debatepedia_pipeline

May 31, 2019

[4]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
import os
import re
from itertools import compress
import time
import random
import math
from pathlib import Path
import shutil
from datetime import datetime
import os
```

1 Split json file

[ ]:
```
df = pd.read_json("../data/debatepedia/debatepedia-preprocessed.json", orient =␣
 ↪"records")
df
```

1.0.1 Drop duplicates

[ ]:
```
df =  df.drop_duplicates(subset="content",keep="first")
df
```

1.0.2 Split by topic

[ ]:
```
topics = df.topic.unique()
n = len(topics)
n
```

[ ]:
```
n_t = math.floor(n*0.8)
n_t
```

[ ]:
```
topics_tr = topics[:n_t]
topics_tr
```

[ ]:
```
topics_te = topics[n_t:]
topics_te
```

[ ]:
```
df_tr = df[df['topic'].isin(topics_tr)]
df_tr
```

[ ]:
```
df_te = df[df['topic'].isin(topics_te)]
df_te
```

[ ]:
```
df_tr.to_json("../data/debatepedia/k-fold/set1/debatepedia-preprocessed-train.
 ↪json", orient = "records")
df_te.to_json("../data/debatepedia/k-fold/set1/debatepedia-preprocessed-test.
 ↪json", orient = "records")
```

2 Split kfold

[ ]:
```
def split_kfold(num_fold,index,split_path):
    """
    Split folders into 80% training and 20% testing based on index
    """
    # walk through folders
    folders = []
    for entry in os.scandir(path):
        if entry.is_dir():
            folders.append(entry.path)
    for f in folders:
        print(f)
    test_ratio = 1/num_fold
    test_len = int(len(folders) * test_ratio)
    t1= int(len(folders) * test_ratio * (index-1))
    t2 = t1+test_len
    # split test set according to t1 and t2
    train = folders[:t1] + folders[t2:]
    test = folders[t1:t2]
    parent = os.path.join(split_path,"set"+str(index))

    # create train path and test path
    train_path = os.path.join(parent, "train")
    test_path = os.path.join(parent, "test")
    print("train path: " + train_path)
    print("test path: " + test_path)
    # copy the folders to train path or test path
    # according to the split

    print(len(train))
    print(len(test))
    for f in train:
        folder_name = f.split("/")[-1]
        write_path = os.path.join(train_path, folder_name)
        copy_folder(f, write_path)
    for f in test:
        folder_name = f.split("/")[-1]
        write_path = os.path.join(test_path, folder_name)
        copy_folder(f, write_path)
```

[ ]:

```
def copy_folder(src, des):
    print(src)
    print(des)
    try:
        shutil.copytree(src, des)
        # Directories are the same
    except shutil.Error as e:
        print('Directory not copied. Error: %s' % e)
        # Any error saying that the directory doesn't exist
    except OSError as e:
        print('Directory not copied. Error: %s' % e)
```

[ ]:

```
path="../data/debatepedia/xmi"
num_fold = 5
#### Write to files
t = datetime.now()
t
```

[ ]:

```
dt = str(t)[:19].replace(' ', '_')
parent = str(Path(path).parent)
split_path = os.path.join(parent, dt)
if not os.path.exists(split_path):
    os.mkdir(split_path)
print(split_path)
for i in range(1,num_fold+1):
    split_kfold(num_fold,i,split_path)
```

3 Scripts to run the whole pipeline

from reading json file, generating xmi, splitting, generating arff to evaluation by Weka

[ ]:

```
# scripts/adu_classification.sh

# ADU Classification
```

```
#./scripts/adu_classification.sh >&1 | tee  "output/$(date +"%Y-%m-%d_%T").log"

# ADU 5 fold Validation
#./scripts/adu_classification.sh kfold >&1 | tee  "output/$(date +"%Y-%m-%d_%T").
 ↪log"

# ADU Random Split Classification
#./scripts/adu_classification.sh random >&1 | tee  "output/$(date␣
 ↪+"%Y-%m-%d_%T").log"
```

4 ADU classification

[ ]:

```
### output from output/2019-05-31_03:42:29.log

Step 1: Split File into Training & Testing
/home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/2019-05-31_03:42:29
....ok

