SENTIMENT ANALYSIS WITH NAIVE BAYES

ON TWITTER DATASET

ANDREA CORVI

WHAT SENTIMENT ANALYSIS IS?

THE PROCESS OF COMPUTATIONALLY IDENTIFYING AND CATEGORIZING OPINIONS EXPRESSED IN A PIECE OF TEXT, IN ORDER TO DETERMINE WHETHER THE WRITER'S ATTITUDE TOWARDS A PARTICULAR TOPIC IS POSITIVE, NEGATIVE, OR NEUTRAL.

WHY IS IT USEFUL?

ALLOWS SOCIAL MEDIA MONITORING IN ORDER TO GET AN OVERVIEW OF THE PUBLIC OPINION BEHIND CERTAIN TOPICS.

SOURCE: ANALYTICS VIDHYA COMPETITION

data - DataFrame

Index	id	label	tweet	clean_tweet
0	1	0	@user when a father is dysfunctional and is so selfish he…	when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.
1	2	0	@user @user thanks for #lyft credit i can't use cause they…	thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #d
2	3	0	bihday your majesty	bihday your majesti
3	4	0	#model i love u take with u all the time in urð ±!!! ð…	#model i love u take with u all the time in urð ±!!! ð ð ð ð ð ¦ð ¦ð ¦
4	5	0	factsguide: society now #motivation	factsguide: society now #motiv
5	6	0	[2/2] huge fan fare and big talking before they leave. cha…	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get
6	7	0	@user camping tomorrow @user @user @user @user @user @use…	camping tomorrow dannyâ ¦
7	8	0	the next school year is the year for exams.ð ¯ can't thi…	the next school year is the year for exams.ð ¯ can't think about that ð #school #exams
8	9	0	we won!!! love the land!!! #allin #cavs #champions #clevel…	we won!!! love the land!!! #allin #cavs #champions #cleveland #clevelandcavaliers â ¦
9	10	0	@user @user welcome here ! i'm it's so #gr8 !	welcome here ! i'm it's so #gr8 !
10	11	0	â ¥ #ireland consumer price index (mom) climbed from pre…	â ¥ #ireland consumer price index (mom) climbed from previous 0.2% to 0.5% in may #blog
11	12	0	we are so selfish. #orlando #standwithorlando #pulseshooti…	we are so selfish. #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproble
12	13	0	i get to see my daddy today!! #80days #gettingfed	i get to see my daddy today!! #80days #gettingf
13	14	1	@user #cnn calls #michigan middle school 'build the wall' …	#cnn calls #michigan middle school 'build the wall' chant '' #tcot
14	15	1	no comment! in #australia #opkillingbay #seashepherd #h…	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove #helpco
15	16	0	ouch...junior is angryð É#got7 #junior #yugyoem #omg	ouch...junior is angryð É#got7 #junior #yugyoem #omg
16	17	0	i am thankful for having a paner. #thankful #positive	i am thankful for having a paner. #thankful #positive
17	18	1	retweet if you agree!	retweet if you agree!
18	19	0	its #friday! ð smiles all around via ig user: @user #c…	its #friday! ð smiles all around via ig user: #cookies make people
19	20	0	as we all know, essential oils are not made of chemicals.	as we all know, essential oils are not made of chemicals.
20	21	0	#euro2016 people blaming ha for conceded goal was it fat r…	#euro2016 people blaming ha for conceded goal was it fat rooney who gave away free kick kno
21	22	0	sad little dude.. #badday #coneofshame #cats #pissed #fu…	sad little dude.. #badday #coneofshame #cats #pissed #funny #laughs
22	23	0	product of the day: happy man #wine tool who's it's the…	product of the day: happy man #wine tool who's it's the #weekend? time to open up &
23	24	1	@user @user lumpy says i am a . prove it lumpy.	lumpy says i am a . prove it lumpy.
24	25	0	@user #tgif #ff to my #gamedev #indiedev #indiegamedev …	#tgif #ff to my #gamedev #indiedev #indiegamedev #squad!
25	26	0	beautiful sign by vendor 80 for $45.00!! #upsideofflorida …	beautiful sign by vendor 80 for $45.00!! #upsideofflorida #shopalyssas #love

TOKENIZATION

BREAKING UP A SENTENCE OR PARAGRAPH INTO SPECIFIC TOKENS OR WORDS. WITHOUT IDENTIFYING THE TOKENS, IT IS DIFFICULT TO IMAGINE EXTRACTING HIGH LEVEL INFORMATION FROM TEXT.

