Political Compass AI

"Fake News," Lies and Propaganda: How to Sort Fact from Fiction

The Liberal Media

Does the media have a liberal bias?

Here's one research-based answer to the question of liberal bias:

The documentary *The Myth of the Liberal Media: The Propaganda Model of News* uses empirical evidence to look at ownership of the mainstream news media, filters that affect what news gets published, and examples of actual news coverage in order to show that conservative political and corporate interests significantly shape news coverage in the United States.

The myth of the liberal media : the propaganda model of news
by Media Education Foundation
Publication Date: 1997
Edward Herman and Noam

New Sources on the Political Spectrum

What news sources are left-leaning, centrist, or right-leaning?

There is no completely clear answer to this question because there is no one exact methodology to measure and rate the partisan bias of news sources.

Here are a couple of resources that can help:

- **AllSides**
  All Sides is a news website that presents multiple sources side by side in order to provide the full scope of news reporting.

The Allsides Bias Ratings page allows you to filter a list of news sources by bias (left, center, right).

AllSides uses a patented bias rating system to classify news sources as left, center, or right leaning. Components of the rating system include crowd-sourcing, surveys, internal research, and use of third party sources such as Wikipedia and research conducted by Groseclose and Milyo at UCLA. Note that while the Groseclose & Milyo results are popular, the methodology it is not without critique.

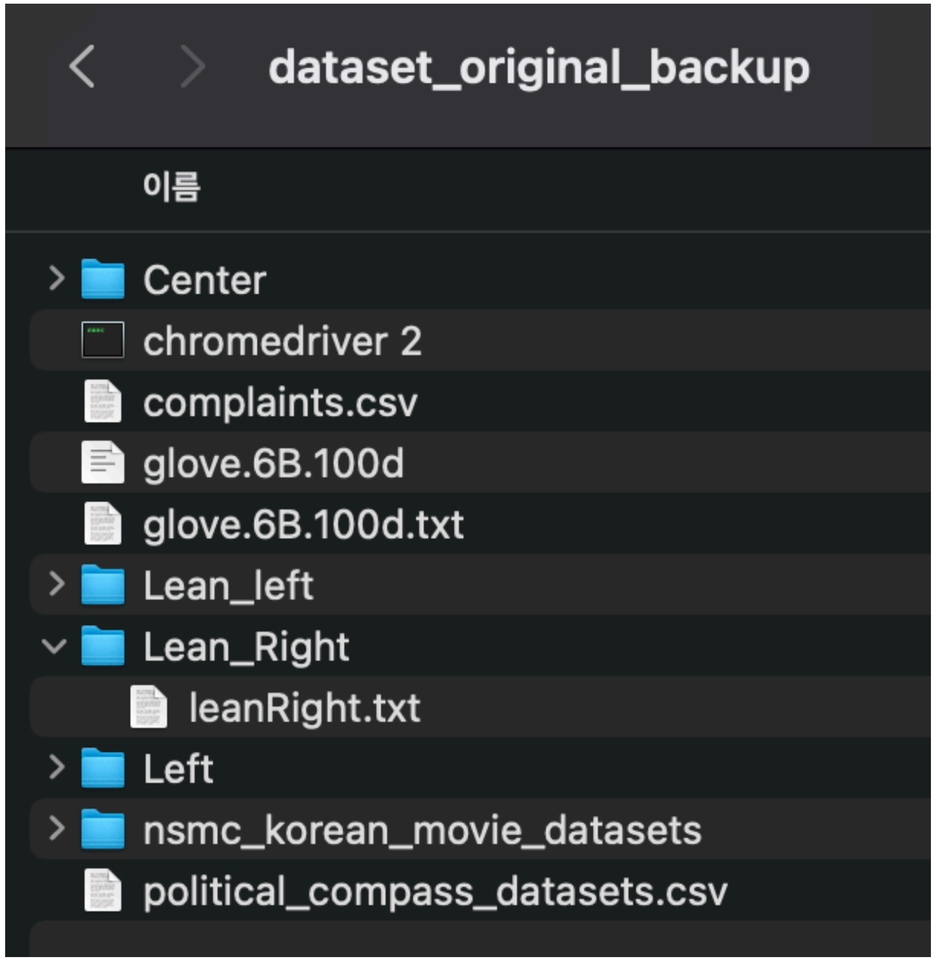
- **Pew Research Center - Political Polarization**
  Survey data reveals the news source favored by people according to their political beliefs.

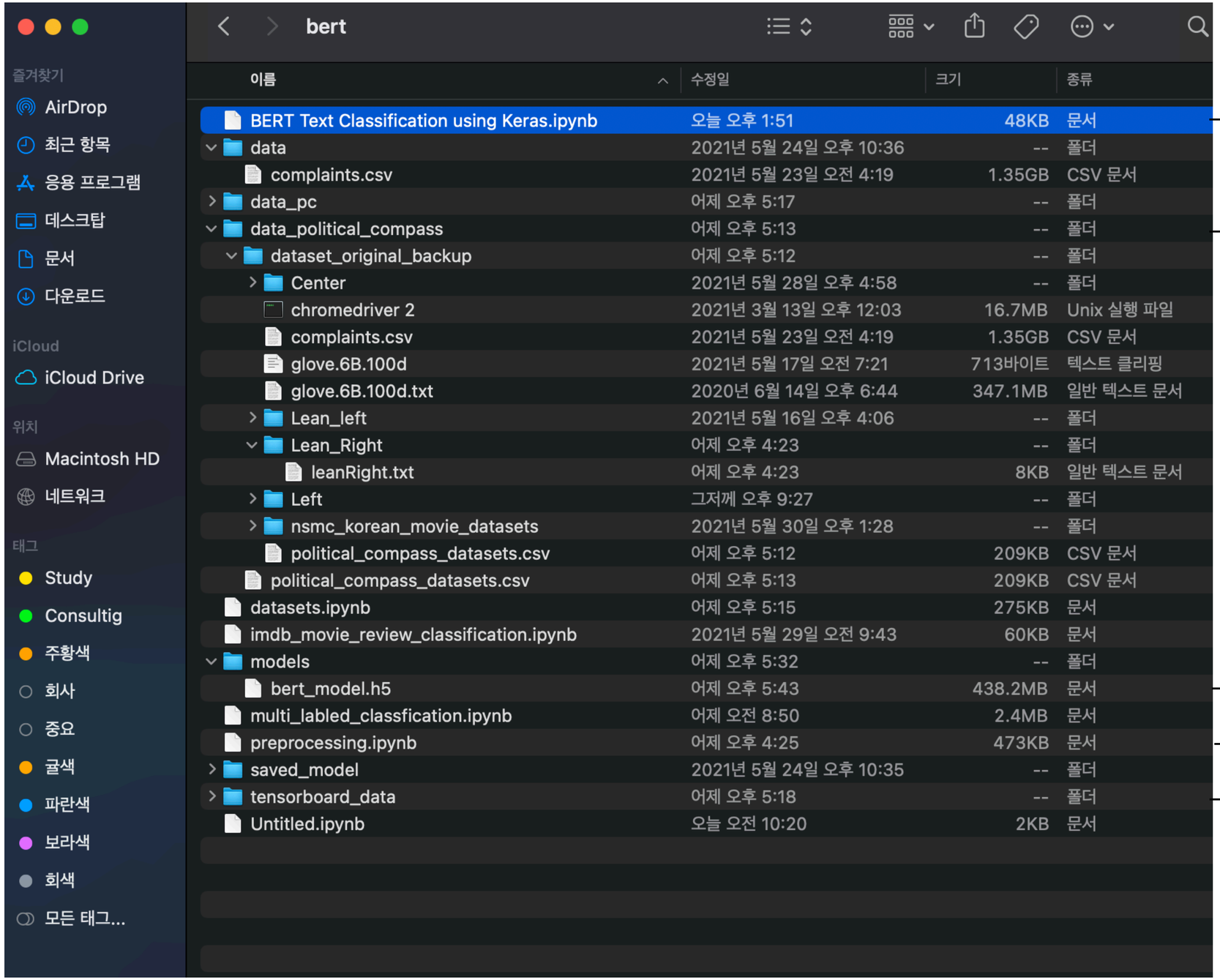
A report based on a 2014 survey shows which news sources are used and considered trustworthy based on individual's political values (liberal or conservative). Note that this report measures the political leanings of the audience rather than the source itself.

Ideological Placement of Each Source's Audience

Average ideological placement on a 10-point scale of ideological consistency of those who got news from each source in the past week...

https://www.allsides.com/media-bias/media-bias-ratings




bert
즐겨찾기
AirDrop
최근 항목
응용 프로그램
데스크탑
문서
다운로드
iCloud
iCloud Drive
위치
Macintosh HD
네트워크
태그
Study
Consultig
주황색
회사
중요
귤색
파란색
보라색
회색
모든 태그...
이름
수정일
크기
종류
BERT Text Classification using Keras.ipynb 오늘 오후 1:51 48KB 문서
data 2021년 5월 24일 오후 10:36 -- 폴더
complaints.csv 2021년 5월 23일 오전 4:19 1.35GB CSV 문서
data_pc 어제 오후 5:17 -- 폴더
data_political_compass 어제 오후 5:13 -- 폴더
dataset_original_backup 어제 오후 5:12 -- 폴더
Center 2021년 5월 28일 오후 4:58 -- 폴더
chromedriver 2 2021년 3월 13일 오후 12:03 16.7MB Unix 실행 파일
complaints.csv 2021년 5월 23일 오전 4:19 1.35GB CSV 문서
glove.6B.100d 2021년 5월 17일 오전 7:21 713바이트 텍스트 클리핑
glove.6B.100d.txt 2020년 6월 14일 오후 6:44 347.1MB 일반 텍스트 문서
Lean_left 2021년 5월 16일 오후 4:06 -- 폴더
Lean_Right 어제 오후 4:23 -- 폴더
leanRight.txt 어제 오후 4:23 8KB 일반 텍스트 문서
Left 그저께 오후 9:27 -- 폴더
nsmc_korean_movie_datasets 2021년 5월 30일 오후 1:28 -- 폴더
political_compass_datasets.csv 어제 오후 5:12 209KB CSV 문서
political_compass_datasets.csv 어제 오후 5:13 209KB CSV 문서
datasets.ipynb 어제 오후 5:15 275KB 문서
imdb_movie_review_classification.ipynb 2021년 5월 29일 오전 9:43 60KB 문서
models 어제 오후 5:32 -- 폴더
bert_model.h5 어제 오후 5:43 438.2MB 문서
multi_labled_classfication.ipynb 어제 오전 8:50 2.4MB 문서
preprocessing.ipynb 어제 오후 4:25 473KB 문서
saved_model 2021년 5월 24일 오후 10:35 -- 폴더
tensorboard_data 어제 오후 5:18 -- 폴더
Untitled.ipynb 오늘 오전 10:20 2KB 문서
핵심개발 코드
데이터셋
학습결과 AI model
데이터 전치리 코드
학습과정 모니터링

pip install -r requirements.txt

jupyter

Files | Running | Clusters

Select items to perform actions on them.

