Lab3 - Assignment Sentiment

This notebook describes the LAB-2 assignment of the Text Mining course. It is about sentiment analysis.

The aims of the assignment are:

- Learn how to run a rule-based sentiment analysis module (VADER)
- Learn how to run a machine learning sentiment analysis module (Scikit-Learn/ Naive Bayes)
- Learn how to run scikit-learn metrics for the quantitative evaluation
- Learn how to perform and interpret a quantitative evaluation of the outcomes of the tools (in terms of Precision, Recall, and F_1)
- Learn how to evaluate the results qualitatively (by examining the data)
- Get insight into differences between the two applied methods
- Get insight into the effects of using linguistic preprocessing
- Be able to describe differences between the two methods in terms of their results
- Get insight into issues when applying these methods across different domains

In this assignment, you are going to create your own gold standard set from 50 tweets. You will the VADER and scikit-learn classifiers to these tweets and evaluate the results by using evaluation metrics and inspecting the data.

We recommend you go through the notebooks in the following order:

- **Read the assignment (see below)**
- **Lab3.2-Sentiment-analysis-with-VADER.ipynb**
- **Lab3.3-Sentiment-analysis.with-scikit-learn.ipynb**
- **Answer the questions of the assignment (see below) using the provided notebooks and submit**

In this assignment you are asked to perform both quantitative evaluations and error analyses:

- a quantitative evaluation concerns the scores (Precision, Recall, and F_1) provided by scikit's classification_report. It includes the scores per category, as well as micro and macro averages. Discuss whether the scores are balanced or not between the different categories (positive, negative, neutral) and between precision and recall. Discuss the shortcomings (if any) of the classifier based on these scores
- an error analysis regarding the misclassifications of the classifier. It involves going through the texts and trying to understand what has gone wrong. It servers to get insight in what could be done to improve the performance of the

classifier. Do you observe patterns in misclassifications? Discuss why these errors are made and propose ways to solve them.

Credits

The notebooks in this block have been originally created by Marten Postma and Isa Maks. Adaptations were made by Filip Ilievski.

Part I: VADER assignments

Preparation (nothing to submit):

To be able to answer the VADER questions you need to know how the tool works.

- Read more about the VADER tool in this blog.
- VADER provides 4 scores (positive, negative, neutral, compound). Be sure to understand what they mean and how they are calculated.
- VADER uses rules to handle linguistic phenomena such as negation and intensification. Be sure to understand which rules are used, how they work, and why they are important.
- VADER makes use of a sentiment lexicon. Have a look at the lexicon. Be sure to understand which information can be found there (lemma?, wordform?, part-of-speech?, polarity value?, word meaning?) What do all scores mean? https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicc

[3.5 points] Question1:

Regard the following sentences and their output as given by VADER. Regard sentences 1 to 7, and explain the outcome **for each sentence**. Take into account both the rules applied by VADER and the lexicon that is used. You will find that some of the results are reasonable, but others are not. Explain what is going wrong or not when correct and incorrect results are produced.

```
INPUT SENTENCE 1 I love apples
VADER OUTPUT {'neg': 0.0, 'neu': 0.192, 'pos': 0.808,
'compound': 0.6369}

INPUT SENTENCE 2 I don't love apples
VADER OUTPUT {'neg': 0.627, 'neu': 0.373, 'pos': 0.0,
'compound': -0.5216}

INPUT SENTENCE 3 I love apples :-)
VADER OUTPUT {'neg': 0.0, 'neu': 0.133, 'pos': 0.867,
'compound': 0.7579}

INPUT SENTENCE 4 These houses are ruins
VADER OUTPUT {'neg': 0.492, 'neu': 0.508, 'pos': 0.0,
'compound': -0.4404}
```

```
INPUT SENTENCE 5 These houses are certainly not considered
ruins
VADER OUTPUT {'neg': 0.0, 'neu': 0.51, 'pos': 0.49,
'compound': 0.5867}

INPUT SENTENCE 6 He lies in the chair in the garden
VADER OUTPUT {'neg': 0.286, 'neu': 0.714, 'pos': 0.0,
'compound': -0.4215}

INPUT SENTENCE 7 This house is like any house
VADER OUTPUT {'neg': 0.0, 'neu': 0.667, 'pos': 0.333,
'compound': 0.3612}
```

Question 1 ANSWER:

• Sentence 1: The sentence "I love apples", is mostly classified as positive (0.808), this is to be expected and due to the positive sentiment rating of the word love in the lexicon used by VADER. • Sentence 2: The sentence "I don't love apples" is classified as negative. This is reasonable since it's the negation of the previous sentence (the same sentence but including "don't") • Sentence 3: The sentence "I love apples :-)" is classified as even more positive than the first sentence. This is due to the smiley emoticon, which is also included in the lexicon with positive sentiment rating. • Sentence 4: "These houses are ruins" is classified between neutral and negative. This is also reasonable. Ruins has a negative sentiment rating but not as low as other words. According to this context, it could however make sense if the sentence was classified as more negative than it did. • Sentence 5: "These houses are certainly not considered ruins" has a similar value for the neutral and positive sentiment ratings (around .5), this is also reasonable since it's the negation of the previous sentence, that was classified between neutral and negative. • Sentence 6: "He lies in the chair in the garden". This sentence is classified as neutral, however it's also partially negative, which is not fitting but it's probably due to the word "lies" having several meanings, some of which are negative. • Sentence 7: "This house is like any house". This sentence is mostly neutral which makes sense. Again, here there is a word with several meanings ("like"), which is skewing the results making them more positive than they otherwise would be.

[Points: 2.5] Exercise 2: Collecting 50 tweets for evaluation

Collect 50 tweets. Try to find tweets that are interesting for sentiment analysis, e.g., very positive, neutral, and negative tweets. These could be your own tweets (typed in) or collected from the Twitter stream.

We will store the tweets in the file **my_tweets.json** (use a text editor to edit). For each tweet, you should insert:

- sentiment analysis label: negative | neutral | positive (this you determine yourself, this is not done by a computer)
- the text of the tweet

- the Tweet-URL

from:

```
        "1": {
            "sentiment_label": "",
            "text_of_tweet": "",
            "tweet_url": "",
```

to:

```
    "1": {
            "sentiment_label": "positive",
            "text_of_tweet": "All across America people chose
    to get involved, get engaged and stand up. Each of us can
    make a difference, and all of us ought to try. So go keep
    changing the world in 2018.",
            "tweet_url" :
    "https://twitter.com/BarackObama/status/946775615893655552",
        },
```

You can load your tweets with human annotation in the following way.

