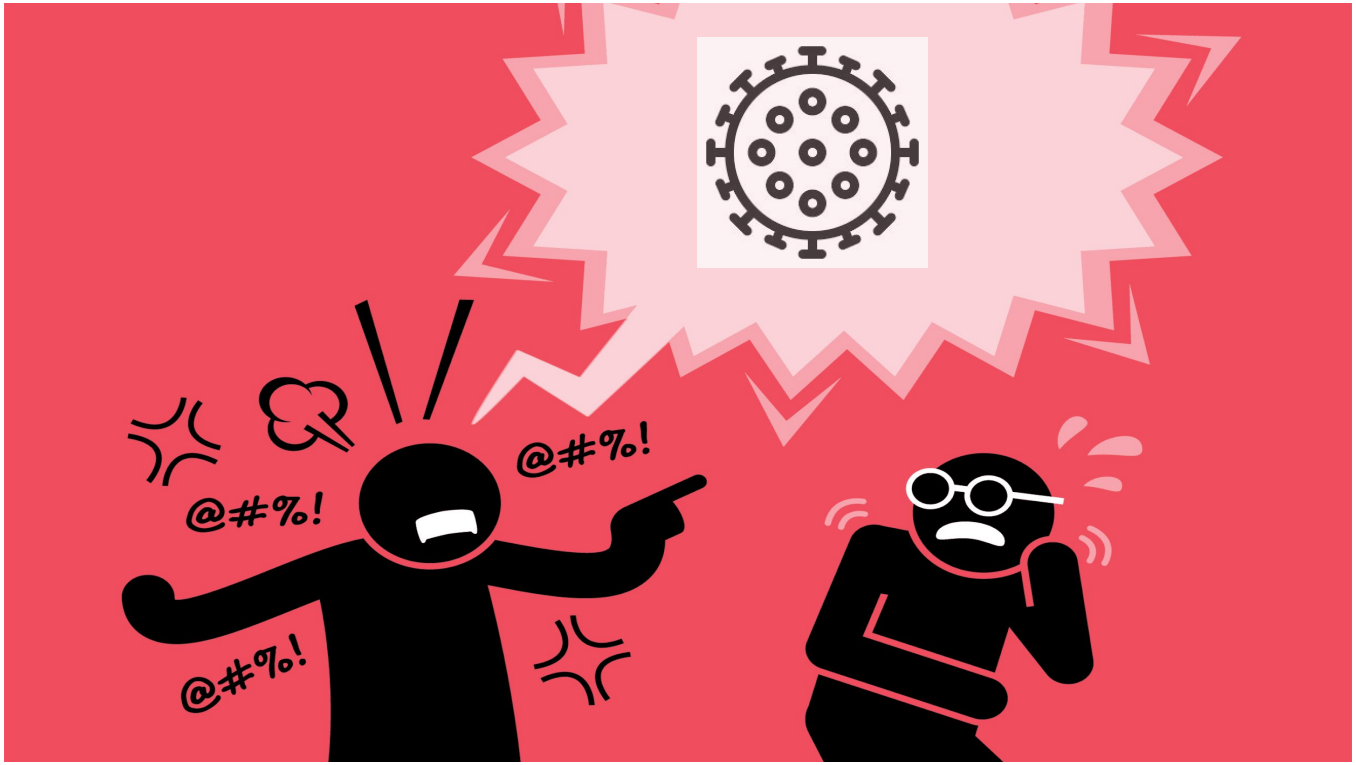
Jedha

Sentiment Analysis on #coronavirus tweets



Motivation

Motivation



Steps

- Scraping tweets en anglais avec mots cles “Coronavirus” et “Italians” avec twitterscraper
- Data preparation des tweets
- Libelisation des tweets avec librarie textblob
- Entrainement modèle de ML sur un dataset de hate speech
- Test du modèle sur les tweets coronavirus

