In [22]:

```
import pandas as pd
import dask.dataframe as dd
import numpy as np
import collections
import numba
from transformers import pipeline
import spacy
```

In [9]:

```
df = pd.read_csv('C:/Users/abdul/text-mining/Project/mpst_full_data.csv')
```

In [10]:

```
df
```

Out[10]:

	imdb_id	title	plot_synopsis	tags	split	synopsis_source
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...	cult, horror, gothic, murder, atmospheric	train	imdb
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	violence	train	imdb
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	romantic	test	imdb
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	inspiring, romantic, stupid, feel-good	train	imdb
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...	cruelty, murder, dramatic, cult, violence, atm...	val	imdb
...	...	...	...	...	...	...
14823	tt0219952	Lucky Numbers	In 1988 Russ Richards (John Travolta), the wea...	comedy, murder	test	wikipedia
14824	tt1371159	Iron Man 2	In Russia, the media covers Tony Stark's discl...	good versus evil, violence	train	wikipedia
14825	tt0063443	Play Dirty	During the North African Campaign in World War...	anti war	train	wikipedia
14826	tt0039464	High Wall	Steven Kenet catches his unfaithful wife in th...	murder	test	wikipedia
14827	tt0235166	Against All Hope	Sometime in the 1950s in Chicago a man, Cecil ...	christian film	test	wikipedia

14828 rows × 6 columns

In [13]:

```
df.isnull().values.any()
```

Out[13]:

```
False
```

In [16]:

```
df = df.drop(columns=["tags", "split", "synopsis_source"])
```

In [17]:

```
df
```

Out[17]:

	imdb_id	title	plot_synopsis
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...
...	...	...	...
14823	tt0219952	Lucky Numbers	In 1988 Russ Richards (John Travolta), the wea...
14824	tt1371159	Iron Man 2	In Russia, the media covers Tony Stark's discl...
14825	tt0063443	Play Dirty	During the North African Campaign in World War...
14826	tt0039464	High Wall	Steven Kenet catches his unfaithful wife in th...
14827	tt0235166	Against All Hope	Sometime in the 1950s in Chicago a man, Cecil ...

14828 rows × 3 columns

In [19]:

```
classifier = pipeline("text-classification",
                      model='bhadresh-savani/distilbert-base-uncased-emotion', # Maybe c
onsider different models.
                      return_all_scores=True)
```

In [29]:

```
nlp = spacy.load("en_core_web_sm")
```

In [7]:

```
# The following function returns the averages of the 6 primary emotions.
# The averages, because we must split the synopsis into sentences before training the mod
el.
def classify_synopsis(sents, classifier):
  preds = np.array([])
  for sent in sents:
    if len(sent) > 512: # maximum accepted by the transformer!
      func = lambda l, x: [l[i:i+x] for i in range(0, len(l), x)]
      return classify_synopsis(func(sent, 512), classifier)
    prediction = classifier(str(sent))
    preds = np.append(prediction, preds)
  x = pd.DataFrame.from_records(preds)
  x = dd.from_pandas(x, npartitions=5).compute()
  r = x.groupby('label').mean()
  return r.score.to_dict()
```

In [27]:

```
df["emotions"] = df["plot_synopsis"].apply(lambda x: classify_synopsis(nlp(x).sents, cla
ssifier))
df.emotions
```

Out[27]:

```
0        {'anger': 0.23212224676995233, 'fear': 0.47384...
1        {'anger': 0.5508819421451977, 'fear': 0.064073...
2        {'anger': 0.21647907543305528, 'fear': 0.22596...
3        {'anger': 0.29966469810739604, 'fear': 0.13232...
4        {'anger': 0.48358979746475456, 'fear': 0.24775...
                               ...
14822    {'anger': 0.3423848313457601, 'fear': 0.331624...
```

```
14823    {'anger': 0.4996576787671074, 'fear': 0.212360...
14824    {'anger': 0.5647137483324984, 'fear': 0.178041...
14825    {'anger': 0.34717418948809303, 'fear': 0.40262...
14826    {'anger': 0.2554710508723344, 'fear': 0.504246...
Name: emotions, Length: 14827, dtype: object
```

In [156]:

```
df.head(10)
```

Out[156]:

	imdb_id	title	plot_synopsis	emotions
0	tt0057603	I tre volti della paura	Note: this synopsis is for the orginal Italian...	{'anger': 0.23212224676995233, 'fear': 0.47384...
1	tt1733125	Dungeons & Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a s...	{'anger': 0.5508819421451977, 'fear': 0.064073...
2	tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the ...	{'anger': 0.21647907543305528, 'fear': 0.22596...
3	tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by anyone'...	{'anger': 0.29966469810739604, 'fear': 0.13232...
4	tt0086250	Scarface	In May 1980, a Cuban man named Tony Montana (A...	{'anger': 0.48358979746475456, 'fear': 0.24775...
5	tt1315981	A Single Man	George Falconer (Colin Firth) approaches a car...	{'anger': 0.36766072945852585, 'fear': 0.24443...
6	tt0249380	Baise-moi	Baise-moi tells the story of Nadine and Manu w...	{'anger': 0.44969198020855683, 'fear': 0.34324...
7	tt0408790	Flightplan	Kyle Pratt (Jodie Foster) is a propulsion engi...	{'anger': 0.2957883257945271, 'fear': 0.501962...
8	tt0021079	Little Caesar	Small-time Italian-American criminals Caesar E...	{'anger': 0.6120028616195279, 'fear': 0.226918...
9	tt1615065	Savages	The movie begins with a video being shot of me...	{'anger': 0.4545921616985204, 'fear': 0.237704...

In [39]:

```
df.iloc[7111]
```

Out[39]:

```
Unnamed: 0                                                 7111
imdb_id                                                tt0062819
title                                                  Commandos
plot_synopsis    It is the middle of World War II, and in the d...
emotions         [As a result, the elderly German soldier fires...
Name: 7111, dtype: object
```

In [40]:

```
df.drop([7111], axis=0, inplace=True)
```

In [41]:

```
df["emotions"] = df["emotions"].apply(lambda x : dict(eval(x)))
```

In [42]:

```
df = pd.concat([df, df["emotions"].apply(pd.Series)], axis = 1)
```

In [43]:

```
df = df.drop(["emotions"],axis = 1)
```

In [44]:

```
df.sort_values(['anger', 'fear'], ascending=[False, False])
```

Out[44]:

	Unnamed: 0	imdb_id	title	plot_synopsis	anger	fear	joy	love	sadness	surprise
11000	11000	tt0043665	I Was a Communist for the FBI	Matt Cvetic (Frank Lovejoy), who works in a Pi...	0.842094	0.087138	0.006485	0.000779	0.062751	0.000754
13722	13722	tt0041142	Awful Orphan	Charlie is showing various thing in the form o...	0.831141	0.070304	0.082511	0.001188	0.013784	0.001072
9919	9919	tt1190536	Black Dynamite	In the 1970s, Black Dynamite, a Vietnam War ve...	0.816960	0.098479	0.062084	0.008011	0.012441	0.002025
13103	13103	tt0038671	Kitty Kornered	The neighborhood's cat owners all (literally) ...	0.814127	0.148266	0.023482	0.001388	0.011547	0.001189
10958	10958	tt0034524	Blitz Wolf	The plot is a parody of the Three Little Pigs,...	0.803907	0.099204	0.064039	0.001850	0.028733	0.002267
...	...	...	...	...	...	...	...	...	...	...
5369	5369	tt0950500	Pen choo kab pee	The story tells of a pregnant village girl, Nu...	0.043965	0.684488	0.090514	0.001247	0.088809	0.090977
13997	13997	tt0816556	Lake Mungo	Sixteen-year-old Alice Palmer drowns while swi...	0.042284	0.753604	0.081386	0.036506	0.081508	0.004711
11896	11896	tt0248661	3 A.M.	The movie starts with few students visiting a ...	0.039836	0.416681	0.381151	0.002937	0.094999	0.064396
10634	10634	tt0382806	Look Both Ways	The film charts the stories of several people ...	0.029845	0.240438	0.409241	0.100514	0.217655	0.002308
633	633	tt1101051	The Waltons: A Decade of the Waltons	The real-life 'John-Boy Walton' (Earl Hamner, ...	0.021421	0.067015	0.731056	0.102624	0.070295	0.007589

14826 rows × 10 columns

In [45]:

```
df.to_csv("movies_emotions.csv")
```