In [1]:
```
import json
```

In [2]:
```
my_tweets = json.load(open('my_tweets.json'))
```

In [3]:
```
for id_, tweet_info in my_tweets.items():
    print(id_, tweet_info)
```

```
1 {'sentiment_label': 'positive', 'text_of_tweet': '“cHiNa cAn’T iNnOvAt
E.” 💥Analysis by ASPI* shows that China leads the USA in whopping 37 ou
t of 44 critical scientific areas such as AI, quantum computing, biotec
h, and advanced materials.<br><br>*funded by U.S. military industrial co
mplex, so no pro-China bias <a href="https://t.co/CgNUmGA0iE"> pic.twitt
er.com/CgNUmGA0iE', 'tweet_url': 'https://twitter.com/Kanthan2030/statu
s/1631622840989675520?ref_src=twsrc%5Etfw'}
2 {'sentiment_label': 'negative', 'text_of_tweet': 'AMERICAN WAR MACHINE
NOW FOCUSED ON CHINA<br> <a href="https://t.co/5zUMGxoXNQ">pic.twitter.c
om/5zUMGxoXNQ</a></p>&mdash; The_Real_Fly (@The_Real_Fly)', 'tweet_url':
'https://twitter.com/The_Real_Fly/status/1631542150675529729?ref_src=tws
rc%5Etfw'}
3 {'sentiment_label': 'negative', 'text_of_tweet': 'China appears to be
requiring foreign law professors to submit their syllabuses to ensure th
ey are following a doctrine pushed by President Xi Jinping <a href="http
s://t.co/SuSWhELiCx">https://t.co/SuSWhELiCx</a></p>&mdash; Bloomberg (@
business)', 'tweet_url': 'https://twitter.com/business/status/1631576391
954169857?ref_src=twsrc%5Etfw'}
4 {'sentiment_label': 'negative', 'text_of_tweet': 'The United States ha
s added two subsidiaries of Chinese genetics company BGI to a trade blac
klist over allegations it conducted genetic analysis and surveillance ac
tivities for Beijing, which Washington says was used to repress ethnic m
inorities in China <a href="https://t.co/siXR57whNs">https://t.co/siXR57
whNs</a></p>&mdash; CNN (@CNN)', 'tweet_url': 'https://twitter.com/CNN/s
tatus/1631622994924544001?ref_src=twsrc%5Etfw'}
5 {'sentiment_label': 'positive', 'text_of_tweet': 'China has a prevalen
t weapon magazine culture which I can’t find in America. There are about
2 dozens of highly professional monthlies published and penned by the MI
C itself covering every branch of the armed forces. You can buy these ma
gazines at every street corner across the <a href="https://t.co/YVNteeP3
Iq">pic.twitter.com/YVNteeP3Iq</a></p>&mdash; Governor General (@manchux
i)', 'tweet_url': 'https://twitter.com/manchuxi/status/16315345834758307
88?ref_src=twsrc%5Etfw'}
6 {'sentiment_label': 'negative', 'text_of_tweet': 'China is building si
x times more new coal plants than the rest of the world combined, new re
search shows <a href="https://t.co/zd7akk1eqV">https://t.co/zd7akk1eqV</
a></p>&mdash; ABC News (@abcnews)', 'tweet_url': 'https://twitter.com/ab
cnews/status/1631450164375478272?ref_src=twsrc%5Etfw'}
7 {'sentiment_label': 'negative', 'text_of_tweet': 'China\'\'s turn towa
rds fascism is accelerating <a href="https://t.co/Bpoey4WnAz">pic.twitte
r.com/Bpoey4WnAz</a></p>&mdash; Chinese History Expert (@chineseciv)',
'tweet_url': 'https://twitter.com/chineseciv/status/1631515516207788033?
ref_src=twsrc%5Etfw'}
8 {'sentiment_label': 'positive', 'text_of_tweet': 'China has a &quot;st
unning lead&quot; in 37 out of 44 critical and emerging technologies as
Western democracies lose a global competition for research output, a sec
urity think tank said on Thursday after tracking defense, space, energy
and biotechnology. <a href="https://t.co/icY1FHvVGK">https://t.co/icY1FH
vVGK</a></p>&mdash; NEWSMAX (@NEWSMAX)', 'tweet_url': 'https://twitter.c
om/NEWSMAX/status/1631523549122007040?ref_src=twsrc%5Etfw'}
9 {'sentiment_label': 'negative', 'text_of_tweet': "I'm just wondering i
f there is any person in Taiwan who thinks that the Biden neocons are pu
mping billions of dollars of weapons onto their Island and antagonizing
China to make them safer?</p>&mdash; Garland Nixon (@GarlandNixon)", 'tw
eet_url': 'https://twitter.com/GarlandNixon/status/1631451970752978947?r
ef_src=twsrc%5Etfw'}
10 {'sentiment_label': 'negative', 'text_of_tweet': 'In response to US a
ctions, China will take retaliatory measures to protect Chinese corporat
ions – Ministry of Commerce of the People&#39;s Republic of China</p>&md
ash; AZ 🛰🌏🌍🌎 (@AZgeopolitics)', 'tweet_url': 'https://twitter.com/AZ
geopolitics/status/1631653133104345088?ref_src=twsrc%5Etfw'}
11 {'sentiment_label': 'negative', 'text_of_tweet': 'Today is March 3, 2
023 and Joe Biden is still an illegitimate President and is owned by Chi
na!</p>&mdash; PISSED OFF PATRIOT HOFFY 🖕 (@PATRIOTGHOFFY)', 'tweet_ur
l': 'https://twitter.com/PATRIOTGHOFFY/status/1631645105684635648?ref_sr
c=twsrc%5Etfw'}
12 {'sentiment_label': 'negative', 'text_of_tweet': 'Let me ask you, how
long would a China Police Station last in the US, Great Britain, Austral
ia, Japan France, New Zealand. And you know if there was a threat of ele
ction interference this would be investigated even before the public dem
and them to do so. 🤔🇨🇳 is so inbedded', 'tweet_url': 'https://t.co/Lfxx
4UD0wg'}
13 {'sentiment_label': 'negative', 'text_of_tweet': 'Wicked cleverness:
China wages border aggression against India and then repeatedly advises
India to not let the border situation come in the way of bilateral coope
ration. China&#39;s latest statement says India should put the border is
sue in &quot;the proper place in bilateral relations.”</p>&mdash; Brahma
Chellaney (@Chellaney)', 'tweet_url': 'https://twitter.com/Chellaney/sta
tus/1631610600781647872?ref_src=twsrc%5Etfw'}
14 {'sentiment_label': 'negative', 'text_of_tweet': "It's fascinating th
at our gov&#39;t suddenly admits all the facts about COVID&#39;s origin,
now that China has decided to side with Russia.</p>&mdash; Shukri Abdira
hman (@ShuForCongress)", 'tweet_url': 'https://twitter.com/ShuForCongres
s/status/1631653770147889153?ref_src=twsrc%5Etfw'}
15 {'sentiment_label': 'negative', 'text_of_tweet': 'The public is inchi
ng closer and closer to the harsh reality.<br>Russia and China’s displea
sure with US biological activity in Ukraine, is because of Covid. <br>We
stern Criminals created SARS-CoV-2, which killed millions of people, and
now the Eastern world is angry.</p>&mdash; D-Bark (@DBark46107258)', 'tw
eet_url': 'https://twitter.com/DBark46107258/status/1631650236279173120?
ref_src=twsrc%5Etfw'}
16 {'sentiment_label': 'negative', 'text_of_tweet': 'Folks, China got wh
at they wanted from Harper. That 31-year trade deal. And they got to exe
cute Canadians.<br><br>Trudeau is less biddable.<br><br>China wants the
CPC back in office, so they&#39;ve set this up. <br><br>That&#39;s what&
#39;s going on here, IMO.<a href="https://twitter.com/hashtag/cdnpoli?sr
c=hash&amp;ref_src=twsrc%5Etfw">#cdnpoli</a></p>&mdash; Timothy Anderson
💉💉💉💉💉🎶 (@AndersonBooz)', 'tweet_url': 'https://twitter.com/Anderso
nBooz/status/1631545345556779009?ref_src=twsrc%5Etfw'}
17 {'sentiment_label': 'negative', 'text_of_tweet': 'Blinken’ trip to Uz
bekistan has only one purpose… to sow the seeds of regime change that wo
uld allow the U.S. Empire to take control of the country in a few years
time and turn it into a dagger on the side of China &amp; Russia.</p>&md
ash; 倪明达 (Ni Mingda) (@NiMingda_GG)', 'tweet_url': 'https://twitter.co
m/NiMingda_GG/status/1631642321933484034?ref_src=twsrc%5Etfw'}
18 {'sentiment_label': 'negative', 'text_of_tweet': 'There is ten times
more evidence of Biden-China collusion than there ever was of Trump-Russ
ia collusion.<br><br>The Hunter Biden laptop is a smoking gun.<br><br>Wh
en have the lamestream media brought this up? Where&#39;s the campaign s
urveillance? When&#39;s a Special Counsel going to investigate?</p>&mdas
h; Kyle Becker (@kylenabecker)', 'tweet_url': 'https://twitter.com/kylen
abecker/status/1631654725367021569?ref_src=twsrc%5Etfw'}
19 {'sentiment_label': 'negative', 'text_of_tweet': '🇨🇳🇺🇸: The heat is t
urning up <br><br>&quot;We strongly oppose the sale of arms to Chinese T
aiwan...<br>We demand that the US cease arms sales to Taiwan and cease m
ilitary ties with the island.&quot; <br>The People&#39;s Liberation Army
of China is always ready to strike back...&quot;<br>-spokesman Tan Kefei
<br>--&gt;👇</p>&mdash; David Roth-Lindberg (@RothLindberg)', 'tweet_ur
l': 'https://twitter.com/RothLindberg/status/1631635667154337794?ref_src
=twsrc%5Etfw'}
```

```
20 {'sentiment_label': 'negative', 'text_of_tweet': 'A report from the A
ustralian Institute for Strategic Policy Research warns that China is ac
hieving a significant advantage over the US and the West in the vast maj
ority of critical and advanced technologies.<br><br>According to the rep
ort, China leads in 37 out of 44 technologies… <a href="https://t.co/nam
ahAiBT2">https://t.co/namahAiBT2</a></p>&mdash; GraphicW (@GraphicW5)',
'tweet_url': 'https://twitter.com/GraphicW5/status/1631634185742868480?r
ef_src=twsrc%5Etfw'}
21 {'sentiment_label': 'negative', 'text_of_tweet': 'Americans falsely a
ssume that a war with China will be fought in China.<br><br>.</p>&mdash;
david kersten (@davidkersten)', 'tweet_url': 'https://twitter.com/davidk
ersten/status/1631469854308827137?ref_src=twsrc%5Etfw'}
22 {'sentiment_label': 'neutral', 'text_of_tweet': 'The boundary issue s
hould be put in the proper place in bilateral relations, Qin said, addin
g that the situation on the borders should be brought under normalized m
anagement as soon as possible: China statement on EAM-China FM meet</p>&
mdash; Sidhant Sibal (@sidhant)', 'tweet_url': 'https://twitter.com/sidh
ant/status/1631601051064467457?ref_src=twsrc%5Etfw'}
23 {'sentiment_label': 'negative', 'text_of_tweet': '#China’s coming for
us. This is war. <a href="https://twitter.com/hashtag/CCP?src=hash&amp;r
ef_src=twsrc%5Etfw">#CCP</a></p>&mdash; Gordon G. Chang (@GordonGChan
g)', 'tweet_url': 'https://twitter.com/GordonGChang/status/1631460454601
043968?ref_src=twsrc%5Etfw'}
24 {'sentiment_label': 'negative', 'text_of_tweet': 'One of the many ong
oing failures of west and particularly the US is this completely flawed
belief that China wants to be a hegemonic power and that this view is sh
ared and demanded by the Chinese people.</p>&mdash; The Sirius Report (@
thesiriusreport)', 'tweet_url': 'https://twitter.com/thesiriusreport/sta
tus/1631558205124771841?ref_src=twsrc%5Etfw'}
25 {'sentiment_label': 'negative', 'text_of_tweet': 'If Australia become
s &quot;Aboriginalia&quot; when we cede sovereignty to the elite militan
t aborigines, how will they defend the country against the Chinese invas
ion when it comes? Will they point sticks and throw stones at China&#39;
s nuclear arsenal? <a href="https://twitter.com/hashtag/voteNO?src=hash&
amp;ref_src=twsrc%5Etfw">#voteNO</a></p>&mdash; Francis_Young (@commonse
nse058)', 'tweet_url': 'https://twitter.com/commonsense058/status/163156
0666103566336?ref_src=twsrc%5Etfw'}
26 {'sentiment_label': 'negative', 'text_of_tweet': '@GordonGChang</a> t
ells One America News China lied about the coronavirus from the beginnin
g. One America’s John Hines has more from CPAC. [VIDEO] <a href="http
s://twitter.com/hashtag/ChinaLiedPeopleDied?src=hash&amp;ref_src=twsrc%5
Etfw">#ChinaLiedPeopleDied</a> <a href="https://twitter.com/hashtag/Chin
aOwnsBiden?src=hash&amp;ref_src=twsrc%5Etfw">#ChinaOwnsBiden</a> <a href
="https://t.co/px1dNsHEeZ">https://t.co/px1dNsHEeZ</a></p>&mdash; Jenny
1776🇺🇸 (@realouMAGAgirl)', 'tweet_url': 'https://twitter.com/realouMAGAg
irl/status/1631638732783730691?ref_src=twsrc%5Etfw'}
27 {'sentiment_label': 'neutral', 'text_of_tweet': 'Chinese aerospace <b
r>engineers used &#13;science developed by an American &#13;hypersonic s
cientist and a National &#13;Aeronautics Space Administration &#13;(NAS
A) project to address an issue with &#13;a proposed hypersonic-speed lau
nch &#13;vehicle meant to intercept hypersonic &#13;missiles.<a href="ht
tps://twitter.com/hashtag/China?src=hash&amp;ref_src=twsrc%5Etfw">#China
</a> <a href="https://t.co/Y8h5OCsQyG">pic.twitter.com/Y8h5OCsQyG</a></p
>&mdash; Hira Bashir (@HiraBK5090)', 'tweet_url': 'https://twitter.com/H