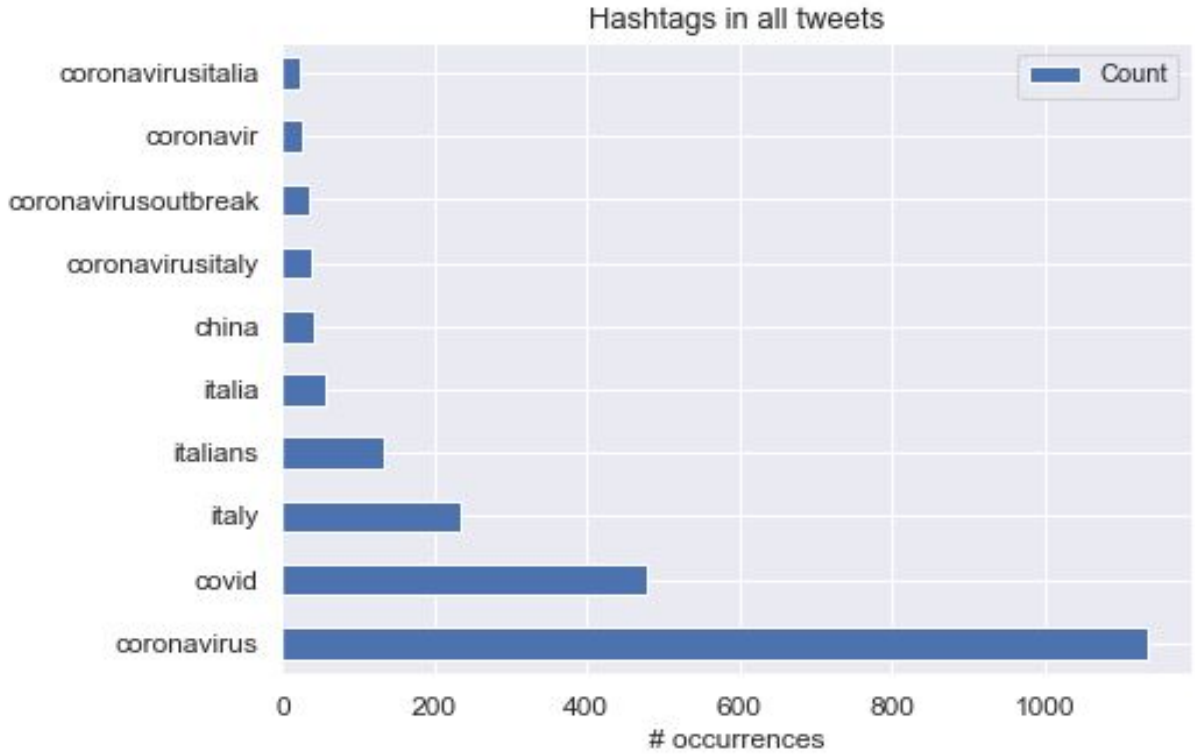
Difficulté avec la data preparation

- Format des tweets “scrapé”: Json avec beaucoup d’info
- Doublons
- Suppression des liens: http..., pic.twitter....

Word Cloud

coronaviru
Italian
Italy
people
Corona
Chinese
viru
country
will
spread
lockdown
now
case
new
world
day
COVID19
quarantine
quarantined
outbreak
China
death
one
via
make
American
making
pasta
Please
stop
Watch
left
look
test
know
good
still
mask
situation
infected
don
nation
much
hard
YouTube
going
Wuhan
hug
travel
media
EU
hand
everyone
go
Covid_19
made
today
free
number
right
friend
ago
put
care
hope
thought
countries
Milan
flu
fight
help
year
think
keep
million
COVID
doctor
Minister
Germany
trying
show
said
dead
face
many
panic
overwhelmed
racist
really
taking
Trump
back
crisis
see
singing
health
got
UK
want
time
German
even
rate
stay
two
live
hit
together
die
around
may
Northern
thousand
saying
patient
sick
seem
life
love
South
already
dying
thing
tested
sing
Italia
send
record
week
died
Asian
say
India
way
BREAKING
due
message
lot
Europe
racism
Africa
far
positive
told
work
first
Iranian
us
confirmed
flight
measure
balconies
home
hospital
take
die
pandemic
video
border
need
amid
fear
social
last
spreading
must
government
well
invented
come
coming