0 / Documents / project / AI / BertMultiClassification / bert_finish

Na

- ..
- data
- data_pc
- data_political_compass
- models
- saved_model
- tensorboard_data
- BERT Text Classification using Keras.ipynb
- datasets.ipynb
- preprocessing.ipynb
- requirements.txt

jupyter requirements.txt a minute ago

Logout

File | Edit | View | Language | Plain Text

```
absl-py==0.12.0
appnope @ file:///opt/concourse/worker/volumes/live/4f734db2-9ca8-4d8b-5b29-6ca15b4b4772/volume/appnope_1606859466979/work
argon2-cffi @ file:///opt/concourse/worker/volumes/live/4afd07c8-7fc3-4a09-6326-d8c70269eb33/volume/argon2-cffi_1613037490059/work
astunparse==1.6.3
async-generator==1.10
attrs @ file:///tmp/build/80754af9/attrs_1620827162558/work
backcall @ file:///home/ktietz/src/ci/backcall_1611930011877/work
bleach @ file:///tmp/build/80754af9/bleach_1612211392645/work
cached-property==1.5.2
cachetools==4.2.2
certifi==2020.12.5
cffi @ file:///opt/concourse/worker/volumes/live/ca3c0fea-9d0f-4489-6889-89b42078ed37/volume/cffi_1613246931142/work
chardet==4.0.0
click==8.0.1
cycler==0.10.0
decorator @ file:///tmp/build/80754af9/decorator_1621259047763/work
defusedxml @ file:///tmp/build/80754af9/defusedxml_1615228127516/work
entrypoints==0.3
filelock==3.0.12
flatbuffers==1.12
gast==0.4.0
google-auth==1.30.0
google-auth-oauthlib==0.4.4
google-pasta==0.2.0
grpcio==1.34.1
h5py==3.1.0
huggingface-hub==0.0.8
idna==2.10
importlib-metadata @ file:///opt/concourse/worker/volumes/live/4e25a0be-45cc-4b73-4314-4604f00e30a4/volume/importlib-metadata_1617877365451/work
ipykernel @ file:///opt/concourse/worker/volumes/live/73e8766c-12c3-4f76-62a6-3dea9a7da5b7/volume/ipykernel_1596206701501/work/dist/ipykernel-5.3.4-py3-none-any.whl
ipython @ file:///opt/concourse/worker/volumes/live/f33fa11b-d908-43e8-693a-f4b58d8c695c/volume/ipython_1617120878195/work
ipython-genutils @ file:///tmp/build/80754af9/ipython_genutils_1606773439826/work
ipywidgets==7.6.3
jedi==0.17.0
Jinja2 @ file:///tmp/build/80754af9/jinja2_1621238361758/work
joblib==1.0.1
jsonschema @ file:///tmp/build/80754af9/jsonschema_1602607155483/work
jupyter-client @ file:///tmp/build/80754af9/jupyter_client_1616770841739/work
jupyter-core @ file:///opt/concourse/worker/volumes/live/a699b83f-e941-4170-5136-bf87e3f37756/volume/jupyter_core_1612213304212/work
jupyterlab-pygments @ file:///tmp/build/80754af9/jupyterlab_pygments_1601490720602/work
jupyterlab-widgets==1.0.0
keras-nightly==2.5.0.dev2021032900
Keras-Preprocessing==1.1.2
kiwisolver==1.3.1
Markdown==3.3.4
MarkupSafe @ file:///opt/concourse/worker/volumes/live/2165c944-5d03-464f-57c1-1df4e51b3ac3/volume/markupsafe_1621528146825/work
matplotlib==3.4.2
mistune==0.8.4
nbclient @ file:///tmp/build/80754af9/nbclient_1614364831625/work
nbconvert @ file:///opt/concourse/worker/volumes/live/d4b0787b-b6c8-4d28-5453-3381885d5b33/volume/nbconvert_1601914848300/work
nbformat @ file:///tmp/build/80754af9/nbformat_1617383369282/work
```

jupyter **datasets** Last Checkpoint: 20시간 전 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [18]:
```
from pandas import DataFrame
```

In [19]:
```
center = ["ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT\n  Mounting trade friction between the\n  U.S. And Japan
 'They told Reuter correspondents in Asian capitals a U.S.\n  Move against Japan might boost protectionist sentiment
 "But some exporters said that while the conflict would hurt\n  them in the long-run, in the short-term Tokyo's loss
 "The U.S. Has said it will impose 300 mln dlrs of tariffs on\n  imports of Japanese electronics goods on April 17,
 'Unofficial Japanese estimates put the impact of the tariffs\n  at 10 billion dlrs and spokesmen for major electron
 '"We wouldn\'t be able to do business," said a spokesman for\n  leading Japanese electronics firm Matsushita Electr
 '"If the tariffs remain in place for any length of time\n  beyond a few months it will mean the complete erosion of
 'In Taiwan, businessmen and officials are also worried.',
 '"We are aware of the seriousness of the U.S.',
 'Threat against\n  Japan because it serves as a warning to us," said a senior\n  Taiwanese trade official who asked
 'Taiwan had a trade trade surplus of 15.6 billion dlrs last\n  year, 95 pct of it with the U.S.',
 "The surplus helped swell Taiwan's foreign exchange reserves\n  to 53 billion dlrs, among the world's largest.",
 '"We must quickly open our markets, remove trade barriers and\n  cut import tariffs to allow imports of U.S. Produc
 'Retaliation," said\n  Paul Sheen, chairman of textile exporters &lt;Taiwan Safe Group>.',
 "A senior official of South Korea's trade promotion\n  association said the trade dispute between the U.S. And Japa
 'Last year South Korea had a trade surplus of 7.1 billion\n  dlrs with the U.S., Up from 4.9 billion dlrs in 1985.'
 'In Malaysia, trade officers and businessmen said tough\n  curbs against Japan might allow hard-hit producers of\n
 'In Hong Kong, where newspapers have alleged Japan has been\n  selling below-cost semiconductors, some electronics\
 'But other businessmen said such\n  a short-term commercial advantage would be outweighed by\n  further U.S. Pressu
 '"That is a very short-term view," said Lawrence Mills,\n  director-general of the Federation of Hong Kong Industry
 '"If the whole purpose is to prevent imports, one day it will\n  be extended to other sources.',
 'Much more serious for Hong Kong\n  is the disadvantage of action restraining trade," he said.',
 "The U.S. Last year was Hong Kong's biggest export market,\n  accounting for over 30 pct of domestically produced e
 'The Australian government is awaiting the outcome of trade\n  talks between the U.S. And Japan with interest and c
 '"This kind of deterioration in trade relations between two\n  countries which are major trading partners of ours i
 "He said Australia's concerns centred on coal and beef,\n  Australia's two largest exports to Japan and also signif
 'Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the\n  trade stand-off continue.',
 "Japan's ruling Liberal Democratic Party yesterday outlined\n  a package of economic measures to boost the Japanese
 'The measures proposed include a large supplementary budget\n  and record public works spending in the first half o
 "They also call for stepped-up spending as an emergency\n  measure to stimulate the economy despite Prime Minister\
 "Deputy U.S. Trade Representative Michael Smith and Makoto\n  Kuroda, Japan's deputy minister of International Trad
 'VIEILLE MONTAGNE SAYS 1986 CONDITIONS UNFAVOURABLE\n  A sharp fall in the dollar price of\n  zinc and the deprecia
 'It said in a statement that the two factors led to a\n  squeeze on refining margins and an 18.24 pct fall in sales
 'Vieille Montagne, which is actively pursuing a\n  restructuring program, reported a 198 mln franc net loss, after\
 'BANKERS TRUST &lt;BT> PUTS BRAZIL ON NON-ACCRUAL\n  Bankers Trust New York Corp said it has\n  placed its approxim
 'Brazil suspended interest payments on its 68 billion dlrs\n  of medium- and long-term debt on February 22.',
 'U.S. banking regulations do not require banks to stop\n  accruing interest on loans until payments are 90 days ove
 'Assuming no cash payments at current interest rates are\n  received for the rest of 1987, Bankers Trust estimated
```

In [20]:
```
center_df = DataFrame(center, columns=['text'])
center_df
```

Out[20]:

	text
0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...
1	They told Reuter correspondents in Asian capit...
2	But some exporters said that while the conflic...
3	The U.S. Has said it will impose 300 mln dlrs ...
4	Unofficial Japanese estimates put the impact o...
...	...
59	cash flow from its low growth tobacco, but the...
60	It can\n therefore use its funds to make acqu...
61	Analysts said the\n National Distillers' spir...
62	"The distilled spirits business has been in a ...
63	REUTER...

