In [152…]:
```
import pandas as pd
import altair as alt
```

In [42]:
```
raw_data = pd.read_csv('clean_tweets.csv')
raw_data
```

Out[42]:

	data.id	data.time	data.lang	data.text	data.fav_count	data.possibly_sensitive	data.source	user.id	user.name	user.screen_name	user.location
0	1.600000e+18	Tue Nov 29 15:31:21 +0000 2022	en	I take 0 action without strategy	0	NaN	<a href="http://twitter.com/download/iphone" r...	2.904332e+09	Massoud	mxssoud	Ottawa, Ontario
1	1.600000e+18	Tue Nov 29 15:31:20 +0000 2022	en	@GuidoDisalle gm Guido!	0	NaN	<a href="https://mobile.twitter.com" rel="nofo...	1.370000e+18	aaronrferguson.eth (he/him) UA | Слава Україні!	aaronrferguson	Ottawa, ON Canada
2	1.600000e+18	Tue Nov 29 15:31:18 +0000 2022	en	@SleuthieGoosie Absolutely agree with you. Not...	0	NaN	<a href="http://twitter.com/download/iphone" r...	2.993299e+08	Middy T	middyt	Ottawa, Ontario, Canada
3	1.600000e+18	Tue Nov 29 15:31:16 +0000 2022	en	@ThisIsKyleR @elonmusk Yes, life must be soooo...	0	NaN	<a href="http://twitter.com/download/iphone" r...	2.798719e+07	GeeGee23	GeeGee23	Ottawa
4	1.600000e+18	Tue Nov 29 15:31:13 +0000 2022	en	@funstonpaleo So looking forward to seeing the...	0	False	<a href="http://twitter.com/download/android" ...	7.250000e+17	Michelle Campbell Mekarski PhD	MichelleCbell	Ottawa, Ontario
...	...	...	...	...	...	...	...	...	...	...	...
721	1.600000e+18	Fri Dec 02 21:09:13 +0000 2022	en	@DrSarteschi City of Ottawa Comms people might...	0	NaN	<a href="http://twitter.com/download/android" ...	1.154079e+08	shawnalucey	Shawnalucey	Ottawa
722	1.600000e+18	Fri Dec 02 21:09:12 +0000 2022	en	@EulanaLebedeva Got any recommendations? Ooh a...	0	NaN	<a href="http://twitter.com/download/android" ...	1.460000e+18	preacher	nithin_zac	Ottawa, Ontario
723	1.600000e+18	Fri Dec 02 21:09:11 +0000 2022	en	*squints at the transphobic trustee candidates...	0	False	<a href="http://twitter.com/download/android" ...	3.997289e+09	Z. Downey	Haligowan	Ottawa, Ontario
724	1.600000e+18	Fri Dec 02 21:09:09 +0000 2022	und	@AdoredTy @seQuenceNyong @Jodykeeling771	0	NaN	<a href="http://twitter.com/download/iphone" r...	1.540000e+18	Jody Keeling	Jodykeeling771	Ottawa, Ontario
725	1.600000e+18	Fri Dec 02 21:09:01 +0000 2022	en	@FBorgal This is considered sub-par in Finland...	0	False	<a href="https://mobile.twitter.com" rel="nofo...	5.012396e+08	Bicycle Seen 🚲 +	frpaul1	Ottawa, Ontario

726 rows × 11 columns

Dataframe metadata containing only :

- data.id
- data.time
- data.lang
- data.text
- data.possibly_sensitive
- user.location
- user.id

In [39]:
```
metadata = raw_data[['data.id', 'data.time', 'data.lang', 'data.text', 'data.possibly_sensitive','user.location', 'user.id' ]].copy()
metadata
```

