Ikhlaq Ahmad

NetID: ixa190000

Assignment: Text Classification

CS 4395 - Dr. Mazidi

Text Classification

The purpose of this notebook is to utilize sklearn libraray to classify a data set using different probablistic approaches. For this notebook, the following techniques were used:

1. Naive Bayes
2. Logistic Regression
3. Neural Networks

Data set: amazon_books_Data.csv

All three techniques use multiclass appraoch and the same data to find out the accuracy impact.

```
# imports
import pandas as pd
import numpy as np

# csv file only gets columns 14 - 16
file_data = pd.read_csv("/content/amazon_books_Data.csv", header=0,
usecols=[14, 16])
print('rows and columns:', file_data.shape)
print(file_data.head())
```

```
rows and columns: (100, 2)
                                         review_body Sentiment_books
0                "I love it and so does my students!"        positive
1  "My wife and I ordered 2 books and gave them a...        positive
2  "Great book just like all the others in the se...        positive
3                                     "So beautiful"        positive
4  "Enjoyed the author's story and his quilts are...        positive
```

```
# set up X and y
X = file_data.review_body
y = file_data.Sentiment_books

# X
X.head()
```

```
0                 "I love it and so does my students!"
1    "My wife and I ordered 2 books and gave them a...
```

```
2    "Great book just like all the others in the se...
3                                        "So beautiful"
4    "Enjoyed the author's story and his quilts are...
Name: review_body, dtype: object
```

```
# y
y[:16]
```

```
0     positive
1     positive
2     positive
3     positive
4     positive
5     negaitve
6     positive
7     positive
8     positive
9     positive
10    positive
11    positive
12    positive
13    positive
14    negaitve
15    positive
Name: Sentiment_books, dtype: object
```

```
# train text
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, train_size=0.8, random_state=1234)
```

```
X_train.shape
```

```
(80,)
```

```
# remove stop words using nltk
import nltk
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer

# initializer tf-idf vectorizor
from sklearn.feature_extraction.text import TfidfVectorizer

stopwords = stopwords.words('english')
vectorizer = TfidfVectorizer(stop_words = stopwords)
vectorizer.fit(X_train)

# apply tfidf vectorizer
X_train = vectorizer.fit_transform(X_train)  # fit and transform the
train data
```

```
X_test = vectorizer.transform(X_test)       # transform only the test
data

# print vocabualry and its length
print("Vocabulary: ", vectorizer.vocabulary_)
print("Lenght: ", len(vectorizer.vocabulary_))
```

```
Vocabulary:  {'quality': 761, 'product': 744, 'fast': 348, 'shipping':
862, 'great': 420, 'like': 563, 'fact': 340, 'small': 879, 'easy':