64 rows × 1 columns

In [21]:
```
# center lable = 0
center_df['label'] = 'center'
center_df
```

Out[21]:

	text	label
0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	center
1	They told Reuter correspondents in Asian capit...	center
2	But some exporters said that while the conflic...	center
3	The U.S. Has said it will impose 300 mln dlrs ...	center
4	Unofficial Japanese estimates put the impact o...	center
...	...	...
59	cash flow from its low growth tobacco, but the...	center
60	It can\n therefore use its funds to make acqu...	center
61	Analysts said the\n National Distillers' spir...	center
62	"The distilled spirits business has been in a ...	center
63	REUTER...	center

64 rows × 2 columns

In [22]:
```
lean_left = ["The impeachment inquiry picks up tomorrow where it left off Friday, with a subpoena sent to a White Ho
 "Just this morning, the lawyer for the whistleblower, whose complaint is at the base of this inquiry, says he's rep
 "There's are a lot of moving parts.",
 "Fortunately, NPR's Mara Liasson is here to help.",
 'Good morning.',
 'Good morning, Lulu.',
 'All right.',
 "What's the latest?",
 'Well, the latest is that the lawyer for the first whistleblower is tweeting that he is now representing a second w
 'The first whistleblower only had second and thirdhand knowledge.',
 'We also know that there are subpoenas for White House documents and State Department documents and personnel.',
 "We don't know how cooperative the administration will be with those
 'We know that the former ambassador to Ukraine, Marie Yovanovitch, is
 "She's was removed because the president's personal lawyer Rudy Giuli
 'And we that the latest bombshell - the text messages between the thr
 'One of the people in those exchanges, Kurt Volker, has already testi
 'A second one, Gordon Sondland, is scheduled to testify on Tuesday.',
 'So lots and lots of action here.',
 'I mean, another whistleblower has got to be bad news for President T
 'It seems like we really just have not heard the whole story yet.',
 'No.',
 'This is kind of like a trail of bread crumbs.',
 'Each revelation leads to something else.',
 "Some of them are pretty dramatic, like those text messages, where we
 'It was U.S. diplomats actively pushing Ukraine to investigate Biden
 'And we have this cast of characters - Kurt Volker, special envoy to
 "We have William Taylor, who we haven't heard from yet, but he was th
 'He was uncomfortable about with the whole arrangement.',
 "And then Gordon Sondland, who's this millionaire donor to Trump and
 'What his happening to his support among Republican lawmakers, Mara?'
 'I mean, has anything changed?',
 'So far, most Republican lawmakers are sheltering in place, which is
 'Some of them, like Marco Rubio, suggested that the president was kid
 'Some of them are saying nothing.',
 'But there are a few cracks.',
 "Ben Sasse, Susan Collins have criticized the president's action.",
 'And Mitt Romney, who has been the voice of conscience in the Senate
 'Trump punched back at Romney.',
 "He said, Mitt Romney was a pompous ass who's been fighting me from t
 "And later, he said, Romney was a fool who's playing right into the h
 "So we don't know yet if Romney and Ben Sasse and Susan Collins are g
 'Or is this the start of something new?',
 "That's NPR political correspondent Mara Liasson.",
```

In [23]:
```
lean_left_df = DataFrame(lean_left, columns=['text'])
# lean_left lable = -1
lean_left_df['label'] = 'lean_left'
lean_left_df
```

Out[23]:

	text	label
0	The impeachment inquiry picks up tomorrow wher...	lean_left
1	Just this morning, the lawyer for the whistleb...	lean_left
2	There's are a lot of moving parts.	lean_left
3	Fortunately, NPR's Mara Liasson is here to help.	lean_left
4	Good morning.	lean_left
...	...	...
741	I appreciate it.	lean_left
742	The economy is humming, but we may be hearing ...	lean_left
743	The stock market fell 3% this past Wednesday b...	lean_left
744	While in the bond market, a key marker of a co...	lean_left
745	Now the payoff for long-term government debt f...	lean_left

746 rows × 2 columns

In [25]:
```
lean_right = ['The Pentagon defended the military's diversity and inclusion training programs amid a torrent of crit
 '"We certainly respect the oversight that Congress provides.',
 'I'm not going to comment on any specific one initiative that members of Congress might be doing\u200b," Kirby said
 '"What I can speak to is what we're really focused on here at the department and that's defending the nation.',
 'And that means putting in place the right resources, the right strategies, the right operational concepts to do th
 '"And the \u200bsecretary has been very clear and fairly unapologetic about the fact that we want to get all the be
 'If you \u200bmeet the standards and you're qualified to be in the military — and you're willing to raise your hand
 '(Sen. Tom Cotton (left) and Rep. Dan Crenshaw (Getty Images))\n\nRep. Dan Crenshaw (R-Texas) and Sen. Tom Cotton (
 'began a campaign last week that seeks to have military members report t
 '"Enough is enough.',
 'We won't let our military fall to woke ideology.',
 'We have just launched a whistleblower webpage where you can submit your
 'Your complaint will be legally protected, and go to my office and @SenT
 'T\u200bhe two GOP lawmakers said they were prompted to act after Cotton
 'CLICK HERE TO GET THE FOX NEWS APP\n\nDefense Secretary Lloyd Austin ad
 'Cruz (R-Texas) accused the Army of portraying soldiers as "pansies" in
 'Austin said America's foes "would like to capitalize on talking points
 'Rep. Alexandria Ocasio-Cortez posted a photo of her grandmother's Puert
 '"Just over a week ago, my abuela fell ill.',
 'I went to Puerto Rico to see her- my 1st time in a year+ bc of COVID.',
 'This is her home.',
 'Hurricane María relief hasn't arrived.',
 'Trump blocked relief $ for PR," the New York Democrat wrote on Twitter.
 '"People are being forced to flee ancestral homes, & developers are taki
 'Conservative commentator Matt Walsh shot back: "Shameful that you live
 '"You don't even have a concept for the role that 1st-gen, first-born da
 '"My abuela is okay.',
 'But instead of only caring for mine & letting others suffer, I'm callin
 'Tuesday the Supreme Court issued two more unanimous decisions in Garlan
 'This follows two unanimous decisions last week.',
 'The weekly display of unanimity is notable given the calls by Democrati
 'This week I wrote on my blog about how the heavy-handed campaigns might
 'As we await important and likely divided decisions on issues like abort
 'In the Garland case, the court ruled (again) unanimously to reverse the
 'In Cooley, the court unanimously ruled in an opinion by Justice Stephen
 'SUPREME COURT PREPARES FINAL PUSH TO RELEASE HOT-BUTTON RULINGS, AMID R
 'Justice Sonia Sotomayor wrote the opinion in United States v. Palomar-S
 'It also ruled unanimously in Territory of Guam v. United States, in an
 'The court ruled in favor of Guam on the collection of funding from the
 'Op-ed calls on Justice Breyer to retire so Biden can choose Supreme Cou
 'Recently, Breyer warned against any move to expand the Supreme Court.',
```

In [30]:
```
lean_right_df = DataFrame(lean_right, columns=['text'])
# lena_right lable = 1
lean_right_df['label'] = 'lean_right'
lean_right_df
```

Out[30]:

	text	label
0	Over Memorial Day weekend, Michael Flynn, who ...	lean_right
1	During a Q&A session, an audience member asked...	lean_right
2	Flynn replied, “No reason.	lean_right
3	I mean, it should happen here.”\n\nThe crowd c...	lean_right
4	I bring this up not to dwell on the fact that ...	lean_right
5	Flynn, a retired general, has been saying loon...	lean_right
6	No, I bring this up to ask a different questio...	lean_right
7	The timing of Flynn’s remarks was darkly fortu...	lean_right
8	Flynn’s comments also coincided with a spirite...	lean_right
9	Some even argue that this is what conservatism...	lean_right
10	It’s certainly true that the Trump era has rev...	lean_right
11	But instead of going down various intellectual...	lean_right
12	The core problem afflicting the right—and to a...	lean_right
13	Definitions of populism vary, but for our purp...	lean_right
14	The defining emotion of populism and mobs alik...	lean_right
15	“Patriotism is when love of your own people co...	lean_right
16	“The people of Nebraska are for free silver, a...	lean_right
17	“I will look up the arguments later.”\n\nBoth ...	lean_right
18	Andrew Jackson, William Jennings Bryan, Huey L...	lean_right
19	Conservatives deserve special criticism for fo...	lean_right
20	But that decision has less to do with conserva...	lean_right
21	Certainly, on paper, intellectual progressivis...	lean_right
22	The surrender to populism was greased by what ...	lean_right

