3_0_Lexical_Complexity_Binary_Classification_Prediction_Baseline_Mode

April 6, 2025

```
[1]: #@title Install Packages
```

```
[2]: !pip install -q transformers
     !pip install -q torchinfo
     !pip install -q datasets
     !pip install -q evaluate
     !pip install -q nltk
     !pip install -q contractions
```

```
                          491.2/491.2 kB
7.9 MB/s eta 0:00:00
                          116.3/116.3 kB
2.2 MB/s eta 0:00:00
                          183.9/183.9 kB
18.8 MB/s eta 0:00:00
                          143.5/143.5 kB
14.7 MB/s eta 0:00:00
                          194.8/194.8 kB
5.3 MB/s eta 0:00:00
                          84.0/84.0 kB
2.5 MB/s eta 0:00:00
                          289.9/289.9 kB
5.0 MB/s eta 0:00:00
                          118.3/118.3 kB
11.8 MB/s eta 0:00:00
```

```
[3]: !sudo apt-get update
     ! sudo apt-get install tree
```

```
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Get:5 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
[3,632 B]
Get:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
```

```
Hit:7 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:8 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:9 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages
[2,775 kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages
[4,148 kB]
Get:11 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,683 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64
Packages [3,978 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages
[1,540 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,092
kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages
[1,241 kB]
Get:16 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,804 kB]
Fetched 28.7 MB in 2s (12.5 MB/s)
Reading package lists… Done
W: Skipping acquire of configured file 'main/source/Sources' as repository
'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to provide
it (sources.list entry misspelt?)
Reading package lists… Done
Building dependency tree… Done
Reading state information… Done
The following NEW packages will be installed:
  tree
0 upgraded, 1 newly installed, 0 to remove and 21 not upgraded.
Need to get 47.9 kB of archives.
After this operation, 116 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tree amd64 2.0.2-1
[47.9 kB]
Fetched 47.9 kB in 0s (355 kB/s)
debconf: unable to initialize frontend: Dialog
debconf: (No usable dialog-like program is installed, so the dialog based
frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 78,
<> line 1.)
debconf: falling back to frontend: Readline
debconf: unable to initialize frontend: Readline
debconf: (This frontend requires a controlling tty.)
debconf: falling back to frontend: Teletype
dpkg-preconfigure: unable to re-open stdin:
Selecting previously unselected package tree.
(Reading database … 122056 files and directories currently installed.)
Preparing to unpack …/tree_2.0.2-1_amd64.deb …
Unpacking tree (2.0.2-1) …
Setting up tree (2.0.2-1) …
Processing triggers for man-db (2.10.2-1) …
```

[4]:

```
#@title Imports
import nltk
from nltk.tokenize import RegexpTokenizer

import evaluate
import transformers

import contractions

from torchinfo import summary
from datasets import load_dataset

from transformers import AutoTokenizer, AutoModel,␣
 ↪AutoModelForSequenceClassification
from transformers import TrainingArguments, Trainer

import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import sklearn

import spacy

from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report
```

[5]:

```
# @title Mount Google Drive
```

[6]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[7]:

```
dir_root = '/content/drive/MyDrive/266-final/'
# dir_data = '/content/drive/MyDrive/266-final/data/'
# dir_data = '/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/'
dir_data = '/content/drive/MyDrive/266-final/data/266-comp-lex-master'
dir_models = '/content/drive/MyDrive/266-final/models/'
dir_results = '/content/drive/MyDrive/266-final/results/'
```