iraBK5090/status/1631545302250299393?ref_src=twsrc%5Etfw'}
28 {'sentiment_label': 'positive', 'text_of_tweet': 'China dominates glo
bal tech race. Beijing has a “stunning lead” over the US.<br><br>China i
s leading the world in 37 out of 44 critical and emerging technologies,
the Australian Strategic Policy Institute (ASPI) said.<br><br>„Beijing i
s the world’s leading science and technology superpower.“</p>&mdash; Mak
```

```
e Peace Now; alternative news (@AlternatNews)', 'tweet_url': 'https://tw
itter.com/AlternatNews/status/1631606264189919232?ref_src=twsrc%5Etfw'}
29 {'sentiment_label': 'negative', 'text_of_tweet': 'It appears as thoug
h as the tables are turning, it will be the west starved for resources w
hile many of the nations with plentiful resources are gravitating to Rus
sia and China...<br><br>Sudan is ready to cooperate with Russia on oil p
roduction issues.<br><br>The head of the Sudan Energy and… <a href="http
s://t.co/HsDWesE4h5">https://t.co/HsDWesE4h5</a></p>&mdash; GraphicW (@G
raphicW5)', 'tweet_url': 'https://twitter.com/GraphicW5/status/163165745
2134440963?ref_src=twsrc%5Etfw'}
30 {'sentiment_label': 'positive', 'text_of_tweet': 'Yuqi's stylist in C
hina is always on point! They never miss! <a href="https://t.co/lSoHJLHx
zP">pic.twitter.com/lSoHJLHxzP</a></p>&mdash; Singer Xiao Song | Little
Giant | Yuqi (@yuqiriiin)', 'tweet_url': 'https://twitter.com/yuqiriiin/
status/1631603822203383809?ref_src=twsrc%5Etfw'}
31 {'sentiment_label': 'neutral', 'text_of_tweet': 'I'm currently workin
g in China. Almost exactly 100 years ago my great grandfather was here.
These are his watercolours he sent home to his son (my grandfather). <a
href="https://twitter.com/hashtag/History?src=hash&amp;ref_src=twsrc%5Et
fw">#History</a> <a href="https://t.co/sipek5usa8">pic.twitter.com/sipek
5usa8</a></p>&mdash; Dr Sam Willis (@DrSamWillis)', 'tweet_url': 'http
s://twitter.com/DrSamWillis/status/1631487477780213760?ref_src=twsrc%5Et
fw'}
32 {'sentiment_label': 'positive', 'text_of_tweet': 'Russia&#39;s energy
policy will rely on reliable partners, including China and India, but no
t the West.<br> Russia will not allow the West to &quot;blow up gas pipe
lines&quot; again -<br> Lavrov</p>&mdash; Enrico60🇨🇳🇷🇺 (互fo) (@enfree19
93)', 'tweet_url': 'https://twitter.com/enfree1993/status/16315694202787
26661?ref_src=twsrc%5Etfw'}
33 {'sentiment_label': 'negative', 'text_of_tweet': 'Iranian opposition:
Iran is too close to China/Russia, and that&#39;s why the US hates us.<b
r><br>Russian opposition: Russia is too close to Iran/China, and that&#3
9;s why the US hates us.<br><br>Chinese opposition: China is too close t
o Russia/Iran, and that&#39;s why the US hates us.<br><br>LOL.</p>&mdas
h; DaiWW (@BeijingDai)', 'tweet_url': 'https://twitter.com/BeijingDai/st
atus/1631569408484323328?ref_src=twsrc%5Etfw'}
34 {'sentiment_label': 'positive', 'text_of_tweet': 'Justin Trudeau has
a level of admiration for China&#39;s money.</p>&mdash; Zachary Tisdale
🇨🇦 (@ztisdale)', 'tweet_url': 'https://twitter.com/ztisdale/status/16314
62637031632898?ref_src=twsrc%5Etfw'}
35 {'sentiment_label': 'negative', 'text_of_tweet': 'It seems that not o
nly does <a href="https://twitter.com/JustinTrudeau?ref_src=twsrc%5Etf
w">@JustinTrudeau</a> have an admiration for the basic dictatorship of C
hina…<br><br>He also has their financing.<a href="https://twitter.com/ha
shtag/ChinaTrudeau?src=hash&amp;ref_src=twsrc%5Etfw">#ChinaTrudeau</a></
p>&mdash; Viva Frei (@thevivafrei)', 'tweet_url': 'https://twitter.com/t
hevivafrei/status/1631466024158519298?ref_src=twsrc%5Etfw'}
36 {'sentiment_label': 'negative', 'text_of_tweet': 'Russia is getting t
heir dick kicked in Ukraine the one thing China and Russia have in commo
n are paper tiger armies that are way over hyped and rife with corruptio
n <a href="https://t.co/A7bnnidRDK">https://t.co/A7bnnidRDK</a></p>&mdas
h; Toriel1one1 (@toriel1one1)', 'tweet_url': 'https://twitter.com/toriel
1one1/status/1631497804345483264?ref_src=twsrc%5Etfw'}
37 {'sentiment_label': 'negative', 'text_of_tweet': '🇺🇸🇨🇳☢⚛&quot;US is t
he main source of the nuclear threat in the world, they are hyping the t
heory of the threat from China in search of an excuse to expand their ar
senal.&quot; - Chinese Foreign Ministry</p>&mdash; AZ 🛰🌏🌍🌎 (@AZgeopo
litics)', 'tweet_url': 'https://twitter.com/AZgeopolitics/status/1631564
558388043776?ref_src=twsrc%5Etfw'}
38 {'sentiment_label': 'negative', 'text_of_tweet': 'Man do I have to st
```

```
op myself from cringing when Lavrov talks.<br><br>Sign of the times real
ly. Outside of energy, parts of defence &amp; a desire to contain China,
there is nothing in the relationship anymore.<br><br>Long term stagnatio
n is best case scenario.</p>&mdash; Yew&#39;s Finest (@FinestYew)', 'twe
et_url': 'https://twitter.com/FinestYew/status/1631660098958540800?ref_s
rc=twsrc%5Etfw'}
39 {'sentiment_label': 'neutral', 'text_of_tweet': '#Flash</a> China has
given a fresh loan of USD 700 million to Pakistan at the rate of 8.9%. T
wo railway stations of Pakistan (Lahore &amp; Sukkur) have been taken by
China as security for 99 years or till the full and final payment of thi
s loan, which is earlier. (Sources)</p>&mdash; Baba Banaras™ (@RealBabab
anaras)', 'tweet_url': 'https://twitter.com/RealBababanaras/status/16314
97938596945920?ref_src=twsrc%5Etfw'}
40 {'sentiment_label': 'positive', 'text_of_tweet': '#China</a> leading
<a href="https://twitter.com/hashtag/US?src=hash&amp;ref_src=twsrc%5Etf
w">#US</a> in technology race in all but a few fields, thinktank finds<b
r><br>Year-long study finds China leads in 37 of 44 areas it tracked, wi
th potential for a monopoly in areas such as nanoscale materials and syn
thetic biology.<a href="https://t.co/IICGKLrDOM">https://t.co/IICGKLrDOM
</a></p>&mdash; Indo-Pacific News - Geo-Politics &amp; Military News (@I
ndoPac_Info)', 'tweet_url': 'https://twitter.com/IndoPac_Info/status/163
1589226478198784?ref_src=twsrc%5Etfw'}
41 {'sentiment_label': 'positive', 'text_of_tweet': 'China&#39;s &#39;Tw
o Sessions&#39; annual legislative body begins, here in Beijing, tomorro
w.<br><br>With all eyes on China&#39;s top law making body, Reuters repo
rts GDP goals may be set as high as 6% growth for 2023.<a href="https://
twitter.com/hashtag/China?src=hash&amp;ref_src=twsrc%5Etfw">#China</a> <
a href="https://twitter.com/hashtag/TwoSessions?src=hash&amp;ref_src=tws
rc%5Etfw">#TwoSessions</a><a href="https://t.co/uZSx67cgRV">https://t.c
o/uZSx67cgRV</a></p>&mdash; Jason - 上官杰文 (@ShangguanJiewen)', 'tweet_u
rl': 'https://twitter.com/ShangguanJiewen/status/1631488736885178370?ref
_src=twsrc%5Etfw'}
42 {'sentiment_label': 'neutral', 'text_of_tweet': 'The Anti-Counterfeit
Authority (ACA) has released goods worth Sh50 million that were seized a
t China Square.<br><br>The quick return of the goods comes a day after t
he Chinese embassy urged the Kenyan government to intervene to protect C
hinese enterprises and citizens.<br><br>— Nation</p>&mdash; Moe (@moneya
cademyKE)', 'tweet_url': 'https://twitter.com/moneyacademyKE/status/1631
512472644632576?ref_src=twsrc%5Etfw'}
43 {'sentiment_label': 'negative', 'text_of_tweet': 'Khan was ousted fro
m power in April after losing a no-confidence vote in his leadership, wh
ich he alleged was part of a US-led conspiracy targeting him because of
his independent foreign policy decisions on Russia, China and Afghanista
n.<a href="https://twitter.com/7n_Star_?ref_src=twsrc%5Etfw">@7n_Star_</
a><a href="https://twitter.com/hashtag/%D8%AA%D8%A8%D8%A7%DB%81%DB%8C_%D
8%B3%D8%B1%DA%A9%D8%A7%D8%B1_%D8%AC%D8%A7%D9%86_%DA%86%DA%BE%D9%88%DA%9
1%D9%88?src=hash&amp;ref_src=twsrc%5Etfw">#تباہی_سرکار_جان_چھوڑو</a></p>
&mdash; Ɲαʝɱυʂ ʂαʅιԋҽҽɳ🙃 (@7n_Star_)', 'tweet_url': 'https://twitter.co
m/7n_Star_/status/1631597034254872577?ref_src=twsrc%5Etfw'}
44 {'sentiment_label': 'negative', 'text_of_tweet': '#China</a> providin
g <a href="https://twitter.com/hashtag/Russia?src=hash&amp;ref_src=twsr
c%5Etfw">#Russia</a> uniforms, weapons and ammunition only prolongs the
war in <a href="https://twitter.com/hashtag/Ukraine?src=hash&amp;ref_src
=twsrc%5Etfw">#Ukraine</a>. Russia has the bodies; China will outfit the
m. Not only will it prolong the war - but it also weakens Russia as wel
l. Is that the plan? Who needs enemies when you <a href="https://t.co/pz
UiyZATEi">https://t.co/pzUiyZATEi</a>… <a href="https://t.co/xVfdrqVlb
y">https://t.co/xVfdrqVlby</a></p>&mdash; Jon Sweet (@JESweet2022)', 'tw
eet_url': 'https://twitter.com/JESweet2022/status/1631630908024401927?re
f_src=twsrc%5Etfw'}
```

```
45 {'sentiment_label': 'negative', 'text_of_tweet': 'Protests in Kenya a
gainst China.<br>People in Kenya think that Chinese projects in Kenya he
lp Chinese companies but not workers in Kenya.<a href="https://twitter.c
om/hashtag/China?src=hash&amp;ref_src=twsrc%5Etfw">#China</a> <a href="h
ttps://twitter.com/hashtag/Chinaprotests?src=hash&amp;ref_src=twsrc%5Etf
w">#Chinaprotests</a> <a href="https://twitter.com/hashtag/Kenya?src=has
h&amp;ref_src=twsrc%5Etfw">#Kenya</a> <a href="https://t.co/qOZI6yyWwI">
pic.twitter.com/qOZI6yyWwI</a></p>&mdash; That is China (@2022_Lockdow
n)', 'tweet_url': 'https://twitter.com/2022_Lockdown/status/163148866538
4779776?ref_src=twsrc%5Etfw'}
46 {'sentiment_label': 'neutral', 'text_of_tweet': 'My latest for <a hre
f="https://twitter.com/dw_hotspotasia?ref_src=twsrc%5Etfw">@dw_hotspotas
ia</a>: As <a href="https://twitter.com/hashtag/China?src=hash&amp;ref_s
rc=twsrc%5Etfw">#China</a>&#39;s rubber-stamp parliament gathers in Beij
ing this weekend, President Xi Jinping is expected to officially kick of
f his third term. China&#39;s Communist party will likely initiate furth
er institutional reform. <a href="https://t.co/8lbe9CJ2SO">https://t.co/
8lbe9CJ2SO</a></p>&mdash; William Yang (@WilliamYang120)', 'tweet_url':
'https://twitter.com/WilliamYang120/status/1631630614549118978?ref_src=t
wsrc%5Etfw'}
47 {'sentiment_label': 'negative', 'text_of_tweet': 'Beijing has critici
zed Canberra for blocking a bid by a Chinese-linked company to boost its
ownership in a rare earths supplier, an episode that underscores the cha
llenges the two nations face repairing ties <a href="https://t.co/1zbM0O
KNgi">https://t.co/1zbM0OKNgi</a></p>&mdash; Bloomberg (@business)', 'tw
eet_url': 'https://twitter.com/business/status/1631602420357758977?ref_s
rc=twsrc%5Etfw'}
48 {'sentiment_label': 'negative', 'text_of_tweet': 'The return of Chin
a’s top basketball league to its normal season format following years of
Covid disruptions has been marred in controversy <a href="https://t.co/u
fVfOYdDO0">https://t.co/ufVfOYdDO0</a></p>&mdash; CNN (@CNN)', 'tweet_ur
l': 'https://twitter.com/CNN/status/1631615109750554624?ref_src=twsrc%5E
tfw'}
49 {'sentiment_label': 'neutral', 'text_of_tweet': 'In meeting with Saud
i FM Prince Faisal bin Farhan Al Saud, Chinese FM <a href="https://twitt
er.com/hashtag/QinGang?src=hash&amp;ref_src=twsrc%5Etfw">#QinGang</a> sa
id <a href="https://twitter.com/hashtag/China?src=hash&amp;ref_src=twsr
c%5Etfw">#China</a> is ready to keep the positive momentum of high-level
exchanges with <a href="https://twitter.com/hashtag/SaudiaArabia?src=has
h&amp;ref_src=twsrc%5Etfw">#SaudiaArabia</a> and work together to advanc
e high-quality Belt and Road Cooperation. <a href="https://t.co/4A5v9ouA
xy">pic.twitter.com/4A5v9ouAxy</a></p>&mdash; Liu Yongfeng (@liupheoni
x)', 'tweet_url': 'https://twitter.com/liupheonix/status/163147342281874
2272?ref_src=twsrc%5Etfw'}
50 {'sentiment_label': 'positive', 'text_of_tweet': 'China ‘Is the Only
One in the Race’ to Make Electric Buses, Taxis and Trucks <a href="http
s://t.co/XF6UkHJ3Ur">https://t.co/XF6UkHJ3Ur</a> by <a href="https://twi
tter.com/Trefor1?ref_src=twsrc%5Etfw">@Trefor1</a> <a href="https://t.c
o/4VpWwZLmV7">pic.twitter.com/4VpWwZLmV7</a></p>&mdash; CHINA (@china)',
'tweet_url': 'https://twitter.com/china/status/1069728152581218305?ref_s
rc=twsrc%5Etfw'}
```