Step 2: Use UIMA to convert to XMI files
input directory: /home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/
 ↪2019-05-31_03:42:29
filename: debatepedia-preprocessed_train.json
output directory: /home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/
 ↪2019-05-31_03:42:29/xmi/debatepedia-preprocessed_train
..................................................................................
 ↪................................................................................
 ↪................................................................................
 ↪................................................................................
 ↪........................................................................
 ↪filename: debatepedia-preprocessed_test.json
output directory: /home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/
 ↪2019-05-31_03:42:29/xmi/debatepedia-preprocessed_test
............................................................................done
....ok

Step 3: Generate Feature Files
---------------------------------------------
Processing corpus in the directory
/home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/2019-05-31_03:42:29/
 ↪xmi/debatepedia-preprocessed_train
---------------------------------------------

Compute feature values on /home/ciso0478/wstud-visit-the-dome-ss19/data/
 ↪debatepedia/2019-05-31_03:42:29/xmi/debatepedia-preprocessed_train
finished in 21.751s
```

```
---------------------------------------------
Processing corpus in the directory
/home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/2019-05-31_03:42:29/
 ↪xmi/debatepedia-preprocessed_test
---------------------------------------------

Compute feature values on /home/ciso0478/wstud-visit-the-dome-ss19/data/
 ↪debatepedia/2019-05-31_03:42:29/xmi/debatepedia-preprocessed_test
finished in 4.205s

....ok

Step 4: Use Weka to train classifier
/home/ciso0478/wstud-visit-the-dome-ss19/data/debatepedia/2019-05-31_03:42:29

Time taken to test model on training data: 18.36 seconds

=== Error on training data ===

Correctly Classified Instances       58624               99.9966 %
Incorrectly Classified Instances         2                0.0034 %
Kappa statistic                          0.9999
Mean absolute error                      0.0274
Root mean squared error                  0.0576
Relative absolute error                  5.4715 %
Root relative squared error             11.5226 %
Total Number of Instances            58626


=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC␣
 ↪Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000␣
 ↪    1.000     conclusion
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000␣
 ↪    1.000     premise
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000␣
 ↪    1.000


=== Confusion Matrix ===

     a     b   <-- classified as
 29313     0 |     a = conclusion
```

```
     2 29311 |     b = premise

Time taken to test model on test data: 2.6 seconds

=== Error on test data ===

Correctly Classified Instances        6916               91.4815 %
Incorrectly Classified Instances       644                8.5185 %
Kappa statistic                          0.7704
Mean absolute error                      0.156
Root mean squared error                  0.2552
Relative absolute error                 31.2096 %
Root relative squared error             51.036  %
Total Number of Instances             7560


=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC␣
 ↪Area  PRC Area  Class
                 0.794    0.044    0.862      0.794    0.827      0.772    0.962␣
 ↪    0.912     conclusion
                 0.956    0.206    0.931      0.956    0.944      0.772    0.962␣
 ↪    0.984     premise
Weighted Avg.    0.915    0.164    0.913      0.915    0.914      0.772    0.962␣
 ↪    0.966


=== Confusion Matrix ===

    a    b   <-- classified as
 1537  398 |    a = conclusion
  246 5379 |    b = premise

....ok
```

5 ADU 5 fold Validation

[ ]:

```
# output from output/2019-05-31_17:00:22.log
```

5.0.1 summary of cross validation

[ ]:

```
set1=90.5621
set2=91.4556
set3=90.6409
set4=90.5261
set5=90.8758
```

5.0.2 adu random classification

[29]:

```
path = "../data/debatepedia/debatepedia-preprocessed.json"

## Cross validation
df = pd.read_json(path, orient='records')

#### Drop duplicates
df = df.drop_duplicates(subset='content', keep='first')
```

[30]:

```
#### Split by topic

df1=df.reindex(np.random.permutation(df.index))
n1 = len(df)
n1_t = math.floor(n1 * 0.8)
```

[27]:

```
df_tr = df1[:n1_t]
df_tr
```

[27]:

```
       argument_id                                            content  \
38739        11182               Increased sales means a greater GDP:
45081        13296  This despite the new technology that has been ...
13710         3759  By the 1990s all gun ban laws were struck down...
29420         8461  This is, of course, assuming that their use do...
27744         7984  Responding to concerns that "Wikipedia will en...
4723          1328              Its just what they are waiting for". 
30444         8759  One strong argument, in this regard, against l...
32534         9355  as supporters of mines here see it, land mines...
31563         9064  11, 1997: "The debate over using marihuana as ...
27104         7814  The biggest marketplace on Earth, China alread...
26434         7624  "Israels geographic vulnerability means that  ...
25291         7274  Climate change is a serious threat to the worl...
33750         9681  With Citi and Bank of America shares down more...
16645         4697  Business defendants in particular overwhelming...
24533         7070  This compares favorably against electric vehic...
38269        11010  Public insurance delivers same quality insuran...
19391         5545  Eastern Washington has information on their si...
25940         7465  ICC will continually expand jurisdiction over ...
21662         6220  But what geoengineering can do is slow the inc...
1217           332  In such cases of medical emergency and in the ...
```

```
10822          2964  State Department spokesman Tom Casey said in M...
26225          7565  Adult courts do not necessarily mean longer se...
27826          8013  Kangaroos are just an inconvenience, and we in...
31853          9152  Teachers should be paid on merit, not seniorit...
33704          9672  Fed chairman said that temporary government ow...
38470         11064  "Puerto Ricans are already considered to be Am...
45064         13294  This proves that even if advanced safety measu...
1692            454  For the same reason it is inadvisable to lift ...
8879           2451  Because of the free market people are free to ...
11739          3199  Quality of decisions matters more than access ...
...             ...                                               ...
42142         12404         LFTR reactor can be operated continuously.
2056            551  If they can single out a part of the spectrum,...
942             265  No individual has rights over another individual.
6049           1689  The desire to have ones own child and to nurt... 
33179          9544  The pretense that "it is just my culture" can ...
45207         13340                       They long for maternal care.
27983          8051  In otherwords, economically, Kangaroos are bet...
44686         13192  Tax deal was win for wealthiest and their lobb...
26366          7606  Blockade unjustly prevents building supplies e...
41058         12040                              Sun light is diffuse.
21307          6104  This technology should, therefore, be used to ...
34980          9999  Nuclear energy is needed to meet growing elect...
29133          8392  Outlawing incest opens slippery slope to other...
17486          4927  US Constitution does not apply to enemy combat...
42706         12575  China is a permanent member of the United Nati...
19689          5645  Free trade increases the purchasing power of c...
19358          5538  And, this suggests that abandoning the filibus...
46258         13704  Issues of quality, equity and environmental st...
27668          7963  Most people only write reliable facts when edi...
10053          2753  It is simply not in their nature to overthrow ...
44623         13171  Obama administration consulted with Congress o...
64               15  "our economy is not a shining example of pure ...
19353          5538  Eliminating filibuster is excessive; reforms a...
45748         13550  While this procedure was fine for routine, cri...
25081          7216  First, there is the question of whether public...
30058          8655  Alcohol consumption not comparable to military...
25334          7285  UN impotent to compel/enforce Annex I climate ...
40110         11732  In general, it seems that voting rights should...
579             147  Coyotes will respond to a border fence with in...
22987          6571                       We intend to win this fight.

                                                   topic    unitType
38739        Reducing value added tax on contraceptives  conclusion
45081                          US offshore oil drilling     premise
13710                                    DC handgun ban     premise
29420                         Legalization of Marijuana     premise
```