[8]:

```
!tree /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/
  fe-test-labels
```

```
      test_multi_df.csv
      test_single_df.csv
   fe-train
      train_multi_df.csv
      train_single_df.csv
   fe-trial-val
      trial_val_multi_df.csv
      trial_val_single_df.csv
   test-labels
      lcp_multi_test.tsv
      lcp_single_test.tsv
   train
      lcp_multi_train.tsv
      lcp_single_train.tsv
   trial
       lcp_multi_trial.tsv
       lcp_single_trial.tsv

6 directories, 12 files
```

```
[9]: !ls -R /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/:
fe-test-labels  fe-train  fe-trial-val  test-labels  train  trial

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels:
test_multi_df.csv  test_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train:
train_multi_df.csv  train_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val:
trial_val_multi_df.csv  trial_val_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/test-labels:
lcp_multi_test.tsv  lcp_single_test.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/train:
lcp_multi_train.tsv  lcp_single_train.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/trial:
lcp_multi_trial.tsv  lcp_single_trial.tsv
```

```
[10]: !tree /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/
   fe-test-labels
      test_multi_df.csv
      test_single_df.csv
```

```
  fe-train
      train_multi_df.csv
      train_single_df.csv
  fe-trial-val
      trial_val_multi_df.csv
      trial_val_single_df.csv
  test-labels
      lcp_multi_test.tsv
      lcp_single_test.tsv
  train
      lcp_multi_train.tsv
      lcp_single_train.tsv
  trial
       lcp_multi_trial.tsv
       lcp_single_trial.tsv

6 directories, 12 files
```

