In [117…

```
#Import all of the packages we will need for this project
#Jupyter Lab 3.44, Python 3
import json
import pandas as pd
import gzip
import numpy as np
import pandas as pd
import matplotlib.pyplot as mpl
import matplotlib.image as mpimg
%matplotlib inline
import seaborn as sb
#This next line was ran in the commmand line
# pip install tensorflow
import tensorflow as tf
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from keras import callbacks
```

In [118…

```
#Parse for json gzip file downloaded from the links provided
def parse(path):
  g = gzip.open(path, 'rb')
  for l in g:
    yield json.loads(l)
```

In [119…

```
# Function to turn json into df
def getDF(path):
  i = 0
  df = {}
  for d in parse(path):
    df[i] = d
    i += 1
  return pd.DataFrame.from_dict(df, orient='index')
```

In [120…

```
#Importing data
df = getDF('Software.json.gz')
#View the dataset
df.head()
```

Out[120]:

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText
0	4.0	True	03 11, 2014	A240ORQ2LF9LUI	0077613252	{'Format:': ' Loose Leaf'}	Michelle W	The materials arrived early and were in excell...
1	4.0	True	02 23, 2014	A1YCCU0YRLS0FE	0077613252	{'Format:': ' Loose Leaf'}	Rosalind White Ames	I am really enjoying this book with the worksh...
2	1.0	True	02 17, 2014	A1BJHRQDYVAY2J	0077613252	{'Format:': ' Loose Leaf'}	Allan R. Baker	IF YOU ARE TAKING THIS CLASS DON"T WASTE YOUR ...
3	3.0	True	02 17, 2014	APRDVZ6QBIQXT	0077613252	{'Format:': ' Loose Leaf'}	Lucy	This book was missing pages!!! Important pages...
4	5.0	False	10 14, 2013	A2JZTTBSLS1QXV	0077775473	NaN	Albert V.	I have used LearnSmart and can officially say ...

In [121…

```
#New data frame with the new selected columns we need
df = df[['overall','reviewText']]
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 459436 entries, 0 to 459435
Data columns (total 2 columns):
 #   Column      Non-Null Count   Dtype  
---  ------      --------------   -----  
 0   overall     459436 non-null  float64
 1   reviewText  459370 non-null  object 
dtypes: float64(1), object(1)
memory usage: 10.5+ MB
```

In [122…

```
# Select the special characters
special_char = ["!",'"',"#","%","&","'","(",")",
 "*","+",",","-",".","/",":",";","<",
 "=",">","?","@","[","\\","]","^","_",
 "`","{","|","}","~","-"]
```

In [123…

```
# Rename Columns
df = df.rename(columns={'overall': 'Score', 'reviewText': 'Review'})
df.head()
```

Out[123]:

	Score	Review
0	4.0	The materials arrived early and were in excell...
1	4.0	I am really enjoying this book with the worksh...
2	1.0	IF YOU ARE TAKING THIS CLASS DON"T WASTE YOUR ...
3	3.0	This book was missing pages!!! Important pages...
4	5.0	I have used LearnSmart and can officially say ...

In [124…

```
# Check for nulls
df.isna().any()
```

Out[124]:

```
Score      False
Review      True
dtype: bool
```

In [125…

```
# Remove nulls from where it says “True” above this cell
df = df.dropna(subset=['Review'])
```

In [126…

```
# Verify we do not have any nulls
df.isna().any()
```

Out[126]:

```
Score      False
Review     False
dtype: bool
```

In [127…

```
# replace the special characters with spaces
for char in special_char:
 df['Review'] = df['Review'].str.replace(char, ' ',regex=True)
```

In [128…

```
# Replace capitals with lowercase letters
df['Review'] = df['Review'].str.lower()
df.head()
```

Out[128]:

	Score	Review
0	4.0	the materials arrived early and were in excell...
1	4.0	i am really enjoying this book with the worksh...
2	1.0	if you are taking this class don t waste your ...
3	3.0	this book was missing pages important pages...
4	5.0	i have used learnsmart and can officially say ...

In [129…

```
# Create sentiments column from the dataset (Changing numeric review answers to positi
# 0 = negative, 1 = positive
df['Sentiments'] = df.Score.apply(lambda x: 0 if x in [1, 2] else 1)
df.head()
```

Out[129]:

	Score	Review	Sentiments
0	4.0	the materials arrived early and were in excell...	1
1	4.0	i am really enjoying this book with the worksh...	1
2	1.0	if you are taking this class don t waste your ...	0
3	3.0	this book was missing pages important pages...	1
4	5.0	i have used learnsmart and can officially say ...	1

In [130…

```
# export the prepared data
df.to_csv('D213_Task2_cleaned.csv', index = False)
```

In [131…

```
# Split data into 80/20 training and testing, standard for testing datasets and lookir
split = round(len(df)*0.8)
train_reviews = df['Review'][:split]
train_label = df['Sentiments'][:split]
test_reviews = df['Review'][split:]
test_label = df['Sentiments'][split:]
```

In [132…

```
#Changing each review into a string after tokenization
training_sentences = []
training_labels = []
testing_sentences = []
testing_labels = []
for row in train_reviews:
    training_sentences.append(str(row))
for row in train_label:
    training_labels.append(row)
for row in test_reviews:
    testing_sentences.append(str(row))
for row in test_label:
    testing_labels.append(row)
```

In [133…

```
#View the max vocab
tokenizer = Tokenizer()
tokenizer.fit_on_texts(df['Review'])
print(len(tokenizer.word_index) + 1)
```

```
130085
```