In [24]:
```
left = ['No \'civil war\' here: Republicans are at peace in their embrace of Trump\nGreg Krieg\nAnalysis by Gregor
 "House Republicans booted Rep. Liz Cheney from their leadership trio this week, replacing the Wyoming conservati
 "Cheney's defenestration and Stefanik's subsequent ascent were an anticlimax, and not just because the switch-a-r
 "Unlike Cheney, Stefanik is a reliable messenger on the one issue, above all else, that unites Republican lawmake
 'After the dust settled on Friday, House Republicans so
 "With Cheney sidelined, they argued, the party would be
 "Those imperatives, as Minority Leader Kevin McCarthy v
 'If the Cheney episode is instructive in any meaningful
 "Rather, to the extent there is conflict within the par
 'The Trumpist forces are on the march, in Congress and
 "Cheney's resistance, for all its sound and fury, attra
 'Instead, it provided yet another platform for Trump lo
 'For her part, Stefanik marched out to put a bow on the
 'The remark, in the context of what had unfolded over t
 'But beneath its apparent absurdity, a bright, glowing
 "Republicans are not only in near-perfect lockstep with
 'Trumpism without Trump\nIn a Friday evening interview
 '"We\'ve got to be able to tell people you can trust us
 "That her replacement, Stefanik, is by any measure a le
 'Those willing to stand by Cheney mostly shared at leas
 'With the exception of Rep. Adam Kinzinger of Illinois,
 'A third bucket exists mostly in the states, where GOP
 "Asked in March why he supported the legislation, given
 'After conflating the two, Raffensperger claimed that t
 'In her interview with Tapper on Friday, Cheney describ
 '"Our system held," Cheney continued, "the institutions
 "That Raffensperger didn't bend to Trump's will is to h
 'That he is now parlaying that credibility to give cove
 "Rep. Claudia Tenney offered up a similar kind of misdi
 '"No one knows about what happened in the election," th
 '"We don\'t know if it was stolen or not, (Cheney) does
 'But what I know is we need to fix it."',
 "As anti-democratic dissembling goes, Tenney's existential journey might seem mild.",
 'But it is, in many ways, the most dangerous kind.',
 'The loudest voices might get the most attention, but it is the measured ones, when peddling an outright lie abou
 "Marjorie Taylor Greene and the new GOP normal\nSome of Tenney's colleagues have been more aggressive -- and less
 "And they are not, as some apologists are quick to suggest, simply acting out of a fear of backlash from Trump's
 'That would be impossible, in a way, because they are, in the most literal sense, representative of them.',
 'None more, of course, than Rep. Marjorie Taylor Greene, a freshman Republican from Georgia.',
 "Earlier on Friday, two days after Greene personally confronted Democratic Rep. Alexandria Ocasio-Cortez of New Y
 "The New Yorker's door was locked, but the group camped out, rattling the letter slots, scribbling nasty messages
 '"We\'re going to go see, we\'re going to visit, Alexandria Ocasio-Cortez.',
 'Crazy eyes.',
 'Crazy eyes.',
 'Nutty.',
 'Cortez," Greene says at one point, mispronouncing "Ocasio," before continuing to taunt and yell at staffers ins
 "By McCarthy's logic -- the one that guided his support for Cheney's ouster -- either of the two episodes involv
```

In [26]:
```
left_df = DataFrame(left, columns=['text'])
# left lable = -2
left_df['label'] = 'left'
left_df
```

Out[26]:

	text	label
0	No 'civil war' here: Republicans are at peace ...	left
1	House Republicans booted Rep. Liz Cheney from ...	left
2	Cheney's defenestration and Stefanik's subsequ...	left
3	Unlike Cheney, Stefanik is a reliable messenge...	left
4	After the dust settled on Friday, House Republ...	left
...	...	...
470	A presidential task force on Covid-19 has alre...	left
471	Last month, Nigeria restricted travel from Bra...	left
472	/\nNigeria's first batch of the AstraZeneca va...	left
473	(Abraham Archiga/Reuters)\n/\nMuslims perform ...	left
474	(Sunday Alamba/AP	left

475 rows × 2 columns

In [27]:
```
right = ["tomorrow that will boost federal spending by 25% - the highest level since WWll\nU.S. is on track to run a
 'POLITICAL EDITOR FOR DAILYMAIL.COM\n\nPUBLISHED: 13:25 BST, 27 May 2021 | UPDATED: 21:49 BST, 27 May 2021\n\n
 'The budget would have the nation continue running deficits of more than $1 trillion, a level it already topped wit
 "It shows the Biden administration's desire to keep spending despite warnings from economists that it will keep dri
 "Taxes are also set to increase by $3trillion over the next decade in line with Biden's proposed hikes in corporate
 "The plan, which will be released in full on Friday, would seek major changes to the economy and welfare systems th
 'The proposal represents a 25 per cent increase over what former President Donald Trump proposed in his final budge
 'The spending levels would continue on an upward trajectory, hitting $8.2 trillion at the end of a decade, accordin
 "President Joe Biden's budget will propose $6 trillion in spending for the next fiscal year\t+6\nPresident Joe Bide
 'She pointed to prior spending laws enacted before and after he took office.',
 "'The president inherited $3 trillion of spending that had already been done to help get the pandemic under control
 "What the budget will reflect is that he is going to continue to deliver on his priorities of dong more work to get
 "Joe Biden's tax hikes targeting the rich: Who will have to pay\nCapital gains tax on investment sales for those ea
 'Investors currently pay 23.8% as the top capital gains rate along with the 3.8% net investment income tax, known a
 'Around 500,000 people in the US, or 0.32% of the population, have recorded a gross income of over $1million.',
 'For those earning more than $1 million in high-tax states, the total rate will be even higher given the combined f
 'In New York it could be as high as 52.22% and for Californians it could be 56.7%\n\nWealthy residents pay Capital
 'They are mainly placed on profitable stock trades and real estate deals.',
 'They can also apply to sales of collectible cars, art, businesses, gold.',
 'Investors are taxed on the difference between what they paid for the asset and what they sold it fo
 'Investments held for at least one year tops out at 20% and those held under a year are taxed the sa
 'An additional 3.8% tax applies to those earning at least $200,000.',
 'The US rate ranks in the middle of countries around the world.',
 'Investors generally support lower capital gains tax because they say it rewards entrepreneurship an
 'Corporate taxes\n\nCurrent top rate: 21%\n\nProposed top rate: 28% \n\nThese hikes have already bee
 "He is also targeting US firms' profits overseas and companies who use offshore businesses.",
 'Biden has still vowed that no one earning under $400,000 a year will pay more taxes in his administ
 'Larry Summers, the former director of the National Economic Council for President Obama and a Clint
 "He told a CoinDesk conference: 'We're taking very substantial risks on the inflation side.",
 "'I think policy is rather overdoing it.",
 "The sense of serenity and complacency being projected by the economic policymakers, that this is al
 'Prices on everything from food to used cars and lumber have risen sharply since Biden took office a
 'The most common nationwide price for a gallon of gas is just below $3, putting them at a seven-year
```

None of the women in the other photos have accused the governor of ina

In [28]:
```
right_df = DataFrame(right, columns=['text'])
# right lable = 2
right_df['label'] = 'right'
right_df
```

Out[28]:

	text	label
0	tomorrow that will boost federal spending by 2...	right
1	POLITICAL EDITOR FOR DAILYMAIL.COM\n\nPUBLISHE...	right
2	The budget would have the nation continue runn...	right
3	It shows the Biden administration's desire to ...	right
4	Taxes are also set to increase by $3trillion o...	right
...	...	...
212	Alyssa McGrath (pictured) is one of two aides ...	right
213	She alleges Cuomo grabbed her face and kissed ...	right
214	Vill, who said she felt uncomfortable at the t...	right
215	The same photos appear on Cuomo's Flickr accou...	right
216	None of the women in the other photos have acc...	right

217 rows × 2 columns

데이터셋 최종 준비 완료!

In [38]:
```
from pandas import DataFrame
import pandas as pd
dataset =pd.concat([center_df, lean_left_df, left_df, lean_right_df, right_df])
dataset
```

Out[38]:

	text	label
0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	center
1	They told Reuter correspondents in Asian capit...	center
2	But some exporters said that while the conflic...	center
3	The U.S. Has said it will impose 300 mln dlrs ...	center
4	Unofficial Japanese estimates put the impact o...	center
...	...	...
212	Alyssa McGrath (pictured) is one of two aides ...	right
213	She alleges Cuomo grabbed her face and kissed ...	right
214	Vill, who said she felt uncomfortable at the t...	right
215	The same photos appear on Cuomo's Flickr accou...	right
216	None of the women in the other photos have acc...	right

1548 rows × 2 columns

In [40]:
```
# 저장하기
dataset.to_csv("political_compass_datasets.csv")
```

Keras를 사용한 BERT 텍스트 분류

In [4]:
```
import tensorflow as tf
import tensorflow_hub as hub
import pandas as pd
from sklearn.model_selection import train_test_split
import numpy as np
import re
import unicodedata
import nltk
from nltk.corpus import stopwords
import keras
from tqdm import tqdm
import pickle
from keras.models import Model
import keras.backend as K
from sklearn.metrics import confusion_matrix,f1_score,classification_report
import matplotlib.pyplot as plt
from keras.callbacks import ModelCheckpoint
import itertools
from keras.models import load_model
from sklearn.utils import shuffle
from transformers import *
from transformers import BertTokenizer, TFBertModel, BertConfig
```

In [5]:
```

def unicode_to_ascii(s):
    return ''.join(c for c in unicodedata.normalize('NFD', s) if unicodedata.category(c) != 'Mn')

def clean_stopwords_shortwords(w):
    stopwords_list=stopwords.words('english')
    words = w.split()
    clean_words = [word for word in words if (word not in stopwords_list) and len(word) > 2]
    return " ".join(clean_words)

def preprocess_sentence(w):
    w = unicode_to_ascii(w.lower().strip())
    w = re.sub(r"([?.!,¿])", r" ", w)
    w = re.sub(r'[" "]+', " ", w)
    w = re.sub(r"[^a-zA-Z?.!,¿]+", " ", w)
    w=clean_stopwords_shortwords(w)
    w=re.sub(r'@\w+', '',w)
    return w
```

In [6]:
```
data = pd.read_csv('./data_political_compass/political_compass_datasets.csv')
```

In [7]:
```
data.head()
```

Out[7]:

	Unnamed: 0	text	label
0	0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	center
1	1	They told Reuter correspondents in Asian capit...	center
2	2	But some exporters said that while the conflic...	center
3	3	The U.S. Has said it will impose 300 mln dlrs ...	center
4	4	Unofficial Japanese estimates put the impact o...	center

In [8]:
```
print('File has {} rows and {} columns'.format(data.shape[0],data.shape[1]))
```

```
File has 1548 rows and 3 columns
```