[11]:

```
#@title Import Data
```

[12]:

```
df_names = [
    "train_single_df",
    "train_multi_df",
    "trial_val_single_df",
    "trial_val_multi_df",
    "test_single_df",
    "test_multi_df"
]

loaded_dataframes = {}

for df_name in df_names:
    if "train" in df_name:
        subdir = "fe-train"
    elif "trial_val" in df_name:
        subdir = "fe-trial-val"
    elif "test" in df_name:
        subdir = "fe-test-labels"
    else:
        subdir = None

    if subdir:
        read_path = os.path.join(dir_data, subdir, f"{df_name}.csv")
        loaded_df = pd.read_csv(read_path)
        loaded_dataframes[df_name] = loaded_df
        print(f"Loaded {df_name} from {read_path}")
```

```python
for df_name, df in loaded_dataframes.items():
    print(f"\n>>> {df_name} shape: {df.shape}")
    if 'binary_complexity' in df.columns:
        print(df['binary_complexity'].value_counts())
        print(df.info())
        print(df.head())

for df_name, df in loaded_dataframes.items():
    globals()[df_name] = df
    print(f"{df_name} loaded into global namespace.")
```

```
Loaded train_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train/train_single_df.csv
Loaded train_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train/train_multi_df.csv
Loaded trial_val_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val/trial_val_single_df.csv
Loaded trial_val_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val/trial_val_multi_df.csv
Loaded test_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels/test_single_df.csv
Loaded test_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels/test_multi_df.csv

>>> train_single_df shape: (7662, 12)
binary_complexity
0    3865
1    3797
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      7662 non-null   object
 1   corpus                  7662 non-null   object
 2   sentence                7662 non-null   object
 3   token                   7655 non-null   object
 4   complexity              7662 non-null   float64
 5   sentence_no_contractions  7662 non-null   object
 6   contraction_expanded    7662 non-null   bool
 7   pos_sequence            7662 non-null   object
 8   dep_sequence            7662 non-null   object
 9   morph_sequence          7662 non-null   object
 10  morph_complexity        7662 non-null   float64
 11  binary_complexity       7662 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 666.1+ KB
```

```
None
                               id corpus  \
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible   
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible   
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible   
3  3BFNCI9LYKQN09BHXHH9CLSX5KP738  bible   
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible   

                                            sentence     token  complexity  \
0  Behold, there came up out of the river seven c…     river    0.000000   
1  I am a fellow bondservant with you and with yo…  brothers    0.000000   
2  The man, the lord of the land, said to us, 'By…  brothers    0.050000   
3  Shimei had sixteen sons and six daughters; but…  brothers    0.150000   
4               "He has put my brothers far from me.  brothers    0.263889   

                            sentence_no_contractions  contraction_expanded  \
0  Behold, there came up out of the river seven c…                 False   
1  I am a fellow bondservant with you and with yo…                 False   
2  The man, the lord of the land, said to us, 'By…                 False   
3  Shimei had sixteen sons and six daughters; but…                  True   
4               "He has put my brothers far from me.                  False   

                                        pos_sequence  \
0  ['ADV', 'PUNCT', 'PRON', 'VERB', 'ADP', 'ADP',…   
1  ['PRON', 'AUX', 'DET', 'ADJ', 'NOUN', 'ADP', '…   
2  ['DET', 'NOUN', 'PUNCT', 'DET', 'PROPN', 'ADP'…   
3  ['PROPN', 'VERB', 'NUM', 'NOUN', 'CCONJ', 'NUM…   
4  ['PUNCT', 'PRON', 'AUX', 'VERB', 'PRON', 'NOUN…   

                                        dep_sequence  \
0  ['advmod', 'punct', 'expl', 'ROOT', 'prt', 'pr…   
1  ['nsubj', 'ROOT', 'det', 'amod', 'attr', 'prep…   
2  ['det', 'nsubj', 'punct', 'det', 'appos', 'pre…   
3  ['nsubj', 'ROOT', 'nummod', 'dobj', 'cc', 'num…   
4  ['punct', 'nsubj', 'aux', 'ROOT', 'poss', 'dob…   

                                      morph_sequence  morph_complexity  \
0  [, PunctType=Comm, , Tense=Past|VerbForm=Fin, …          1.041667   
1  [Case=Nom|Number=Sing|Person=1|PronType=Prs, M…          1.461538   
2  [Definite=Def|PronType=Art, Number=Sing, Punct…          1.354167   
3  [Number=Sing, Tense=Past|VerbForm=Fin, NumType…          1.275862   
4  [PunctSide=Ini|PunctType=Quot, Case=Nom|Gender…          2.500000   

   binary_complexity  
0                  0  
1                  0  
2                  0  
3                  0  
```

```
4                  0

>>> train_multi_df shape: (1517, 12)
binary_complexity
0    759
1    758
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       1517 non-null   object
 1   corpus                   1517 non-null   object
 2   sentence                 1517 non-null   object