```
27744                         Is Wikipedia valuable?     premise
4723                        Bailout of US automakers     premise
30444   Mandatory labeling of genetically modified foods     premise
32534                Mine Ban Treaty (Ottawa Treaty)     premise
31563                 Medical marijuana dispensaries     premise
27104     Is China a threat to international stability?     premise
26434                       Israeli blockade of Gaza     premise
25291  Increase UN Annex I aid for climate change ada...     premise
33750    Nationalization of banks during economic crisis     premise
16645                             Election of judges     premise
24533                                Hybrid vehicles     premise
38269                 Public health insurance option  conclusion
19391                          First year dorm rooms     premise
25940                   International Criminal Court  conclusion
21662                                 Geoengineering     premise
1217                                        Abortion     premise
10822                               Cluster bomb ban     premise
26225  In some cases juveniles should be tried as adults  conclusion
27826                  Kangaroo culling in Australia     premise
31853                         Merit pay for teachers  conclusion
33704    Nationalization of banks during economic crisis     premise
38470               Puerto Rico statehood in America     premise
45064                       US offshore oil drilling     premise
1692                         Adult male circumcision     premise
8879                         Capitalism vs socialism     premise
11739                              Compulsory voting  conclusion
...                                              ...         ...
42142                   Thorium based nuclear energy  conclusion
2056                              Affirmative action     premise
942                                         Abortion     premise
6049               Ban on human reproductive cloning  conclusion
33179              Multiculturalism vs. assimilation  conclusion
45207                                           Veal     premise
27983                  Kangaroo culling in Australia     premise
44686                           US debt ceiling deal  conclusion
26366                       Israeli blockade of Gaza  conclusion
41058                                   Solar energy     premise
21307                              Genetic screening     premise
34980                                 Nuclear energy  conclusion
29133                   Legalization of adult incest  conclusion
17486              Enhanced interrogation techniques  conclusion
42706                             Tibet independence     premise
19689                                     Free trade  conclusion
19358                                     Filibuster     premise
46258                            Water privatization     premise
27668                         Is Wikipedia valuable?     premise
10053               China is headed for a revolution     premise
```

```
44623                 US and NATO intervention in Libya  conclusion
64                     $700 billion US economic bailout     premise
19353                                        Filibuster  conclusion
45748      Warrantless wiretapping in the United States     premise
25081         Immunity from prosecution for politicians     premise
30058            Lowering US drinking age from 21 to 18  conclusion
25334  Increase UN Annex I aid for climate change ada...  conclusion
40110  Should the minimum age of candidacy for politi...     premise
579                     700 mile US Mexico border fence  conclusion
22987                   Guantanamo Bay detention center     premise

[37185 rows x 4 columns]
```

[28]:

```
df_te = df1[n1_t:]
df_te
```

[28]:

```
       argument_id                                            content  \
20465         5877  "Don't ask don't tell" made US military policy...
34017         9749                    Natural gas is a non-toxic gas.
18823         5371  Jorge Luis Borges (an Argentinean writer) like...
25141         7231  Should we ignore the fact that the higher the ...
14426         3949  Contrary to popular belief, it is economic pro...
44357        13099  Nuclear deal places safeguards on India's exis...
37579        10802  It sends out a clear policy signal that income...
27087         7801  Patrick Clawson of The Washington Institute fo...
9163          2520  The American Civil Liberties Union has opposed...
3775          1066  Nothing, however, unites adversaries like a co...
40007        11683  The constitutional restriction springs from th...
38460        11061  "We cannot continue to operate a colony, forci...
150             38  In a USA Today/Gallup Poll conducted on Septem...
40942        12005                    This is inefficient and costly.
24650         7099                  Dams can destroy marine fisheries
35350        10096  Obama's position on meeting hostile leaders is...
42493        12512  They should not be rewarded with a right to se...
39318        11415  In 1965, in Maryland v. United States, 381 U.S...
11563         3159                        give very positive results.
26331         7597          Because it imposes 'collective punishment.
42205        12421  Reservoirs worldwide are being more or less be...
43269        12755  Trying terrorists shows confidence in US syste...
17498         4929  Second, even if the Fifth Amendment applied to...
2154           581  Those that contributed to AIG's collapse have ...
13831         3789           Love seeks healing, peace and wholeness.
17621         4956  Male athletes have received unrepresentative s...
33074         9519  ' He said, 'Indeed ye have been in manifest er...
1183           324  Adoption does not spare a women the pains/risk...
33424         9594  This process of public exposure leading to the...
16524         4665  The inclusion in 1998 of religious private sch...
...            ...                                                ...
```