In [134…

```
# Getting the word index (vocabulary size)
word_index = tokenizer.word_index
word_index
```

```
Out[134]: {'the': 1,
           'i': 2,
           'to': 3,
           'and': 4,
           'it': 5,
           'a': 6,
           'of': 7,
           'is': 8,
           'for': 9,
           'this': 10,
           'you': 11,
           'that': 12,
           'my': 13,
           'in': 14,
           'with': 15,
           'have': 16,
           'on': 17,
           'not': 18,
           'was': 19,
           'but': 20,
           't': 21,
           'as': 22,
           'software': 23,
           's': 24,
           'so': 25,
           'be': 26,
           'are': 27,
           'can': 28,
           'product': 29,
           'if': 30,
           'use': 31,
           'all': 32,
           'from': 33,
           'program': 34,
           'or': 35,
           'me': 36,
           'they': 37,
           'had': 38,
           'your': 39,
           'do': 40,
           'an': 41,
           'will': 42,
           'version': 43,
           'very': 44,
           'no': 45,
           'one': 46,
           'at': 47,
           'get': 48,
           'just': 49,
           'would': 50,
           'up': 51,
           'has': 52,
           'time': 53,
           'like': 54,
           'windows': 55,
           'when': 56,
           'there': 57,
           'more': 58,
           'computer': 59,
           'what': 60,
           'great': 61,
           'out': 62,
           'been': 63,
           'work': 64,
           'good': 65,
           'used': 66,
           'new': 67,
           'only': 68,
           'years': 69,
           'using': 70,
           'easy': 71,
           'some': 72,
           'which': 73,
           'than': 74,
           'other': 75,
           'about': 76,
           'after': 77,
           'don': 78,
           'am': 79,
           'now': 80,
           'by': 81,
           'their': 82,
           'any': 83,
           'then': 84,
           'well': 85,
           'even': 86,
           'year': 87,
           've': 88,
           'does': 89,
           'also': 90,
           'because': 91,
           'much': 92,
           'them': 93,
           'did': 94,
           'support': 95,
           'quicken': 96,
           'm': 97,
           'works': 98,
           'really': 99,
           'need': 100,
           'many': 101,
           'could': 102,
           'how': 103,
           'we': 104,
           'install': 105,
           'back': 106,
           'tax': 107,
           'still': 108,
           'over': 109,
           'way': 110,
           'want': 111,
           'buy': 112,
           'better': 113,
           'download': 114,
           'money': 115,
           'problem': 116,
           'first': 117,
           'find': 118,
           'make': 119,
           'mac': 120,
           'since': 121,
           'system': 122,
           '2': 123,
           'user': 124,
           'amazon': 125,
           'go': 126,
           'pc': 127,
           'upgrade': 128,
           'microsoft': 129,
           'were': 130,
           'norton': 131,
           'problems': 132,
           '1': 133,
           'price': 134,
           'old': 135,
           '3': 136,
           'never': 137,
           'into': 138,
           'bought': 139,
           'file': 140,
           'got': 141,
           '7': 142,
           'help': 143,
           'its': 144,
           'free': 145,
           'again': 146,
           'able': 147,
           'most': 148,
           'know': 149,
           'before': 150,
           '10': 151,
           'found': 152,
           'features': 153,
           'who': 154,
           'same': 155,
           'installed': 156,
           'every': 157,
           '5': 158,
           'tried': 159,
           'without': 160,
           'office': 161,
           'too': 162,
           '8': 163,
           'love': 164,
           'best': 165,
           'recommend': 166,
           'through': 167,
           'lot': 168,
           'should': 169,
           'files': 170,
           'another': 171,
           'see': 172,
           'data': 173,
           'update': 174,
           'programs': 175,
           'few': 176,
           'didn': 177,
           'game': 178,
           'two': 179,
           'think': 180,
           'doesn': 181,
           're': 182,
           'video': 183,
           'worked': 184,
           'purchased': 185,
           'turbotax': 186,
           'far': 187,
           'everything': 188,
           'home': 189,
           'run': 190,
           'little': 191,
           'try': 192,
           'however': 193,
           'things': 194,
           'these': 195,
           'while': 196,
           'down': 197,
           'business': 198,
           'always': 199,
           '4': 200,
           'purchase': 201,
           'online': 202,
           'say': 203,
           'hard': 204,
           'several': 205,
           'different': 206,
           'off': 207,
           'customer': 208,
           'drive': 209,
           'going': 210,
           'word': 211,
           'right': 212,
           'xp': 213,
           'something': 214,
           'having': 215,
           'last': 216,
           'people': 217,
           'where': 218,
           'pro': 219,
           'service': 220,
           'internet': 221,
           'return': 222,
           'running': 223,
           'intuit': 224,
           'made': 225,
           'take': 226,
           'bit': 227,
           'cd': 228,
           'our': 229,
           'state': 230,
           'give': 231,
           'security': 232,
           'sure': 233,
           'once': 234,
           'thing': 235,
           'working': 236,
           'needed': 237,
           'fine': 238,
           'products': 239,
           'own': 240,
           'trying': 241,
           'simple': 242,
           'd': 243,
           'each': 244,
           'seems': 245,
           'read': 246,
           'may': 247,
           'he': 248,
           'long': 249,
           'hours': 250,
           'being': 251,
           'getting': 252,
           'start': 253,
           'information': 254,
           'both': 255,
           'll': 256,
           'issues': 257,
           'learn': 258,
           'versions': 259,
           'why': 260,
           'screen': 261,
           'those': 262,
           'though': 263,
           'keep': 264,
           'company': 265,
           'times': 266,
           'learning': 267,
           'ever': 268,
           'laptop': 269,
           'taxes': 270,
           'etc': 271,
           'worth': 272,
           'dvd': 273,
           'done': 274,
           'reviews': 275,
           'virus': 276,
           'turbo': 277,
           'set': 278,
           'os': 279,
           'vista': 280,
           'feature': 281,
           'said': 282,
           'review': 283,
           'took': 284,
           'pay': 285,
           'day': 286,
           'installation': 287,
           'interface': 288,
           'days': 289,
           'e': 290,
           'look': 291,
           'less': 292,
           'went': 293,
           'anything': 294,
           'looking': 295,
           'nothing': 296,
           'computers': 297,
           'available': 298,
           'such': 299,
           'doing': 300,
           'here': 301,
           'makes': 302,
           'next': 303,
           'nice': 304,
           'actually': 305,
           'number': 306,
           'yet': 307,
           'happy': 308,
           'wanted': 309,
           'pretty': 310,
           'around': 311,
           'bad': 312,
           'previous': 313,
           'open': 314,
           'put': 315,
           'won': 316,
           'until': 317,
           'full': 318,
           'process': 319,
           'already': 320,
           'users': 321,
           'tech': 322,
           'excellent': 323,
           'account': 324,
           'web': 325,
           'finally': 326,
           'updates': 327,
           'thought': 328,
           'issue': 329,
           'cannot': 330,
           'card': 331,
           'copy': 332,
           'minutes': 333,
           'enough': 334,
           'play': 335,
           'key': 336,
           'save': 337,
           'phone': 338,
           'app': 339,
           'music': 340,
           'website': 341,
           'downloaded': 342,
           'came': 343,
           '6': 344,
           'desktop': 345,