STOPWORDS REMOVAL

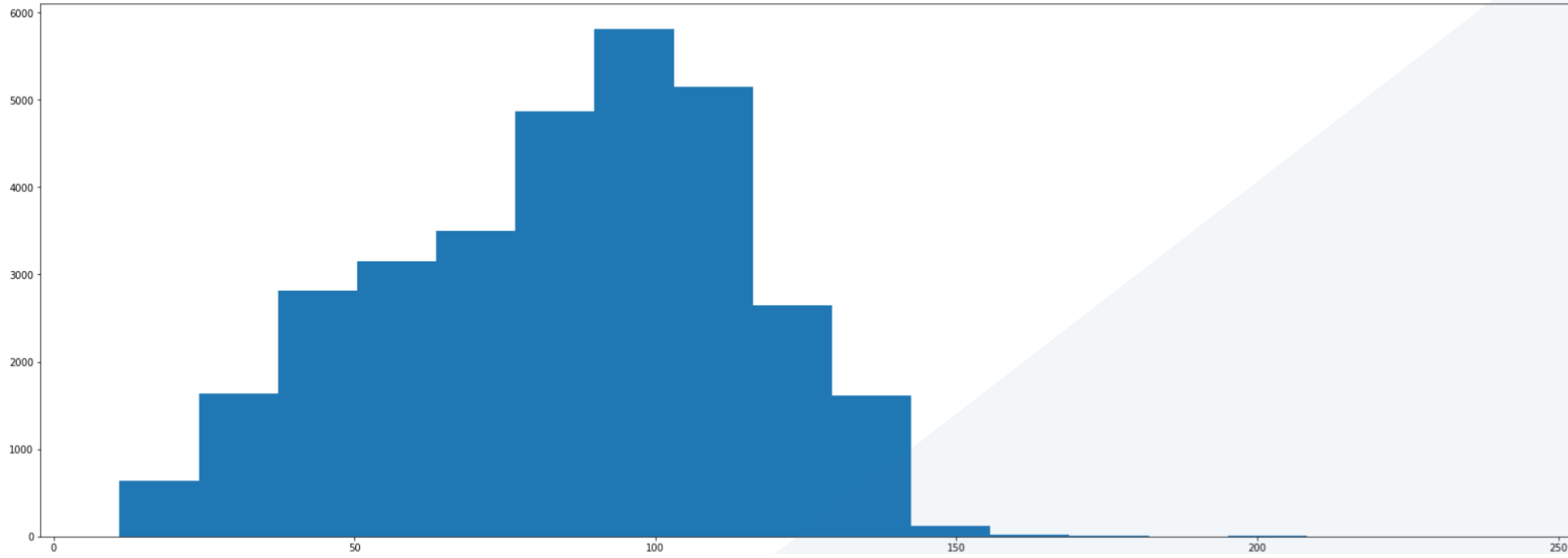
REMOVING CONJUNCTIONS, ADJECTIVE, ETC, IN ORDER TO MAKE THE INPUT DATA MORE MEANINGFUL TO THE ALGORITHM