Out[39]:

	data.id	data.time	data.lang	data.text	data.possibly_sensitive	user.location	user.id
0	1.600000e+18	Tue Nov 29 15:31:21 +0000 2022	en	I take 0 action without strategy	NaN	Ottawa, Ontario	2.904332e+09
1	1.600000e+18	Tue Nov 29 15:31:20 +0000 2022	en	@GuidoDisalle gm Guido!	NaN	Ottawa, ON Canada	1.370000e+18
2	1.600000e+18	Tue Nov 29 15:31:18 +0000 2022	en	@SleuthieGoosie Absolutely agree with you. Not...	NaN	Ottawa, Ontario, Canada	2.993299e+08
3	1.600000e+18	Tue Nov 29 15:31:16 +0000 2022	en	@ThisIsKyleR @elonmusk Yes, life must be soooo...	NaN	Ottawa	2.798719e+07
4	1.600000e+18	Tue Nov 29 15:31:13 +0000 2022	en	@funstonpaleo So looking forward to seeing the...	False	Ottawa, Ontario	7.250000e+17
...	...	...	...	...	...	...	...
721	1.600000e+18	Fri Dec 02 21:09:13 +0000 2022	en	@DrSarteschi City of Ottawa Comms people might...	NaN	Ottawa	1.154079e+08
722	1.600000e+18	Fri Dec 02 21:09:12 +0000 2022	en	@EulanaLebedeva Got any recommendations? Ooh a...	NaN	Ottawa, Ontario	1.460000e+18
723	1.600000e+18	Fri Dec 02 21:09:11 +0000 2022	en	*squints at the transphobic trustee candidates...	False	Ottawa, Ontario	3.997289e+09
724	1.600000e+18	Fri Dec 02 21:09:09 +0000 2022	und	@AdoredTy @seQuenceNyong @Jodykeeling771	NaN	Ottawa, Ontario	1.540000e+18
725	1.600000e+18	Fri Dec 02 21:09:01 +0000 2022	en	@FBorgal This is considered sub-par in Finland...	False	Ottawa, Ontario	5.012396e+08

726 rows × 7 columns

Dataframe tweets only containing:

- data.text
- data.lang

In [45]:
```
tweets = raw_data[['data.text', 'data.lang']].copy()
tweets.rename(columns={'data.text': 'text', 'data.lang': 'language'}, inplace=True)
tweets
```

Out[45]:

	text	language
0	I take 0 action without strategy	en
1	@GuidoDisalle gm Guido!	en
2	@SleuthieGoosie Absolutely agree with you. Not...	en
3	@ThisIsKyleR @elonmusk Yes, life must be soooo...	en
4	@funstonpaleo So looking forward to seeing the...	en
...	...	...
721	@DrSarteschi City of Ottawa Comms people might...	en
722	@EulanaLebedeva Got any recommendations? Ooh a...	en
723	*squints at the transphobic trustee candidates...	en
724	@AdoredTy @seQuenceNyong @Jodykeeling771	und
725	@FBorgal This is considered sub-par in Finland...	en

726 rows × 2 columns

In [ ]:

Exploring data using the tweets dataframe

List of the different languages

In [70]:
```
len(tweets['language'].unique())
```

Out[70]: 22

Distribution of languages

In [50]:
```
eng = tweets[tweets['language']== 'en']
fr = tweets[tweets['language']== 'fr']
other = tweets[(tweets['language']!= 'en')& (tweets['language']!= 'fr')]
```

Ratio of english tweets

In [57]:
```
print( f' {(len(eng)/len(tweets))*100:.2f} %')
```

79.89 %

Ratio of french tweets

In [58]:
```
print( f' {(len(fr)/len(tweets))*100:.2f} %')
```

6.89 %

Ratio of tweets in other languages

In [59]:
```
print( f' {(len(other)/len(tweets))*100:.2f} %')
```

13.22 %

In [107…
```
toChart = tweets.copy()

class Lang(dict):
    def __missing__(self, key):
        return "other"

lang = { 'en':'en', 'fr':'fr'}

toChart['language'] = toChart['language'].map(Lang(lang))
toChart
```

Out[107]:

	text	language
0	I take 0 action without strategy	en
1	@GuidoDisalle gm Guido!	en
2	@SleuthieGoosie Absolutely agree with you. Not...	en
3	@ThisIsKyleR @elonmusk Yes, life must be soooo...	en
4	@funstonpaleo So looking forward to seeing the...	en
...	...	...
721	@DrSarteschi City of Ottawa Comms people might...	en
722	@EulanaLebedeva Got any recommendations? Ooh a...	en
723	*squints at the transphobic trustee candidates...	en
724	@AdoredTy @seQuenceNyong @Jodykeeling771	other
725	@FBorgal This is considered sub-par in Finland...	en