In [9]:
```
# Select required columns
data = data[['text', 'label']]
```

In [10]:
```
data.head()
```

Out[10]:

	text	label
0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	center
1	They told Reuter correspondents in Asian capit...	center
2	But some exporters said that while the conflic...	center
3	The U.S. Has said it will impose 300 mln dlrs ...	center
4	Unofficial Japanese estimates put the impact o...	center

In [11]:
```
# Remove rows, where the label is present only ones (can't be split)
data = data.groupby('label').filter(lambda x : len(x) > 1)
#data = data.groupby('Product').filter(lambda x : len(x) > 1)
```

In [12]:
```
data.head()
```

Out[12]:

	text	label
0	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	center
1	They told Reuter correspondents in Asian capit...	center
2	But some exporters said that while the conflic...	center
3	The U.S. Has said it will impose 300 mln dlrs ...	center
4	Unofficial Japanese estimates put the impact o...	center

In [13]:
```
# shaffle datasets
data = data.sample(frac=1).reset_index(drop=True)
data
```

Out[13]:

	text	label
0	In 2019, New York saw 265 million visitors wit...	right
1	Yes.	lean_left

In [14]:
```
data=data.dropna()                                                    # Drop NaN valuues, if any
data=data.reset_index(drop=True)                                      # Reset index after dropping the column
data = shuffle(data)                                                  # Shuffle the dataset
print('Available labels: ',data.label.unique())                      # Print all the unique labels in the da
data['text']=data['text'].map(preprocess_sentence)                    # Clean the text column using preproces
```

```
Available labels:  ['left' 'lean_left' 'right' 'lean_right' 'center']
```

In [15]:
```
print('File has {} rows and {} columns'.format(data.shape[0],data.shape[1]))
data.head()
```

```
File has 1548 rows and 2 columns
```

Out[15]:

	text	label
1252	tianwen mars orbiter relay signal rover missio...	left
1407	know lulu every chance situation could increas...	lean_left
1394	last month south african health minister zweli...	left
936	recently also sat board log cabin republicans ...	lean_left
999	woman told colleague winter alleged encounter ...	right

In [16]:
```
num_classes=len(data.label.unique())
num_classes
```

Out[16]: 5

In [17]:
```
bert_tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
bert_model = TFBertForSequenceClassification.from_pretrained('bert-base-uncased',num_labels=num_classes)
```

```
All model checkpoint layers were used when initializing TFBertForSequenceClassification.

Some layers of TFBertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased
and are newly initialized: ['classifier']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

In [18]:
```
sent= 'how to train the model, lets look at how a trained model calculates its prediction.'
tokens=bert_tokenizer.tokenize(sent)
print(tokens)
```

```
['how', 'to', 'train', 'the', 'model', ',', 'lets', 'look', 'at', 'how', 'a', 'trained', 'model', 'calculate', '##s
', 'its', 'prediction', '.']
```

In [19]:
```
tokenized_sequence= bert_tokenizer.encode_plus(sent,add_special_tokens = True,max_length =30,pad_to_max_length = Tru
return_attention_mask = True)
```

```
Truncation was not explicitly activated but `max_length` is provided a specific value, please use `truncation=True`
to explicitly truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pai
rs of sequences (GLUE-style) with the tokenizer you can select this strategy more precisely by providing a specific
strategy to `truncation`.
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/transformers/tokenization_utils_base.p
y:2104: FutureWarning: The `pad_to_max_length` argument is deprecated and will be removed in a future version, use
`padding=True` or `padding='longest'` to pad to the longest sequence in the batch, or use `padding='max_length'` to
pad to a max length. In this case, you can give a specific length with `max_length` (e.g. `max_length=45`) or leave
max_length to None to pad to the maximal input size of the model (e.g. 512 for Bert).
  warnings.warn(
```

In [20]:
```
tokenized_sequence
```

Out[20]:
```
{'input_ids': [101, 2129, 2000, 3345, 1996, 2944, 1010, 11082, 2298, 2012, 2129, 1037, 4738, 2944, 18422, 2015, 2049, 17547, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]}
```

In [21]:
```
bert_tokenizer.decode(tokenized_sequence['input_ids'])
```

Out[21]:
```
'[CLS] how to train the model, lets look at how a trained model calculates its prediction. [SEP] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]'
```

In [22]:
```
data['gt'] = data['label'].map({'left':0,'lean_left':1, 'center':2,'lean_right':3, 'right':4})
data.head()
```

Out[22]:

	text	label	gt
1252	tianwen mars orbiter relay signal rover missio...	left	0
1407	know lulu every chance situation could increas...	lean_left	1
1394	last month south african health minister zweli...	left	0
936	recently also sat board log cabin republicans ...	lean_left	1
999	woman told colleague winter alleged encounter ...	right	4

In [23]:
```
sentences=data['text']
labels=data['gt']
len(sentences),len(labels)
```

Out[23]:
```
(1548, 1548)
```

In [24]:
```
input_ids=[]
attention_masks=[]

for sent in sentences:
    bert_inp=bert_tokenizer.encode_plus(sent,add_special_tokens = True,max_length =64,pad_to_max_length = True,retur
    input_ids.append(bert_inp['input_ids'])
    attention_masks.append(bert_inp['attention_mask'])

input_ids=np.asarray(input_ids)
attention_masks=np.array(attention_masks)
labels=np.array(labels)
```

```
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/transformers/tokenization_utils_base.py:2104: FutureWarning: The `pad_to_max_length` argument is deprecated and will be removed in a future version, use `padding=True` or `padding='longest'` to pad to the longest sequence in the batch, or use `padding='max_length'` to pad to a max length. In this case, you can give a specific length with `max_length` (e.g. `max_length=45`) or leave max_length to None to pad to the maximal input size of the model (e.g. 512 for Bert).
  warnings.warn(
```

In [25]:
```
len(input_ids),len(attention_masks),len(labels)
```

Out[25]:
```
(1548, 1548, 1548)
```

In [26]:
```
print('Preparing the pickle file.....')

pickle_inp_path='./data_pc/bert_inp.pkl'
pickle_mask_path='./data_pc/bert_mask.pkl'
pickle_label_path='./data_pc/bert_label.pkl'

pickle.dump((input_ids),open(pickle_inp_path,'wb'))
pickle.dump((attention_masks),open(pickle_mask_path,'wb'))
pickle.dump((labels),open(pickle_label_path,'wb'))


print('Pickle files saved as ',pickle_inp_path,pickle_mask_path,pickle_label_path)
```

```
Preparing the pickle file.....
Pickle files saved as  ./data_pc/bert_inp.pkl ./data_pc/bert_mask.pkl ./data_pc/bert_label.pkl
```

In [27]:
```
print('Loading the saved pickle files..')

input_ids=pickle.load(open(pickle_inp_path, 'rb'))
attention_masks=pickle.load(open(pickle_mask_path, 'rb'))
labels=pickle.load(open(pickle_label_path, 'rb'))

print('Input shape {} Attention mask shape {} Input label shape {}'.format(input_ids.shape,attention_masks.shape,lab
```

```
Loading the saved pickle files..
Input shape (1548, 64) Attention mask shape (1548, 64) Input label shape (1548,)
```

In [28]:
```
train_inp,val_inp,train_label,val_label,train_mask,val_mask=train_test_split(input_ids,labels,attention_masks,test_s

print('Train inp shape {} Val input shape {}\nTrain label shape {} Val label shape {}\nTrain attention mask shape {}
```

```
Train inp shape (1238, 64) Val input shape (310, 64)
Train label shape (1238,) Val label shape (310,)
Train attention mask shape (1238, 64) Val attention mask shape (310, 64)
```

In [29]:
```
log_dir='tensorboard_data/tb_bert'
model_save_path='./models/bert_model.h5'

callbacks = [tf.keras.callbacks.ModelCheckpoint(filepath=model_save_path,save_weights_only=True,monitor='val_loss',m

print('\nBert Model',bert_model.summary())

loss = tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True)
metric = tf.keras.metrics.SparseCategoricalAccuracy('accuracy')
optimizer = tf.keras.optimizers.Adam(learning_rate=2e-5,epsilon=1e-08)