[5 points] Question 3:

Run VADER on your own tweets (see function **run_vader** from notebook **Lab2-Sentiment-analysis-using-VADER.ipynb**). You can use the code snippet below this explanation as a starting point.

- [2.5 points] a. Perform a quantitative evaluation. Explain the different scores, and explain which scores are most relevant and why.
- [2.5 points] b. Perform an error analysis: select 10 positive, 10 negative and 10 neutral tweets that are not correctly classified and try to understand why. Refer to the VADER-rules and the VADER-lexicon. Of course, if there are less than 10 errors for a category, you only have to check those. For example, if there are only 5 errors for positive tweets, you just describe those.

In [4]:

```
import spacy
from nltk.sentiment import vader
from nltk.sentiment.vader import SentimentIntensityAnalyzer
vader_model = SentimentIntensityAnalyzer()
```

In [5]:

```
def run_vader(textual_unit,
              lemmatize=False,
              parts_of_speech_to_consider=None,
              verbose=0):
    """
    Run VADER on a sentence from spacy

    :param str textual unit: a textual unit, e.g., sentence, sentences (c
    (by looping over doc.sents)
    :param bool lemmatize: If True, provide lemmas to VADER instead of wo
    :param set parts_of_speech_to_consider:
    -None or empty set: all parts of speech are provided
    -non-empty set: only these parts of speech are considered.
    :param int verbose: if set to 1, information is printed
    about input and output

    :rtype: dict
    :return: vader output dict
    """
    nlp = spacy.load('en_core_web_sm')
    doc = nlp(textual_unit)

    input_to_vader = []

    for sent in doc.sents:
        for token in sent:

            to_add = token.text

            if lemmatize:
                to_add = token.lemma_

                if to_add == '-PRON-':
                    to_add = token.text

            if parts_of_speech_to_consider:
                if token.pos_ in parts_of_speech_to_consider:
                    input_to_vader.append(to_add)
            else:
                input_to_vader.append(to_add)

    scores = vader_model.polarity_scores(' '.join(input_to_vader))
    return scores
```

In [6]:

```
def vader_output_to_label(vader_output):
    """
    map vader output e.g.,
    {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compound': 0.4215}
    to one of the following values:
    a) positive float -> 'positive'
    b) 0.0 -> 'neutral'
    c) negative float -> 'negative'

    :param dict vader_output: output dict from vader

    :rtype: str
    :return: 'negative' | 'neutral' | 'positive'
    """
    compound = vader_output['compound']

    if compound < 0:
        return 'negative'
    elif compound == 0.0:
        return 'neutral'
    elif compound > 0.0:
        return 'positive'

assert vader_output_to_label( {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compo
assert vader_output_to_label( {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compo
assert vader_output_to_label( {'neg': 0.0, 'neu': 0.0, 'pos': 1.0, 'compo
```

In [7]:

```
tweets = []
all_vader_output = []
gold = []

# settings (to change for different experiments)
to_lemmatize = True
pos = set()

for id_, tweet_info in my_tweets.items():
    the_tweet = tweet_info['text_of_tweet']
    vader_output = run_vader(the_tweet)
    vader_label = vader_output_to_label(vader_output)# convert vader outp
    tweets.append(the_tweet)
    all_vader_output.append(vader_label)
    gold.append(tweet_info['sentiment_label'])


# use scikit-learn's classification report
from sklearn.metrics import classification_report
print(classification_report(gold, all_vader_output))
```

```
              precision    recall  f1-score   support

    negative       0.84      0.48      0.62        33
     neutral       0.20      0.29      0.24         7
    positive       0.19      0.40      0.26        10

    accuracy                           0.44        50
   macro avg       0.41      0.39      0.37        50
weighted avg       0.62      0.44      0.49        50
```

Question 3a Answer Quantitative evaluation:

Precision: The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0. Recall: The recall is intuitively the ability of the classifier to find all the positive samples. The best value is 1 and the worst value is 0. F1-score: The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. Support: The support is the number of occurrences of each class in y_true. Accuracy: The accuracy is the number of correctly classified samples divided by the total number of samples. The best value is 1 and the worst value is 0. Macro avg: Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account. Weighted avg: Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro' to account for label imbalance; it can result in an F-score that is not between precision and recall. Micro avg: Calculate metrics globally by counting the total true positives, false negatives and false positives. This is a better metric when we have class imbalance. Samples avg: Calculate metrics for each instance, and find their average (only meaningful for multilabel classification where this differs from accuracy_score). According to the classification report generated previously, it can be seen that the model has a high precision for the negative tweets, but low for the neutral and positive ones. This means that most things classified as negative are indeed negative, but that's not the case for the positive and neutral tweets. Recall indicates how many relevant items are retrieved (e.g. how many of the negative items where classified as negative), the recall is low for all labels, being slightly higher for the negative (.48) and the lowest for the neutral (.39). The f1 score is relatively high for the negative, which makes sense since it had high precision and the highest recall out of the three, however the f1 is low for the negative and the neutral since both the precision and recall were low. Macro average for the precision is 0.41, while the weighted average is 0.69, the difference is due to the macro average not taking label imbalance into account.

In [8]:

```
# error analysis
misclassified_pos = []
misclassified_neg = []
misclassified_neu = []

for i, (tweet, vader_label, gold_label) in enumerate(zip(tweets, all_vade
    if vader_label != gold_label:
        if gold_label == 'positive':
            misclassified_pos.append((i, tweet, vader_label, gold_label))
        elif gold_label == 'negative':
            misclassified_neg.append((i, tweet, vader_label, gold_label))
        elif gold_label == 'neutral':
            misclassified_neu.append((i, tweet, vader_label, gold_label))

print('Number of misclassified positive tweets: {}'.format(len(misclassif
print('Number of misclassified negative tweets: {}'.format(len(misclassif
print('Number of misclassified neutral tweets: {}'.format(len(misclassifi
```

```
Number of misclassified positive tweets: 6
Number of misclassified negative tweets: 17
Number of misclassified neutral tweets: 5
```