Hashtags les plus populaires



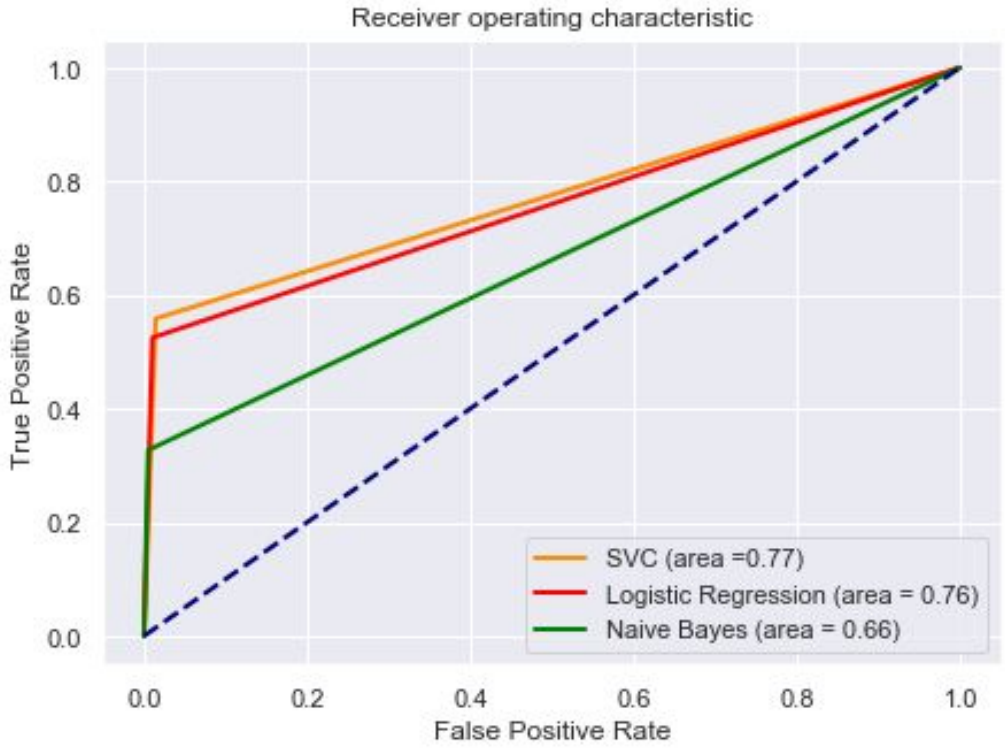
Comment convertir données textuelles en numériques?

- **TF*IDF: Term Frequency – Inverse Document Frequency**

$$TF(t,d) = \frac{number\ of\ times\ term(t)\ appears\ in\ document(d)}{total\ number\ of\ terms\ in\ document(d)}$$

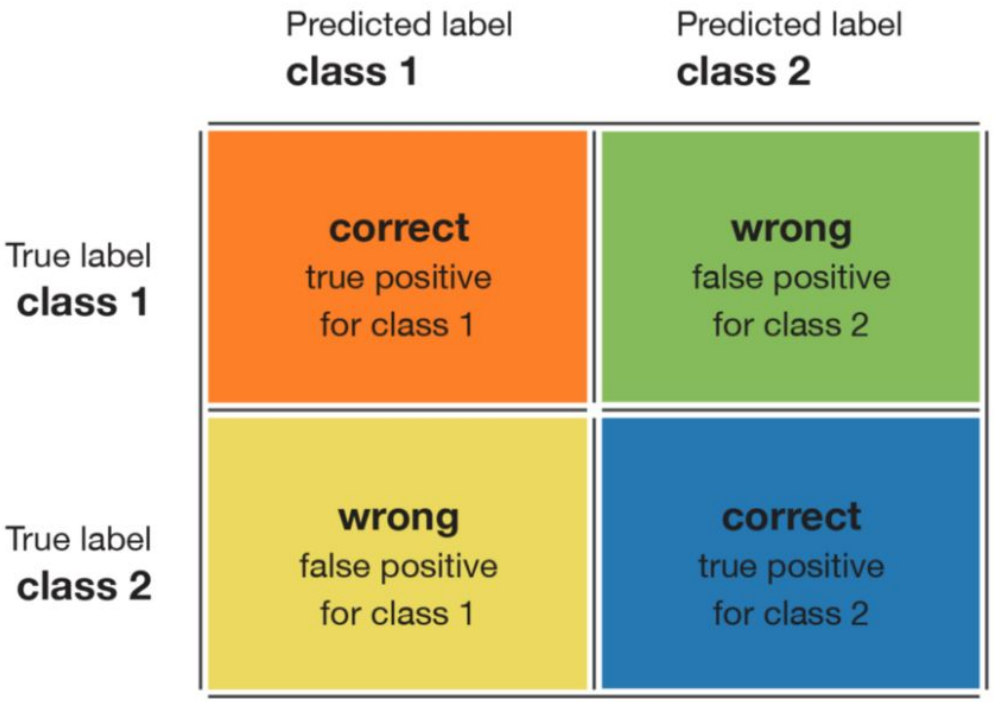
$$IDF(t,D) = \log\left(\frac{total\ number\ of\ documents(D)}{number\ of\ documents\ with\ the\ term(t)\ in\ it}\right)$$