bert_model.compile(loss=loss,optimizer=optimizer,metrics=[metric])
```

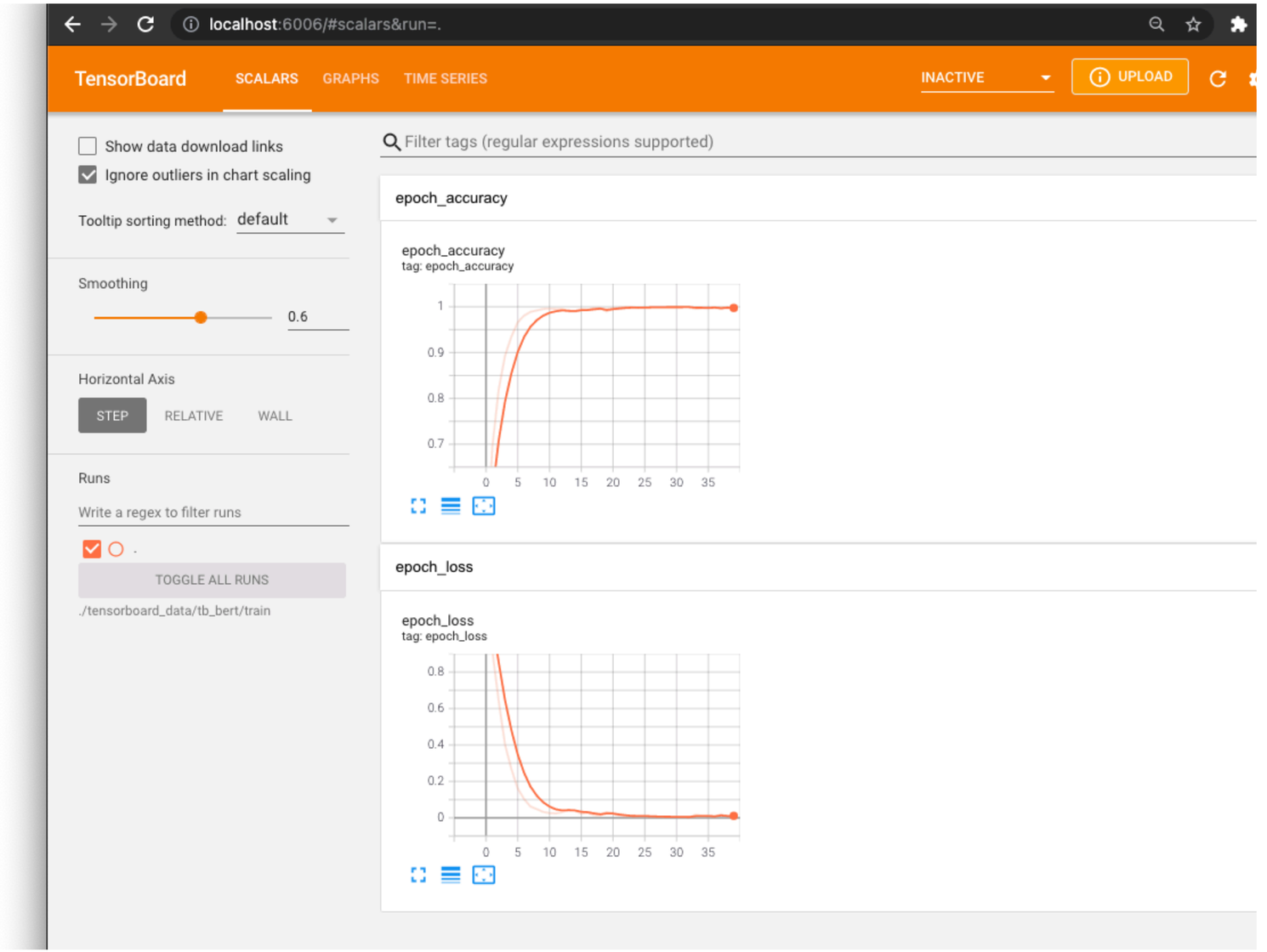
```
Model: "tf_bert_for_sequence_classification"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bert (TFBertMainLayer)       multiple                  109482240
_________________________________________________________________
dropout_37 (Dropout)         multiple                  0
_________________________________________________________________
classifier (Dense)           multiple                  3845
=================================================================
Total params: 109,486,085
Trainable params: 109,486,085
Non-trainable params: 0
_________________________________________________________________

Bert Model None
```

In [30]:

```
history=bert_model.fit([train_inp,train_mask],train_label,batch_size=32,epochs=20,validation_data=([val_inp,val_mask
```

```
Epoch 1/20
WARNING:tensorflow:The parameters `output_attentions`, `output_hidden_states` and `use_cache` cannot be updated whe
n calling a model.They have to be set to True/False in the config object (i.e.: `config=XConfig.from_pretrained('na
me', output_attentions=True)`).
WARNING:tensorflow:AutoGraph could not transform <bound method Socket.send of <zmq.sugar.socket.Socket object at 0x
7f8e52456460>> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux, `export AUTOGRAP
H_VERBOSITY=10`) and attach the full output.
Cause: module, class, method, function, traceback, frame, or code object was expected, got cython_function_or_metho
d
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform <bound method Socket.send of <zmq.sugar.socket.Socket object at 0x7f8e524564
60>> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux, `export AUTOGRAP
H_VERBOSITY=10`) and attach the full output.
Cause: module, class, method, function, traceback, frame, or code object was expected, got cython_function_or_metho
d
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING:tensorflow:The parameter `return_dict` cannot be set in graph mode and will always be set to `True`.
WARNING:tensorflow:From /Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/tensorflow/pyt
hon/ops/array_ops.py:5043: calling gather (from tensorflow.python.ops.array_ops) with validate_indices is deprecate
d and will be removed in a future version.
Instructions for updating:
The `validate_indices` argument has no effect. Indices are always validated on CPU and never validated on GPU.
WARNING:tensorflow:The parameters `output_attentions`, `output_hidden_states` and `use_cache` cannot be updated whe
n calling a model.They have to be set to True/False in the config object (i.e.: `config=XConfig.from_pretrained('na
me', output_attentions=True)`).
WARNING:tensorflow:The parameter `return_dict` cannot be set in graph mode and will always be set to `True`.
39/39 [==============================] - ETA: 0s - loss: 1.2539 - accuracy: 0.4790WARNING:tensorflow:The parameters
`output_attentions`, `output_hidden_states` and `use_cache` cannot be updated when calling a model.They have to be
set to True/False in the config object (i.e.: `config=XConfig.from_pretrained('name', output_attentions=True)`).
WARNING:tensorflow:The parameter `return_dict` cannot be set in graph mode and will always be set to `True`.
39/39 [==============================] - 177s 4s/step - loss: 1.2539 - accuracy: 0.4790 - val_loss: 0.9331 - val_ac
curacy: 0.6871
Epoch 2/20
39/39 [==============================] - 161s 4s/step - loss: 0.8863 - accuracy: 0.7173 - val_loss: 0.7600 - val_ac
curacy: 0.7742
Epoch 3/20
39/39 [==============================] - 161s 4s/step - loss: 0.5830 - accuracy: 0.8393 - val_loss: 0.5819 - val_ac
curacy: 0.8290
Epoch 4/20
39/39 [==============================] - 161s 4s/step - loss: 0.3653 - accuracy: 0.9144 - val_loss: 0.6034 - val_ac
curacy: 0.8097
Epoch 5/20
39/39 [==============================] - 161s 4s/step - loss: 0.2438 - accuracy: 0.9410 - val_loss: 0.6258 - val_ac
curacy: 0.8323
Epoch 6/20
39/39 [==============================] - 161s 4s/step - loss: 0.1565 - accuracy: 0.9580 - val_loss: 0.7288 - val_ac
curacy: 0.8452
Epoch 7/20
39/39 [==============================] - 161s 4s/step - loss: 0.1148 - accuracy: 0.9725 - val_loss: 0.6449 - val_ac
curacy: 0.8419
Epoch 8/20
39/39 [==============================] - 161s 4s/step - loss: 0.0926 - accuracy: 0.9806 - val_loss: 0.6404 - val_ac
curacy: 0.8452
Epoch 9/20
39/39 [==============================] - 161s 4s/step - loss: 0.0559 - accuracy: 0.9903 - val_loss: 0.7248 - val_ac
curacy: 0.8355
Epoch 10/20
39/39 [==============================] - 161s 4s/step - loss: 0.0383 - accuracy: 0.9960 - val_loss: 0.7361 - val_ac
curacy: 0.8290
Epoch 11/20
39/39 [==============================] - 161s 4s/step - loss: 0.0278 - accuracy: 0.9976 - val_loss: 0.7259 - val_ac
```


localhost:6006/#scalars&run=.
TensorBoard
SCALARS
GRAPHS
TIME SERIES
INACTIVE
UPLOAD
Show data download links
Ignore outliers in chart scaling
Tooltip sorting method: default
Smoothing
0.6
Horizontal Axis
STEP
RELATIVE
WALL
Runs
Write a regex to filter runs
TOGGLE ALL RUNS
./tensorboard_data/tb_bert/train
Filter tags (regular expressions supported)
epoch_accuracy
epoch_accuracy
tag: epoch_accuracy
1
0.9
0.8
0.7
0 5 10 15 20 25 30 35
epoch_loss
epoch_loss
tag: epoch_loss
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35

TensorBoard | 원격 액세스 - Chrome 원격 데스 | [텐서플로우] 텐서보드 (TensorB | TensorBoard

localhost:6006/#scalars&run=.&_smoothingWeight=0.006

TensorBoard SCALARS GRAPHS TIME SERIES INACTIVE UPLOAD

Show data download links

Ignore outliers in chart scaling

Tooltip sorting method: descending

Smoothing

0.006

Horizontal Axis

STEP RELATIVE WALL

Runs

Write a regex to filter runs

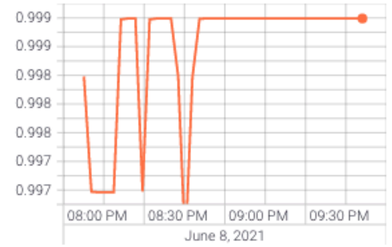
TOGGLE ALL RUNS

./tensorboard_data/tb_bert/train

Filter tags (regular expressions supported)

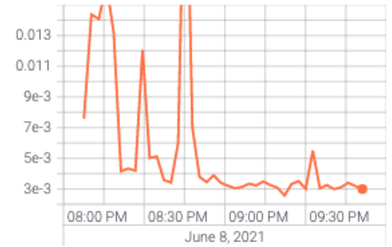
epoch_accuracy

epoch_accuracy
tag: epoch_accuracy



epoch_loss

epoch_loss
tag: epoch_loss



In [59]:

```
model_save_path='./models/bert_model.h5'

trained_model = TFBertForSequenceClassification.from_pretrained('bert-base-uncased',num_labels=5)
trained_model.compile(loss=loss,optimizer=optimizer, metrics=[metric])
trained_model.load_weights(model_save_path)

preds = trained_model.predict([val_inp,val_mask],batch_size=32)
print('preds:', preds)
print(type(preds))
preds = preds['logits']
print('preds_:', preds)
# import numpy as np
# pred_labels = np.argmax(preds)
pred_labels = preds.argmax(axis=1)
print('pred_labels', pred_labels)
print('val_label', val_label)

f1 = f1_score(val_label,pred_labels, average='micro')
print('F1 score',f1)
print('Classification Report')
target_names = ['left', 'lean_left', 'center', 'lean_right', 'right']
print(classification_report(val_label,pred_labels,target_names=target_names))

print('Training and saving built model.....')
```

```
All model checkpoint layers were used when initializing TFBertForSequenceClassification.

Some layers of TFBertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased
and are newly initialized: ['classifier']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

```
WARNING:tensorflow:The parameters `output_attentions`, `output_hidden_states` and `use_cache` cannot be updated whe
n calling a model.They have to be set to True/False in the config object (i.e.: `config=XConfig.from_pretrained('na
me', output_attentions=True)`).
WARNING:tensorflow:The parameter `return_dict` cannot be set in graph mode and will always be set to `True`.
preds: TFSequenceClassifierOutput(loss=None, logits=array([[-0.49090445,  3.2869735 , -1.013723  , -1.1928202 , -0.
56186223],
       [ 2.469226  ,  0.00602326, -1.1767939 , -0.8681813 , -0.42022336],
       [-0.17980987,  0.7214129 , -0.22669274, -1.2690276 ,  2.09991   ],
       ...,
       [ 0.6358856 ,  2.750293  , -1.5966513 , -1.0756427 , -0.36031324],
       [ 2.7467015 , -1.3513494 , -0.7086528 , -0.6813501 , -0.1758186 ],
       [ 2.412053  , -0.24241978, -1.2128685 , -0.7496565 , -0.6367665 ]],
      dtype=float32), hidden_states=None, attentions=None)
<class 'transformers.modeling_tf_outputs.TFSequenceClassifierOutput'>
preds_: [[-0.49090445  3.2869735  -1.013723   -1.1928202  -0.56186223]
 [ 2.469226    0.00602326 -1.1767939  -0.8681813  -0.42022336]
 [-0.17980987  0.7214129  -0.22669274 -1.2690276   2.09991   ]
 ...
 [ 0.6358856   2.750293   -1.5966513  -1.0756427  -0.36031324]
 [ 2.7467015  -1.3513494  -0.7086528  -0.6813501  -0.1758186 ]
 [ 2.412053   -0.24241978 -1.2128685  -0.7496565  -0.6367665 ]]
pred_labels [1 0 4 1 0 0 0 1 1 0 1 1 1 1 4 1 0 0 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1
 1 1 0 0 1 0 0 1 1 1 1 1 1 0 0 1 4 1 1 1 4 1 1 1 1 4 1 1 2 1 0 1 0 1 0 4 0
 1 0 1 4 0 4 1 4 0 1 4 1 1 1 1 1 1 0 0 1 0 0 1 0 1 1 0 1 0 1 0 0 1 1 4 1 1 0
 1 1 0 1 1 4 1 1 1 4 1 1 1 0 1 0 0 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 1 0 0 0 0
 1 1 1 1 0 1 1 4 1 1 1 1 4 0 4 1 1 1 1 4 0 1 0 4 1 4 0 1 1 0 1 1 1 4 1 1 1
 1 0 2 1 1 0 4 1 1 1 0 0 2 0 1 1 1 0 1 1 1 0 0 0 0 1 0 1 0 1 1 1 4 1 1 4 1
 1 1 1 1 1 1 1 4 1 1 1 1 0 4 0 1 1 1 0 1 1 0 0 1 1 1 1 0 1 1 1 1 0 0 1 4 1
 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 1 4 1 0 4 1 0 0 1 4 1 1 0 1 0 0 0 0
 1 1 4 1 1 1 1 0 4 1 1 1 0 0]
val_label [1 0 4 1 0 0 0 2 1 1 4 1 1 4 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 3 1 1 1 0 1
 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0 1 4 1 1 1 0 1 4 1 0 4 1 1 2 1 0 1 0 1 0 4 0
 1 0 1 4 0 4 1 4 0 1 4 1 1 1 1 1 1 0 1 1 0 0 1 3 1 1 0 0 0 1 0 0 1 1 4 1 1 0
 1 1 0 1 1 0 1 1 1 4 0 1 1 3 1 0 1 1 1 0 0 1 1 1 2 0 0 1 1 1 1 2 1 0 0 0 0
```

인공지능 AI 의 political compass accuracysms 20회 반복 학습한 결과 83% 달성!

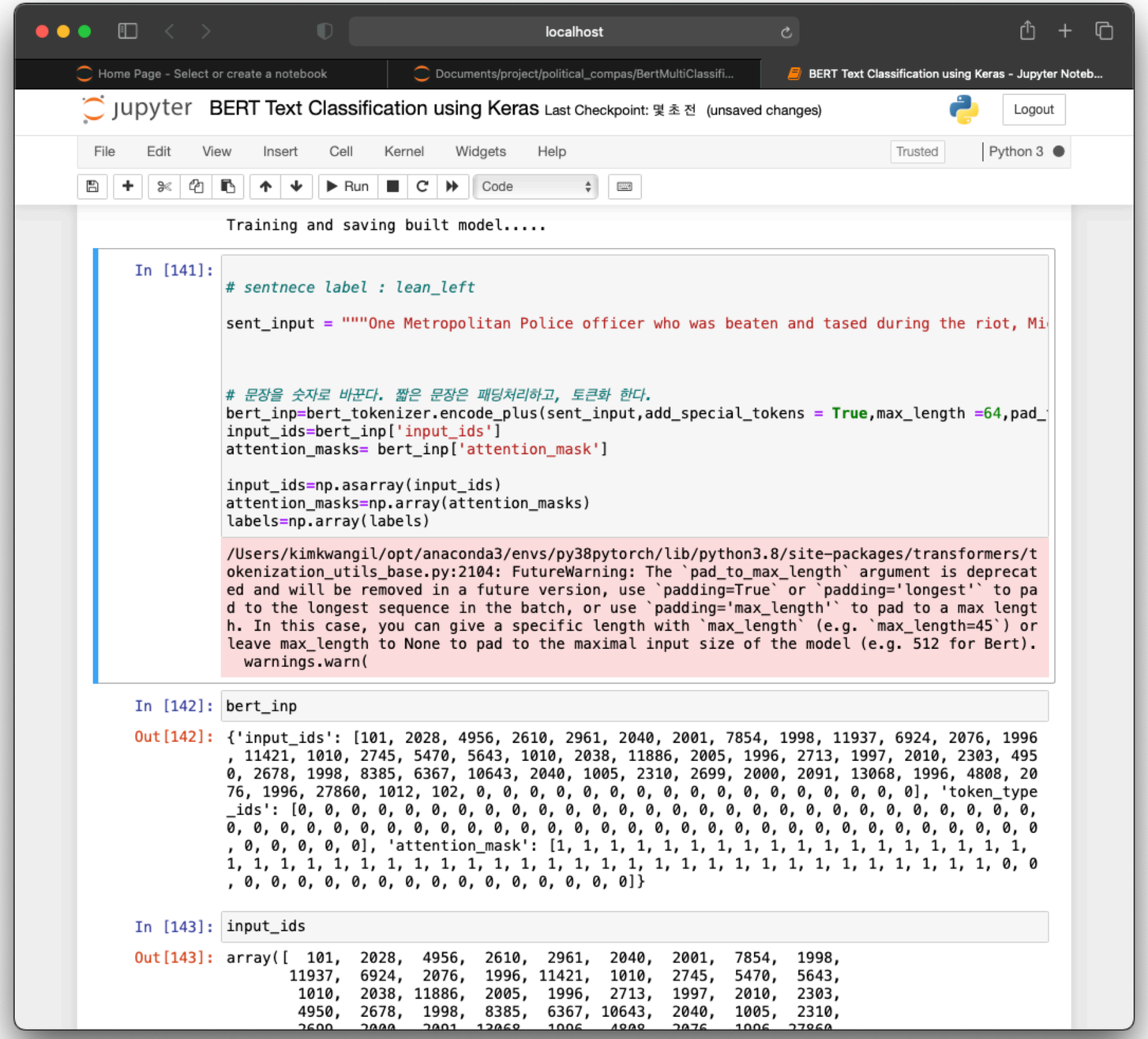
```
F1 score 0.8290322580645161
Classification Report
              precision    recall  f1-score   support

        left       0.84      0.75      0.79        99
   lean_left       0.81      0.95      0.88       161
      center       1.00      0.38      0.55         8
  lean_right       0.00      0.00      0.00        10
       right       0.87      0.84      0.86        32

    accuracy                           0.83       310
   macro avg       0.71      0.58      0.61       310
weighted avg       0.81      0.83      0.81       310

Training and saving built model.....
```

```
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/sklearn/metrics/_classification.py:124
8: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/sklearn/metrics/_classification.py:124
8: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/sklearn/metrics/_classification.py:124
8: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```


localhost
Home Page - Select or create a notebook
Documents/project/political_compas/BertMultiClassifi...
BERT Text Classification using Keras - Jupyter Noteb...
jupyter BERT Text Classification using Keras Last Checkpoint: 몇 초 전 (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Training and saving built model.....
In [141]:
# sentnece label : lean_left
sent_input = """One Metropolitan Police officer who was beaten and tased during the riot, Mi
# 문장을 숫자로 바꾼다. 짧은 문장은 패딩처리하고, 토큰화 한다.
bert_inp=bert_tokenizer.encode_plus(sent_input,add_special_tokens = True,max_length =64,pad_
input_ids=bert_inp['input_ids']
attention_masks= bert_inp['attention_mask']
input_ids=np.asarray(input_ids)
attention_masks=np.array(attention_masks)
labels=np.array(labels)
/Users/kimkwangil/opt/anaconda3/envs/py38pytorch/lib/python3.8/site-packages/transformers/tokenization_utils_base.py:2104: FutureWarning: The `pad_to_max_length` argument is deprecated and will be removed in a future version, use `padding=True` or `padding='longest'` to pad to the longest sequence in the batch, or use `padding='max_length'` to pad to a max length. In this case, you can give a specific length with `max_length` (e.g. `max_length=45`) or leave max_length to None to pad to the maximal input size of the model (e.g. 512 for Bert).
warnings.warn(
In [142]:
bert_inp
Out[142]:
{'input_ids': [101, 2028, 4956, 2610, 2961, 2040, 2001, 7854, 1998, 11937, 6924, 2076, 1996, 11421, 1010, 2745, 5470, 5643, 1010, 2038, 11886, 2005, 1996, 2713, 1997, 2010, 2303, 4950, 2678, 1998, 8385, 6367, 10643, 2040, 1005, 2310, 2699, 2000, 2091, 13068, 1996, 4808, 2076, 1996, 27860, 1012, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'token_type_ids': [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]}
In [143]:
input_ids
Out[143]:
array([101, 2028, 4956, 2610, 2961, 2040, 2001, 7854, 1998,
11937, 6924, 2076, 1996, 11421, 1010, 2745, 5470, 5643,
1010, 2038, 11886, 2005, 1996, 2713, 1997, 2010, 2303,
4950, 2678, 1998, 8385, 6367, 10643, 2040, 1005, 2310,

localhost

Home Page - Select or create a notebook | Documents/project/political_compas/BertMultiClassifi... | BERT Text Classification using Keras - Jupyter Noteb...

jupyter BERT Text Classification using Keras Last Checkpoint: 2분 전 (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [144]:
```
attention_masks
```

Out[144]:
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

In [145]:
```
# val_inp : input_ids
# val_mask : attention_masks