 3   token                    1517 non-null   object
 4   complexity               1517 non-null   float64
 5   sentence_no_contractions  1517 non-null   object
 6   contraction_expanded     1517 non-null   bool
 7   pos_sequence             1517 non-null   object
 8   dep_sequence             1517 non-null   object
 9   morph_sequence           1517 non-null   object
 10  morph_complexity         1517 non-null   float64
 11  binary_complexity        1517 non-null   int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 132.0+ KB
None
                               id corpus  \
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible

                                            sentence            token  \
0  but the seventh day is a Sabbath to Yahweh you…      seventh day
1  But let each man test his own work, and then h…         own work
2  To him who by understanding made the heavens; …  loving kindness
3  Remember to me, my God, this also, and spare m…  loving kindness
4  Because your loving kindness is better than li…  loving kindness

   complexity                         sentence_no_contractions  \
0    0.027778  but the seventh day is a Sabbath to Yahweh you…
1    0.050000  But let each man test his own work, and then h…
2    0.050000  To him who by understanding made the heavens; …
3    0.050000  Remember to me, my God, this also, and spare m…
4    0.075000  Because your loving kindness is better than li…
```

```
   contraction_expanded                                       pos_sequence  \
0                  False  ['CCONJ', 'DET', 'ADJ', 'NOUN', 'AUX', 'DET', …
1                  False  ['CCONJ', 'VERB', 'DET', 'NOUN', 'VERB', 'PRON…
2                  False  ['ADP', 'PRON', 'PRON', 'ADP', 'VERB', 'VERB',…
3                  False  ['VERB', 'ADP', 'PRON', 'PUNCT', 'PRON', 'PROP…
4                  False  ['SCONJ', 'PRON', 'ADJ', 'NOUN', 'AUX', 'ADJ',…

                                        dep_sequence  \
0  ['cc', 'det', 'amod', 'nsubj', 'ccomp', 'det',…
1  ['cc', 'ROOT', 'det', 'nsubj', 'ccomp', 'poss'…
2  ['prep', 'pobj', 'nsubj', 'prep', 'pcomp', 'ad…
3  ['ROOT', 'prep', 'pobj', 'punct', 'poss', 'npa…
4  ['mark', 'poss', 'amod', 'nsubj', 'advcl', 'ac…

                                      morph_sequence  morph_complexity  \
0  [ConjType=Cmp, Definite=Def|PronType=Art, Degr…          1.341772
1  [ConjType=Cmp, VerbForm=Inf, , Number=Sing, Ve…          1.608696
2  [, Case=Acc|Gender=Masc|Number=Sing|Person=3|P…          1.562500
3  [VerbForm=Inf, , Case=Acc|Number=Sing|Person=1…          1.590909
4  [, Person=2|Poss=Yes|PronType=Prs, Degree=Pos,…          1.600000

   binary_complexity
0                  0
1                  0
2                  0
3                  0
4                  0

>>> trial_val_single_df shape: (421, 12)
binary_complexity
0    229
1    192
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        421 non-null    object
 1   corpus                    421 non-null    object
 2   sentence                  421 non-null    object
 3   token                     421 non-null    object
 4   complexity                421 non-null    float64
 5   sentence_no_contractions  421 non-null    object
 6   contraction_expanded      421 non-null    bool
 7   pos_sequence              421 non-null    object
 8   dep_sequence              421 non-null    object
 9   morph_sequence            421 non-null    object
 10  morph_complexity          421 non-null    float64
 11  binary_complexity         421 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 36.7+ KB
None
                               id corpus  \
0  3QI9WAYOGQB8GQIR4MDIEF0D2RLS67  bible
1  3T8DUCXY0N6WD9X4RTLK8UN1U929TF  bible
2  3I7KR83SNADXAQ7HXK7S7305BYB9KD  bible
3  3BO3NEOQM0HK9ERYPN0GQIWCPC4IAQ  bible
4  3Y3CZJSZ9KT0W7I0KE38WZHHKSW5RH  bible

                                            sentence token  complexity  \
0  They will not hurt nor destroy in all my holy …   sea    0.000000
1  that sends ambassadors by the sea, even in ves…   sea    0.102941
2  and they entered into the boat, and were going…   sea    0.109375
3  Joseph laid up grain as the sand of the sea, v…   sea    0.160714
4  There will be a highway for the remnant that i…  land    0.000000

                            sentence_no_contractions  contraction_expanded  \
0  They will not hurt nor destroy in all my holy …                 False
1  that sends ambassadors by the sea, even in ves…                 False
2  and they entered into the boat, and were going…                 False
3  Joseph laid up grain as the sand of the sea, v…                 False
4  There will be a highway for the remnant that i…                 False

                                        pos_sequence  \
0  ['PRON', 'AUX', 'PART', 'VERB', 'CCONJ', 'VERB…
1  ['PRON', 'VERB', 'NOUN', 'ADP', 'DET', 'NOUN',…
2  ['CCONJ', 'PRON', 'VERB', 'ADP', 'DET', 'NOUN'…
3  ['PROPN', 'VERB', 'ADP', 'NOUN', 'ADP', 'DET',…
4  ['PRON', 'AUX', 'AUX', 'DET', 'NOUN', 'ADP', '…

                                        dep_sequence  \
0  ['nsubj', 'aux', 'neg', 'ccomp', 'cc', 'conj',…
1  ['nsubj', 'ROOT', 'dobj', 'prep', 'det', 'pobj…
2  ['cc', 'nsubj', 'ROOT', 'prep', 'det', 'pobj',…