283, 'save': 828, 'store': 915, 'wedding': 1060, 'day': 229, 'wait':
1045, 'use': 1031, 'helpful': 448, 'med': 603, 'surg': 941, 'class':
175, 'used': 1032, 'study': 926, 'exams': 326, 'addition': 22, 'test':
971, 'books': 123, 'really': 776, 'summarized': 932, 'system': 951,
'focused': 371, 'nclex': 640, 'thought': 984, 'important': 476,
'still': 913, 'working': 1076, 'john': 512, 'fitzgerald': 366,
'interest': 497, 'full': 390, 'disclosure': 266, 'worked': 1075,
'together': 994, 'red': 783, 'hen': 452, 'press': 739, 'would': 1081,
'nudge': 659, 'say': 830, 'book': 122, 'mind': 610, '34': 7, 'skynet':
876, 'becomes': 100, 'self': 846, 'aware': 85, '14': 1, 'eastern':
282, 'time': 991, 'august': 76, '29th': 6, 'sense': 847, 'long': 575,
'solipsistic': 884, 'narcissistic': 637, 'way': 1056, 'keen': 516,
'observer': 664, 'consumer': 197, 'origins': 681, 'fine': 363,
'distinctions': 268, 'continua': 204, 'grand': 418, 'schemes': 832,
'minute': 612, 'details': 249, 'likely': 564, 'began': 105,
'observing': 665, 'contemplating': 199, 'information': 484, 'moment':
617, 'experienced': 333, 'glare': 407, 'light': 561, 'delivery': 242,
'room': 822, 'never': 646, 'stopped': 914, 'interestingly': 500,
'remarkable': 791, 'thinks': 981, 'speaks': 895, 'larger': 537,
'questions': 764, 'think': 979, 'came': 147, 'sapient': 826, 'first':
365, 'place': 717, 'develop': 252, 'thinking': 980, 'souls': 890,
'space': 892, 'keeping': 518, 'language': 536, 'prose': 749,
'tercets': 968, 'basic': 94, 'unadorned': 1020, 'free': 387,
'flowing': 369, 'accomplishes': 16, 'poetry': 728, 'significance':
867, 'elemental': 289, 'beauty': 97, 'left': 551, 'brain': 131,
'contemplation': 200, 'structure': 922, 'systems': 952, 'aligns': 36,
'right': 814, 'wonder': 1070, 'whimsy': 1063, 'neither': 644,
'hemisphere': 451, 'dominates': 274, 'work': 1074, 'reader': 770,
'expect': 329, 'unexpected': 1022, 'rewards': 809, 'poems': 727,
'curiosity': 222, 'orientation': 680, 'universe': 1026, 'sorrow': 888,
'finding': 362, 'center': 156, 'surprising': 943, 'hilarity': 455,
'make': 589, 'idea': 469, 'rocks': 818, 'funny': 391, 'teaching': 959,
'encourage': 295, 'students': 924, 'examine': 321, 'masterful': 599,
'skill': 875, 'personification': 705, 'philosophy': 709, 'wrestle':
1082, 'experiences': 334, 'phenomena': 708, 'ask': 69, 'psychology':
753, 'neuro': 645, 'biology': 116, 'candidates': 148, 'experience':
332, 'inside': 490, 'physics': 711, 'explore': 337, 'process': 742,
'era': 307, 'concepts': 191, 'continually': 205, 'challenged': 160,
'updated': 1030, 'divinity': 269, 'consider': 196, 'creation': 215,
'point': 729, 'view': 1044, 'created': 214, 'weighs': 1061, 'many':
594, 'approaches': 62, 'devour': 256, 'one': 673, 'two': 1019,
'sittings': 873, 'read': 769, 'genesis': 397, 'hawking': 438,
```