           'job': 346,
           'email': 347,
           'block': 348,
           'let': 349,
           'create': 350,
           'fast': 351,
           'least': 352,
           'change': 353,
           'needs': 354,
           'anyone': 355,
           'r': 356,
           'experience': 357,
           'add': 358,
           'must': 359,
           'editing': 360,
           'fix': 361,
           'basic': 362,
           'she': 363,
           'error': 364,
           'easier': 365,
           'access': 366,
           'line': 367,
           'three': 368,
           'box': 369,
           'ms': 370,
           'h': 371,
           'past': 372,
           'refund': 373,
           'ago': 374,
           'instead': 375,
           'end': 376,
           'older': 377,
           'print': 378,
           'slow': 379,
           'disk': 380,
           'highly': 381,
           'started': 382,
           'fun': 383,
           'us': 384,
           'cost': 385,
           '0': 386,
           'package': 387,
           'games': 388,
           'come': 389,
           'small': 390,
           'site': 391,
           'quite': 392,
           'complete': 393,
           'probably': 394,
           'tool': 395,
           'almost': 396,
           'link': 397,
           'either': 398,
           'order': 399,
           'machine': 400,
           'backup': 401,
           'quickbooks': 402,
           'deluxe': 403,
           'longer': 404,
           'click': 405,
           'stars': 406,
           'buying': 407,
           'real': 408,
           'might': 409,
           'easily': 410,
           'takes': 411,
           'big': 412,
           'point': 413,
           'called': 414,
           'code': 415,
           'option': 416,
           'reason': 417,
           'others': 418,
           'page': 419,
           'part': 420,
           'couldn': 421,
           'decided': 422,
           'away': 423,
           'wish': 424,
           'haven': 425,
           'told': 426,
           'protection': 427,
           'check': 428,
           'comes': 429,
           'months': 430,
           'understand': 431,
           'feel': 432,
           'call': 433,
           'received': 434,
           'difficult': 435,
           'professional': 436,
           'waste': 437,
           'making': 438,
           'load': 439,
           'installing': 440,
           'else': 441,
           'options': 442,
           'friendly': 443,
           'tell': 444,
           'ok': 445,
           'accounts': 446,
           'figure': 447,
           'isn': 448,
           'type': 449,
           '20': 450,
           '2013': 451,
           'mcafee': 452,
           'gave': 453,
           'someone': 454,
           'overall': 455,
           'quickly': 456,
           'upgraded': 457,
           'simply': 458,
           'book': 459,
           'couple': 460,
           'deal': 461,
           'useful': 462,
           'especially': 463,
           'fact': 464,
           'disappointed': 465,
           'quick': 466,
           'family': 467,
           'paid': 468,
           'antivirus': 469,
           '30': 470,
           'included': 471,
           'wrong': 472,
           'plus': 473,
           'her': 474,
           '2007': 475,
           'quality': 476,
           'helpful': 477,
           'corel': 478,
           'spent': 479,
           'although': 480,
           '9': 481,
           'perfect': 482,
           'tools': 483,
           'operating': 484,
           '2014': 485,
           'language': 486,
           'application': 487,
           'says': 488,
           'apple': 489,
           'instructions': 490,
           'photo': 491,
           'later': 492,
           'transactions': 493,
           'course': 494,
           'reading': 495,
           'between': 496,
           'based': 497,
           'seem': 498,
           'step': 499,
           'forms': 500,
           'second': 501,
           'completely': 502,
           'maybe': 503,
           'words': 504,
           'edition': 505,
           'enter': 506,
           '2010': 507,
           'anti': 508,
           'level': 509,
           'format': 510,
           'multiple': 511,
           'bank': 512,
           'allow': 513,
           'suite': 514,
           'disc': 515,
           'wouldn': 516,
           'personal': 517,
           'subscription': 518,
           'kaspersky': 519,
           'import': 520,
           'ability': 521,
           'thanks': 522,
           'adobe': 523,
           'expected': 524,
           'lost': 525,
           'image': 526,
           'rather': 527,
           'changes': 528,
           'case': 529,
           'text': 530,
           'form': 531,
           'unless': 532,
           'intuitive': 533,
           'star': 534,
           'item': 535,
           'wasn': 536,
           'faster': 537,
           'his': 538,
           'high': 539,
           'graphics': 540,
           'wait': 541,
           'videos': 542,
           'within': 543,
           'mail': 544,
           'itself': 545,
           'trial': 546,
           'com': 547,
           'updated': 548,
           '15': 549,
           'search': 550,
           'month': 551,
           'allows': 552,
           'message': 553,
           'design': 554,
           'list': 555,
           'person': 556,
           'definitely': 557,
           'extra': 558,
           'credit': 559,
           'photoshop': 560,
           'expensive': 561,
           'spend': 562,
           'believe': 563,
           'top': 564,
           'whole': 565,
           'latest': 566,
           'sometimes': 567,
           'win': 568,
           'menu': 569,
           'stuff': 570,
           'myself': 571,
           'studio': 572,
           'hour': 573,
           'date': 574,
           'place': 575,
           'thank': 576,
           'under': 577,
           'questions': 578,
           'edit': 579,
           'guess': 580,
           'name': 581,
           'kids': 582,
           'photos': 583,
           'including': 584,
           'weeks': 585,
           'exactly': 586,
           'due': 587,
           'yes': 588,
           'compatible': 589,
           'trouble': 590,
           'given': 591,
           'show': 592,
           'excel': 593,
           'books': 594,
           'useless': 595,
           'documents': 596,
           'manual': 597,
           'hope': 598,
           'audio': 599,
           'runs': 600,
           'additional': 601,
           'current': 602,
           'move': 603,
           'window': 604,
           'correct': 605,
           'left': 606,
           'often': 607,
           '2011': 608,
           'non': 609,
           'example': 610,
           'track': 611,
           'customers': 612,
           'gets': 613,
           'speed': 614,
           'class': 615,
           '12': 616,
           'value': 617,
           'pleased': 618,
           'continue': 619,
           'manually': 620,
           'pdf': 621,
           'bugs': 622,
           'live': 623,
           'federal': 624,
           'week': 625,
           'downloading': 626,
           'major': 627,
           'scan': 628,
           'extremely': 629,
           'automatically': 630,
           'write': 631,
           'sent': 632,
           'filing': 633,
           'control': 634,
           'lots': 635,
           'license': 636,
           'during': 637,
           'info': 638,
           'settings': 639,
           'x': 640,
           'pages': 641,
           'saved': 642,
           'movie': 643,
           'please': 644,
           'media': 645,
           'offer': 646,
           'images': 647,
           'uninstall': 648,
           'life': 649,
           'gps': 650,
           'technical': 651,
           'applications': 652,
           'send': 653,
           'gives': 654,
           'performance': 655,
           'original': 656,
           'memory': 657,
           'absolutely': 658,
           '2012': 659,
           'function': 660,
           'outlook': 661,
           'results': 662,
           'recommended': 663,
           'expect': 664,
           'boot': 665,
           'turn': 666,
           'clean': 667,
           '11': 668,
           'along': 669,
           'looks': 670,
           'loaded': 671,
           'recently': 672,
           'piece': 673,
           'ram': 674,
           'cards': 675,
           'mode': 676,
           'financial': 677,