726 rows × 2 columns

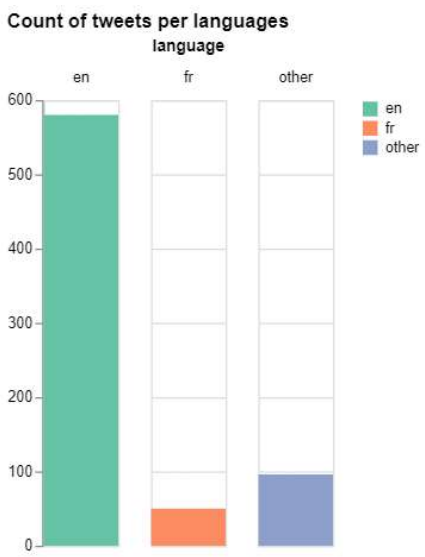
Visualization

In [158…
```
mixed_lang=alt.Chart(toChart).encode(
    y= alt.Y("count(language):N", title = ""),
    color = alt.Color('language', title = "", scale=alt.Scale(scheme='set2'))

).mark_bar().properties(
    width=50,
    height=300
).facet(
'language:N',
    columns = 3,
    title="Count of tweets per languages"
)
```

In [145…
```
mixed_lang
```

Out[145]:



Since there are so many tweets that are not in english nor french,

Here is a dataframe containing only english and french tweets

Dataframe en_fr_tweets containing only english and french tweets

In [63]:
```
en_fr_tweets = tweets[(tweets['language']== 'en') | (tweets['language']== 'fr') ].copy()
en_fr_tweets
```

Out[63]:

	text	language
0	I take 0 action without strategy	en
1	@GuidoDisalle gm Guido!	en
2	@SleuthieGoosie Absolutely agree with you. Not...	en
3	@ThisIsKyleR @elonmusk Yes, life must be soooo...	en
4	@funstonpaleo So looking forward to seeing the...	en
...	...	...
720	@roulinski Maybe mass report as something that...	en
721	@DrSarteschi City of Ottawa Comms people might...	en
722	@EulanaLebedeva Got any recommendations? Ooh a...	en
723	*squints at the transphobic trustee candidates...	en
725	@FBorgal This is considered sub-par in Finland...	en

630 rows × 2 columns

Distribution of languages

Ratio of english tweets

In [66]:
```
print( f' {(len(eng)/len(en_fr_tweets))*100:.2f} %')
```

```
 92.06 %
```

Ratio of french tweets

In [64]:
```
print( f' {(len(fr)/len(en_fr_tweets))*100:.2f} %')
```

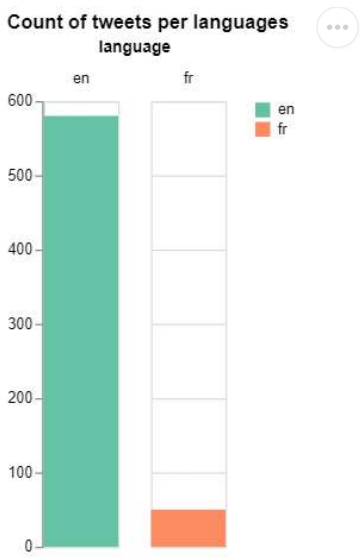
```
 7.94 %
```

Visualization

In [146…
```
en_fr = alt.Chart(en_fr_tweets).encode(
    y= alt.Y("count(language):N", title = ""),
    color = alt.Color('language', title = "", scale=alt.Scale(scheme='set2'))
).mark_bar().properties(
    width=50,
    height=300
).facet(
'language:N',
    columns = 2,
    title="Count of tweets per languages"
)
```

In [147…
```
en_fr
```

Out[147]:



Sentiment analysis

Language	Accuracy (exact)	Accuracy (off-by-1)
English	67%	95%
Dutch	57%	93%
German	61%	94%
French	59%	94%
Italian	59%	95%
Spanish	58%	95%

Since the accuracy of the model is better when considered off-by-1, I will redefine the categories as such:

- 1 (Negative) : group 1-star and 2-stars
- 2 (Neutral) : 3-stars
- 3 (Positive) : group 4-stars and 5-stars

In [159…
```
categories = {"1 star": 1,
         "2 stars": 1,
         "3 stars": 2,
         "4 stars": 3,
         "5 stars": 3}
```

Preparing the model and the data for analysis

In [160…
```
from transformers import pipeline

classifier = pipeline("text-classification", model = "nlptown/bert-base-multilingual-uncased-sentiment")
```

```
C:\Users\Noura\anaconda3\envs\SDSProject\lib\site-packages\tqdm\auto.py:22: TqdmWarning: IProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

In [161…
```
text = tweets['text'].head(50).tolist()
text
```

Out[161]:
```
['I take 0 action without strategy',
 '@GuidoDisalle gm Guido!',
 '@SleuthieGoosie Absolutely agree with you. Not normal.',
 '@ThisIsKyleR @elonmusk Yes, life must be soooooo hard for you. Being able to murder whoever you want isn’t enough,… https://t.co/Mzp3Li38s0',
 '@funstonpaleo So looking forward to seeing the results that come out of this!! https://t.co/o9Esvh5PXw',
 '@HugoFaz gm Hugo!',
 "💥Donner en GRAND à votre communauté aujourd'hui💥\nNous lançons 🚤Bateaux pour Tous🚤 - une campagne pour s'assurer que… https://t.co/gkZVQKP99X",
 '💥Give back to your community in a BIG way today!💥\n\nWe’re launching 🚤Boats for All🚤 – a campaign to ensure our races… https://t.co/Sdr7UnX1fC',
 'y’all wouldn’t get it https://t.co/9rzxQRwdO6',
 '@WashedSummon OH MY GOODNESS WHAT A BUILD IM ABOUT TO BE FRICKING T’D',
 '@_BeFootball Il va rien faire du tout',
 '“The problem is that #BC isn’t planning to generate enough electricity to meet its #2030ClimateTargets,” before eve… https://t.co/Ri7S1j7xfW',
 'The Ottawa Police Services Board abruptly ended its meeting Monday as public delegates held a protest demanding act… https://t.co/WtF4D8qxiD',
 '@alexandrepratt Ton per diem ne te fournira pas 2000 calories par jour 😉',
 "@GraemeNichols I've been of the opinion that the team must know at least something but trading an asset under these… https://t.co/fB2PuvLi2J",
 '@_benui thanks, all done in a custom game engine',
 "@118Thumper118 @ElseSlayer What's that got to do with the price of butter? 😂\n\nBut seriously, yes, yes and yes, except for Scrappy Doo!",
 '@NatashaZimmers As a Canadian student and teacher I know what it is. My daughter denies having heard it called that… https://t.co/oUxMmYpv3t',
 'Pain is starting to subside. https://t.co/B85iuW36lQ',
 '@JenStewartOtt  will you help me with a RT? \n\nhttps://t.co/rpv5ikYyWa https://t.co/iy7e0tnm7d',
 '@TheJFreakinC Traumatic head injuries are mandatory to be elected to the Republican Party.',
 'Seeds Canada Members are about to hear all about what we have been busy working on from our senior management team… https://t.co/dAy15557wS',
 'Being part of a community that shares your views on important ... More for Aries https://t.co/bhzrBI9zZh',
 "@DPatel_PharmD And yet his common sense trumps your opponents so-called compassion, why don't you argue the issues… https://t.co/xhg1Bgw4Jd",
 "CORRECTION! Watson est là. Je ne l'avais pas vu.",
 'Pratique des Sénateurs ce matin, Anton Forsberg est présent.',
 'New collaboration between CAGBC &amp;amp; @UofT to research solutions for embodied carbon reductions from building material… https://t.co/LkMROdkNiv',
 '@jbru11 @MikeDrucker @GeraldLeroy6 He left after the fall of apartheid. So yes, to avoid military service, but not… https://t.co/9UwFgNirNF',