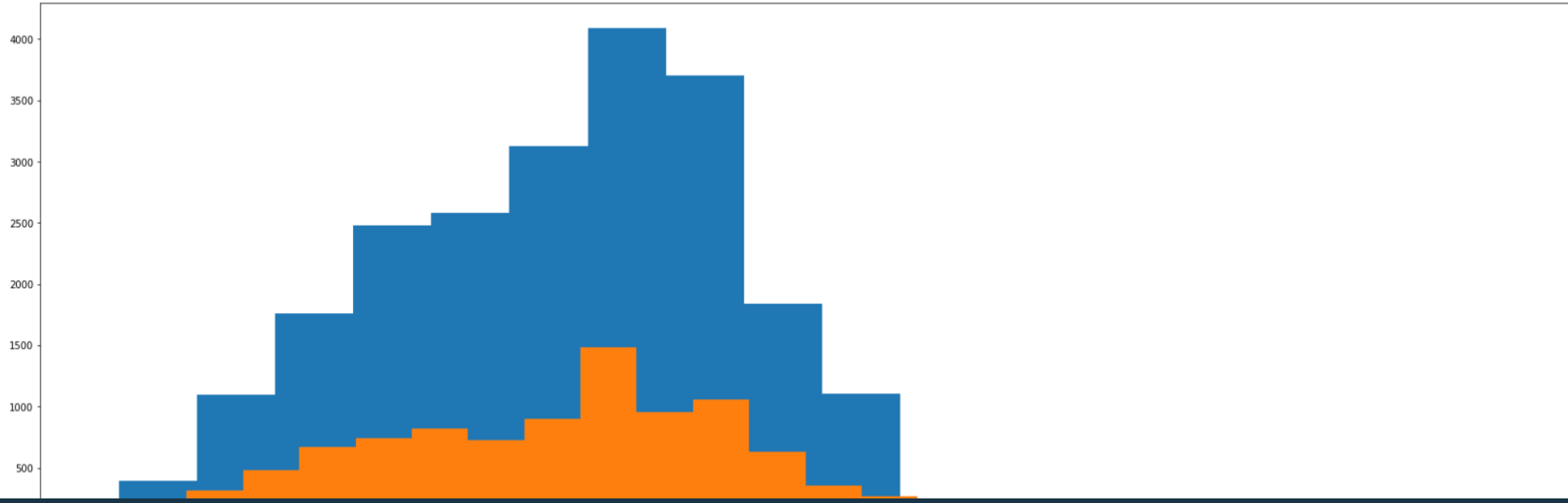
STEMMING

REDUCING INFLECTED (OR SOMETIMES DERIVED) WORDS TO THEIR WORD STEM, BASE OR ROOT FORM

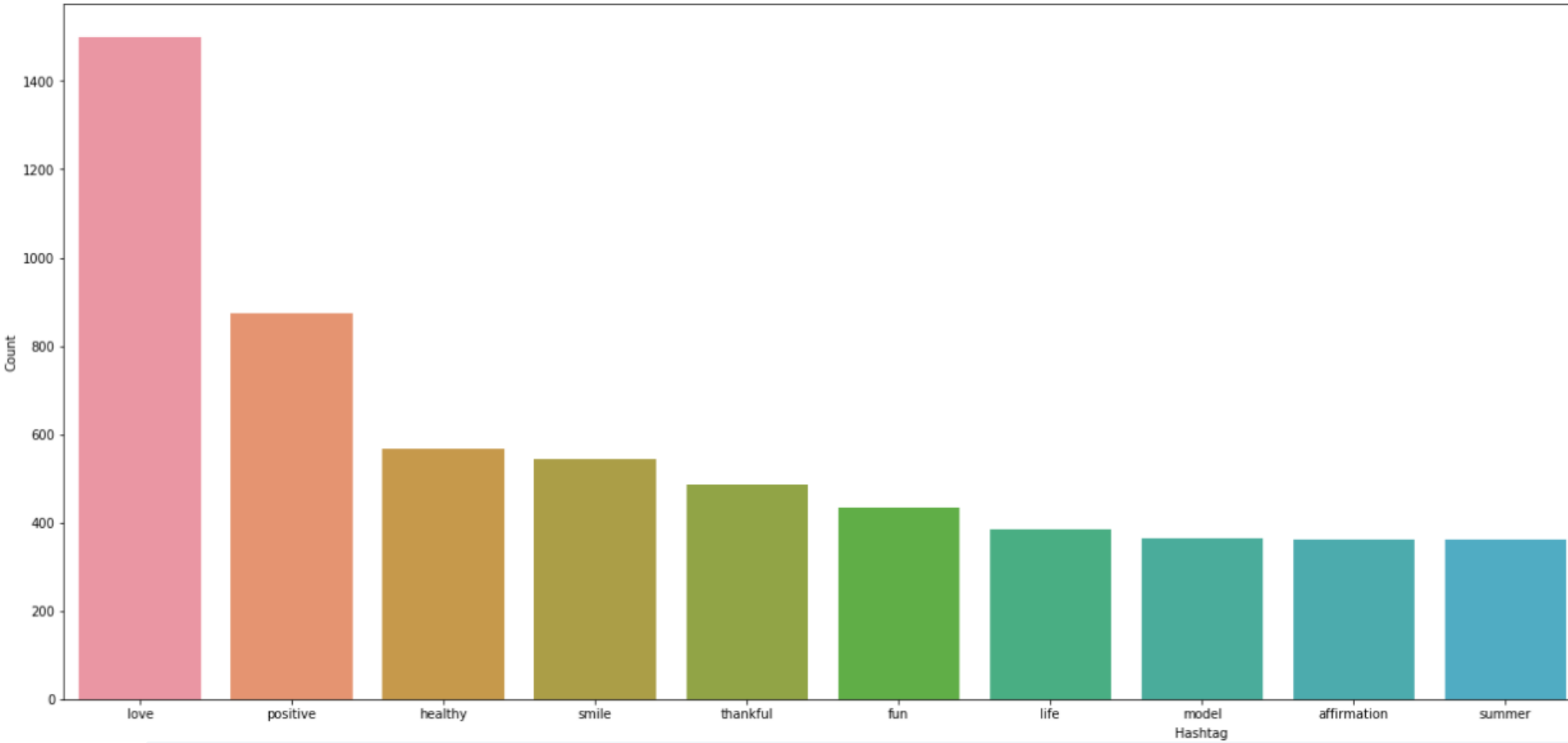
PUNCTUATIONS, NUMBERS, AND SPECIAL CHARACTERS REMOVAL