           'plan': 678,
           'devices': 679,
           'crashes': 680,
           'asked': 681,
           'short': 682,
           'kind': 683,
           'half': 684,
           'store': 685,
           'functionality': 686,
           'normal': 687,
           '100': 688,
           'properly': 689,
           'contact': 690,
           'total': 691,
           'button': 692,
           'amount': 693,
           'writing': 694,
           '64': 695,
           'playing': 696,
           'changed': 697,
           'single': 698,
           'pictures': 699,
           'webroot': 700,
           'note': 701,
           'entire': 702,
           'digital': 703,
           'dragon': 704,
           'sound': 705,
           '2008': 706,
           'c': 707,
           'solution': 708,
           'device': 709,
           'standard': 710,
           'answer': 711,
           'release': 712,
           'unfortunately': 713,
           'errors': 714,
           'kept': 715,
           'annoying': 716,
           'view': 717,
           '2015': 718,
           'required': 719,
           'switch': 720,
           'added': 721,
           'systems': 722,
           'directly': 723,
           'apps': 724,
           'ordered': 725,
           'son': 726,
           'clear': 727,
           'choose': 728,
           'returns': 729,
           'map': 730,
           'fixed': 731,
           'worst': 732,
           'via': 733,
           'idea': 734,
           'power': 735,
           'possible': 736,
           'cloud': 737,
           'limited': 738,
           'important': 739,
           'side': 740,
           'various': 741,
           'purchasing': 742,
           'remove': 743,
           'ask': 744,
           'parallels': 745,
           'newer': 746,
           'school': 747,
           'spanish': 748,
           'yourself': 749,
           'choice': 750,
           'satisfied': 751,
           'maps': 752,
           'functions': 753,
           'today': 754,
           'follow': 755,
           'similar': 756,
           'perfectly': 757,
           'reports': 758,
           'hand': 759,
           'stop': 760,
           'true': 761,
           'earlier': 762,
           'project': 763,
           'created': 764,
           'items': 765,
           'large': 766,
           'space': 767,
           'advertised': 768,
           'saying': 769,
           'keeps': 770,
           'seen': 771,
           'except': 772,
           'close': 773,
           'address': 774,
           'unable': 775,
           'future': 776,
           'poor': 777,
           'provide': 778,
           'cheaper': 779,
           'transfer': 780,
           'liked': 781,
           'five': 782,
           'uses': 783,
           'main': 784,
           'world': 785,
           'totally': 786,
           'background': 787,
           'glad': 788,
           'everyone': 789,
           'care': 790,
           'awesome': 791,
           'ran': 792,
           'linux': 793,
           'learned': 794,
           'garmin': 795,
           'mind': 796,
           'complicated': 797,
           'prior': 798,
           'market': 799,
           'together': 800,
           'ie': 801,
           'register': 802,
           'huge': 803,
           '2009': 804,
           '99': 805,
           'viruses': 806,
           'fairly': 807,
           'upgrading': 808,
           'firewall': 809,
           'offers': 810,
           'document': 811,
           'worse': 812,
           'paper': 813,
           'kindle': 814,
           'supposed': 815,
           'restore': 816,
           'premier': 817,
           'means': 818,
           'correctly': 819,
           'soon': 820,
           'player': 821,
           'frustrating': 822,
           'response': 823,
           'premium': 824,
           'hardware': 825,
           'cut': 826,
           'amazing': 827,
           'seemed': 828,
           'picture': 829,
           'missing': 830,
           'ease': 831,
           'remember': 832,
           'include': 833,
           'loved': 834,
           'immediately': 835,
           '2003': 836,
           'paying': 837,
           'tt': 838,
           'enjoy': 839,
           'typing': 840,
           'charge': 841,
           'setup': 842,
           'income': 843,
           'rosetta': 844,
           'anyway': 845,
           'daughter': 846,
           'thinking': 847,
           'ones': 848,
           'usually': 849,
           'stay': 850,
           'appears': 851,
           'reinstall': 852,
           'ready': 853,
           'usb': 854,
           'voice': 855,
           'built': 856,
           'compared': 857,
           'symantec': 858,
           'provided': 859,
           'forward': 860,
           'him': 861,
           'requires': 862,
           'goes': 863,
           'password': 864,
           'four': 865,
           '50': 866,
           'onto': 867,
           'otherwise': 868,
           'crash': 869,
           'advanced': 870,
           'forced': 871,
           'numbers': 872,
           'server': 873,
           'accounting': 874,
           'select': 875,
           'negative': 876,
           'certain': 877,
           'question': 878,
           'google': 879,
           '360': 880,
           'mobile': 881,
           'roxio': 882,
           'delete': 883,
           'gone': 884,
           'wonderful': 885,
           'final': 886,
           'designed': 887,
           'become': 888,
           'giving': 889,
           'improved': 890,
           'taking': 891,
           'stone': 892,
           'helps': 893,
           'pop': 894,
           'network': 895,
           'curve': 896,
           'addition': 897,
           'english': 898,
           'test': 899,
           'matter': 900,
           'hate': 901,
           'basically': 902,
           'creating': 903,
           'mouse': 904,
           'failed': 905,
           'written': 906,
           'convert': 907,
           'shows': 908,
           'safe': 909,
           'none': 910,
           'terrible': 911,
           'payroll': 912,
           'student': 913,
           'wants': 914,
           'size': 915,
           'color': 916,
           'looked': 917,
           'chat': 918,
           '3d': 919,
           'loves': 920,
           'powerful': 921,
           'provides': 922,
           'training': 923,
           '00': 924,
           'ended': 925,
           'effects': 926,
           'crashed': 927,
           'nero': 928,
           '2000': 929,
           'difference': 930,
           'malware': 931,
           'lessons': 932,
           'removed': 933,
           'contacted': 934,
           'rating': 935,
           'friends': 936,
           'activation': 937,
           'investment': 938,
           'speak': 939,
           'twice': 940,
           'waiting': 941,
           'bottom': 942,
           'low': 943,
           'trend': 944,
           'downloads': 945,
           'section': 946,
           'shop': 947,
           'drivers': 948,
           'sync': 949,
           'opinion': 950,
           'includes': 951,
           'tablet': 952,
           'impressed': 953,
           'apparently': 954,
           'acronis': 955,
           'report': 956,
           '2006': 957,
           'library': 958,
           'linked': 959,
           'practice': 960,
           '2016': 961,
           'wasted': 962,
           'taxcut': 963,
           'uninstalled': 964,
           'switched': 965,
           'folder': 966,
           'content': 967,
           'helped': 968,
           'whether': 969,
           '95': 970,
           'qb': 971,
           'luck': 972,
           'mistake': 973,
           'transaction': 974,
           'offered': 975,
           'turned': 976,
           'nearly': 977,
           'arrived': 978,
           'starting': 979,
           'connection': 980,
           'database': 981,
           'hold': 982,
           'accurate': 983,
           'improvement': 984,
           'third': 985,
           'trust': 986,
           'setting': 987,
           'art': 988,
           'stopped': 989,
           'schedule': 990,
           'necessary': 991,
           'spending': 992,
           'research': 993,
           'upgrades': 994,
           'consider': 995,
           'favorite': 996,
           'brand': 997,
           'wife': 998,
           '2004': 999,
           'hook': 1000,
           ...}
```