In [9]:

```
# print misclassified positive tweets
for i, tweet, vader_label, gold_label in misclassified_pos:
    print('Tweet: {}'.format(tweet))
    print('Vader label: {}'.format(vader_label))
    print('Gold label: {}'.format(gold_label))
    print('-----------------------------')
```

```
Tweet: “cHiNa cAn’T iNnOvAtE.” 💥Analysis by ASPI∗ shows that China lead
s the USA in whopping 37 out of 44 critical scientific areas such as AI,
quantum computing, biotech, and advanced materials.<br><br>∗funded by U.
S. military industrial complex, so no pro-China bias <a href="https://t.
co/CgNUmGA0iE"> pic.twitter.com/CgNUmGA0iE
Vader label: negative
Gold label: positive
------------------------------
Tweet: China has a prevalent weapon magazine culture which I can’t find
in America. There are about 2 dozens of highly professional monthlies pu
blished and penned by the MIC itself covering every branch of the armed
forces. You can buy these magazines at every street corner across the <a
href="https://t.co/YVNteeP3Iq">pic.twitter.com/YVNteeP3Iq</a></p>&mdash;
Governor General (@manchuxi)
Vader label: negative
Gold label: positive
------------------------------
Tweet: China has a &quot;stunning lead&quot; in 37 out of 44 critical an
d emerging technologies as Western democracies lose a global competition
for research output, a security think tank said on Thursday after tracki
ng defense, space, energy and biotechnology. <a href="https://t.co/icY1F
HvVGK">https://t.co/icY1FHvVGK</a></p>&mdash; NEWSMAX (@NEWSMAX)
Vader label: neutral
Gold label: positive
------------------------------
Tweet: Russia&#39;s energy policy will rely on reliable partners, includ
ing China and India, but not the West.<br> Russia will not allow the Wes
t to &quot;blow up gas pipelines&quot; again -<br> Lavrov</p>&mdash; Enr
ico60🇨🇳🇷🇺 (互fo) (@enfree1993)
Vader label: negative
Gold label: positive
------------------------------
Tweet: #China</a> leading <a href="https://twitter.com/hashtag/US?src=ha
sh&amp;ref_src=twsrc%5Etfw">#US</a> in technology race in all but a few
fields, thinktank finds<br><br>Year-long study finds China leads in 37 o
f 44 areas it tracked, with potential for a monopoly in areas such as na
noscale materials and synthetic biology.<a href="https://t.co/IICGKLrDO
M">https://t.co/IICGKLrDOM</a></p>&mdash; Indo-Pacific News - Geo-Politi
cs &amp; Military News (@IndoPac_Info)
Vader label: neutral
Gold label: positive
------------------------------
Tweet: China ‘Is the Only One in the Race’ to Make Electric Buses, Taxis
and Trucks <a href="https://t.co/XF6UkHJ3Ur">https://t.co/XF6UkHJ3Ur</a>
by <a href="https://twitter.com/Trefor1?ref_src=twsrc%5Etfw">@Trefor1</a
> <a href="https://t.co/4VpWwZLmV7">pic.twitter.com/4VpWwZLmV7</a></p>&m
dash; CHINA (@china)
Vader label: neutral
Gold label: positive
------------------------------
```

Question 3b Answer

Error Analysis on Positive Tweets: We found 6 positive tweets that were missclassified by VADER.

For instance, the tweet contains information about China being powerful in certain areas of scientific research. The content is mostly possitive, but VADER classifies it as negative. This could be due to VADER just taking into account the words present on it's lexicon or the sarcastic comment in the beginning that says "China can't innovate", can't being possibly seen as negative. The tweet contain words such as bias and no that have a negative sentiment rating.

The second tweet VADER classifies as negative, while Gold as positive, it's ambiguos by the text itself whether it's positive or negative but could be classified as negative due to the use of the word weapon and can't. We argued it was positive about China as they had something that was apparently desired by the person that they missed while being in the US.

The third tweet is classified as negative mostly because it contains words such as "lose" to make a comparison. Arguably the text could be negative based on perspective but we chose to focus on the sentiment about China instead of Western disappointment at Chinese success.

In four and five there is a combination of positive words with negations in complex sentence structures so that might explain by the tweets were missclassified.

The last tweet is classifies as neutral, but then again the meaning of the text itself is ambiguous. Probably the words in the text are just neither positive nor negative in the VADER lexicon.

In [10]:

```
# print misclassified negative tweets
for i, tweet, vader_label, gold_label in misclassified_neg:
    print('Tweet: {}'.format(tweet))
    print('Vader label: {}'.format(vader_label))
    print('Gold label: {}'.format(gold_label))
    print('-----------------------------')
```

```
Tweet: China appears to be requiring foreign law professors to submit th
eir syllabuses to ensure they are following a doctrine pushed by Preside
nt Xi Jinping <a href="https://t.co/SuSWhELiCx">https://t.co/SuSWhELiCx
</a></p>&mdash; Bloomberg (@business)
Vader label: positive
Gold label: negative
------------------------------
Tweet: The United States has added two subsidiaries of Chinese genetics
company BGI to a trade blacklist over allegations it conducted genetic a
nalysis and surveillance activities for Beijing, which Washington says w
as used to repress ethnic minorities in China <a href="https://t.co/siXR
57whNs">https://t.co/siXR57whNs</a></p>&mdash; CNN (@CNN)
Vader label: positive
Gold label: negative
------------------------------
Tweet: China is building six times more new coal plants than the rest of
the world combined, new research shows <a href="https://t.co/zd7akk1eq
V">https://t.co/zd7akk1eqV</a></p>&mdash; ABC News (@abcnews)
Vader label: neutral
Gold label: negative
------------------------------
Tweet: China''s turn towards fascism is accelerating <a href="https://t.
co/Bpoey4WnAz">pic.twitter.com/Bpoey4WnAz</a></p>&mdash; Chinese History
Expert (@chineseciv)
Vader label: neutral
Gold label: negative
------------------------------
Tweet: In response to US actions, China will take retaliatory measures t
o protect Chinese corporations – Ministry of Commerce of the People&#39;
s Republic of China</p>&mdash; AZ 🛰🌏🌍🌎 (@AZgeopolitics)
Vader label: positive
Gold label: negative
------------------------------
Tweet: Let me ask you, how long would a China Police Station last in the
US, Great Britain, Australia, Japan France, New Zealand. And you know if
there was a threat of election interference this would be investigated e
ven before the public demand them to do so. 🤔🇨🇳 is so inbedded
Vader label: positive
Gold label: negative
------------------------------
Tweet: It's fascinating that our gov&#39;t suddenly admits all the facts
about COVID&#39;s origin, now that China has decided to side with Russi
a.</p>&mdash; Shukri Abdirahman (@ShuForCongress)
Vader label: positive
Gold label: negative
------------------------------
Tweet: Folks, China got what they wanted from Harper. That 31-year trade
deal. And they got to execute Canadians.<br><br>Trudeau is less biddabl
e.<br><br>China wants the CPC back in office, so they&#39;ve set this u
p. <br><br>That&#39;s what&#39;s going on here, IMO.<a href="https://twi
tter.com/hashtag/cdnpoli?src=hash&amp;ref_src=twsrc%5Etfw">#cdnpoli</a>
</p>&mdash; Timothy Anderson 💉💉💉💉💉🎶 (@AndersonBooz)
Vader label: neutral
Gold label: negative
------------------------------
Tweet: Blinken’ trip to Uzbekistan has only one purpose… to sow the seed
s of regime change that would allow the U.S. Empire to take control of t
he country in a few years time and turn it into a dagger on the side of
China &amp; Russia.</p>&mdash; 倪明达 (Ni Mingda) (@NiMingda_GG)
Vader label: positive
```

```
Gold label: negative
------------------------------
Tweet: There is ten times more evidence of Biden-China collusion than th
ere ever was of Trump-Russia collusion.<br><br>The Hunter Biden laptop i
s a smoking gun.<br><br>When have the lamestream media brought this up?
Where&#39;s the campaign surveillance? When&#39;s a Special Counsel goin
g to investigate?</p>&mdash; Kyle Becker (@kylenabecker)
Vader label: positive
Gold label: negative
------------------------------
Tweet: 🇨🇳🇺🇸: The heat is turning up <br><br>&quot;We strongly oppose the
sale of arms to Chinese Taiwan...<br>We demand that the US cease arms sa
les to Taiwan and cease military ties with the island.&quot; <br>The Peo
ple&#39;s Liberation Army of China is always ready to strike back...&quo
t;<br>-spokesman Tan Kefei<br>--&gt;👇</p>&mdash; David Roth-Lindberg (@
RothLindberg)
Vader label: positive
Gold label: negative
------------------------------
Tweet: A report from the Australian Institute for Strategic Policy Resea
rch warns that China is achieving a significant advantage over the US an
d the West in the vast majority of critical and advanced technologies.<b
r><br>According to the report, China leads in 37 out of 44 technologies…
<a href="https://t.co/namahAiBT2">https://t.co/namahAiBT2</a></p>&mdash;
GraphicW (@GraphicW5)
Vader label: positive
Gold label: negative
------------------------------
Tweet: If Australia becomes &quot;Aboriginalia&quot; when we cede sovere
ignty to the elite militant aborigines, how will they defend the country
against the Chinese invasion when it comes? Will they point sticks and t
hrow stones at China&#39;s nuclear arsenal? <a href="https://twitter.co
m/hashtag/voteNO?src=hash&amp;ref_src=twsrc%5Etfw">#voteNO</a></p>&mdas
h; Francis_Young (@commonsense058)
Vader label: neutral
Gold label: negative
------------------------------
Tweet: It appears as though as the tables are turning, it will be the we
st starved for resources while many of the nations with plentiful resour
ces are gravitating to Russia and China...<br><br>Sudan is ready to coop
erate with Russia on oil production issues.<br><br>The head of the Sudan
Energy and… <a href="https://t.co/HsDWesE4h5">https://t.co/HsDWesE4h5</a
></p>&mdash; GraphicW (@GraphicW5)
Vader label: neutral
Gold label: negative
------------------------------
Tweet: It seems that not only does <a href="https://twitter.com/JustinTr
udeau?ref_src=twsrc%5Etfw">@JustinTrudeau</a> have an admiration for the
basic dictatorship of China…<br><br>He also has their financing.<a href
="https://twitter.com/hashtag/ChinaTrudeau?src=hash&amp;ref_src=twsrc%5E
tfw">#ChinaTrudeau</a></p>&mdash; Viva Frei (@thevivafrei)
Vader label: positive
Gold label: negative
------------------------------
Tweet: Man do I have to stop myself from cringing when Lavrov talks.<br>
<br>Sign of the times really. Outside of energy, parts of defence &amp;
a desire to contain China, there is nothing in the relationship anymore.
<br><br>Long term stagnation is best case scenario.</p>&mdash; Yew&#39;s
Finest (@FinestYew)
Vader label: positive
```

```
Gold label: negative
-----------------------------
Tweet: Protests in Kenya against China.<br>People in Kenya think that Ch
inese projects in Kenya help Chinese companies but not workers in Kenya.
<a href="https://twitter.com/hashtag/China?src=hash&amp;ref_src=twsrc%5E
tfw">#China</a> <a href="https://twitter.com/hashtag/Chinaprotests?src=h
ash&amp;ref_src=twsrc%5Etfw">#Chinaprotests</a> <a href="https://twitte
r.com/hashtag/Kenya?src=hash&amp;ref_src=twsrc%5Etfw">#Kenya</a> <a href
="https://t.co/qOZI6yyWwI">pic.twitter.com/qOZI6yyWwI</a></p>&mdash; Tha
t is China (@2022_Lockdown)
Vader label: positive
Gold label: negative
-----------------------------
```

Question 3b Answer

Error Analysis on Negative Tweets:

NOTE: since we found more than 10 negative missclassified tweets we'll try to explain the results for 10 of them.