TWITTER @USER REMOVAL

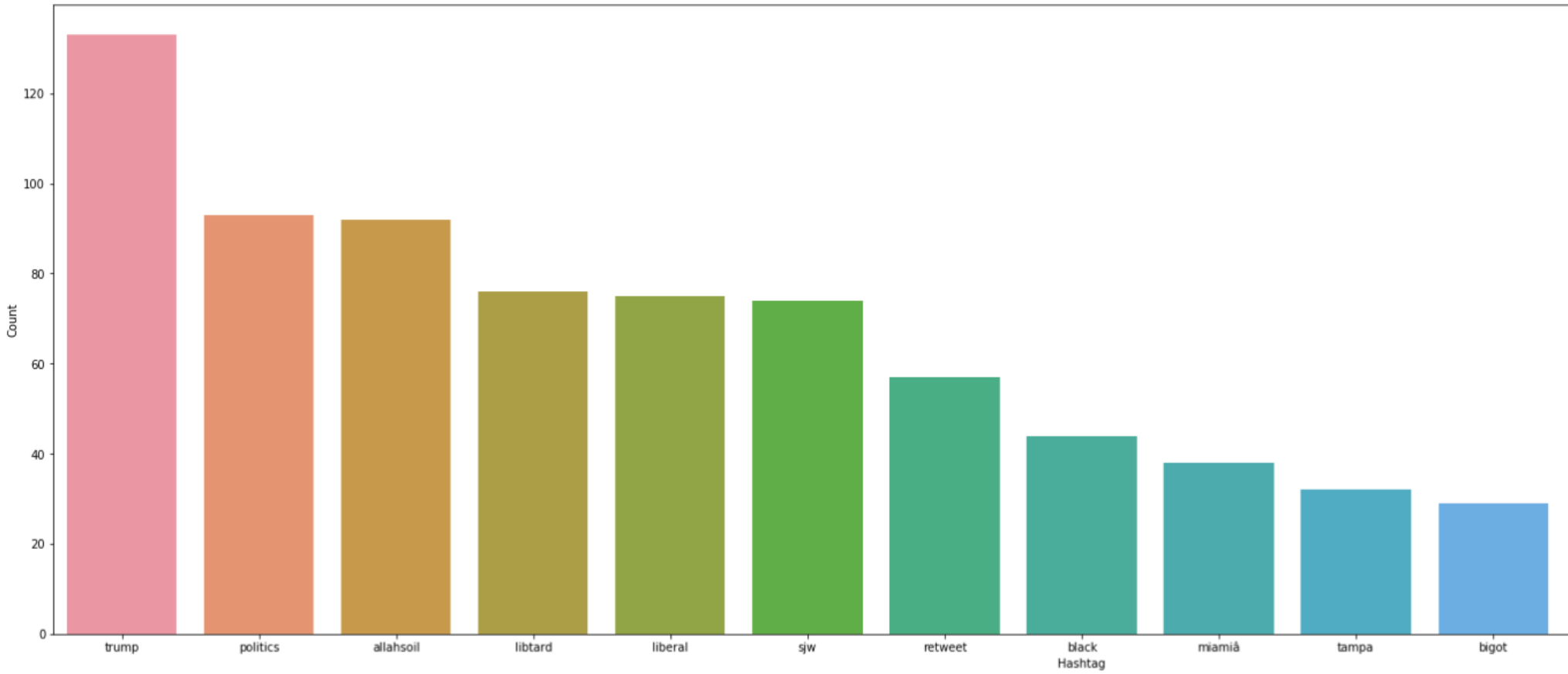

6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250


4000
3500
3000
2500
2000
1500
1000
500

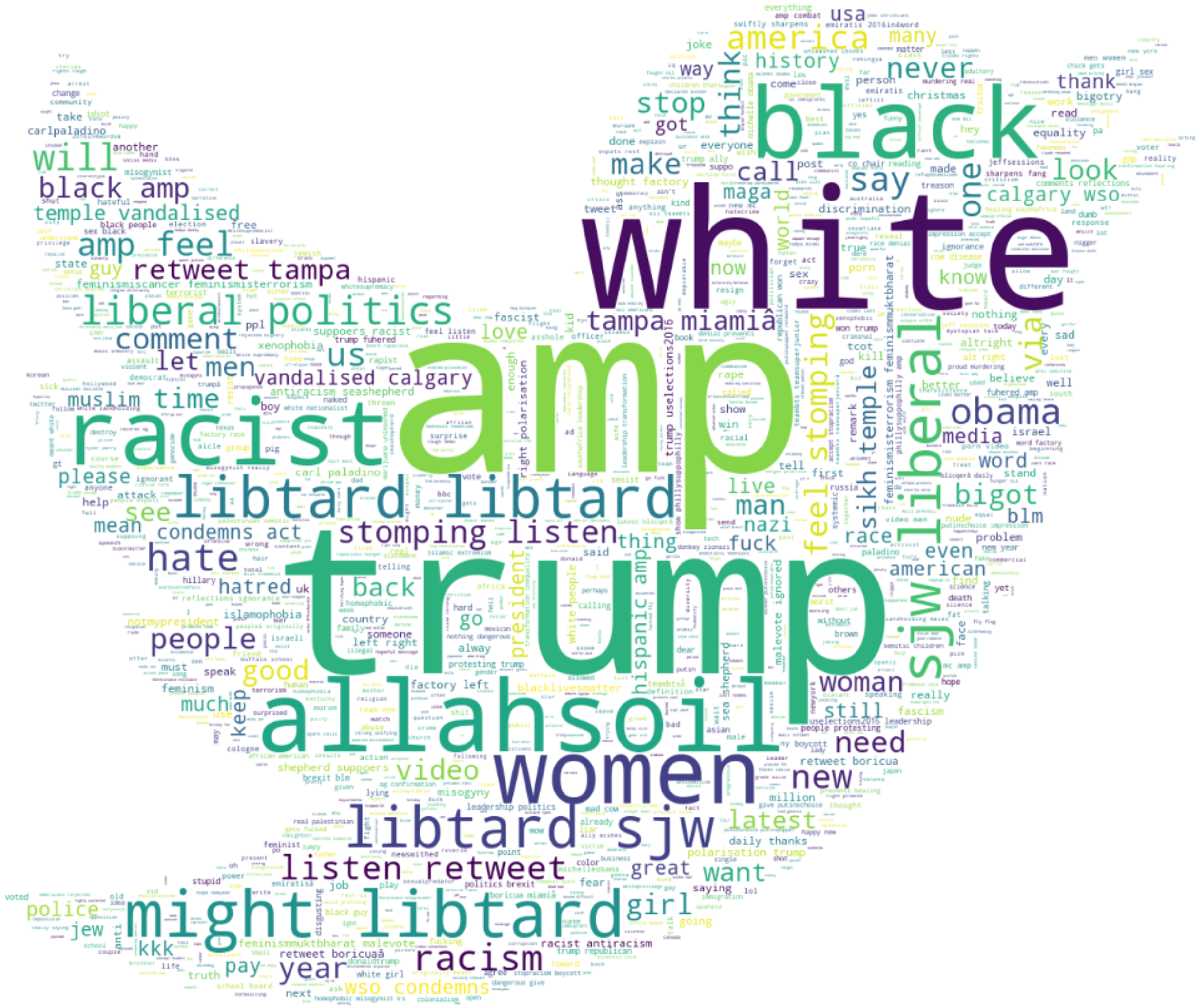
POSITIVE #HASHTAGS



NEGATIVE #HASHTAGS




love
amp
happy
day
today
new
life
smile
people
make
friend
bihday
thank
weekend
now
got
one
see
girl
thankful positive
summer
beautiful
time
tonight
best
sad
year
thankful
sunday
positive affirmation
i_am positive
father day
tomorrow
going
back
friday
model love
week
world
family
first
wait
orlando
selfie
great
good
go
look
still
live
come
even
feel
happiness
will
want
love take
fun
always
us
silver gold
gold forex
blog silver
take time
bull will
morning
show
time urð
thing
hate
kid
know
say
really
work
amazing
fathers day
dance
book
quote
everyone
believe
sunshine
america
fathersday
guy
fashion
meet
thursday
daughter
mom
wednesday
blessed
photo
music
children
lot
attack bull
cute
rip
already
polar bear
loved
home
read
stay
word
black
game
wish
lost
person
getting
big
couple
money
truth
proud
change
bear climb
picture
man
hope
mean
relax
full
pretty