In [135…

```
#Setting vocab size to 130086, aka taking 130086 words to train
vocab_size = 130086
# View the embedding length
max_sequence_embedding = int(round(np.sqrt(np.sqrt(vocab_size)), 0))
max_sequence_embedding
```

Out[135]:

```
19
```

In [136…

```
# Embedding to 19 dimensions, given from the above code
embedding_dim = 19
# Max length of 150 words per review as a cut off
max_length = 150
#If review is bigger than 150 words, it will be truncated "post" or after the 150th wo
trunc_type = 'post'
oov_tok = '<OOV>'
# Padding type "post" meaning each word will receive padding after, not before
padding_type = 'post'
```

In [137…

```
#Starting the tokenizer
tokenizer = Tokenizer(num_words=vocab_size, oov_token=oov_tok)
# Fitting the tokenizer
tokenizer.fit_on_texts(training_sentences)
```

In [138…

```
#Setting the sequences
sequences = tokenizer.texts_to_sequences(training_sentences)
#Setting the padding
padded = pad_sequences(sequences, maxlen=max_length, truncating=trunc_type)
testing_sentences = tokenizer.texts_to_sequences(testing_sentences)
testing_padded = pad_sequences(testing_sentences, maxlen=max_length)
```

In [139…

```
# Let’s check out the padded array
padded
print(df)
```

```
        Score                                             Review  Sentiments
0         4.0  the materials arrived early and were in excell...           1
1         4.0  i am really enjoying this book with the worksh...           1
2         1.0  if you are taking this class don t waste your ...           0
3         3.0  this book was missing pages    important pages...           1
4         5.0  i have used learnsmart and can officially say ...           1
...       ...                                                ...         ...
459431    2.0  no instructions     no help unless you want to...           0
459432    1.0                                        it s a joke           0
459433    5.0  i have multiple licenses of the antivirus  i h...           1
459434    5.0                                         good value           1
459435    5.0                     very nice designs easy to use            1

[459370 rows x 3 columns]
```

In [140…

```
#Time to create the model
model = tf.keras.Sequential([
    tf.keras.layers.Embedding(vocab_size, embedding_dim, input_length=max_length),
    tf.keras.layers.GlobalAveragePooling1D(),
    tf.keras.layers.Dense(10, activation='relu'),
    tf.keras.layers.Dense(6, activation='relu'),
    tf.keras.layers.Dense(1, activation='relu')
])
```

In [141…

```
#Compile the model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

```
Model: "sequential_5"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 embedding_6 (Embedding)     (None, 150, 19)           2471634   
                                                                 
 global_average_pooling1d_6   (None, 19)               0         
 (GlobalAveragePooling1D)                                        
                                                                 
 dense_17 (Dense)            (None, 10)                200       
                                                                 
 dense_18 (Dense)            (None, 6)                 66        
                                                                 
 dense_19 (Dense)            (None, 1)                 7         
                                                                 
=================================================================
Total params: 2,471,907
Trainable params: 2,471,907
Non-trainable params: 0
_________________________________________________________________
```

In [142…

```
#Train the model
training_final = np.array(training_labels)
testing_final = np.array(testing_labels)
# Run the model for 25 epochs
model1 = model.fit(padded, training_final, epochs=25, validation_data=(testing_padded,
testing_final))
# We need to stop the model Early to prevent overfitting of the model
earlystopping = callbacks.EarlyStopping(monitor ="val_loss",
                                        mode ="min", patience = 2,
                                        restore_best_weights = True)
```