'introduction': 502, 'meditation': 606, 'ever': 315, 'reading': 772, 'good': 414, 'seen': 845, 'videos': 1043, 'youtube': 1093, 'web': 1059, 'page': 686, 'learningbeekeeping': 546, 'com': 181, 'info': 482, 'beekeeping': 104, 'valuable': 1037, 'tool': 998, 'library': 557, 'recommend': 782, 'new': 647, 'beekeepers': 103, 'want': 1048, 'get': 400, 'bee': 102, 'content': 201, 'given': 405, 'stars': 908, 'binding': 115, 'break': 132, 'pages': 687, 'falling': 341, 'months': 620, 'got': 416, 'bought': 126, 'quite': 767, 'tech': 962, 'far': 347, 'cheapest': 166, 'put': 759, 'writing': 1084, 'star': 907, 'reviews': 808, 'dreck': 277, 'either': 288, 'part': 694, 'publisher': 755, 'marketing': 597, 'department': 245, 'fanboys': 345, 'author': 78, 'actually': 20, 'sleeping': 877, 'writer': 1083, 'br': 130, 'main': 587, 'character': 163, 'starts': 910, 'annoying': 52, 'immature': 473, 'centered': 157, 'little': 570, 'bore': 124, 'rides': 813, 'wave': 1055, 'background': 87, 'story': 916, 'seemed': 844, 'irrelevant': 504, 'occupying': 667, 'pack': 685, 'additional': 23, 'thickness': 976, 'spine': 899, 'development': 254, 'actual': 19, 'house': 464, 'analysis': 48, 'events': 314, 'doctor': 273, 'leading': 542, 'research': 795, 'laughable': 540, 'whole': 1064, 'rehashed': 787, 'haunted': 437, 'works': 1077, 'screamed': 834, 'done': 275, 'spoiler': 902, 'alert': 35, 'matters': 600, 'hopefully': 462, 'waste': 1052, 'anyway': 57, 'enjoyable': 300, 'committing': 186, 'suicide': 930, 'second': 838, 'last': 538, 'happened': 431, 'chapter': 162, 'maybe': 602, 'developed': 253, 'something': 886, 'marginally': 596, 'palatable': 689, 'bottom': 125, 'line': 566, 'unfortunate': 1023, 'enough': 303, 'cut': 225, 'rabbit': 768, 'cage': 146, 'arrived': 68, 'estimated': 312, 'date': 227, 'advertised': 28, 'enjoyed': 301, 'quilts': 766, 'incredible': 480, 'plans': 719, 'three': 987, 'held': 445, 'attention': 72, 'informative': 485, 'well': 1062, 'entertaining': 305, 'happy': 434, 'source': 891, 'going': 412, 'olympic': 672, 'peninsula': 699, 'need': 641, 'lots': 582, 'suspense': 945, 'action': 18, 'fear': 351, 'terror': 970, 'also': 42, 'includes': 479, 'conclusion': 192, 'authors': 79, 'pull': 757, 'surprise': 942, 'end': 296, 'comes': 183, 'crime': 218, 'happens': 432, 'absolutely': 13, 'love': 583, 'perfect': 701, 'balance': 89, 'romance': 821, 'twists': 1018, 'keeps': 519, 'intensity': 495, 'level': 555, 'high': 453, 'seth': 857, 'tonya': 996, 'serious': 854, 'issues': 505, 'dealt': 233, 'ex': 319, 'girlfriend': 403, 'walked': 1046, 'injured': 489, 'leaving': 548, 'disillusion': 267, 'broken': 139, 'heart': 441, 'stalker': 905, 'boyfriend': 128, 'prison': 741, 'trying': 1016, 'kill': 523, 'died': 258, 'bullfighting': 142, 'accident': 15, 'road': 816, 'ahead': 32, 'attraction': 74, 'feel': 353, 'strong': 921, 'connection': 195, 'go': 410, 'deeper': 238, 'respect': 800, 'protectiveness': 750, 'kindness': 527, 'towards': 1003, 'things': 978, 'common': 187, 'seamless': 835, 'team': 960, 'independent': 481, 'loved': 584, 'natural': 638, 'affection': 30, 'genuine': 399, 'sweet': 948, 'took': 997, 'care': 152, 'charming': 165, 'likable': 562, 'characters': 164, 'loving': 585, 'families': 342, 'makes': 590, 'smile': 880, 'string': 920, 'crimes': 219, 'mysteries': 634, 'presented': 738, 'solid': 883, 'plot': 725, 'entertain': 304, 'touch': 1001, 'four': 385, 'spoons': 903, 'teaspoon': 961, 'side': 866, 'authentic': 77, 'mexican': 608, 'recipes': 781, 'directions': 262, 'food': 376, 'preparations': 736, 'wonderful': 1071, 'bring': 135, 'family': 343, 'table': 953, 'chef': 167, 'cross': 220, 'referenced': 785, 'throughout': 988, 'gave': 394, 'serving': 856, 'suggestions': 929, 'specified': 897, 'alternate': 43, 'ingredients': 486, 'vary': 1039, 'kick': 521, 'picture': 712, 'worth': 1080, 'thousand': 986, 'words': 1073, 'know': 531, 'recipe': 780, 'supposed': 939, 'look': 576, 'ricardo': 810, 'knocked': 530, 'ball': 90, 'park': 693, 'thank': 973, 'favorite': 350, 'jeffrey': 510, 'archer': 63, 'friend': 388, 'unique': 1025, 'blessed': 119, 'anything': 56, 'bikes': 114, 'learned': 544, 'lot': 581, 'daughter': 228, 'mals': 592, 'someone': 885, 'knows': 533, 'jaynie': 508, 'journey': 514, 'odds': 669, 'demonstrates': 243, 'environment': 306, 'determination': 251, 'greater': 421, 'healer': 439, 'areas': 64, 'medicine': 605, 'currently': 223, 'capable': 150, 'jeff': 509, 'edgecombe': 284, 'football': 377, 'seasons': 836, 'restarts': 801, 'northwest': 657, 'university': 1027, 'labor': 534, 'decision': 237, 'players': 722, 'lost': 580, 'handed': 430, 'topical': 1000, 'reads': 773, 'combination': 182, 'sports': 904, 'finance': 360, 'section': 840, 'newspages': 649, 'college': 180, 'remade': 790, 'administrators': 25, 'ncaa': 639, 'nfl': 650, 'play': 721, 'role': 820, 'watching': 1054, 'simulation': 870, 'although': 44, 'could': 209, 'real': 774, 'secrets': 839, 'shallows': 860, 'monastery': 618, 'murders': 630, 'series': 853, 'grabbed': 417, 'drew': 278, 'devil': 255, 'detals': 250, 'tightened': 990, 'hold': 458, 'addicted': 21, 'mother': 621, 'superior': 935, 'murder': 628, 'catholic': 154, 'school': 833, 'boys': 129, 'murderer': 629, 'found': 382, 'caught': 155, 'beautiful': 95, 'golden': 413, 'tressed': 1008, 'nun': 661, 'boy': 127, 'apparent': 58, 'ten': 967, 'years': 1090, 'earlier': 279, 'tie': 989, 'recent': 779, 'spirit': 900, 'answers': 54, 'revealed': 805, 'mystery': 635, 'broadens': 138, 'arise': 65, 'gripping': 424, 'immensely': 475, 'must': 633, 'purchase': 758, 'exceptionally': 328, 'written': 1085, 'researched': 796, 'review': 806, 'eye': 339, 'opener': 675, 'interested': 498, 'nutrition': 662, 'politicalization': 732, 'health': 440, 'novel': 658, 'lester': 553, 'resonates': 798, 'levels': 556, 'anne': 51, 'rice': 811, 'least': 547, 'times': 992, 'buy': 144, 'value': 1038, 'unfortunately': 1024, 'realize': 775, 'buying': 145, 'ordered': 679, 'item': 506, 'huge': 466, 'fan': 344, 'richard': 812, 'paul': 696, 'evens': 313, 'give': 404, 'order': 678, 'much': 625, 'however': 465, 'customers': 224, 'essay': 309, 'barely': 92, '17': 2, 'heavily': 444, 'spaced': 893, 'length': 552, 'publishers': 756, 'tried': 1009, 'stretch': 918, 'appears': 60, 'triple': 1012, 'spacing': 894, 'rip': 815, 'mr': 624, 'believe': 109, 'bit': 117, 'published': 754, 'alone': 39, 'included': 478, 'collection': 178, 'short': 864, 'pieces': 715, 'emp': 293, 'music': 632, 'museum': 631, 'seattle': 837, 'washington': 1051, 'definitely': 240, 'shows': 865, 'another': 53, 'perspective': 706, 'band': 91, 'since': 871, 'remember': 792, 'complaint': 188, 'runner': 823, 'former': 379, 'autobiography': 81, 'insight': 491, 'pop': 733,

'history': 457, 'greatest': 422, 'utmost': 1036, 'nile': 653,
'rogers': 819, 'anyone': 55, 'wants': 1049, '70': 11, 'genre': 398,
'include': 477, 'studies': 925, 'pleased': 724, 'inspired': 492,
'nly': 654, 'looke': 577, 'life': 558, 'minister': 611, 'others': 682,
'helping': 449, 'become': 99, 'disciples': 265, 'examines': 322,
'ways': 1057, 'people': 700, 'progress': 746, 'relationship': 789,
'practical': 734, 'steps': 912, 'take': 954, 'along': 40, 'thanks':
974, 'opportunity': 676, 'focuses': 372, 'special': 896, 'needs': 643,
'autism': 80, 'hopes': 463, 'understand': 1021, 'son': 887,
'spectrum': 898, 'help': 446, 'cope': 208, 'brent': 134, 'thoroughly':
982, 'explained': 335, 'stuffs': 927, 'body': 120, 'growing': 425,
'contains': 198, 'advice': 29, 'parents': 692, 'guardians': 426,
'properly': 748, 'provide': 751, 'physical': 710, 'emotional': 292,
'kids': 522, 'learning': 545, 'illustrations': 471, 'enjoy': 299,
'keep': 517, 'suspected': 944, '10': 0, '20': 3, 'sure': 940, 'best':
110, 'okay': 671, 'unnecessarily': 1028, 'taken': 955, 'trouble':
1014, 'craft': 213, 'suspenseful': 946, 'turn': 1017, 'mixed': 614,
'feelings': 354, 'vibrant': 1041, 'energetic': 297, 'assertive': 70,
'woman': 1069, 'yet': 1092, 'became': 98, 'weak': 1058, 'submissive':
928, 'manipulative': 593, 'husband': 467, 'hard': 435, 'identiify':
470, 'additionally': 24, 'interesting': 499, 'attracted': 73,
'getting': 401, 'build': 141, 'heat': 443, '500': 9, '250': 5,
'feels': 355, 'stretched': 919, 'sequences': 851, 'unnecessary': 1029,
'dialogue': 257, 'superfluous': 934, 'scenes': 831, 'establish': 310,
'already': 41, 'established': 311, 'wondering': 1072, 'nora': 656,
'roberts': 817, 'days': 230, 'number': 660, 'seem': 843, 'diluted':
261, 'behind': 106, 'burning': 143, 'goes': 411, 'glad': 406, 'paid':
688, 'steamed': 911, 'wrote': 1087, 'fewer': 358, 'year': 1089,
'better': 111, 'ones': 674, 'happier': 433, 'aircam': 33, 'airwar':
34, 'serie': 852, '24': 4, 'items': 507, '48': 8, 'pictures': 713,
'beautitful': 96, 'drawings': 276, 'amazon': 47, 'dealers': 231,
'kind': 525, 'regard': 786, 'excellent': 327, 'easier': 280, 'google':
415, 'elisa': 290, 'find': 361, 'gluten': 409, 'listed': 568, 'globe':
408, 'trotting': 1013, 'gazillionaire': 395, 'talk': 957, 'tough':
1002, 'personal': 703, 'chefs': 168, 'trainers': 1005,
'nutritionists': 663, 'whomever': 1065, 'else': 291, 'pays': 698,
'lift': 560, 'finger': 364, 'seriously': 855, 'literally': 569,
'tries': 1010, 'relate': 788, 'lifestyle': 559, 'readers': 771,
'though': 983, 'living': 574, 'fantasy': 346, 'world': 1078, 'see':
842, 'results': 803, 'certainly': 158, 'helps': 450, 'journaling':
513, 'bible': 112, 'deborah': 235, 'harness': 436, 'trilogy': 1011,
'hope': 461, 'movie': 622, 'continue': 206, 'sort': 889, 'lame': 535,
'folds': 373, 'clothes': 177, 'joy': 515, 'zero': 1094, 'received':
778, 'wrong': 1086, 'teen': 964, 'expressed': 338, 'sirens': 872,
'decided': 236, 'highly': 454, 'presbyterian': 737, 'confessions':
194, 'multiple': 627, 'creeds': 217, 'denomination': 244, 'guidebook':
428, 'lay': 541, 'person': 702, 'context': 203, 'today': 993, 'av':
82, 'guide': 427, 'average': 84, 'admired': 26, 'kimberly': 524,
'guilfoyle': 429, 'pre': 735, 'fox': 386, 'news': 648, 'expecting':
331, 'debit': 234, 'credit': 216, 'card': 151, 'making': 591,

'friends': 389, 'assuming': 71, 'target': 958, 'student': 923,
'disappointing': 264, 'said': 825, 'useful': 1033, 'certification':
159, 'exam': 320, 'depth': 246, 'evident': 318, 'always': 45,
'accurate': 17, 'reputable': 794, 'shipper': 861, 'gift': 402, 'ok':
670, 'transported': 1006, 'back': 86, 'war': 1050, 'forties': 380,
'lived': 572, 'limbo': 565, 'every': 316, 'hearts': 442, 'deeply':
239, 'continuously': 207, 'breaking': 133, 'complex': 189,
'believable': 108, 'sweeps': 947, 'lives': 573, 'plotlines': 726,
'eras': 308, 'intertwine': 501, 'around': 67, 'theme': 975, 'return':
804, 'sadie': 824, 'forced': 378, 'beliefs': 107, 'shakes': 859,
'examining': 323, 'question': 763, 'fortunately': 381, 'via': 1040,
'kindle': 526, 'program': 745, 'absolute': 12, 'garbage': 393, 'soft':
882, 'mpd': 623, 'feminist': 357, 'trash': 1007, 'losers': 579,
'otherwise': 683, 'female': 356, 'accept': 14, 'dealing': 232,
'worst': 1079, 'feeders': 352, 'biggest': 113, 'piece': 714, 'arktos':
66, 'generally': 396, '66': 10, 'nonsense': 655, 'made': 586, 'laugh':
539, 'cry': 221, 'playing': 723, 'game': 392, 'hygiene': 468, 'teeth':
965, 'everyone': 317, 'tooth': 999, 'paste': 695, 'brush': 140, 'kit':
529, 'maya': 601, 'marello': 595, 'amazing': 46, 'job': 511,
'bringing': 136, 'sweetwater': 949, 'county': 210, 'inhabitants': 487,
'enormous': 302, 'bed': 101, 'children': 171, 'strange': 917,
'editorial': 285, 'clinical': 176, 'psychologist': 752, 'mom': 616,
'divorce': 270, 'separated': 849, 'divorced': 271, 'different': 259,
'marriage': 598, 'without': 1068, 'blame': 118, 'fault': 349, 'told':
995, 'parable': 690, 'appeal': 59, 'queen': 762, 'king': 528,
'cannot': 149, 'live': 571, 'home': 459, 'case': 153, 'reassuring':
777, 'maintain': 588, 'child': 169, 'despite': 248, 'divorcing': 272,
'separately': 850, 'five': 367, 'bonsai': 121, 'admiring': 27,
'flower': 368, 'shop': 863, 'commitment': 185, 'pet': 707, 'pine':
716, 'sapling': 827, 'yard': 1088, 'try': 1015, 'start': 909,
'following': 375, 'simple': 869, 'stand': 906, 'disappointed': 263,
'learn': 543, 'plant': 720, 'trained': 1004, 'planning': 718, 'term':
969, 'missions': 613, 'examples': 325, 'expected': 330, 'saved': 829,
'money': 619, 'pretty': 740, 'quick': 765, 'resource': 799, 'almost':
38, 'night': 652, 'thoughts': 985, 'afterwards': 31, 'brings': 137,
'childhood': 170, 'grew': 423, 'ocean': 668, 'cute': 226, 'covering':
212, 'walls': 1047, 'spiritually': 901, 'based': 93, 'several': 858,
'backgrounds': 88, 'appreciate': 61, 'informaripn': 483, 'collector':
179, 'folks': 374, 'commented': 184, 'text': 972, 'medical': 604,
'concur': 193, 'introductory': 503, 'obtain': 666, 'foundation': 383,
'education': 286, 'instead': 493, 'focus': 370, 'lecture': 549,
'supplement': 936, 'easily': 281, 'overlooked': 684, 'let': 554,
'tell': 966, 'designing': 247, 'producing': 743, 'interactive': 496,
'educational': 287, 'multimedia': 626, 'takes': 956, 'video': 1042,
'supplements': 938, 'nice': 651, 'summaries': 931, 'emphasis': 294,
'points': 731, 'supplementing': 937, 'enhance': 298, 'memorization':
607, 'choice': 172, 'immbedded': 474, 'sections': 841, 'reminders':
793, 'pointers': 730, 'key': 520, 'explanation': 336, 'choices': 173,
'concept': 190, 'similar': 868, 'uses': 1034, 'sensory': 848,
'modalities': 615, 'graphs': 419, 'animations': 50, 'images': 472,

```
'audio': 75, 'example': 324, 'anesthesia': 49, 'historic': 456,
'figures': 359, 'instruments': 494, 'helped': 447, 'link': 567,
'names': 636, 'projects': 747, 'digital': 260, 'contents': 202,
'available': 83, 'inherent': 488, 'restrictions': 802, 'technical':
963, 'parameters': 691, 'allowed': 37, 'definitions': 241,
'smoothness': 881, 'qualities': 760, 'thing': 977, 'switch': 950,
'slides': 878, 'lecturer': 550, 'windows': 1067, 'might': 609,
'option': 677, 'change': 161, 'sizes': 874, 'personally': 704,
'window': 1066, 'utilize': 1035, 'watch': 1053, 'pause': 697, 'refer':
784, 'clarification': 174, 'needed': 642, 'cover': 211, 'reviewing':
807, 'summary': 933, 'yes': 1091, 'resident': 797, 'looking': 578,
'foundational': 384, 'knowledge': 532, 'honest': 460}
Lenght:  1095
```

```python
# sparse matrix

print('train size:', X_train.shape)
print(X_train.toarray()[:5])

print('\ntest size:', X_test.shape)
print(X_test.toarray()[:5])
```

```
train size: (80, 1095)
[[0.         0.         0.         ... 0.         0.         0.
]
 [0.         0.         0.         ... 0.         0.         0.
]
 [0.         0.         0.         ... 0.         0.         0.
]
 [0.         0.05641973 0.         ... 0.         0.         0.
]
 [0.         0.         0.         ... 0.         0.         0.
]]

test size: (20, 1095)
[[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.]]
```

```python
#X and Y training using mulitnomial NB
from sklearn.naive_bayes import MultinomialNB

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