3  ['nsubj', 'ROOT', 'prt', 'dobj', 'prep', 'det'…
4  ['expl', 'aux', 'ROOT', 'det', 'attr', 'prep',…

                                      morph_sequence  morph_complexity  \
0  [Case=Nom|Number=Plur|Person=3|PronType=Prs, V…          1.129032
1  [PronType=Rel, Number=Sing|Person=3|Tense=Pres…          1.263158
2  [ConjType=Cmp, Case=Nom|Number=Plur|Person=3|P…          1.437500
3  [Number=Sing, Tense=Past|VerbForm=Fin, , Numbe…          1.400000
4  [, VerbForm=Fin, VerbForm=Inf, Definite=Ind|Pr…          1.277778

   binary_complexity
```

```
0                  0
1                  0
2                  0
3                  0
4                  0

>>> trial_val_multi_df shape: (99, 12)
binary_complexity
1    51
0    48
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        99 non-null     object
 1   corpus                    99 non-null     object
 2   sentence                  99 non-null     object
 3   token                     99 non-null     object
 4   complexity                99 non-null     float64
 5   sentence_no_contractions  99 non-null     object
 6   contraction_expanded      99 non-null     bool
 7   pos_sequence              99 non-null     object
 8   dep_sequence              99 non-null     object
 9   morph_sequence            99 non-null     object
 10  morph_complexity          99 non-null     float64
 11  binary_complexity         99 non-null     int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 8.7+ KB
None
                               id corpus  \
0  31HLTCK4BLVQ5BO1AUR91TX9V9IVGH  bible
1  389A2A304OIXVY7G5B71Q9M43LE0CL  bible
2  31N9JPQXIPIRX2A3S9N0CCFXO6TNHR  bible
3  3JVP4ZJHDPSO81TGXL3N1CKZGQY0IN  bible
4  3JAOYN9IHL25ZQAUV5EJZ4GH0KL33L  bible

                                            sentence          token  \
0  The name of one son was Gershom, for Moses sai…   foreign land
1  unleavened bread, unleavened cakes mixed with …    wheat flour
2  However the high places were not taken away; t…  burnt incense
3  and he burnt incense of sweet spices on it, as…  burnt incense
4  The same day the king made the middle of the c…   bronze altar

   complexity                           sentence_no_contractions  \
0    0.000000  The name of one son was Gershom, for Moses sai…
1    0.157895  unleavened bread, unleavened cakes mixed with …
```

```
2    0.200000  However the high places were not taken away; t…
3    0.250000  and he burnt incense of sweet spices on it, as…
4    0.214286  The same day the king made the middle of the c…

   contraction_expanded                                        pos_sequence  \
0                 False  ['DET', 'NOUN', 'ADP', 'NUM', 'NOUN', 'AUX', '…
1                 False  ['ADJ', 'NOUN', 'PUNCT', 'ADJ', 'NOUN', 'VERB'…
2                 False  ['ADV', 'DET', 'ADJ', 'NOUN', 'AUX', 'PART', '…
3                 False  ['CCONJ', 'PRON', 'VERB', 'NOUN', 'ADP', 'ADJ'…
4                 False  ['DET', 'ADJ', 'NOUN', 'DET', 'NOUN', 'VERB', …

                                         dep_sequence  \
0  ['det', 'nsubj', 'prep', 'nummod', 'pobj', 'RO…
1  ['amod', 'dep', 'punct', 'amod', 'appos', 'acl…
2  ['advmod', 'det', 'amod', 'nsubjpass', 'auxpas…
3  ['cc', 'nsubj', 'ROOT', 'dobj', 'prep', 'amod'…
4  ['det', 'amod', 'npadvmod', 'det', 'nsubj', 'c…

                                       morph_sequence  morph_complexity  \
0  [Definite=Def|PronType=Art, Number=Sing, , Num…           1.520000
1  [Degree=Pos, Number=Sing, PunctType=Comm, Degr…           1.200000
2  [, Definite=Def|PronType=Art, Degree=Pos, Numb…           1.190476
3  [ConjType=Cmp, Case=Nom|Gender=Masc|Number=Sin…           1.466667
4  [Definite=Def|PronType=Art, Degree=Pos, Number…           1.352113

   binary_complexity
0                  0
1                  0
2                  0
3                  0
4                  0

>>> test_single_df shape: (917, 12)
binary_complexity
0    476
1    441
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       917 non-null    object
 1   corpus                   917 non-null    object
 2   sentence                 917 non-null    object
 3   token                    917 non-null    object
 4   complexity               917 non-null    float64
 5   sentence_no_contractions 917 non-null    object
 6   contraction_expanded      917 non-null    bool
 7   pos_sequence              917 non-null    object
 8   dep_sequence              917 non-null    object
 9   morph_sequence            917 non-null    object
 10  morph_complexity          917 non-null    float64
 11  binary_complexity         917 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 79.8+ KB
None
                               id corpus  \
0  3K8CQCU3KE19US5SN890DFPK3SANWR  bible
1  3Q2T3FD0ON86LCI41NJYV3PN0BW3MV  bible
2  3ULIZ0H1VA5C32JJMK0TQ8Z4GUS51B  bible
3  3BFF0DJK8XCEIOT30ZLBPPSRMZQTSD  bible
4  3QREJ3J433XSBS8QMHAICCR0BQ1LKR  bible

                                            sentence     token  complexity  \
0  But he, beckoning to them with his hand to be …      hand    0.000000
1  If I forget you, Jerusalem, let my right hand …      hand    0.197368
2  the ten sons of Haman the son of Hammedatha, t…      hand    0.200000