holiday
season
sweet
long
real
bad
june
yay
success
sta
coffee
tired
dominate bull
business
enjoy
making
perfect
night
keep
anything
grateful
many
team
place
event
please
right
way
remember
never
suppo
lol
trump
right now
find useful
actually
everything
trying
play
cat
pay
hour
young
monday
thought
last
face
bing bong
school
made
next
food
euro2016
ready
month
happy father
waiting
looking
left
much
movie
help
fathers day
flower
need tear
watching
awesome
dream
climb racing
hard
crazy
shit
little
stop
said
baby
drink
house
another
true
watch
find
excited
mood
may
yet
looking forward
yesterday
good morning
direct whatever
feeling
moment
son will dominate
alwaystoheal healthy
saturday
twitter
reason
two
try
finally
beauty
follow
im
tell
london
video
funny
women
alone
let
miss
pic
gay
job
working
gone
tear city
nice
need adapt
away
hea
fan
nothing
simulator need
gt gt
better
happy_bihday
done
win
free
end
put
adapt environment need
gorilla simulator
country
every
love instagood
travel think
gonna
oh
dog
beach
boy
dad
wow
tuesday
whatever_want
next week
post
soon
will direct
ppl
something
follower
check
fitness
wedding
really
someone
pa
zoo


white
amp
trump
black
racist
allahsoil
women
libtard libtard
libtard sjw
might libtard
sjw liberal
liberal politics
america
will
black amp
temple vandalised
amp feel
retweet tampa
comment
tampa miamiâ
obama
hate
stomping listen
feel stomping
listen retweet
racism
people
good
video
new
need
woman
bigot
make
say
stop
think
history
never
look
calgary wso
call
maga
us
men
let
time
muslim
vandalised calgary
back
hatred
year
girl
great
want
latest
police
jew
kkk
pay
wso condemns
condemns act
sikh temple
media
word
american
man
nazi
race
fuck
live
see
mean
please
love
know
now
one
thank
way
many
usa
got
world
via
hispanic amp
president
much
keep
still
even
blm

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

- TF-IDF IS AN INFORMATION RETRIEVAL TECHNIQUE THAT WEIGHS A TERM'S FREQUENCY (TF) AND ITS INVERSE DOCUMENT FREQUENCY (IDF).
- EACH WORD OR TERM HAS ITS RESPECTIVE TF AND IDF SCORE. THE PRODUCT OF THE TF AND IDF SCORES OF A TERM IS CALLED THE TF*IDF WEIGHT OF THAT TERM.
- THE HIGHER THE TF*IDF SCORE (WEIGHT), THE RARER THE TERM AND VICE VERSA.