 'Insurance can get complicated. Our latest blog post explains the basics. If you have questions, we can help.… https://t.co/0kyHcdT1Wj',
 'Vote for the #LionsCelly for #BudLightCelly of the Week!',
 'Balancing recent family and career commitments could be a chal... More for Virgo https://t.co/GCmNxAND1W',
 '@aureliusraines2 May tomorrow be so boring it almost brings you to tears',
 '@PaulChampLaw You know I agree with this. They could probably hold an entirely separate inquiry on this point thoug… https://t.co/OdhJUT8hAr',
 'Worth reading. https://t.co/AkKkunbXfr via @nationalpost',
 '“Hi, my name is ***David Frey***. I don’t usually endorse other people’s products, but in this case I’m making an excepti… https://t.co/r2Ggqidh2v',
 'The SSI journey begins.\n#owasp #ottawa @BsidesOttawa #bsidesottawa #identity #SelfSovereignIdentity https://t.co/Je2MYOWQvb',
 '@lordsweetpotato @StephenPunwasi Lot braver than me looking that far out. Climate is playing a bigger role every ye… https://t.co/FLmZt2ComQ',
 '@shortstack_dan @jamiezjohnson @QuantumNFT gm Dan! Best wishes to your sick computer &amp;lt;3',
 "@tersthebear So so happy this post came to you at the right time! Hope you're feeling a bit better after reading it",
 '@RoughChopOttawa Gotta love the memorial graphic that Sgt Maria Keen (@sgtonpatrol) shared recently. In addition to… https://t.co/5EARj3J4kB',
 'Do you want to win a NYC Holiday Getaway to see @mariahcarey at Madison Square Garden and $10,000!? \n\nTune in to MO… https://t.co/0Mj9JRhG5V',
 '@Partyof3blog You too, my friend. 💜',
 '@Sethalos @TorontoStar About what ?',
 '@smwgilbert will you help me with a RT? \n\nhttps://t.co/rpv5ikYyWa https://t.co/iy7e0tnm7d',
 'The #K4DM22 Marketplace continues this week with policy-relevant research updates on #ClimateChange, gender equalit… https://t.co/rsvWTvZ2SW',
 'Le marché #K4DM22 continue avec des mises à jour de recherches pertinentes pour les politiques sur les changements… https://t.co/KghRg2Syzd',
 '@StephenRWade3 @TrueNorthCentre They fought each other.  That convoy should have been way more effective but too ma… https://t.co/ScB9LtDfgD',
 'PODCAST: Wake Up with Rob Snow - November 29, 2022 https://t.co/1sRZR2hSBn',
 '@maritstiles If your budget is dependent on one time development fees to provide service, then your tax rate is too… https://t.co/7abBZj56he',
 "You're probably eager to share a good time with friends today.... More for Aquarius https://t.co/xN5uNaZsre"]
```

In [163…
```
raw_list = classifier(text)
```

Raw results

In [164…
```
raw_result = pd.DataFrame.from_dict(raw_list)
raw_result
```

Out[164]:

	label	score
0	1 star	0.924041
1	5 stars	0.498549
2	1 star	0.283423
3	1 star	0.315153
4	5 stars	0.428461
5	5 stars	0.478129
6	5 stars	0.445332
7	5 stars	0.677167
8	1 star	0.607453
9	1 star	0.256044
10	5 stars	0.354743
11	1 star	0.324010
12	1 star	0.613772
13	1 star	0.480750
14	4 stars	0.272111
15	5 stars	0.569570
16	3 stars	0.264587
17	4 stars	0.288386
18	2 stars	0.359314
19	1 star	0.242647
20	1 star	0.459253
21	5 stars	0.364008
22	4 stars	0.366059
23	5 stars	0.403067
24	3 stars	0.328802
25	4 stars	0.474071
26	5 stars	0.450861
27	1 star	0.518113
28	4 stars	0.368586
29	5 stars	0.520244
30	3 stars	0.367253
31	1 star	0.583530
32	3 stars	0.287752
33	4 stars	0.462365
34	3 stars	0.298967
35	1 star	0.225022
36	5 stars	0.424655
37	5 stars	0.409506
38	5 stars	0.705695
39	5 stars	0.555300
40	1 star	0.414112
41	5 stars	0.643906
42	1 star	0.399027
43	1 star	0.253801
44	4 stars	0.442195
45	4 stars	0.441337
46	3 stars	0.314041
47	1 star	0.283709
48	1 star	0.330353
49	3 stars	0.363568

In [169…

```
label = {"1 star": 1,
         "2 stars": 2,
         "3 stars": 3,
         "4 stars": 4,
         "5 stars": 5}

raw_result["label"] = raw_result["label"].map(label)

raw_result
```

Out[169]:

	label	score
0	1	0.924041
1	5	0.498549
2	1	0.283423
3	1	0.315153
4	5	0.428461
5	5	0.478129
6	5	0.445332
7	5	0.677167
8	1	0.607453
9	1	0.256044
10	5	0.354743
11	1	0.324010
12	1	0.613772
13	1	0.480750
14	4	0.272111
15	5	0.569570
16	3	0.264587
17	4	0.288386
18	2	0.359314
19	1	0.242647
20	1	0.459253
21	5	0.364008
22	4	0.366059
23	5	0.403067
24	3	0.328802
25	4	0.474071
26	5	0.450861
27	1	0.518113
28	4	0.368586
29	5	0.520244
30	3	0.367253
31	1	0.583530
32	3	0.287752
33	4	0.462365
34	3	0.298967
35	1	0.225022
36	5	0.424655
37	5	0.409506
38	5	0.705695
39	5	0.555300
40	1	0.414112
41	5	0.643906
42	1	0.399027
43	1	0.253801
44	4	0.442195
45	4	0.441337
46	3	0.314041
47	1	0.283709
48	1	0.330353
49	3	0.363568

Sentiment mean

In [170…
```
raw_result['label'].mean()
```

Out[170]:
```
3.06
```

Results with the new category

In [166…
```
result = raw_result.copy()
result["label"] = result["label"].map(categories)

result
```

Out[166]:

	label	score
0	1	0.924041
1	3	0.498549
2	1	0.283423
3	1	0.315153
4	3	0.428461
5	3	0.478129
6	3	0.445332
7	3	0.677167
8	1	0.607453
9	1	0.256044
10	3	0.354743
11	1	0.324010
12	1	0.613772
13	1	0.480750
14	3	0.272111
15	3	0.569570
16	2	0.264587
17	3	0.288386
18	1	0.359314
19	1	0.242647
20	1	0.459253
21	3	0.364008
22	3	0.366059
23	3	0.403067
24	2	0.328802
25	3	0.474071
26	3	0.450861
27	1	0.518113
28	3	0.368586
29	3	0.520244
30	2	0.367253
31	1	0.583530
32	2	0.287752
33	3	0.462365
34	2	0.298967
35	1	0.225022
36	3	0.424655
37	3	0.409506
38	3	0.705695
39	3	0.555300
40	1	0.414112
41	3	0.643906
42	1	0.399027
43	1	0.253801
44	3	0.442195
45	3	0.441337
46	2	0.314041
47	1	0.283709
48	1	0.330353
49	2	0.363568

Sentiment mean

In [167…
```
result['label'].mean()
```

Out[167]:
```
2.1
```

Note:

The general tone of the tweets is neutral in both cases ({1,2,3,4,5} scale and {1,2,3} scale)

Note:

A better approach for converting the label from the 1-5 scale to the 1-3 scale would be:

- 1, 2 are considered negative -> 1
- 3 is considered neutral -> 2
- 4,5 are considered negative -> 3

If label = 1: -> negative (1)

If label = 5 : -> positive (3)

```
If label in {2,3,4}:
    left = label-1
    right = label+1
    left_prob = prob (left)+ prob(label)
    right_prob = prob (right)+ prob(label)
```

```
if left_prob > right_prob:
    use left to convert label
else use right
```