3  Let your hand be lifted up above your adversar…      hand    0.267857
4  Abimelech chased him, and he fled before him, …  entrance    0.000000

                            sentence_no_contractions  contraction_expanded  \
0  But he, beckoning to them with his hand to be …                 False
1  If I forget you, Jerusalem, let my right hand …                 False
2  the ten sons of Haman the son of Hammedatha, t…                  True
3  Let your hand be lifted up above your adversar…                 False
4  Abimelech chased him, and he fled before him, …                 False

                                        pos_sequence  \
0  ['CCONJ', 'PRON', 'PUNCT', 'VERB', 'ADP', 'PRO…
1  ['SCONJ', 'PRON', 'VERB', 'PRON', 'PUNCT', 'PR…
2  ['DET', 'NUM', 'NOUN', 'ADP', 'PROPN', 'DET', …
3  ['VERB', 'PRON', 'NOUN', 'AUX', 'VERB', 'ADP',…
4  ['PROPN', 'VERB', 'PRON', 'PUNCT', 'CCONJ', 'P…

                                        dep_sequence  \
0  ['cc', 'nsubj', 'punct', 'advcl', 'prep', 'pob…
1  ['mark', 'nsubj', 'advcl', 'dobj', 'punct', 'n…
2  ['det', 'nummod', 'ROOT', 'prep', 'pobj', 'det…
3  ['ROOT', 'poss', 'nsubjpass', 'auxpass', 'ccom…
4  ['nsubj', 'ROOT', 'dobj', 'punct', 'cc', 'nsub…

                                      morph_sequence  morph_complexity  \
0  [ConjType=Cmp, Case=Nom|Gender=Masc|Number=Sin…          1.703704
1  [, Case=Nom|Number=Sing|Person=1|PronType=Prs,…          1.800000
2  [Definite=Def|PronType=Art, NumType=Card, Numb…          1.269231
```

```
3  [VerbForm=Inf, Person=2|Poss=Yes|PronType=Prs,…          1.250000
4  [Number=Sing, Tense=Past|VerbForm=Fin, Case=Ac…          1.652174

   binary_complexity
0                  0
1                  0
2                  0
3                  0
4                  0

>>> test_multi_df shape: (184, 12)
binary_complexity
1    99
0    85
Name: count, dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        184 non-null    object
 1   corpus                    184 non-null    object
 2   sentence                  184 non-null    object
 3   token                     184 non-null    object
 4   complexity                184 non-null    float64
 5   sentence_no_contractions  184 non-null    object
 6   contraction_expanded      184 non-null    bool
 7   pos_sequence              184 non-null    object
 8   dep_sequence              184 non-null    object
 9   morph_sequence            184 non-null    object
 10  morph_complexity          184 non-null    float64
 11  binary_complexity         184 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 16.1+ KB
None
                               id corpus  \
0  3UXQ63NLAAMRIP4WG4XPD98AOYOBLX  bible
1  3FJ2RVH25Z62TA3R8E1O77EBUYU92W  bible
2  3YO4AH2FPDK1PZHZAT8WAEBL70EQ0F  bible
3  3X52SWXE0X5Q3O81YI0MX4V84QTCWZ  bible
4  32K26U12DN0NTREA84Q1V8UCIH2VD7  bible

                                            sentence           token  \
0  for he had an only daughter, about twelve year…   only daughter
1  All these were cities fortified with high wall…      high walls
2  In the morning, 'It will be foul weather today…   weather today
3  Her young children also were dashed in pieces …  young children
4  All king Solomon's drinking vessels were of go…       pure gold
```

```
   complexity                           sentence_no_contractions  \
0    0.025000  for he had an only daughter, about twelve year…
1    0.100000  All these were cities fortified with high wall…
2    0.125000  In the morning, 'It will be foul weather today…
3    0.160714  Her young children also were dashed in pieces …
4    0.178571  All king Solomon's drinking vessels were of go…

   contraction_expanded                                     pos_sequence  \
0                 False  ['SCONJ', 'PRON', 'VERB', 'DET', 'ADJ', 'NOUN'…
1                 False  ['DET', 'PRON', 'AUX', 'NOUN', 'VERB', 'ADP', …
2                 False  ['ADP', 'DET', 'NOUN', 'PUNCT', 'PUNCT', 'PRON…
3                 False  ['PRON', 'ADJ', 'NOUN', 'ADV', 'AUX', 'VERB', …
4                 False  ['DET', 'NOUN', 'PROPN', 'PART', 'NOUN', 'NOUN…

                                      dep_sequence  \
0  ['mark', 'nsubj', 'ROOT', 'det', 'amod', 'dobj…
1  ['predet', 'nsubj', 'ROOT', 'attr', 'acl', 'pr…
2  ['prep', 'det', 'pobj', 'punct', 'punct', 'nsu…
3  ['poss', 'amod', 'nsubjpass', 'advmod', 'auxpa…
4  ['det', 'compound', 'poss', 'case', 'compound'…

                                    morph_sequence  morph_complexity  \
0  [, Case=Nom|Gender=Masc|Number=Sing|Person=3|P…          1.722222
1  [, Number=Plur|PronType=Dem, Mood=Ind|Tense=Pa…          1.136364
2  [, Definite=Def|PronType=Art, Number=Sing, Pun…          1.476190
3  [Gender=Fem|Number=Sing|Person=3|Poss=Yes|Pron…          1.514286
4  [, Number=Sing, Number=Sing, , Number=Sing, Nu…          1.162791

   binary_complexity
0                  0
1                  0
2                  0
3                  0
4                  0
train_single_df loaded into global namespace.
train_multi_df loaded into global namespace.
trial_val_single_df loaded into global namespace.
trial_val_multi_df loaded into global namespace.
test_single_df loaded into global namespace.
test_multi_df loaded into global namespace.
```

- Functional tests pass, we can proceed with Baseline Modeling

[23]:
```
#@title Experiment 1: Baseline Modeling
```

0.0.1 Reminders:

- Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- F1 Score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Cosine Similarity

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|}$$

- Jaccard Similarity

$$\text{Jaccard Similarity} = \frac{|A \cap B|}{|A \cup B|}$$

- Overlap Similarity (Overlap Coefficient)

$$\text{Overlap Similarity} = \frac{|A \cap B|}{\min(|A|, |B|)}$$

- Dice Coefficient

$$\text{Dice Coefficient} = \frac{2\,|A \cap B|}{|A| + |B|}$$

0.1 Naive Bayes

0.1.1 X = Sentence: contractions and no contractions

- sentence no contractions

[20]:

```
train_df = train_single_df
val_df = trial_val_single_df

vectorizer = TfidfVectorizer()  # just on 'sentence_no_contractions'
X_train = vectorizer.fit_transform(train_df['sentence_no_contractions'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['sentence_no_contractions'])
y_val = val_df['binary_complexity']

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

```
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.58      0.74      0.65       229
           1       0.55      0.38      0.44       192

    accuracy                           0.57       421
   macro avg       0.57      0.56      0.55       421
weighted avg       0.57      0.57      0.56       421
```

- sentence with contractions

[26]:

```
train_df = train_single_df
val_df = trial_val_single_df

vectorizer = TfidfVectorizer()  # just on 'sentence'
X_train = vectorizer.fit_transform(train_df['sentence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['sentence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.58      0.74      0.65       229
           1       0.55      0.38      0.44       192

    accuracy                           0.57       421
   macro avg       0.57      0.56      0.55       421
weighted avg       0.57      0.57      0.56       421
```

- sentence no contractions

[25]:

```
train_df = train_multi_df
val_df = trial_val_multi_df

vectorizer = TfidfVectorizer()  # just on 'sentence_no_contractions'
X_train = vectorizer.fit_transform(train_df['sentence_no_contractions'])
y_train = train_df['binary_complexity']
```

```
X_val = vectorizer.transform(val_df['sentence_no_contractions'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.52      0.67      0.58        48
           1       0.57      0.41      0.48        51

    accuracy                           0.54        99
   macro avg       0.54      0.54      0.53        99
weighted avg       0.54      0.54      0.53        99
```

- sentence with contractions

[27]:

```
train_df = train_multi_df
val_df = trial_val_multi_df

vectorizer = TfidfVectorizer()  # just on 'sentence'
X_train = vectorizer.fit_transform(train_df['sentence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['sentence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.52      0.67      0.58        48
           1       0.57      0.41      0.48        51

    accuracy                           0.54        99
   macro avg       0.54      0.54      0.53        99
weighted avg       0.54      0.54      0.53        99
```

- Score is higher than expected for a Naive Bayes model
- There is no difference in performance when using the input sequence of the sentence with and without contractions

0.1.2 X = pos_sequence: Part-of-Speech Tags

- POS Tags: Extracts the part-of-speech (POS) tags for each token (e.g., “DET”, “NOUN”, “VERB”).

[29]:

```
train_df = train_single_df
val_df = trial_val_single_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['pos_sequence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['pos_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.60      0.67      0.63       229
           1       0.54      0.46      0.50       192

    accuracy                           0.57       421
   macro avg       0.57      0.57      0.56       421
weighted avg       0.57      0.57      0.57       421
```

[32]:

```
train_df = train_multi_df
val_df = trial_val_multi_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['pos_sequence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['pos_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.58      0.54      0.56        48
           1       0.59      0.63      0.61        51
```

```
    accuracy                           0.59        99
   macro avg       0.59      0.58      0.58        99
weighted avg       0.59      0.59      0.59        99
```

- Part of Speech tags outperform raw input sequence

X = dep_sequence: Dependency Tags

- Dependency Tags: Extracts the syntactic dependency labels for each token (e.g., "det", "nsubj", "ROOT").

[30]:

```
train_df = train_single_df
val_df = trial_val_single_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['dep_sequence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['dep_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.61      0.60      0.60       229
           1       0.53      0.54      0.54       192

    accuracy                           0.57       421
   macro avg       0.57      0.57      0.57       421
weighted avg       0.57      0.57      0.57       421
```

[35]:

```
train_df = train_multi_df
val_df = trial_val_multi_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['dep_sequence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['dep_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
```

```
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.51      0.46      0.48        48
           1       0.54      0.59      0.56        51

    accuracy                           0.53        99
   macro avg       0.52      0.52      0.52        99
weighted avg       0.52      0.53      0.52        99
```

X = morph_sequence: Morphological Features

- For each token, the morphological attributes have been retrieved for each token

[33]:

```
train_df = train_single_df
val_df = trial_val_single_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['morph_sequence'])
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['morph_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.62      0.59      0.60       229
           1       0.53      0.57      0.55       192

    accuracy                           0.58       421
   macro avg       0.58      0.58      0.58       421
weighted avg       0.58      0.58      0.58       421
```

[34]:

```
train_df = train_multi_df
val_df = trial_val_multi_df

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(train_df['morph_sequence'])
```

```
y_train = train_df['binary_complexity']

X_val = vectorizer.transform(val_df['morph_sequence'])
y_val = val_df['binary_complexity']

clf = MultinomialNB()
clf.fit(X_train, y_train)
preds = clf.predict(X_val)
print(classification_report(y_val, preds))
```

```
              precision    recall  f1-score   support

           0       0.62      0.52      0.57        48
           1       0.61      0.71      0.65        51

    accuracy                           0.62        99
   macro avg       0.62      0.61      0.61        99
weighted avg       0.62      0.62      0.61        99
```

0.2 Transformers Models

0.3 BERT

[14]:

```

```