NAIVE BAYES CLASSIFIER

SUPERVISED LEARNING ALGORITHMS BASED ON APPLYING BAYES' THEOREM WITH THE "NAIVE" ASSUMPTION OF CONDITIONAL INDEPENDENCE BETWEEN EVERY PAIR OF FEATURES GIVEN THE VALUE OF THE CLASS VARIABLE.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

THE GENERAL IDEA OF NAIVE BAYES:

- REPRESENT A DOCUMENT X AS A SET OF (W, A FREQUENCY OF W) PAIRS.
- FOR EACH LABEL Y, BUILD A PROBABILISTIC MODEL P(X| Y = Y) OF DOCUMENTS IN CLASS Y.
- TO CLASSIFY, SELECT LABEL Y WHICH IS MOST LIKELY TO GENERATE X:

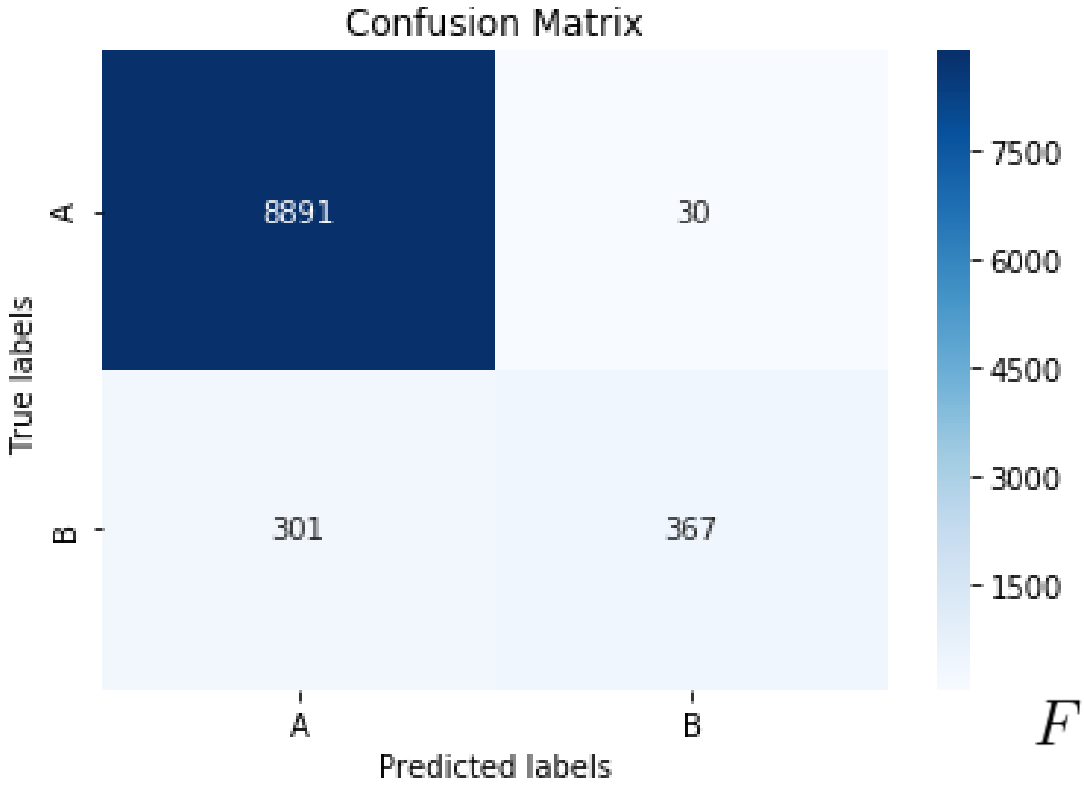
$$\hat{y} = \underset{y}{argmax} P(X|y) * P(y)$$

ASSUMPTIONS:

- THE ORDER OF THE WORDS IN DOCUMENT X MAKES NO DIFFERENCE BUT REPETITIONS OF WORDS DO.
- WORDS APPEAR INDEPENDENTLY OF EACH OTHER, GIVEN THE DOCUMENT CLASS.