preds = trained_model.predict([input_ids,attention_masks],batch_size=32)
preds = preds['logits'] # 시그모이드함수 적용하여 데이터의 결과를 두 그룹으로 분류해본다. 0 or 1
print("prieds'logit':", preds)
pred_labels = preds.argmax(axis=1)
print('pred_labels', pred_labels)
```

```
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33990204e-01  1.41494155e+00 -1.26178539e+00 -1.31466374e-01
 -9.77089882e-01]
 [ 5.33997416e-01  1.41493905e+00 -1.26178646e+00 -1.31467432e-01
 -9.77093399e-01]
 [ 5.33997416e-01  1.41493905e+00 -1.26178646e+00 -1.31467432e-01
 -9.77093399e-01]]
pred_labels [1 4 1 4 0 1 1 0 4 1 0 1 0 1 1 0 4 1 1 1 1 1 0 3 0 0 1 0 0 4 1 1 0 1 0 1 1
 1 4 1 0 1 1 0 0 0 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
```

In [146]:
```
type(pred_labels)
```

Out[146]:
```
numpy.ndarray
```

In [147]:
```
plt.hist(pred_labels)

plt.title("Political Compass AI")


plt.show()
```

localhost

Home Page - Select or create a notebook | Documents/project/political_compas/BertMultiClassifi... | BERT Text Classification using Keras - Jupyter Noteb...

jupyter BERT Text Classification using Keras Last Checkpoint: 3분 전 (unsaved changes) Logout

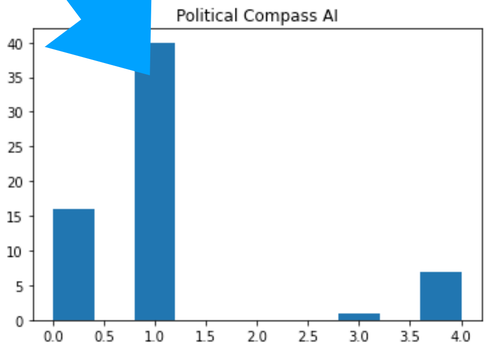
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [147]:
```
plt.hist(pred_labels)

plt.title("Political Compass AI")


plt.show()
```



In [148]:
```
result = np.bincount(pred_labels)
result = np.argmax(result)
result
```

Out[148]: 1

In [149]:
```
def get_political_compass(result):
    if result == 0:
        print('Left')
        result = 'Left'
    elif result == 1:
        print('Lean_Left')
        result = 'Lean_Left'
    elif result == 2:
        print('Center')
        result = 'Center'
    elif reulst == 3:
        print('Lean_Right')
        result = 'Lean_Right'
    else:
        print('Right')
        result = 'Right'
```

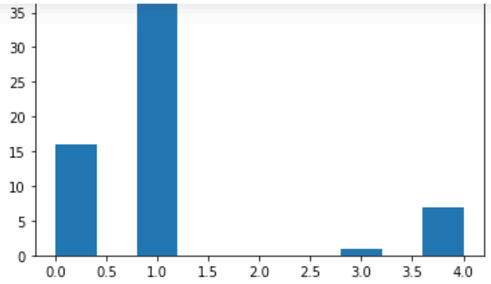
localhost

Home Page - Select or create a notebook | Documents/project/political_compas/BertMultiClassifi... | BERT Text Classification using Keras - Jupyter Noteb...

jupyter BERT Text Classification using Keras Last Checkpoint: 4분 전 (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code



In [148]:
```
result = np.bincount(pred_labels)
result = np.argmax(result)
result
```

Out[148]: 1

In [149]:
```
def get_political_compass(result):
    if result == 0:
        print('Left')
        result = 'Left'
    elif result == 1:
        print('Lean_Left')
        result = 'Lean_Left'
    elif result == 2:
        print('Center')
        result = 'Center'
    elif reulst == 3:
        print('Lean_Right')
        result = 'Lean_Right'
    else:
        print('Right')
        result = 'Right'
    return result
```

In [150]:
```
Final_Result = get_political_compass(result)
```

```
Lean_Left
```

In [ ]:

Precision(정밀도)

정밀도란 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율입니다. 즉, 아래와 같은 식으로 표현할 수 있습니다.

$$(Precision) = \frac{TP}{TP + FP}$$

Recall(재현율)

재현율이란 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율입니다.

Positive 정답률, **PPV(Positive Predictive Value), 예를 들어** 날씨 예측 모델이 맑다로 예측했는데, 실제 날씨가 맑았는지를 살펴보는 지표

통계학에서는 **sensitivity**으로, 그리고 다른 분야에서는 **hit rate**라는 용어로도 사용합니다. 실제 날씨가 맑은 날 중에서 모델이 맑다고 예측한 비율을 나타낸 지표인데, 정밀도(Precision)와 True Positive의 경우를 다르게 바라보는 것입니다. 즉, Precision이나 Recall은 모두 실제 True인 정답을 모델이 True라고 예측한 경우에 관심이 있으나, 바라보고자 하는 관점만 다릅니다. Precision은 모델의 입장에서, 그리고 Recall은 실제 정답(data)의 입장에서 정답을 정답이라고 맞춘 경우를 바라보고 있습니다. 다음의 경우를 생각해보겠습니다.

$$(Recall) = \frac{TP}{TP + FN}$$ *"어떤 요소에 의해, 확실히 맑은 날을 예측할 수 있다면 해당하는 날에만 맑은 날이라고 예측하면 되겠다."*

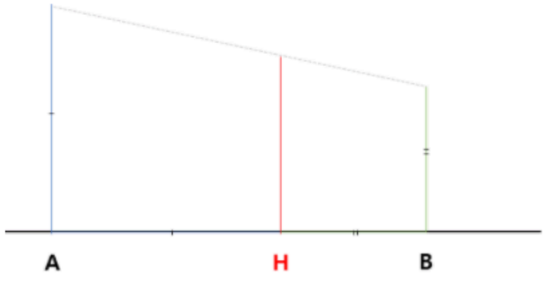
이 경우에는 확실하지 않은 날에는 아에 예측을 하지 않고 보류하여 FP의 경우의 수를 줄여, Precision을 극도로 끌어올리는 일종의 편법입니다. 예를 들어 한달 30일 동안 맑은 날이 20일이었는데, 확실한 2일만 맑다고 예측한다면, 당연히 맑다고 한 날 중에 실제 맑은 날(Precision)은 100%가 나오게 됩니다. 하지만 과연, 이러한 모델이 이상적인 모델일까요?

따라서, 우리는 실제 맑은 20일 중에서 예측한 맑은 날의 수도 고려해 보아야합니다. 이 경우에는 Precision만큼 높은 결과가 나오지 않습니다. Precision과 함께 Recall을 함께 고려하면 실제 맑은 날들(즉, 분류의 대상이 되는 정의역, 실제 data)의 입장에서 우리의 모델이 맑다고 예측한 비율을 함께 고려하게 되어 제대로 평가할 수 있습니다. Precision과 Recall은 상호보완적으로 사용할 수 있으며, 두 지표

$$(F1\text{-}score) = 2\times\frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\times\frac{Precision \times Recall}{Precision + Recall}$$

F1 score는 Precision과 Recall의 조화평균입니다.

조화평균은 기하학적으로 다음과 같이 표현할 수 있습니다. 서로 다른 길이의 A, B와 이 두 길이의 합만큼 떨어진 변(AB)으로 이루어진 사다리꼴을 생각해봅시다. 이 AB에서 각 변의 길이가 만나는 지점으로부터 맞은 편의 사다리꼴의 변으로 내린 선분이 바로 조화평균을 나타냅니다.



<Fig5. 조화평균의 기하학적 의미>

기하학적으로 봤을 때, 단순 평균이라기보다는 작은 길이 쪽으로 치우치게 된, 그러면서 작은 쪽과 큰 쪽의 사이의 값을 가진 평균이 도출됩니다. 이렇게 조화평균을 이용하면 산술평균을 이용하는 것보다, 큰 비중이 끼치는 bias가 줄어든다고 볼 수 있습니다. 즉, F1-score는 아래와 같이 생각할 수 있습니다.