```
Epoch 1/25
11485/11485 [==============================] - 333s 29ms/step - loss: 0.3631 - accura
cy: 0.8483 - val_loss: 0.3563 - val_accuracy: 0.8652
Epoch 2/25
11485/11485 [==============================] - 332s 29ms/step - loss: 0.2890 - accura
cy: 0.8870 - val_loss: 0.3501 - val_accuracy: 0.8645
Epoch 3/25
11485/11485 [==============================] - 333s 29ms/step - loss: 0.3462 - accura
cy: 0.8606 - val_loss: 0.3660 - val_accuracy: 0.8699
Epoch 4/25
11485/11485 [==============================] - 335s 29ms/step - loss: 0.3162 - accura
cy: 0.8705 - val_loss: 0.3476 - val_accuracy: 0.8701
Epoch 5/25
11485/11485 [==============================] - 333s 29ms/step - loss: 0.2852 - accura
cy: 0.8905 - val_loss: 0.3577 - val_accuracy: 0.8763
Epoch 6/25
11485/11485 [==============================] - 333s 29ms/step - loss: 0.2901 - accura
cy: 0.8869 - val_loss: 0.3408 - val_accuracy: 0.8718
Epoch 7/25
11485/11485 [==============================] - 332s 29ms/step - loss: 0.2922 - accura
cy: 0.8905 - val_loss: 0.3554 - val_accuracy: 0.8757
Epoch 8/25
11485/11485 [==============================] - 335s 29ms/step - loss: 0.2653 - accura
cy: 0.8990 - val_loss: 0.3866 - val_accuracy: 0.8716
Epoch 9/25
11485/11485 [==============================] - 335s 29ms/step - loss: 0.2991 - accura
cy: 0.8880 - val_loss: 0.3708 - val_accuracy: 0.8694
Epoch 10/25
11485/11485 [==============================] - 337s 29ms/step - loss: 0.2531 - accura
cy: 0.9062 - val_loss: 0.3382 - val_accuracy: 0.8712
Epoch 11/25
11485/11485 [==============================] - 339s 30ms/step - loss: 0.3060 - accura
cy: 0.8821 - val_loss: 0.3430 - val_accuracy: 0.8733
Epoch 12/25
11485/11485 [==============================] - 336s 29ms/step - loss: 0.2589 - accura
cy: 0.9029 - val_loss: 0.3516 - val_accuracy: 0.8683
Epoch 13/25
11485/11485 [==============================] - 338s 29ms/step - loss: 0.2593 - accura
cy: 0.9029 - val_loss: 0.3505 - val_accuracy: 0.8719
Epoch 14/25
11485/11485 [==============================] - 340s 30ms/step - loss: 0.2513 - accura
cy: 0.9087 - val_loss: 0.3589 - val_accuracy: 0.8688
Epoch 15/25
11485/11485 [==============================] - 341s 30ms/step - loss: 0.2468 - accura
cy: 0.9100 - val_loss: 0.3671 - val_accuracy: 0.8650
Epoch 16/25
11485/11485 [==============================] - 341s 30ms/step - loss: 0.2368 - accura
cy: 0.9113 - val_loss: 0.4051 - val_accuracy: 0.8535
Epoch 17/25
11485/11485 [==============================] - 340s 30ms/step - loss: 0.2350 - accura
cy: 0.9111 - val_loss: 0.4411 - val_accuracy: 0.8644
Epoch 18/25
11485/11485 [==============================] - 339s 30ms/step - loss: 0.2387 - accura
cy: 0.9106 - val_loss: 0.3927 - val_accuracy: 0.8723
Epoch 19/25
11485/11485 [==============================] - 332s 29ms/step - loss: 0.2628 - accura
cy: 0.8968 - val_loss: 0.3807 - val_accuracy: 0.8689
Epoch 20/25
11485/11485 [==============================] - 337s 29ms/step - loss: 0.2334 - accura
cy: 0.9137 - val_loss: 0.3791 - val_accuracy: 0.8715
```

```
Epoch 21/25
11485/11485 [==============================] - 338s 29ms/step - loss: 0.3121 - accura
cy: 0.8522 - val_loss: 0.5409 - val_accuracy: 0.7144
Epoch 22/25
11485/11485 [==============================] - 337s 29ms/step - loss: 0.2870 - accura
cy: 0.8977 - val_loss: 0.4150 - val_accuracy: 0.8668
Epoch 23/25
11485/11485 [==============================] - 337s 29ms/step - loss: 0.2521 - accura
cy: 0.9124 - val_loss: 0.3855 - val_accuracy: 0.8716
Epoch 24/25
11485/11485 [==============================] - 339s 30ms/step - loss: 0.2347 - accura
cy: 0.9150 - val_loss: 0.4049 - val_accuracy: 0.8637
Epoch 25/25
11485/11485 [==============================] - 341s 30ms/step - loss: 0.2338 - accura
cy: 0.9148 - val_loss: 0.3883 - val_accuracy: 0.8663
```

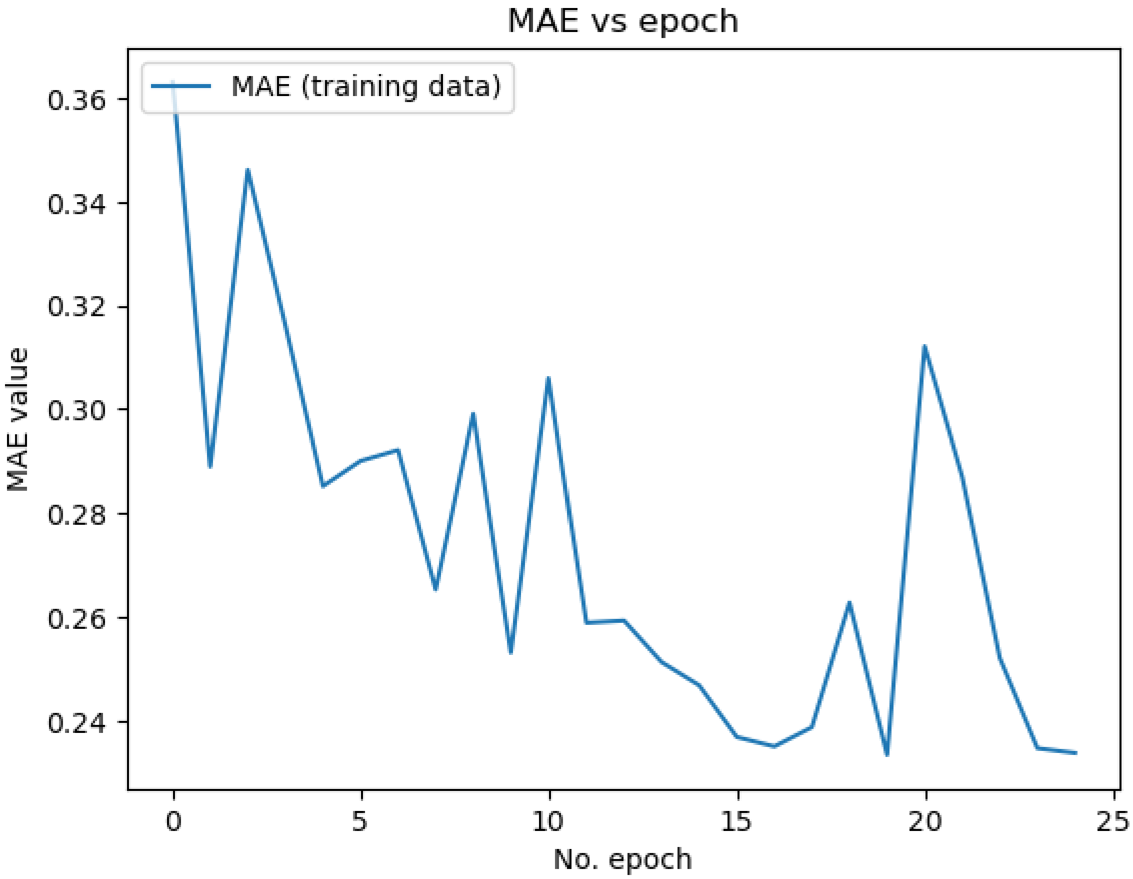
In [143…