```
39796        11614  And as Nichols and McChesney point out, our go...
25963         7471  It has been enshrined in American law from the...
35408        10125  Face-to-face communications allow for mis-comm...
22485         6453  Should Greece default, the whole EU would be s...
10290         2830  But, because "separate can never be equal", ci...
43029        12668  If some terrorist have inflicted a massive amo...
32819         9457  Why not do it on the International Space Stati...
15824         4467                            You have kids don't you?
26177         7545  (...) 100% of murders and 50% of manslaughters...
28349         8180  And thats without even touching upon the many...
11397         3119  In fact, the nuclear weapon states' commitment...
9471          2606                     Priestly celibacy is unnatural.
200             51  Ensuring against no-bid contracts, enforcing t...
9636          2649  Gangsta rap is especially pernicious because i...
14071         3847  Because Capital Punishment is resolute and irr...
38576        11085  Not only that any discrimination is inherently...
17714         4976  And when the associated $225 billion in higher...
36190        10405        Wind power cannot provide this flexibility".
34029         9753  Odorless natural gas presents greater risk of ...
29238         8416  Consumption is wrong and should never be autho...
32941         9494  As well as this colossal mass murder, the US h...
1833           485  Opponents of circumcision have adopted nefario...
13644         3743  This goes to show that even though the law has...
15839         4471   DREAM Act will increase govt revenue and budgets.
21335         6125  "In the noise and misinformation about gene pa...
18986         5422  Potential medical benefits claimed to exist be...
23102         6609        Congressional Research Service Report 4/6/06
2912           811  Jacques Deval, Afin de vivre bel et bien - "Go...
42258        12437  The economic gains were seen as of primary imp...
11777         3210  If students are having sex, they need to know ...

                                                   topic    unitType
20465                            Gays in the US military     premise
34017                                        Natural gas     premise
18823                        Falkland Islands, return of     premise
25141                             Rebuilding New Orleans     premise
14426                                          Democracy     premise
44357                             US-Indian nuclear deal  conclusion
37579                       Progressive tax vs. flat tax     premise
27087                     Is a nuclear Iran intolerable?     premise
9163                         Castration of sex offenders     premise
3775                         Assassination of a Dictator     premise
40007   Should Japan remove limitations on its military?  conclusion
38460                   Puerto Rico statehood in America     premise
150                     $700 billion US economic bailout     premise
40942                                       Solar energy     premise
24650                                 Hydroelectric dams  conclusion
```

```
35350  Obama, meeting with hostile foreign leaders wi...  conclusion
42493                                 Tibet independence     premise
39318                       Right to bear arms in the US     premise
11563                             Compulsory vaccination     premise
26331                           Israeli blockade of Gaza     premise
42205                                   Three Gorges Dam     premise
43269          Trying 9/11 terror suspects in NYC courts  conclusion
17498                  Enhanced interrogation techniques     premise
2154                                         AIG bonuses     premise
13831                                      Death penalty     premise
17621     Equal prize money for male and female athletes  conclusion
33074                      Muhammad cartoons controversy     premise
1183                                            Abortion  conclusion
33424                              Myspace is your space     premise
16524                                 Education vouchers     premise
...                                                  ...         ...
39796             Should governments bailout journalism?     premise
25963                       International Criminal Court     premise
35408                         Online debate and dialogue     premise
22485                                     Greece bailout     premise
10290                      Civil unions vs. gay marriage     premise
43029                                            Torture     premise
32819                       Mission to the Moon or Mars?     premise
15824                                          DREAM Act     premise
26177  In some cases juveniles should be tried as adults     premise
28349                                         Law school     premise
11397             Comprehensive Nuclear-Test-Ban Treaty     premise
9471                            Catholic priest celibacy  conclusion
200                            2009 US economic stimulus     premise
9636                           Censorship of gangsta rap  conclusion
14071                                      Death penalty     premise
38576               Quotas for women in corporate boards     premise
17714                    Estate tax in the United States     premise
36190                             Pickens US energy plan     premise
34029                                        Natural gas  conclusion
29238                              Legalization of drugs     premise
32941             More troops to Afghanistan under Obama     premise
1833                             Adult male circumcision  conclusion
13644                                     DC handgun ban     premise
15839                                          DREAM Act  conclusion
21335                                       Gene patents     premise
18986   Federal funding for embryonic stem cell research  conclusion
23102                 Guest workers in the United States  conclusion
2912                                      Animal testing     premise
42258                                   Three Gorges Dam     premise
11777                                 Condoms in schools  conclusion
```

```
[9297 rows x 4 columns]
```

[ ]:

```
92.9762
```