Comparaison algorithmes (hate speech data)

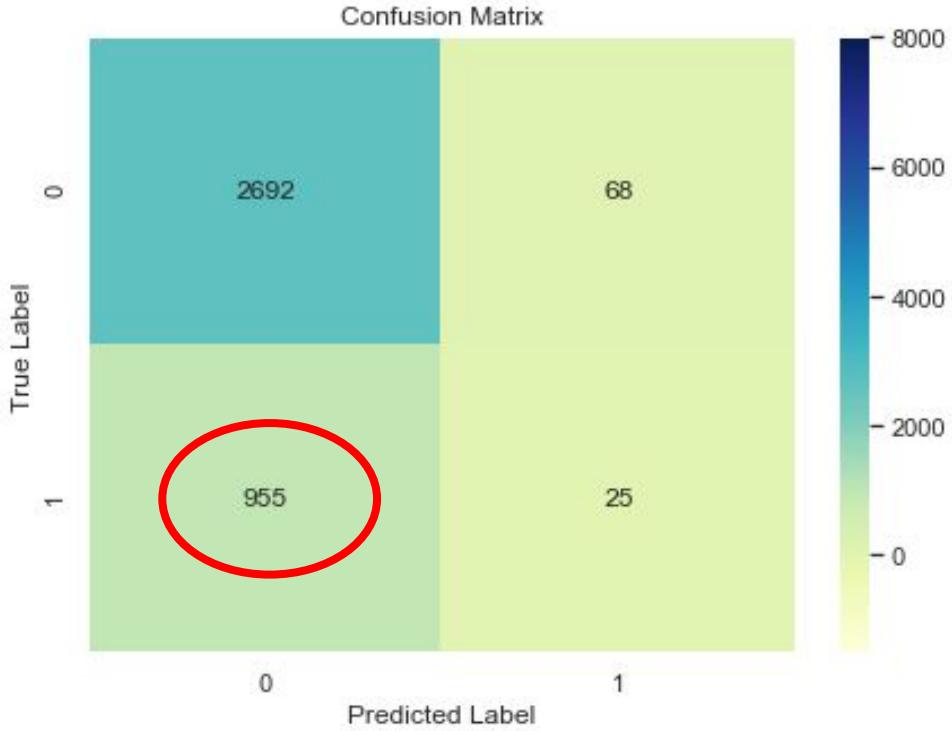


Algo	F-1 score
SVC	0.64
Logistic	0.64
NB	0.53

Matrice de confusion



Matrice de confusion (coronavirus data)



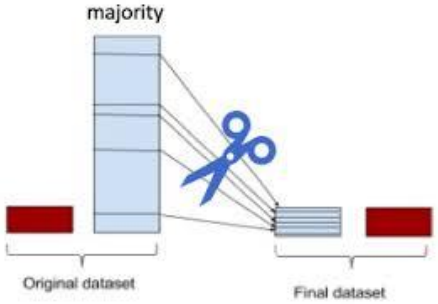
Matrice de confusion (coronavirus data)

Limitations. Pourquoi?

1. Mauvaise classification de TextBlob des tweets négatifs (des tweets normaux étaient classifié comme péjoratifs).
2. Les deux datasets (hate speech et coronavirus) n'avaient pas la même taille.
3. Le modèle prends en considération "high signal words". Mais qu'est ce que il se passe avec des nouveaux mots?
4. Le dataset initial de hate speech (mais aussi du coronavirus) sont imbalanced.

Conclusions et améliorations

Word2Vec



Downsampling/
Upsampling

Human classification


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

Data Science Bootcamp

Des questions ?

Jedha