```
# Time to run the model again and see our accuracy
model2 = model.fit(padded, training_final, epochs=25, validation_data=(testing_padded,
                                                                       testing_final),
                                                                       callbacks = [ea
```

```
Epoch 1/25
11485/11485 [==============================] - 327s 28ms/step - loss: 0.2312 - accura
cy: 0.9153 - val_loss: 0.3715 - val_accuracy: 0.8639
Epoch 2/25
11485/11485 [==============================] - 327s 28ms/step - loss: 0.2261 - accura
cy: 0.9165 - val_loss: 0.3883 - val_accuracy: 0.8708
Epoch 3/25
11485/11485 [==============================] - 327s 28ms/step - loss: 0.2349 - accura
cy: 0.9138 - val_loss: 0.3749 - val_accuracy: 0.8652
```

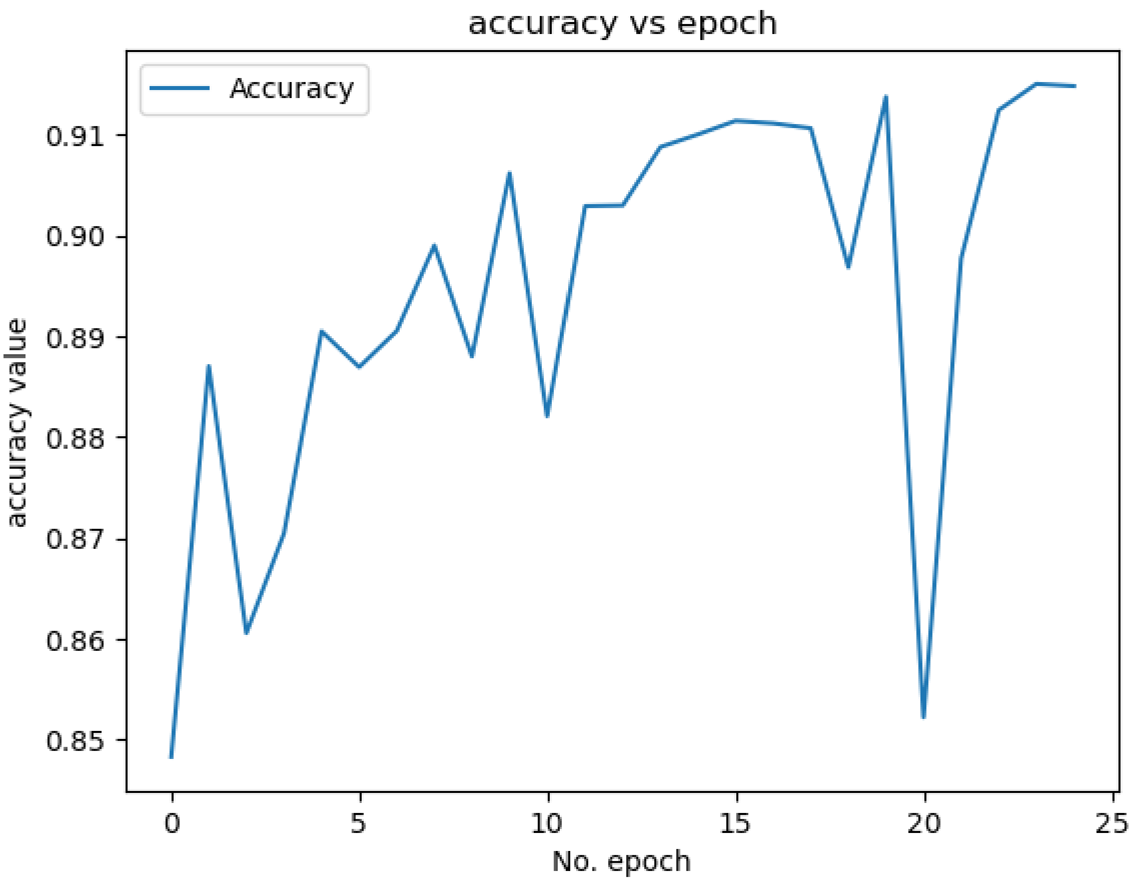
In [144…

```
# epoch vs Mean Absolute Error (MAE)
mpl.plot(model1.history['loss'], label='MAE (training data)')
mpl.title('MAE vs epoch')
mpl.ylabel('MAE value')
mpl.xlabel('No. epoch')
mpl.legend(loc="upper left")
mpl.show()
```



In [145…