OBJECTIVE:

CLASSIFY TWEET AS NEGATIVE OR POSITIVE



$$Accuracy = \frac{TP + TN}{Total} = 0.9654$$

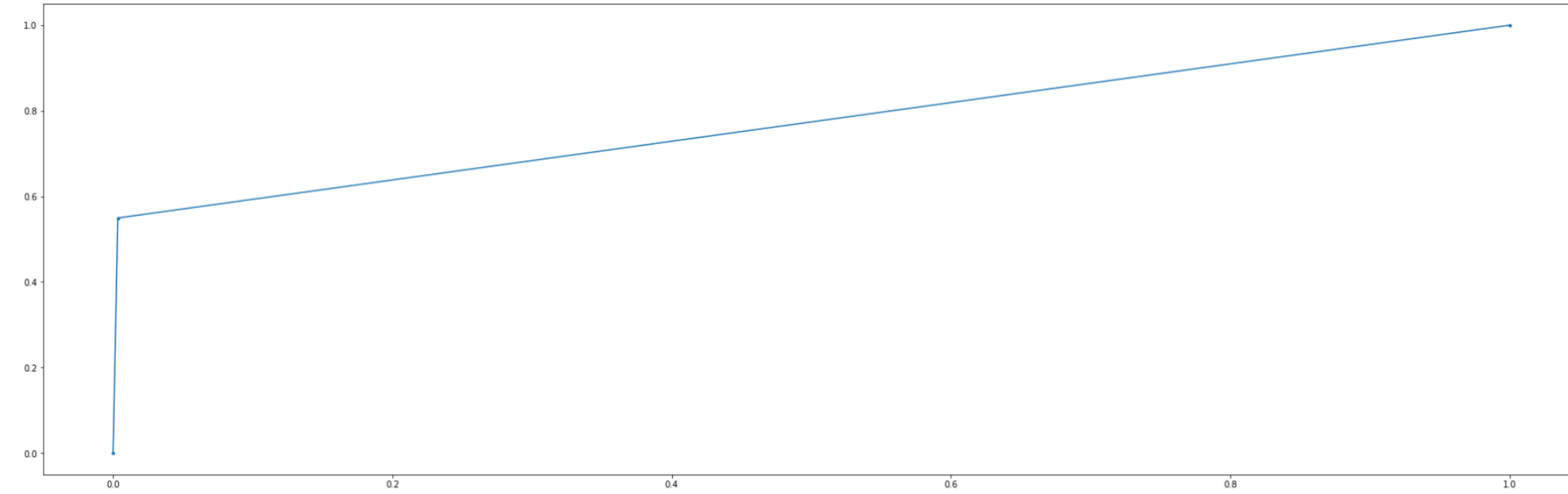
$$Precision = \frac{TP}{TP + FP} = 0.9244$$

$$Recall = \frac{TP}{TP + FN} = 0.5494$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} = 0.6892$$

ROC CURVE:

AREA UNDER THE CURVE=0.7730



CONCLUSIONS

NEVER STOP AT ACCURACY. PERMORMANCE IS STILL GOOD EVEN WHEN EVALUATED WITH OTHER MEASURES BUT GIVEN SOME NEGATIVE TWEETS NOT RECOGNISED, IT IS NOT THAT GOOD.

THE PROBLEM IS THE RECOGNITION OF NEGATIVE TWEETS. POSSIBLE DEVELOPEMENTS ARE:

- BUILDING A MORE BALANCED DATASET
- EXPLOITING MORE ADVANCED ALGORITHM FOR CLASSIFICATION

THANK YOU ALL FOR THE ATTENTION

SOURCES:

- ANALYTICS VIDHA, HTTPS://DATAHACK.ANALYTICSVIDHYA.COM/CONTEST/PRACTICE-PROBLEM-TWITTER-SENTIMENT-ANALYSIS, 2018
- GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE AND ROBERT TIBSHIRANI, AN INTRODUCTION TO STATISTICAL LEARNING, SPRINGER, LONDON, 2017.
- GITHUB, HTTPS://GITHUB.COM/PRATEEKJOSHI565/TWITTER_SENTIMENT_ANALYSIS/BLOB/MASTER/CODE_SENTIMENT_ANALYSIS.IPYNB, 2018
- KAGGLE, HTTPS://WWW.KAGGLE.COM/IMRANDUDE/BASIC-NLP-TUTORIAL-WITH-NAIVE-BAYES, 2018
- ONELY, HTTPS://WWW.ONELY.COM/BLOG/WHAT-IS-TF-IDF/, 2018
- SCIKIT-LEARN, DOCUMENTATION
- TOWARDS DATA SCIENCE, HTTPS://TOWARDSDATASCIENCE.COM/ALGORITHMS-FOR-TEXT-CLASSIFICATION-PART-1-NAIVE-BAYES-3FF1D116FDD8, 2019
- COURSE SLIDES