```
# Upload the dataset.
from google.colab import files
uploaded = files.upload()
```

```
<IPython.core.display.HTML object>

Saving Imdb_data.csv to Imdb_data.csv
```

```
!pip install --upgrade --no-cache-dir numpy==1.23.5 scipy==1.10.1
pandas==1.5.3 scikit-learn==1.2.2 gensim==4.3.1
```

```
Requirement already satisfied: numpy==1.23.5 in
/usr/local/lib/python3.11/dist-packages (1.23.5)
Collecting scipy==1.10.1
  Downloading scipy-1.10.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (58 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 58.9/58.9 kB 10.7 MB/s eta
0:00:00
ent already satisfied: pandas==1.5.3 in
/usr/local/lib/python3.11/dist-packages (1.5.3)
Traceback (most recent call last):
  File
"/usr/local/lib/python3.11/dist-packages/pip/_internal/cli/base_comman
d.py", line 179, in exc_logging_wrapper
^C
```

1. Data Exploration and Preprocessing (5 Marks)

- Analyze the dataset for trends, missing values, and outliers.

o Perform basic data exploration, such as checking for missing values, identifying imbalanced classes (positive/negative), and analyzing the length of reviews.

```
# Print the top 5 rows of the datset.
import pandas as pd

df = pd.read_csv('Imdb_data.csv')
df.head()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 50000,\n  \"fields\":
[\n    {\n      \"column\": \"review\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 49581,\n
\"samples\": [\n          \"Poorly done political actioner. Badly
photographed, acted, and directed. Every single scene is underlighted,
including those very few that are shot during the daytime. It doesn't
matter what the location is. At an important conference in the White
House, no lights are on, and the only available lighting is a gloomy
blue that is filtered through a few windows. The primier of China
conducts an earth-shattering phone conversation under conditions of
such intense chiaroscuro that he should be contemplating a bust of
Homer in a Rembrandt painting. Honest. It's as if he had a tiny
```

spotlight on his face and was otherwise in total darkness. The slow motion deaths are by now obligatory in any ill-thought-out movie.

Roy Scheider and Maria Conchita Alonzo do well by their roles, but Scheider is rarely on screen. The other performances are dismissable. There is a pretty Oriental woman in a short tight skirt who totes a gun and is right out of a Bond movie who's accent suggests a childhood spent in Basset, Nebraska, and who should have remained the model she probably started out as. Whoever plays the surviving Secret Service agent aboard the cruise ship was probably picked for the part because he looked most like Johnny Depp, not because of any display of talent. The Chinese villains, representing both Taiwan and mainland China, hiss and grin as they threaten the heroes.

The script is pretty awful, recycled from other, better films. There is a lot of shooting aboard the ship and practically everyone winds up mincemeat. Two thirds of the way through, the ship explodes into the expected series of fireballs. Then the movie splits into two related parts. Part one, another shootout, this time in a waterfront warehouse. Part two, an exchange between the Vice President, now acting president, and the oily Chinese premiere, lifted out of both \\\"Dr. Strangelove\\\" and \\\"Fail Safe.\\\" We unwittingly launch our missiles. They launch theirs in retaliation. We cannot convince them that our launch was accidental, even though we offer to help them destroy our own missiles. There is even the George C. Scott/ Walter Matthau general who argues that their \\\"nucular\\\" armory can't match ours so we should hit them with everything we've got. More fireballs.

The end comes none too soon.\",\n\"In Sri Lanka, a country divided by religion and language, the civil war between the pro-Sinhalese government and the Liberation Tigers of Tamil Eelam (LTTE), a separatist organization, has claimed an estimated 68,000 lives since 1983. Human rights groups have said that, as a result of the war, more than one million people have been displaced, homeless or living in camps. The impact on children and families caught in the conflict is sensitively dramatized by acclaimed Tamil director Mani Ratnam in his 2002 film A Peck on the Cheek, winner of several awards at the National Film Awards in India. While the civil war is merely a backdrop for the story of a young girl's voyage of discovery, the human cost of war is made quite clear and Ratnam gives the fighting a universal context, pointing the finger at global arms traffickers as the source of wrongdoing.

Beautifully photographed in Southern India by cinematographer Ravi K Chandran in a setting mirroring the terrain of Sri Lanka, the film tells a moving story about an adopted 9-year old girl who sets out to find her real mother in the middle of the fighting in Sri Lanka. Played with deep feeling and expressiveness by P.S. Keerthana in a memorable performance, Amudha is brought up by a loving middle class family with two younger brothers after her natural parents Shyama (Nandita Das) and Dileepan (J.D. Chakravarthi) were forced to flee when the fighting broke out, leaving her in a Red Cross camp. In a loving flashback, we see Amudha's adoptive parents, father Thiru

(Madhavan) a prominent Tamil writer, and mother Indra (Simran) a TV personality, marry to facilitate their adoption of the darker-skinned little girl.

Young Amudha has no idea that she is adopted until it is sprung upon her abruptly on her ninth birthday, according to the parents' prior agreement. While she is playing, Thiru tells her almost in a matter of fact tone that \\\"you are not our daughter\\\" and the response is predictable. Distraught, she questions who her father was, what her mother's name was, why she gave her up, and so forth but few answers are forthcoming. Amudha runs away several times until her parents agree to go to Sri Lanka to help her find her true mother, now a fighter for the Tamil separatists. The family's immersion in the reality of the civil war leads to some traumatic moments and difficult decisions, handled mostly with skill by Ratnam, though a sequence where the family was caught in a crossfire felt amateurish.

A Peck on the Cheek is of course a Bollywood-style film and that means tons of music and melodrama. The melodrama did not get in the way because of the strong performances by the lead actors; however, I found the musical dramatizations of songs by A. R. Rahman counter to the mood of the film with their slick, high production techniques and fast-paced music video-style editing. Yet the compelling nature of the story and the honesty in which it is told transcend the film's limitations. Tamil cinema has been criticized by many, even within the country as being too clich\\u00e9d and commercial, yet A Peck on the Cheek is both a film of entertainment and one that tackles serious issues. That it successfully straddles the line between art and commerce is not a rejection but a tribute.\",\n   \"FUTZ is the only show preserved from the experimental theatre movement in New York in the 1960s (the origins of Off Off Broadway). Though it's not for everyone, it is a genuinely brilliant, darkly funny, even more often deeply disturbing tale about love, sex, personal liberty, and revenge, a serious morality tale even more relevant now in a time when Congress wants to outlaw gay marriage by trashing our Constitution. The story is not about being gay, though -- it's about love and sex that don't conform to social norms and therefore must be removed through violence and hate. On the surface, it tells the story of a man who falls in love with a pig, but like any great fable, it's not really about animals, it's about something bigger -- stifling conformity in America.

The stage version won international acclaim in its original production, it toured the U.S. and Europe, and with others of its kind, influenced almost all theatre that came after it. Luckily, we have preserved here the show pretty much as it was originally conceived, with the original cast and original director, Tom O'Horgan (who also directed HAIR and Jesus Christ Superstar on Broadway).

This is not a mainstream, easy-to-take, studio film -- this is an aggressive, unsettling, glorious, deeply emotional, wildly imaginative piece of storytelling that you'll never forget. And it just might change the way you see the world...\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\":

```
\"sentiment\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"negative\",\n          \"positive\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
# Check for missing values
df.info()
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   review     50000 non-null  object
 1   sentiment  50000 non-null  object
dtypes: object(2)
memory usage: 781.4+ KB
```

```
review       0
sentiment    0
dtype: int64
```

```
# Check class distribution
df['sentiment'].value_counts()
```

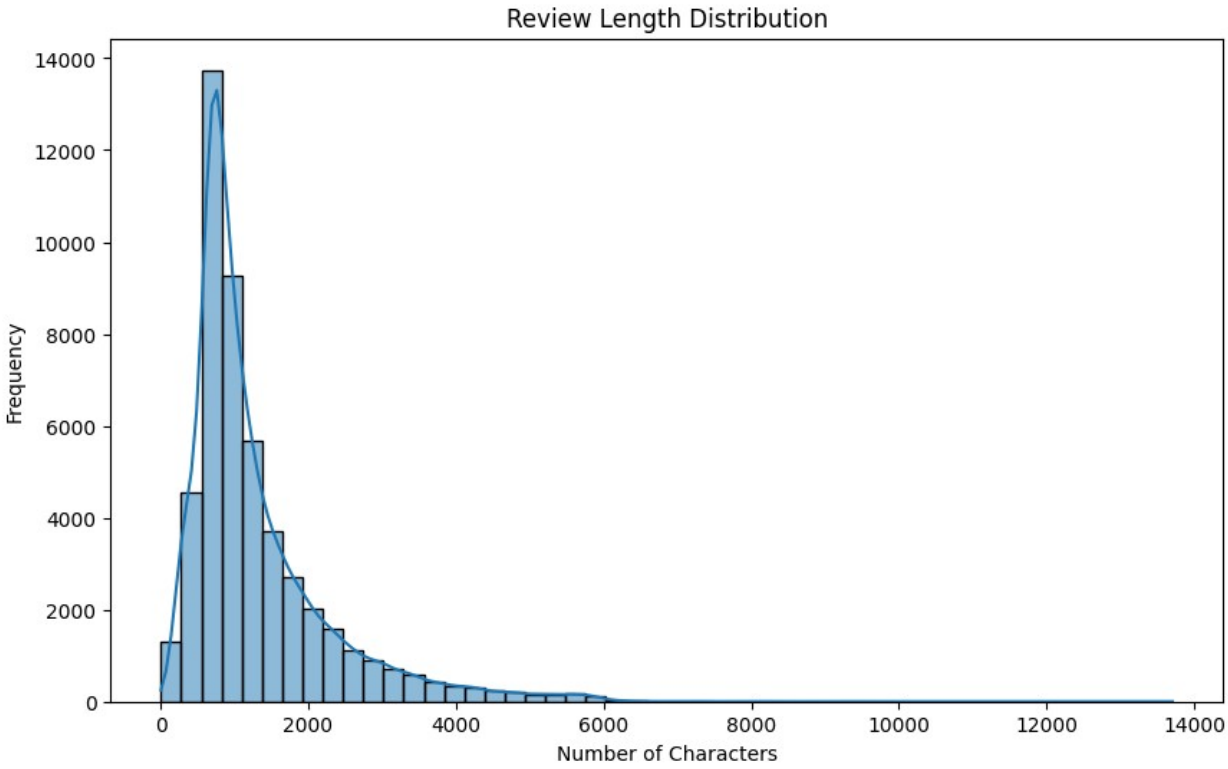
```
positive    25000
negative    25000
Name: sentiment, dtype: int64
```

```
# Add a new column for Review Length
df['review_length'] = df['review'].apply(len)
df['review_length'].describe()
```

```
count    50000.000000
mean      1309.367720
std        989.759532
min          7.000000
25%        699.000000
50%        970.000000
75%       1590.000000
max      13704.000000
Name: review_length, dtype: float64
```

```
# Visualize Review Length Distribution
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10,6))
sns.histplot(df['review_length'], bins=50, kde=True)
```

```
plt.title('Review Length Distribution')
plt.xlabel('Number of Characters')
plt.ylabel('Frequency')
plt.show()
```




```
# Identify Outliers
outliers = df[df['review_length'] > 3000]
print(f'Number of outliers: {len(outliers)}')
outliers.head(10)
```

```
Number of outliers: 3492
```


```
{"summary":"{\n  \"name\": \"outliers\",\n  \"rows\": 3492,\n
\"fields\": [\n    {\n      \"column\": \"review\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 3475,\n        \"samples\": [\n          \"I
went into a Video Store and looked around to find some Horror Movies,
after about 30 minutes I just rushed and picked out a few. I stumbled
upon \\\"Masters of Horror\\\" which contained \\\"Pro-Life\\\" and
\\\"Right to Die\\\". They seemed OK, same-old cheesy Horror crap, but
I was interested for some reason. It said about Pro-Life on the case
about being a classic, a return to form for John Carpenter (I loved
his \\\"The Thing\\\", so I thought this would be good) and all that.
So I turned it on thinking it would be something great and
interesting, I was very wrong... It started off casual, just a girl
```

running through a Forest, scared of something. A car stops and picks her up (just so being the people she needed to see, amazing?) They take her back to some Clinic and examine her, at the sametime all this is happening her Father appears at the gates and they don't allow him in, he isn't aloud near the area. Most likely from something he would of done in the past, but you don't know of any of this at the moment. He really does not want his Daughter in this place, an Abortion center. He is very strongly against such acts, believing it's sickening and not what \\\"God\\\" would want. He \\\"supports\\\" what I heard is called \\\"Pro-Life\\\". Acting against Abortions and going to extremes to allow the Babies to be born, they are sick. They don't like the Life of an unborn being taken, yet they've killed Humans in the past to allow the Birth? Justice is only a figment of the mind. Anyway, back on track, after the girl is examined they find out shes pregnant, but far ahead than what she should be. She is only a few weeks pregnant, but is months ahead. She keeps telling them they wont understand her, and that she wants an Abortion and all, but finally tells the truth that she was raped by a Demon from Hell, and that her Father wants this baby (but believes \\\"God\\\" wants this baby, not who truly does). He gets his 3 Sons (they arm themselves with Pistols and Shotguns), and begin to make they're way into the Clinic (shooting down anyone who won't co-operate). The head of the Clinic, who must of had trouble with them in the past, is well prepared this time. Ends up killing one the Fathers Sons, but in the end gets shot a few times (wearing a bullet proof jacket). The Father then performs what he believes is done to the Women. He cuts a hole, where the Vagina would be if he we're a Female, and sticks some sort of sucking thing up there and sucks out all this blood. Whilst all this is happening, the girl gives birth to some Demonic baby with many legs, and some Demon raises from beneath the Earth (not in the same room) and starts looking for its child. The Father sees this later on, and starts questioning why this happened, he did what he was told to do, and doesn't understand why it's like this. The Demon had killed both of his Sons earlier, and now goes for Father... Whilst the girl kills the baby, and the Demon carries it away (not in the same scene).

Yeah, it probably sounds pretty cool, and a thrilling Horror Movie, but it isn't. The acting is horrible and lacks enthusiasm, the script is boring and not even creative, they choose the wrong characters and don't even build on them; just everything put together, all the small parts, don't even add up to something great, a waste of time. I wouldn't classify this as a Horror, though it has elements of Horror, they ultimately fail at what they try to succeed. It felt more like a \\\"Beginners\\\" Short-Movie, than by John Carpenter.

Sorry for my lack of information, and detailed review, I just didn't have the time to waste to write something exciting. Also sorry if my spelling and details are incorrect, I couldn't really be bothered to research anything.\",\n   \"Uhhh ... so, did they even have writers for this? Maybe I'm picky, but I like a little dialog with my movies. And, as far as slasher films go,

just a sliver of character development will suffice.

Unfortunately, The Prey provides neither\\u0097and if you think I'm being hyperbolic, you'll just have to see it for yourself. Scene after scene, we just get actors standing around, looking forlorn and awkward, abandoned by any sense of a script. