F21AA Applied Text Analytics: Coursework 2

Applying Sentiment Analysis on Twitter

Ismail Marashi

BSc (Hons) Computer Science

Heriot-Watt University
School of Mathematical and Computer Sciences
Department of Computer Science
F21AA: Applied Text Analytics
Instructor: Neamat Elgayar

March 21, 2021

1. **Introduction**

For this coursework our task is to collect twitter data related to ”covid-19 online classes”, explore the data and apply sentiment analysis and gain insights on students towards the covid situation.

2. **Data Collection**

First we had to gather some twitter data related to COVID-19 and online classes, we queried twitter the tweepy api to return any tweet that contained the keywords ”online classes” and ”Covid” in the tweet. The query used was:

#covid-19 OR #covid_19 OR #covid19 OR #covid OR covid OR covid19 OR covid_19 OR covid-19 AND online AND classes

We intentionally did not include ”corona” as it may also refer to a beverage and we did not limit our data collection to a geographical location as we could not retrieve a satisfactory number of tweets when we queried UAEs twitter alone. Applying this we collected 1516 unique tweets and were saved and loaded to a CSV to avoid losing our collected data and expanded upon automatically when we ran the code. To label the sentiment of our data we opted for using textblob as a baseline since sentiment classifier is a widely tested, reliable, and this approach lets us experiment with large data sets which is an advantage over hand labelling a limited data set. Before using TextBlob we did preprocess the text to avoid any unwanted artefacts, details of this is mentioned in the section below.

Inspection of TextBlobs Classifications:

Looking at TextBlobs classification we saw that it classified sentiments with 3 categories, neutral as 0, positive as anything greater than 0 to 1 and negative as any value less than 0 to -1 in a continuous manner. Since we plan to use this as classification targets we need to remap the continuous confidence values to categorical values for training later on.

3. Tweets Text Pipeline

Preprocessing:

Twitter text comes with a unique set of challenges since individuals do not also use proper english and they also contain features that are not completely part of the text for example Hashtags ”#Covid” or user tags ”@David1_24”, therefore quite a lot of preprocessing is required to minimize the feature counts and improve model accuracy. We applied the following preprocessing steps using regexs:

1. Lower-casing: All sentences were lower-cased to minimize so that machine learning algorithms can recognise words with different casings as the same thus minimising feature counts.
2. Links: All links were replaced with a space since they don’t contribute much to the sentiment and increase feature counts when left in.
3. Retweets Marker: if a tweet is a retweet it begins with a retweet marker with the following pattern ”RT @USERNAME:” these were also removed as they do not contain any useful information.
4. User-tag: Users can be tagged like so ”@USERNAME” user-tags are named entities and do not contain sentiment information therefore they were removed as well.
5. Hashtags: Hashtags (aka any word that starts with a ”#”) were also removed. They give useful topic information however they do not provide anything useful for sentiment.
6. Special-Chars: Special characters are useful in large text classification since it helps separate one sentence from another. However twitter only allows 280 characters(including spaces) making tweets fairly short in length which means that they aren’t particularly helpful in this case, therefore, it was better to remove them. In addition, in some cases unwanted special characters are inserted into the text fields, those are also removed.
7. Numbers: Numbers are not really helpful in sentiment classification so they are also removed.
8. Shortening Beginnings: Repeating characters in word beginnings are shorted to minimize duplicated features.(eg. tttall becomes tall)
9. Shortening Repeats: Repeating characters anywhere in the word are shortened to 2 repetitions just like before, (eg. taaaaallll becomes taaaall)
10. Stop Words: stop words are removed since they do contribute much to the sentiment
11. Lemmatizing: words are then lemmatized using the WordNet lemmatizer since it maintains word integrity while minimizing feature counts.

The result of these steps is a sentence without any unnecessary features which helps ensure better generalization.

Experiment Setup:
For our experiments we decided to vary and experiment with different models. The configuration of each of the models is as follows:

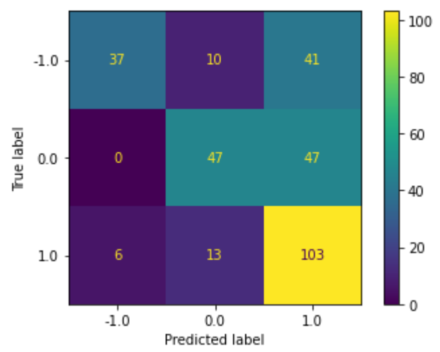
- N-Grams: Unigrams and bigrams
- Vector Representation:
  - TF-IDF: Logistic Regression, and Multinomail NB.
  - Word Embbedings: Bidirectional LSTM

Each of the models were trained on a training set of 1212 samples and evaluated on test set of 304 samples labelled by TextBlob and converted into categorical values for negative, neutral, and positive sentiment, the task for our classifiers was to learn the categorical sentiment values.

Test Set Results:

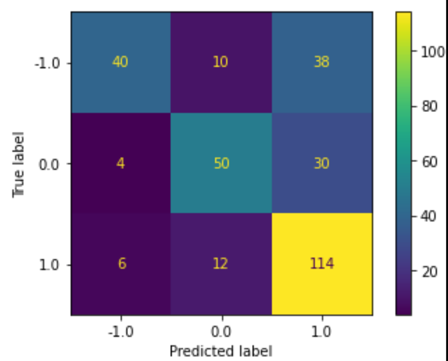
Logistic Regression

	Precision	Recall	F1-Score	Support
-1	0.86	0.42	0.56	88
0	0.67	0.50	0.57	84
1	0.54	0.84	0.66	132
Accuracy			0.62	304
Macro Avg	0.69	0.59	0.60	304
Weighted Avg	0.67	0.62	0.60	304



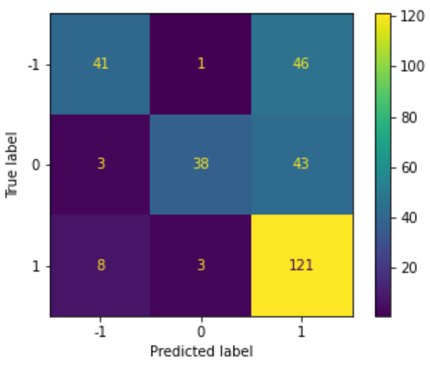
Multinomial Naive Bayes

	Precision	Recall	F1-Score	Support
-1	0.88	0.33	0.48	88
0	0.68	0.36	0.47	84
1	0.52	0.94	0.67	132
Accuracy			0.59	304
Macro Avg	0.69	0.54	0.54	304
Weighted Avg	0.67	0.59	0.55	304



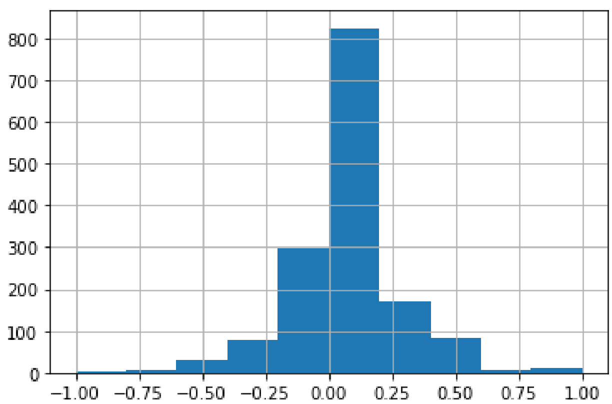
<u>**Bidirectional LSTM**</u>

	Precision	Recall	F1-Score	Support
-1	0.81	0.39	0.52	88
0	0.85	0.48	0.61	84
1	0.54	0.93	0.68	132
Accuracy			0.63	304
Macro Avg	0.73	0.60	0.61	304
Weighted Avg	0.71	0.63	0.61	304



Similar to the previous coursework with amazon reviews, these tweets are short therefore the performance of LSTMs, Logistic Regression, and MultinomailNB are comparable. However, the bidirectional LSTM performed the best while MultinomailNB performed the worst.

4. Insights:



Surprisingly there seems to be more positive tweets in the dataset when compared to the negative tweets further investigation using Latent Dirichlet Allocation topic analysis hints at topics relating to exams, assignments, returning to face to face lessons, and education. Also looking at logistic regressions coefficients it seems like words related to covid, exams, stress, and distance learning are positively correlated with negative sentiment and positive words are positive like new better and safe are positively correlated with positive sentiment and since we treated our classification task as 3 categories the 3rd is neutral and contains more informative terms.

Conclusion:

To summarize, we applied sentiment analysis on the initial data using TextBlob and converted its sentiment outputs to categorical values to train our own classifier. We visually inspected TextBlobs classification accuracy by looking at tweets with negative, positive and neutral classification to ensure its accuracy. We then split the data into a test and training set to use for our experiments. Our results show that the majority of people have a negative sentiment towards online classes and are uneasy about the upcoming exams. Applying this pipeline is very practical to find out what people think about a certain thing, for example companies could use a similar method to figure out what people think about their products or recent events.

Figures:

LDA:

topic 0	topic 1	topic 2	topic 3	topic 4
that	cases	with	year	have
have	as	please	ago	that
no	many	schools	one	they
it	be	as	was	so
with	ktrtrs	about	today	during
has	officials	shut	it	onlin
still	kindly	down	all	model
due	on	would	my	on
at	this	more	on	at
out	sabithaindratrsour		break	not

topic 5	topic 6	topic 7	topic 8	topic 9
school	but	not	that	varshaegaikwad
because	our	on	know	not
not	services	our	you	francis_joseph
colleges	ms	who	not	now
schools	please	it	our	student
take	new	that	being	many
what	an	could	re	move
about	covid19	still	why	after
cases	due	like	were	just
me	has	be	seen	was

topic 10	topic 11	topic 12	topic 13	topic 14
me	be	not	one	sir
my	school	were	with	day
was	my	exams	as	exam
when	not	it	due	by
had	last	us	year	please
school	new	its	all	give
it	then	year	will	not
have	you	this	that	open
year	on	enough	wouldn	because
time	does	our	shit	on

topic 15	topic 16	topic 17	topic 18	topic 19
they	our	year	return	it
were	so	after	medical	off
forced	it	more	foreign	at
during	pandemic	this	await	10
exams	college	than	varsities	covid19
although	mine	by	with	wisconsin
give	honestly	have	grapple	missing
pandemic	worst	now	poor	436
demanding	other	on	indiainkyrgyz	fined
be	stuff	not	quality	hour

topic 20	topic 21	topic 22	topic 23	topic 24
have	nimmasuresh	with	face	days
it	be	school	rajasthan	get

all	can	us	year	offline
my	conducted	learning	this	pte
education	exams	amp	time	cases
that	cbse	on	top	deepti_classes
will	final	please	class	overall
you	puc	year	also	study_abroad
with	attending	your	announced	british_council
be	almost	management	country	idp

topic 25	topic 26	topic 27	topic 28	topic 29
up	dhirajreddy47	will	kids	closed
with	drrpnishank	all	as	schools
my	ddnewslive	have	compulsory	11
all	narendramodi	not	no	covid19
class	mib_india	my	class	continue
before	mygovindia	even	have	march
never	do	on	please	orders
now	pib_india	cases	sir	10
year	eduminofindia	it	your	gt
that	or	our	but	further

topic 30	topic 31	topic 32	topic 33	topic 34
all	this	no	this	sir
on	face	with	me	delhi
have	return	it	with	want
please	as	when	school	due
at	week	be	pandemic	pmoindia
with	take	like	who	class
learning	that	keep	had	exam
toronto	learning	was	have	us
don	more	covid19	aur	please
it	on	more	positive	schools

topic 35	topic 36	topic 37	topic 38	topic 39
this	schools	have	on	exams
with	only	am	some	you
china	orders	at	teaching	drrpnishank
during	colleges	my	this	why
digital	barring	laptop	march	exam
international	examinations	can	spring	offline
take	function	no	break	sir
all	tuition	please	before	board
want	again	that	after	please
go	all	but	completely	us

topic 40	topic 41	topic 42	topic 43	topic 44
be	my	it	can	class
due	year	science	who	not
why	no	school	me	art
person	be	when	take	how
10	it	during	re	that
will	university	kids	through	have
these	have	with	covid19	should
not	our	your	all	education
don	how	postpone	going	world
college	from	cases	schools	propelled

topic 45	topic 46	topic 47	topic 48	topic 49
you	will	day	positive	sir
covid19	cases	cases	school	it
close	be	increasing	tested	this
have	year	as	please	from
on	they	so	staff	have
how	from	covid19	teachers	was
time	have	all	at	but
year	all	be	can	college
last	this	from	few	my
who	one	schools	hyderabad	now

topic 50	topic 51	topic 52	topic 53	topic 54
on	college	schools	my	no
exams	only	be	school	amp
offline	time	colleges	year	who
education	lack	go	tn	school
covid19	they	that	can	also
due	this	has	cm	because
rajasthan	at	due	request	all
have	pandemic	an	have	can
with	that	can	due	even
any	was	all	teachers	those

topic 55	topic 56	topic 57	topic 58	topic 59
amp	amp	my	an	have
cases	from	that	from	again
have	why	only	have	my
my	our	from	with	due
not	this	last	can	this
as	schools	all	even	physical
been	had	me	time	our
parents	would	on	my	at
submit	move	it	spring	be
assignments	se	since	continue	if

topic 60	topic 61	topic 62	topic 63	topic 64
have	shall	year	not	it
this	covid19	had	this	due
all	digital	post	due	that
they	mode	them	education	use
f2f	schools	first	it	against
or	get	due	fees	action
no	education	be	how	rising
but	continue	face	has	they
day	be	help	still	be
had	with	not	here	see

topic 65	topic 66	topic 67	topic 68	topic 69
wave	go	as	can	our
will	school	take	than	schools
that	my	dear	more	sir
due	delhi	schools	still	only
stop	time	cases	there	as
at	on	day	less	with
gt	video	request	our	today
level	with	sir	so	am

its	campus	has	but	have
go	back	kids	how	situation

topic 70	topic 71	topic 72	topic 73	topic 74
not	this	with	have	sir
fees	year	face	then	hi
it	pandemic	their	they	se
school	rather	children	this	koi
this	problems	will	do	her
as	faced	have	it	or
only	educational	school	during	university
pay	took	do	be	shafqatcancelcaies
covid19	many	you	safe	plz
go	due	can	all	there

topic 75	topic 76	topic 77	topic 78	topic 79
00	on	that	it	week
000	it	they	my	me
colleges	you	been	government	my
has	have	colleges	back	they
not	because	no	up	all
15	if	school	some	have
cases	they	should	with	campus
this	can	exams	schools	get
take	that	due	colleges	this
but	at	college	get	but

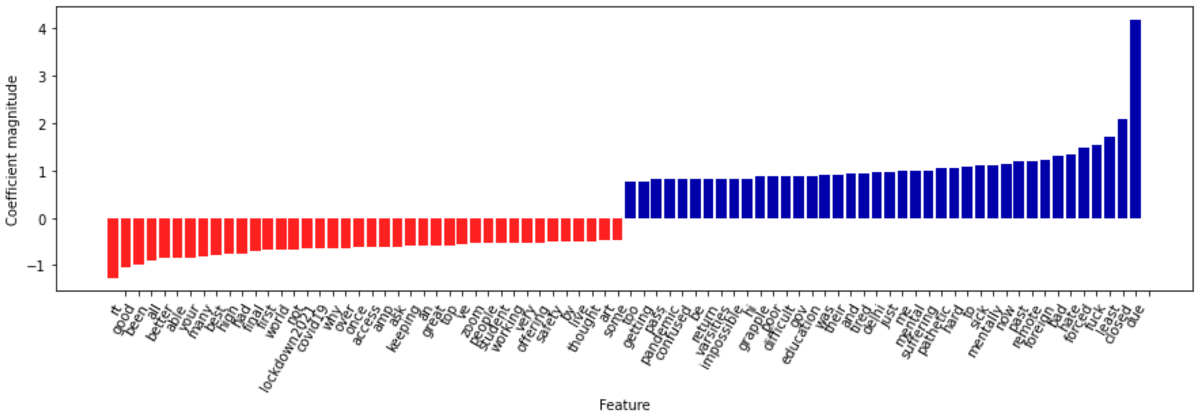
topic 80	topic 81	topic 82	topic 83	topic 84
uni	it	year	us	vs
has	syllabus	being	college	our
college	exams	less	making	school
covid19	don	cancelcaies2021be		plz
amp	with	was	our	schls
as	year	can	as	meetings
but	mental	as	this	day
was	nsut_endsem_onlineso		want	small
cases	march	all	cmofkarnataka	was
mein	an	that	let	have