```
#Check our accuracy per Epoch
mpl.plot(model1.history['accuracy'], label='Accuracy')
mpl.title('accuracy vs epoch')
mpl.ylabel('accuracy value')
mpl.xlabel('No. epoch')
mpl.legend(loc="upper left")
mpl.show()
```

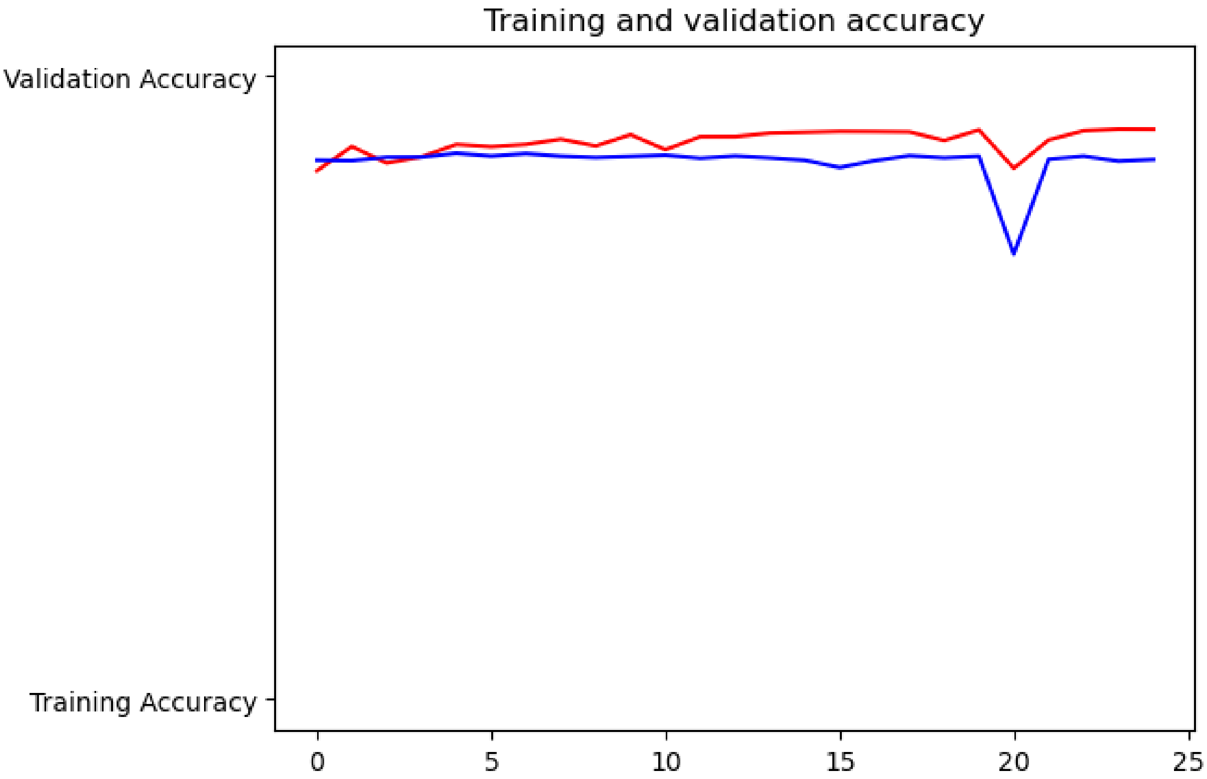


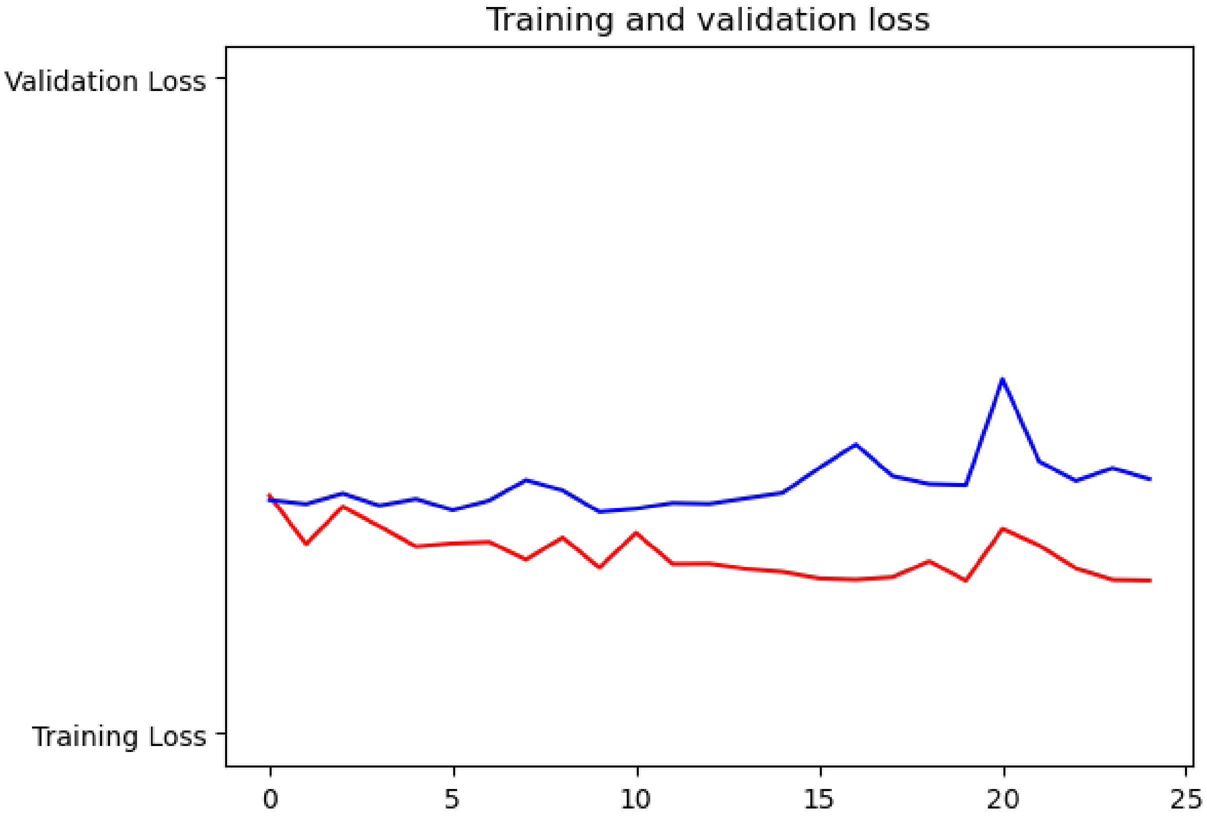
In [146…

```
#Training versus Accuracy
acc = model1.history['accuracy']
val_acc = model1.history['val_accuracy']
loss = model1.history['loss']
val_loss = model1.history['val_loss']
epochs=range(len(acc))
mpl.plot(epochs, acc, 'r', 'Training Accuracy')
mpl.plot(epochs, val_acc, 'b', 'Validation Accuracy')
mpl.title('Training and validation accuracy')
mpl.figure()
mpl.plot(epochs, loss, 'r', 'Training Loss')
mpl.plot(epochs, val_loss, 'b', 'Validation Loss')
mpl.title('Training and validation loss')
mpl.figure()
```

Out[146]:

```
<Figure size 640x480 with 0 Axes>
```





```
<Figure size 640x480 with 0 Axes>
```

In [152…

```
#Exports
train_reviews.to_csv('D213_Task2_train_reviews.csv', index = False)
train_label.to_csv('D213_Task2_train_label.csv', index = False)
test_reviews.to_csv('D213_Task2_test_reviews.csv', index = False)
test_label.to_csv('D213_Task2_test_label.csv', index = False)
```

In [153…

```
df.to_csv('D213_Task2_df_Ready_to_Run.csv', index = False)
```

In [154…

```
#Saving our model
model.save('D213_Task2_Model.h5')
```