- 1. Words have a negative meaning because of the context, VADER misses out on that. (These words don't have a negative sentiment rating in the lexicon)
- 2. Again, probably misses out on the context interpretation of words.
- 3. The tweet contains words with negative sentiment rating such as "repress", but other such as "ethnical" are positive. VADER is not able to gauge the overall meaning of the sentence in this tweet.
- 5. The tweet contains words that are neutral according to the VADER lexicon
- 6. Lexicon contains word "fascist" but not fascism, since we are using the words and not the lemmas it might be that it doesn't recognize it as negative.
- 9. "retaliatory" not in lexicon, protect is positive.
- 11. Sentence is ambiguous, VADER just looks at the valence of each word.
- 13. "fascinating" has a positive sentiment rating. The words in the tweet just have a positive sentiment rating, VADER is not able to gauge the context.
- 16. Negative due to political context, not to separate words, so it's missclassified by vader.
- 17. Similar to 16, meaning depends on context and knowledge about the world and politics, which vader doesn't have.

In [11]:

```python
# print misclassified neutral tweets
for i, tweet, vader_label, gold_label in misclassified_neu:
    print('Tweet: {}'.format(tweet))
    print('Vader label: {}'.format(vader_label))
    print('Gold label: {}'.format(gold_label))
    print('-----------------------------')
```

```
Tweet: I’m currently working in China. Almost exactly 100 years ago my g
reat grandfather was here. These are his watercolours he sent home to hi
s son (my grandfather). <a href="https://twitter.com/hashtag/History?src
=hash&amp;ref_src=twsrc%5Etfw">#History</a> <a href="https://t.co/sipek5
usa8">pic.twitter.com/sipek5usa8</a></p>&mdash; Dr Sam Willis (@DrSamWil
lis)
Vader label: positive
Gold label: neutral
-----------------------------
Tweet: #Flash</a> China has given a fresh loan of USD 700 million to Pak
istan at the rate of 8.9%. Two railway stations of Pakistan (Lahore &am
p; Sukkur) have been taken by China as security for 99 years or till the
full and final payment of this loan, which is earlier. (Sources)</p>&mda
sh; Baba Banaras™ (@RealBababanaras)
Vader label: positive
Gold label: neutral
-----------------------------
Tweet: The Anti-Counterfeit Authority (ACA) has released goods worth Sh5
0 million that were seized at China Square.<br><br>The quick return of t
he goods comes a day after the Chinese embassy urged the Kenyan governme
nt to intervene to protect Chinese enterprises and citizens.<br><br>— Na
tion</p>&mdash; Moe (@moneyacademyKE)
Vader label: positive
Gold label: neutral
-----------------------------
Tweet: My latest for <a href="https://twitter.com/dw_hotspotasia?ref_src
=twsrc%5Etfw">@dw_hotspotasia</a>: As <a href="https://twitter.com/hasht
ag/China?src=hash&amp;ref_src=twsrc%5Etfw">#China</a>&#39;s rubber-stamp
parliament gathers in Beijing this weekend, President Xi Jinping is expe
cted to officially kick off his third term. China&#39;s Communist party
will likely initiate further institutional reform. <a href="https://t.c
o/8lbe9CJ2SO">https://t.co/8lbe9CJ2SO</a></p>&mdash; William Yang (@Will
iamYang120)
Vader label: positive
Gold label: neutral
-----------------------------
Tweet: In meeting with Saudi FM Prince Faisal bin Farhan Al Saud, Chines
e FM <a href="https://twitter.com/hashtag/QinGang?src=hash&amp;ref_src=t
wsrc%5Etfw">#QinGang</a> said <a href="https://twitter.com/hashtag/Chin
a?src=hash&amp;ref_src=twsrc%5Etfw">#China</a> is ready to keep the posi
tive momentum of high-level exchanges with <a href="https://twitter.com/
hashtag/SaudiaArabia?src=hash&amp;ref_src=twsrc%5Etfw">#SaudiaArabia</a>
and work together to advance high-quality Belt and Road Cooperation. <a
href="https://t.co/4A5v9ouAxy">pic.twitter.com/4A5v9ouAxy</a></p>&mdash;
Liu Yongfeng (@liupheonix)
Vader label: positive
Gold label: neutral
-----------------------------
```

Question 3b Answer

Error Analysis on Neutral Tweets:

NOTE: since we found more than 10 negative missclassified tweets we'll try to explain the results for 10 of them.

- 30. Classified as positive due to the word "great" before grandfather.

- 38. Missclassified as positive due to words such as "fresh" that are positive in the lexicon.
- 41. Meaning depends on the context or possible missclassification as positive because of the use of word goods.
- 45. Missclassified as positive due to words that are positive in the lexicon maybe the word could be likely or the possible argument that "kick off" is a positive or "exciting" word.
- 48. Missclassified as positive due to words that are positive in the lexicon ("positive")

[4 points] Question 4:

Run VADER on the set of airline tweets with the following settings:

- Run VADER (as it is) on the set of airline tweets
- Run VADER on the set of airline tweets after having lemmatized the text
- Run VADER on the set of airline tweets with only adjectives
- Run VADER on the set of airline tweets with only adjectives and after having lemmatized the text
- Run VADER on the set of airline tweets with only nouns
- Run VADER on the set of airline tweets with only nouns and after having lemmatized the text
- Run VADER on the set of airline tweets with only verbs
- Run VADER on the set of airline tweets with only verbs and after having lemmatized the text
- [1 point] a. Generate for all separate experiments the classification report, i.e., Precision, Recall, and F_1 scores per category as well as micro and macro averages. **Use a different code cell (or multiple code cells) for each experiment.**
- [3 points] b. Compare the scores and explain what they tell you.
-
  - Does lemmatisation help? Explain why or why not.
-
  - Are all parts of speech equally important for sentiment analysis? Explain why or why not.

In [12]:
```
import pathlib
from sklearn.datasets import load_files
cwd = pathlib.Path.cwd()
airline_tweets_folder = cwd.joinpath('airlinetweets')
airline_tweets_train = load_files(str(airline_tweets_folder))
```

In [13]:

```
# run vader on the set of airline tweets
tweets = []
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


from sklearn.metrics import classification_report
print("VADER (as it is) on the set of airline tweets Classification Repor
print(classification_report(gold, all_vader_output))
```

```
VADER (as it is) on the set of airline tweets Classification Report
              precision    recall  f1-score   support

    negative       0.86      0.49      0.62        39
     neutral       0.79      0.63      0.70        30
    positive       0.52      0.90      0.66        31

    accuracy                           0.66       100
   macro avg       0.72      0.67      0.66       100
weighted avg       0.74      0.66      0.66       100
```

In [14]:

```
# run vader on the set of airline tweets after having lemmatized the text
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


print("VADER on the set of airline tweets after having lemmatized the tex
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets after having lemmatized the text Clas
sification Report
              precision    recall  f1-score   support

    negative       0.80      0.51      0.62        39
     neutral       0.74      0.57      0.64        30
    positive       0.54      0.90      0.67        31

    accuracy                           0.65       100
   macro avg       0.69      0.66      0.65       100
weighted avg       0.70      0.65      0.65       100
```

In [15]:

```
# run vader on the set of airline tweets with only adjectives
all_vader_output = []
gold = []
```

```
for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


print("VADER on the set of airline tweets with only adjectives Classifica
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only adjectives Classification R
eport
              precision    recall  f1-score   support

    negative       1.00      0.21      0.34        39
     neutral       0.39      0.93      0.55        30
    positive       0.80      0.52      0.63        31

    accuracy                           0.52       100
   macro avg       0.73      0.55      0.51       100
weighted avg       0.75      0.52      0.49       100
```

In [16]:

```
# run vader on the set of airline tweets with only adjectives and after h
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


print("VADER on the set of airline tweets with only adjectives and after 
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only adjectives and after having
lemmatized the text Classification Report
              precision    recall  f1-score   support

    negative       1.00      0.21      0.34        39
     neutral       0.39      0.93      0.55        30
    positive       0.80      0.52      0.63        31

    accuracy                           0.52       100
   macro avg       0.73      0.55      0.51       100
weighted avg       0.75      0.52      0.49       100
```

In [17]:

```
# run vader on the set of airline tweets with only nouns
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


print("VADER on the set of airline tweets with only nouns Classification 
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only nouns Classification Report
              precision    recall  f1-score   support

    negative       0.83      0.13      0.22        39
     neutral       0.35      0.87      0.50        30
    positive       0.45      0.29      0.35        31

    accuracy                           0.40       100
   macro avg       0.54      0.43      0.36       100
weighted avg       0.57      0.40      0.35       100
```

In [18]:

```
# run vader on the set of airline tweets with only nouns and after having
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta


print("VADER on the set of airline tweets with only nouns and after havin
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only nouns and after having lemm
atized the text Classification Report
              precision    recall  f1-score   support

    negative       0.83      0.13      0.22        39
     neutral       0.35      0.87      0.50        30
    positive       0.45      0.29      0.35        31

    accuracy                           0.40       100
   macro avg       0.54      0.43      0.36       100
weighted avg       0.57      0.40      0.35       100
```

In [19]:

```
# run vader on the set of airline tweets with only verbs
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta
print("VADER on the set of airline tweets with only verbs Classification
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only verbs Classification Report
              precision    recall  f1-score   support

    negative       0.93      0.33      0.49        39
     neutral       0.39      0.90      0.55        30
    positive       0.65      0.35      0.46        31

    accuracy                           0.51       100
   macro avg       0.66      0.53      0.50       100
weighted avg       0.68      0.51      0.50       100
```

In [20]:

```
# run vader on the set of airline tweets with only verbs and after having
all_vader_output = []
gold = []

for i in range(100):
    tweets.append(airline_tweets_train.data[i].decode('UTF-8'))
    vader_output = run_vader(airline_tweets_train.data[i].decode('UTF-8')
    vader_label = vader_output_to_label(vader_output)
    all_vader_output.append(vader_label)
    gold.append(airline_tweets_train.target_names[airline_tweets_train.ta

print("VADER on the set of airline tweets with only verbs and after havin
print(classification_report(gold, all_vader_output))
```

```
VADER on the set of airline tweets with only verbs and after having lemm
atized the text Classification Report
              precision    recall  f1-score   support

    negative       0.84      0.41      0.55        39
     neutral       0.37      0.83      0.52        30
    positive       0.79      0.35      0.49        31

    accuracy                           0.52       100
   macro avg       0.67      0.53      0.52       100
weighted avg       0.68      0.52      0.52       100
```

Question 4 Answer

If we compare the results of the first two experiments we can see that where all parts of speech are considered the difference between accuracy is minimal however precision for negative and neutral tweets are higher in this case without lemmatization. This is because lemmatization could be removing some of the context of the word and therefore the sentiment of the word. In negative tweet recall, positive tweet precision and the positive tweet f1-score the lemmatized data prodced higher scores but also only by amounts between 0.1-0.2. Lemmatization makes the most difference in scores when considering only the verb part of speech. One can assume becasue this removes the verbs conjugation. One can see the precision for negative tweets drops by almost 0.10 when the data is lemmatized, along with the precision and recall for neutral tweets. However, the F1-score for negative tweets and positive tweets increase after lemmatization so if one considers the weighted averages as a metric then the lemmatized data is marginally better.