```
MultinomialNB()
```

```python
# Log NB
naive_bayes.class_log_prior_[1]
```

```
-0.14792013007662153
```

```python
# Log pob using NB
naive_bayes.feature_log_prob_
```

```
array([[-6.96885799, -6.92137959, -6.91163916, ..., -6.96885799,
        -7.05021421, -6.65774295],
       [-7.18281505, -7.12792947, -7.18281505, ..., -7.18281505,
        -6.99315153, -7.18281505]])
```

```python
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score, confusion_matrix

# make predictions on the test data
pred = naive_bayes.predict(X_test)

# print confusion matrix
print(confusion_matrix(y_test, pred))
```

```
[[ 0  5]
 [ 0 15]]
```

```
accuracy score:  0.75

precision score (positive):  0.75

recall score: (positive):  1.0

f1 score:  0.8571428571428571
```

```python
# Stats
print('positive(s) in test data:',y_test[y_test=="positive"].shape[0])
print('negative(s) in test data:',y_test[y_test=="negaitve"].shape[0])
print('test size: ', len(y_test))

baseline = y_test[y_test=="positive"].shape[0] / y_test.shape[0]
print("Positive %: ", baseline)

baseline = y_test[y_test=="negaitve"].shape[0] / y_test.shape[0]
print("Negative %: ", baseline)


print('accuracy score: ', accuracy_score(y_test, pred))
print('precision score (positive): ', precision_score(y_test, pred,
pos_label="positive"))
print('recall score: (positive): ', recall_score(y_test, pred,
pos_label="positive"))
```

```
print("f1 score: ", f1_score(y_test, pred, pos_label="positive"))
```

```
positive(s) in test data: 15
negative(s) in test data: 5
test size:  20
Positive %:  0.75
Negative %:  0.25
accuracy score:  0.75
precision score (positive):  0.75
recall score: (positive):  1.0
f1 score:  0.8571428571428571
```

```
# Missed
y_test[y_test != pred]
```

```
42    negaitve
33    negaitve
59    negaitve
94    negaitve
96    negaitve
Name: Sentiment_books, dtype: object
```

Logic Regression

```
# imports
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score

# Making the logistic regression model
logistic_model = LogisticRegression()

# Training the model on the training data and labels using the same
data
logistic_model.fit(X_train, y_train)
```

```
LogisticRegression()
```

```
# Using the model to predict the labels of the test data
y_pred = logistic_model.predict(X_test)

# Evaluating the accuracy of the model using the sklearn functions
accuracy = accuracy_score(y_test,y_pred)*100
confusion_mat = confusion_matrix(y_test,y_pred)

# Printing the results
print("Accuracy: ",accuracy)
print("Confusion Matrix")
print(confusion_mat)
```

```
Accuracy is 75.0
Confusion Matrix
```

```
[[ 0  5]
 [ 0 15]]
```

Neural Networks

```python
# imports
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix, accuracy_score

# Neural Network Classifier
NN = MLPClassifier()

# Training the model on the training data and labels using the same data
NN.fit(X_train, y_train)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/neural_network/
_multilayer_perceptron.py:686: ConvergenceWarning: Stochastic
Optimizer: Maximum iterations (200) reached and the optimization
hasn't converged yet.
  warnings.warn(

MLPClassifier()
```

```python
# predicting the labels of the test data.
y_pred = NN.predict(X_test)

# Step 5
# Evaluating the results of the model
accuracy = accuracy_score(y_test,y_pred)*100
confusion_mat = confusion_matrix(y_test,y_pred)

# Step 6
# Printing the Results
print("Accuracy for Neural Network is:",accuracy)
print("Confusion Matrix")
print(confusion_mat)
```

```
Accuracy for Neural Network is: 75.0
Confusion Matrix
[[ 0  5]
 [ 0 15]]
```

#**Analysis Naive Bayes:**

Naive Bayes is a simple probabilistic algorithm that works well on datasets with many features and is relatively insensitive to irrelevant features. It is fast and requires a small amount of training data. Naive Bayes assumes that the features are conditionally independent.

Logistic Regression:

Logistic Regression is a statistical method that models the probability of a binary outcome based on one or more predictor variables. It works well with linearly separable data. It also provides a measure of feature importance, which can be useful in feature selection. However, logistic regression may not perform well with highly non-linear data and may suffer from overfitting or underfitting if the model is not properly tuned.

Neural Networks:

Neural networks are highly flexible and powerful for complex, non-linear relationships between input and output variables. It can handle large amounts of data and can automatically extract relevant features from raw data. Neural networks have been highly successful in a wide range of applications, including computer vision, natural language processing, and speech recognition. However, they require a large amount of data to train, and their complexity makes them difficult to interpret and debug. Neural networks are also computationally expensive, requiring specialized hardware for training and inference.

In summary, the choice of algorithm depends on the nature of the problem, the size and complexity of the dataset, and the resources available for training and inference. Naive Bayes is a good choice for simple problems with many features and limited training data, while logistic regression is a good choice for problems with linearly separable data and a small number of predictors. Neural networks are a good choice for complex problems with large amounts of data and require more computational resources.