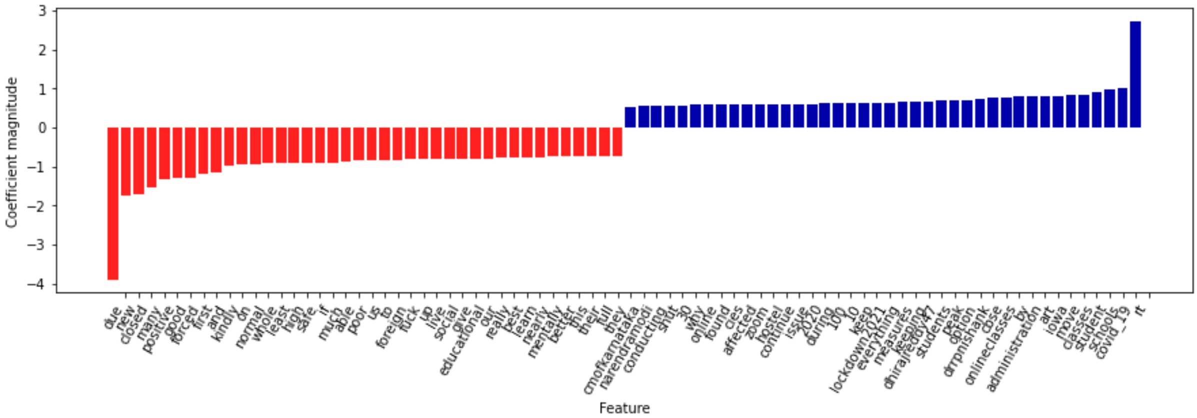
topic 85	topic 86	topic 87	topic 88	topic 89
they	be	now	with	education
have	that	has	covid19	had
mam	as	at	at	has
on	has	that	amp	school
school	edu	covid19	people	at
now	some	been	week	me
be	just	gt	this	world
but	schools	campus	that	my
how	will	kids	most	dscork
where	this	let	during	university

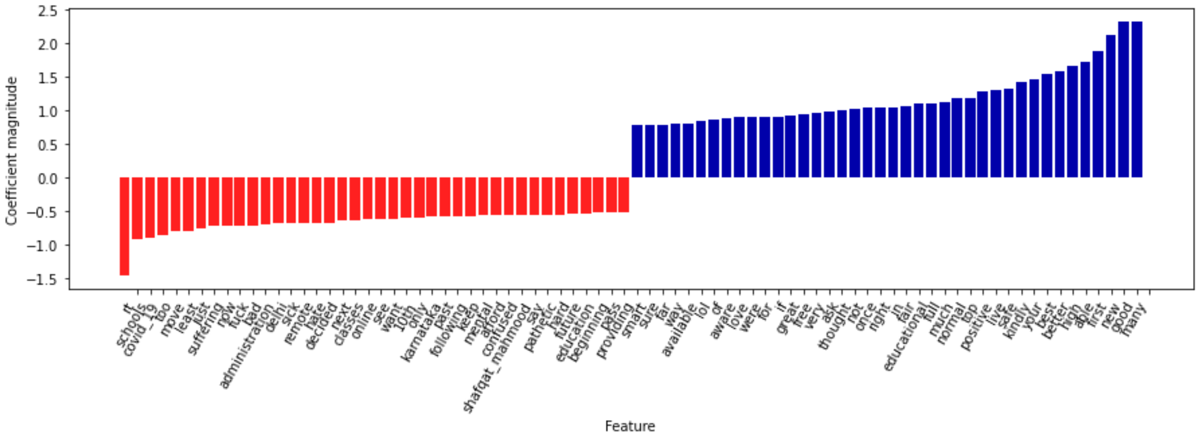
topic 90	topic 91	topic 92	topic 93	topic 94
with	after	at	you	you
more	our	be	our	that
due	they	school	have	can
that	what	risk	if	this

he	about	have	your	just
this	tourism	inadequate	but	have
pandemic	course	connectivity	us	when
at	person	resurgence	need	get
fee	us	network	do	offline
year	one	high	it	why

topic 95	topic 96	topic 97	topic 98	topic 99
--------	--------	--------	--------	--------
my	been	have	at	it
at	since	cases	cases	my
not	but	all	by	on
year	my	due	on	have
was	it	be	gse	do
now	have	semester	2021	amp
all	ve	year	will	time
but	with	may	school	some
school	at	will	week	us
have	all	if	six	as






Coefficient magnitude
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
rt
schools
covid_19
too
move
least
just
suffering
now
fuck
bad
administration
delhi
sick
remote
hate
decided
next
classes
online
see
want
10th
only
karnataka
past
following
keep
mental
afford
confused
shafqat_mahmood
say
pathetic
hard
future
education
beginning
pass
providing
smart
sure
far
way
available
lol
of
aware
love
were
for
if
great
free
very
ask
thought
not
once
right
in
fair
educational
full
much
normal
top
positive
live
safe
kindly
your
best
better
high
able
first
new
good
many
Feature