In terms of the importance of the different parts of speech one could consider the accuracy and macro and weighted averages. When considering all parts of speech the overall accuracy is 0.66 and the weighted average is 0.74. When considering only verbs the accuracy is 0.52 and the weighted average is 0.68. When considering only nouns the accuracy is 0.40 and the weighted average is 0.57. When considering only adjectives the accuracy is 0.52 and the weighted average is 0.75. Therefore, one can see that the most important part of speech is the adjective. This is because the accuracy and weighted average are the highest besides when filtering for a part of speech. This is because adjectives are often used to describe the sentiment of a tweet. For example, if a tweet is positive it will often contain words such as "great" or "amazing". If a tweet is negative it will often contain words such as "terrible" or "awful". Therefore, adjectives are often used to describe the sentiment of a tweet and therefore are the most important part of speech for sentiment analysis. In some regards considering only adjectives performed better than all parts of speech but not in overall accuracy which is very interesting. What one could continue to do is consider parts of speech in combination with one another. For example, one could consider only nouns and adjectives or only verbs and adjectives. This could be interesting to see if the accuracy and weighted average increase or decrease to form a more concrete ranking of part of speech importance.

Part II: scikit-learn assignments

[4 points] Question 5

Train the scikit-learn classifier (Naive Bayes) using the airline tweets.

- Train the model on the airline tweets with 80% training and 20% test set and default settings (TF-IDF representation, min_df=2)
- Train with different settings:
  - with respect to vectorizing: TF-IDF ('airline_tfidf') vs. Bag of words representation ('airline_count')
  - with respect to the frequency threshold (min_df). Carry out experiments with increasing values for document frequency (min_df = 2; min_df = 5; min_df =10)

- [1 point] a. Generate a classification_report for all experiments
- [3 points] b. Look at the results of the experiments with the different settings and try to explain why they differ:
  - which category performs best, is this the case for any setting?
  - does the frequency threshold affect the scores? Why or why not according to you?

In [27]:

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from nltk.corpus import stopwords
import nltk
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransfo

airline_vec = CountVectorizer(min_df=2, # If a token appears fewer times 
                              tokenizer=nltk.word_tokenize, # we use the n
                              stop_words=stopwords.words('english')) # sto
#  bag of words representation of the airline tweets
airline_counts = airline_vec.fit_transform(airline_tweets_train.data)

docs_train, docs_test, y_train, y_test = train_test_split(
    airline_counts, # the bag of words representation of the tweets
    airline_tweets_train.target, # the category values for each tweet
    test_size = 0.20 # we use 80% for training and 20% for development
    )

clf = MultinomialNB().fit(docs_train, y_train)
y_pred = clf.predict(docs_test)

print("Classification report for the Naive Bayes classifier on the airlin
print(classification_report(y_test, y_pred, target_names=airline_tweets_t
```

```
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:528: UserWarning: The parameter 'token_pattern' will not
be used since 'tokenizer' is not None'
  warnings.warn(
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:409: UserWarning: Your stop_words may be inconsistent wi
th your preprocessing. Tokenizing the stop words generated tokens ["'d",
"'ll", "'re", "'s", "'ve", 'could', 'might', 'must', "n't", 'need', 'sh
a', 'wo', 'would'] not in stop_words.
  warnings.warn(
Classification report for the Naive Bayes classifier on the airline twee
ts with 80% training and 20% test set and default settings (Bag of words
representation, min_df=2)
              precision    recall  f1-score   support

    negative       0.85      0.89      0.87       339
     neutral       0.86      0.74      0.80       309
    positive       0.81      0.88      0.85       303

    accuracy                           0.84       951
   macro avg       0.84      0.84      0.84       951
weighted avg       0.84      0.84      0.84       951
```

In [22]:
```
# TF-IDF representation of the airline tweets
tfidf_transformer = TfidfTransformer()
airline_tfidf = tfidf_transformer.fit_transform(airline_counts)
docs_train2, docs_test2, y_train2, y_test2 = train_test_split(
    airline_tfidf, # the tf-idf model
    airline_tweets_train.target, # the category values for each tweet
    test_size = 0.20 # we use 80% for training and 20% for development
    )
clf2 = MultinomialNB().fit(docs_train2, y_train2)
y_pred2 = clf2.predict(docs_test2)

print("Classification report for the Naive Bayes classifier on the airlin
print(classification_report(y_test2, y_pred2, target_names=airline_tweets
Classification report for the Naive Bayes classifier on the airline twee
ts with 80% training and 20% test set and default settings (TF-IDF repre
sentation, min_df=2)
              precision    recall  f1-score   support

    negative       0.80      0.89      0.84       345
     neutral       0.84      0.65      0.73       313
    positive       0.79      0.87      0.83       293

    accuracy                           0.80       951
   macro avg       0.81      0.80      0.80       951
weighted avg       0.81      0.80      0.80       951
```

In [23]:

```
# TF-IDF representation of the airline tweets with min_df=5
airline_vec = CountVectorizer(min_df=5, # If a token appears fewer times
                             tokenizer=nltk.word_tokenize, # we use the n
                             stop_words=stopwords.words('english')) # sto
#  bag of words representation of the airline tweets
airline_counts = airline_vec.fit_transform(airline_tweets_train.data)

tfidf_transformer = TfidfTransformer()
airline_tfidf = tfidf_transformer.fit_transform(airline_counts)
docs_train3, docs_test3, y_train3, y_test3 = train_test_split(
    airline_tfidf, # the tf-idf model
    airline_tweets_train.target, # the category values for each tweet
    test_size = 0.20 # we use 80% for training and 20% for development
    )
clf3 = MultinomialNB().fit(docs_train3, y_train3)
y_pred3 = clf3.predict(docs_test3)

print("Classification report for the Naive Bayes classifier on the airlin
print(classification_report(y_test3, y_pred3, target_names=airline_tweets
```

```
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:528: UserWarning: The parameter 'token_pattern' will not
be used since 'tokenizer' is not None'
  warnings.warn(
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:409: UserWarning: Your stop_words may be inconsistent wi
th your preprocessing. Tokenizing the stop words generated tokens ["'d",
"'ll", "'re", "'s", "'ve", 'could', 'might', 'must', "n't", 'need', 'sh
a', 'wo', 'would'] not in stop_words.
  warnings.warn(
Classification report for the Naive Bayes classifier on the airline twee
ts with 80% training and 20% test set and default settings (TF-IDF repre
sentation, min_df=5)
              precision    recall  f1-score   support

    negative       0.79      0.91      0.85       339
     neutral       0.81      0.70      0.75       316
    positive       0.86      0.84      0.85       296

    accuracy                           0.82       951
   macro avg       0.82      0.82      0.82       951
weighted avg       0.82      0.82      0.82       951
```

In [24]:

```
# TF-IDF representation of the airline tweets with min_df=10
airline_vec = CountVectorizer(min_df=10, # If a token appears fewer times
                              tokenizer=nltk.word_tokenize, # we use the n
                              stop_words=stopwords.words('english')) # sto
#  bag of words representation of the airline tweets
airline_counts = airline_vec.fit_transform(airline_tweets_train.data)

tfidf_transformer = TfidfTransformer()
airline_tfidf = tfidf_transformer.fit_transform(airline_counts)
docs_train4, docs_test4, y_train4, y_test4 = train_test_split(
    airline_tfidf, # the tf-idf model
    airline_tweets_train.target, # the category values for each tweet
    test_size = 0.20 # we use 80% for training and 20% for development
    )
clf4 = MultinomialNB().fit(docs_train4, y_train4)
y_pred4 = clf4.predict(docs_test4)

print("Classification report for the Naive Bayes classifier on the airlin
print(classification_report(y_test4, y_pred4, target_names=airline_tweets
```

```
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:528: UserWarning: The parameter 'token_pattern' will not
be used since 'tokenizer' is not None'
  warnings.warn(
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:409: UserWarning: Your stop_words may be inconsistent wi
th your preprocessing. Tokenizing the stop words generated tokens ["'d",
"'ll", "'re", "'s", "'ve", 'could', 'might', 'must', "n't", 'need', 'sh
a', 'wo', 'would'] not in stop_words.
  warnings.warn(
```

```
Classification report for the Naive Bayes classifier on the airline twee
ts with 80% training and 20% test set and default settings (TF-IDF repre
sentation, min_df=10)
              precision    recall  f1-score   support

    negative       0.85      0.89      0.87       349
     neutral       0.83      0.73      0.78       327
    positive       0.79      0.85      0.82       275

    accuracy                           0.83       951
   macro avg       0.83      0.83      0.82       951
weighted avg       0.83      0.83      0.83       951
```

Question 5 Answer:

When comparing the two tweet prepresentations (bag of words and TF-IDF) we see that suprisingly the bag of word representation performs better. One woudl expect the TF-IDF representaiton to perform better because the TF-IDF representation takes into account the frequency and importance of the words in the tweets. However, one can see as the frequency threshold increases so does the accuracy of the sentiment analysis for the tweets in TF-IDF. Perhaps if we would continue increasing this threshold TF-IDF would be more effective than the bag of words representation. Bag of words ended with an accuracy of 0.84 while TF-IDF ended with an accuracy of 0.80. This is a difference of 0.04. This is not a large difference but it is still a difference when both had a frequency threshold of 2. As we increased the frequency threshold from 2 to 5 to 10 the accuracy increased from 0.80 to 0.82 and then to 0.83 which is comparable to the bag of words representation accuracy. The scores seem to be at least a bit effected by the frequency threshold but not very significantly. For further investigation one should most like increase the frequency threshold to see if the TF-IDF representation would outperform the bag of words representation and increase the frequency with bag of words and see what happens.

[4 points] Question 6: Inspecting the best scoring features

- Train the scikit-learn classifier (Naive Bayes) model with the following settings (airline tweets 80% training and 20% test; Bag of words representation ('airline_count'), min_df=2)

- [1 point] a. Generate the list of best scoring features per class (see function **important_features_per_class** below) [1 point]
- [3 points] b. Look at the lists and consider the following issues:
  - [1 point] Which features did you expect for each separate class and why?
  - [1 point] Which features did you not expect and why ?
  - [1 point] The list contains all kinds of words such as names of airlines, punctuation, numbers and content words (e.g., 'delay' and 'bad'). Which words would you remove or keep when trying to improve the model and why?

In [25]:
```
def important_features_per_class(vectorizer,classifier,n=80):
    class_labels = classifier.classes_
    feature_names =vectorizer.get_feature_names_out()
    topn_class1 = sorted(zip(classifier.feature_count_[0], feature_names)
    topn_class2 = sorted(zip(classifier.feature_count_[1], feature_names)
    topn_class3 = sorted(zip(classifier.feature_count_[2], feature_names)
    print("Important words in negative documents")
    for coef, feat in topn_class1:
        print(class_labels[0], coef, feat)
    print("-----------------------------------------")
    print("Important words in neutral documents")
    for coef, feat in topn_class2:
        print(class_labels[1], coef, feat)
    print("-----------------------------------------")
    print("Important words in positive documents")
    for coef, feat in topn_class3:
        print(class_labels[2], coef, feat)

# example of how to call from notebook:

airline_vec = CountVectorizer(min_df=2, # If a token appears fewer times
                             tokenizer=nltk.word_tokenize, # we use the n
                             stop_words=stopwords.words('english')) # sto
#  bag of words representation of the airline tweets
airline_counts = airline_vec.fit_transform(airline_tweets_train.data)

docs_train, docs_test, y_train, y_test = train_test_split(
    airline_counts, # the bag of words model
    airline_tweets_train.target, # the category values for each tweet
    test_size = 0.20 # we use 80% for training and 20% for development
    )

clf = MultinomialNB().fit(docs_train, y_train)
```

```
y_pred = clf.predict(docs_test)
important_features_per_class(airline_vec, clf)
```

```
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:528: UserWarning: The parameter 'token_pattern' will not
be used since 'tokenizer' is not None'
  warnings.warn(
/Users/bella/TextMining/lib/python3.10/site-packages/sklearn/feature_ext
raction/text.py:409: UserWarning: Your stop_words may be inconsistent wi
th your preprocessing. Tokenizing the stop words generated tokens ["'d",
"'ll", "'re", "'s", "'ve", 'could', 'might', 'must', "n't", 'need', 'sh
a', 'wo', 'would'] not in stop_words.
  warnings.warn(
Important words in negative documents
0 1521.0 @
0 1401.0 united
0 1264.0 .
0 426.0 ``
0 407.0 flight
0 385.0 ?
0 373.0 !
0 311.0 #
0 230.0 n't
0 159.0 ''
0 138.0 's
0 117.0 service
0 104.0 virginamerica
0 100.0 :
0 98.0 get
0 96.0 customer
0 95.0 cancelled
0 91.0 delayed
0 91.0 bag
0 80.0 time
0 79.0 plane
0 79.0 'm
0 74.0 hours
0 74.0 ...
0 69.0 still
0 68.0 -
0 66.0 gate
0 66.0 ;
0 65.0 http
0 65.0 hour
0 64.0 late
0 64.0 airline
0 61.0 would
0 59.0 &
0 56.0 help
0 54.0 one
0 54.0 2
0 53.0 delay
0 53.0 ca
0 53.0 amp
0 52.0 like
0 50.0 $
0 49.0 worst
0 47.0 flights
0 46.0 waiting
0 46.0 never
0 45.0 flightled
0 44.0 us
0 43.0 fly
0 43.0 3
0 42.0 've
0 40.0 wait
0 39.0 really
0 39.0 lost
0 39.0 ever
0 39.0 (
0 38.0 back
0 37.0 thanks
0 37.0 due
0 37.0 bags
0 36.0 u
0 36.0 check
0 35.0 ticket
0 35.0 day
0 34.0 trying
0 34.0 seat
0 34.0 people
0 34.0 )
0 33.0 crew
0 33.0 another
0 32.0 luggage
0 32.0 even
0 32.0 airport
0 32.0 4
0 31.0 problems
0 30.0 staff
0 30.0 seats
0 29.0 last
0 28.0 today
0 28.0 phone
-----------------------------------------
Important words in neutral documents
1 1406.0 @
1 498.0 ?
1 490.0 .
1 313.0 jetblue
1 294.0 :
1 258.0 southwestair
1 253.0 united
1 253.0 ``
1 239.0 flight
1 220.0 #
1 191.0 americanair
1 186.0 http
1 177.0 !
1 160.0 usairways
1 133.0 's
1 87.0 get
1 77.0 virginamerica
1 77.0 ''
1 71.0 -
1 63.0 flights
1 62.0 please
1 62.0 )
1 54.0 need
1 53.0 (
1 52.0 help
1 49.0 n't
1 45.0 dm
1 43.0 would
1 43.0 ;
1 41.0 us
1 40.0 ...
1 38.0 ”
1 37.0 “
1 36.0 fleet
1 36.0 fleek
1 36.0 &
1 35.0 tomorrow
```

```
1 35.0 flying
1 34.0 way
1 34.0 hi
1 34.0 'm
1 33.0 thanks
1 33.0 know
1 30.0 change
1 30.0 cancelled
1 29.0 one
1 29.0 number
1 29.0 like
1 27.0 fly
1 26.0 time
1 26.0 could
1 26.0 amp
1 25.0 today
1 25.0 check
1 24.0 new
1 23.0 travel
1 23.0 see
1 23.0 guys
1 23.0 destinationdragons
1 23.0 airport
1 22.0 go
1 20.0 tickets
1 20.0 sent
1 20.0 next
1 20.0 going
1 20.0 back
1 19.0 use
1 19.0 ceo
1 18.0 want
1 18.0 ticket
1 18.0 follow
1 18.0 add
1 18.0 2
1 17.0 weather
1 17.0 trying
1 17.0 question
1 17.0 passengers
1 17.0 make
1 16.0 start
1 16.0 service
-----------------------------------------
Important words in positive documents
2 1324.0 @
2 1046.0 !
2 772.0 .
2 310.0 #
2 302.0 southwestair
2 283.0 thanks
2 283.0 jetblue
2 245.0 united
2 242.0 thank
2 227.0 ``
2 174.0 americanair
2 169.0 flight
2 168.0 :
2 136.0 usairways
2 133.0 great
```

```
2 94.0 )
2 87.0 service
2 75.0 virginamerica
2 73.0 guys
2 72.0 http
2 70.0 love
2 66.0 much
2 65.0 best
2 64.0 's
2 59.0 awesome
2 59.0 ;
2 58.0 customer
2 52.0 –
2 48.0 time
2 48.0 amazing
2 47.0 good
2 42.0 got
2 41.0 n't
2 41.0 airline
2 40.0 us
2 40.0 help
2 39.0 &
2 36.0 get
2 36.0 ...
2 35.0 crew
2 34.0 today
2 34.0 gate
2 33.0 appreciate
2 33.0 amp
2 31.0 fly
2 30.0 ''
2 28.0 see
2 28.0 home
2 27.0 made
2 27.0 flying
2 26.0 response
2 26.0 first
2 26.0 ever
2 26.0 'm
2 25.0 work
2 25.0 back
2 25.0 always
2 25.0 (
2 24.0 new
2 24.0 like
2 23.0 well
2 23.0 tonight
2 23.0 nice
2 23.0 day
2 22.0 would
2 22.0 u
2 22.0 team
2 22.0 ?
2 22.0 'll
2 21.0 southwest
2 21.0 know
2 21.0 job
2 21.0 flights
2 20.0 yes
2 20.0 please
```

```
2 19.0 plane
2 19.0 follow
2 19.0 agent
2 18.0 staff
2 18.0 really
```

Question 6 Answer:

which features did you expect for each separate class and why? Which features did you not expect and why ? The list contains all kinds of words such as names of airlines, punctuation, numbers and content words (e.g., 'delay' and 'bad'). Which words would you remove or keep when trying to improve the model and why?

In the section important words in negative documents we see a high feature count of punctuation and the tweet @ symbol and some expected airplane related words like united(the airline) and flight. What one sees that is also to be expected in the negative is the contraction n't, a negation and words like cancelled, delayed, late which are flight and domain negative concepts. We also see negative adjectives and adverbs like worst, and never. What one would not usually xpect in this section is the "0 37.0 thanks" we see but that could be because thanks can be used in a condescending, sarcastic or negative tone such as "this was the worst service thanks to incompetent staff" or something. In the section important words in positive tweets we see also a lot of punctuation and the @ symbol. Interestingly like the negative tweet sections thanks is one of the words high in feature count but unsurprisingly with a much higher count than in the negetive tweet section, almost 10 times higher. We also see positive adjectives and adverbs like great, best, and good. We also see positive words like love, thanks, and yes. These are all words that are used in a positive context and were thus to be expected. In this section there are not really any words that do not fit expectations. In the section important words in neutral tweets we see a lot of punctuation and the @ symbol. We also see words like flight, united, jetblue and other airline related words. These are all words that are used in a neutral context and were thus to be expected. However, in this section we can also see some words that one would expect more in the other two sections such as cancelled, and thanks. These words are not necessarily neutral but they could possibly be used in a neutral context and are not counted as high as in the negative and positve sections. For example, "I cancelled my flight because it was delayed" is a negative tweet but "I cancelled my flight because I had to go to the hospital" is a neutral tweet. The list contains all kinds of words such as names of airlines, punctuation, numbers and content words and some can definitely be removed without too much effect we believe. For instance, all the punctuation and @ symbol specifically that are very high in count in all three sections would most likely make no difference to the sentiment analysis if removed. Possibly with the only exception being ! which can be used to express emotion arguably more than a lot of other punctuation. We also believe that the numbers could be removed as they are not really words and are not really used in a context that would be relevant to sentiment analysis. We also believe that possibly if we are not searching for sentiments in regard to or in connection with specific airlines the names of airlines could be removed as they are not as relevant and appear in all three sections. One must however be careful about removing contextual flight related words that have sentiment attached like cancelled.

[Optional! (will not be graded)] Question 7

Train the model on airline tweets and test it on your own set of tweets

- Train the model with the following settings (airline tweets 80% training and 20% test; Bag of words representation ('airline_count'), min_df=2)
- Apply the model on your own set of tweets and generate the classification report

- [1 point] a. Carry out a quantitative analysis.
- [1 point] b. Carry out an error analysis on 10 correctly and 10 incorrectly classified tweets and discuss them
- [2 points] c. Compare the results (cf. classification report) with the results obtained by VADER on the same tweets and discuss the differences.

[Optional! (will not be graded)] Question 8: trying to improve the model

- [2 points] a. Think of some ways to improve the scikit-learn Naive Bayes model by playing with the settings or applying linguistic preprocessing (e.g., by filtering on part-of-speech, or removing punctuation). Do not change the classifier but continue using the Naive Bayes classifier. Explain what the effects might be of these other settings

- [1 point] b. Apply the model with at least one new setting (train on the airline tweets using 80% training, 20% test) and generate the scores

- [1 point] c. Discuss whether the model achieved what you expected.

End of this notebook

```
In [ ]:
```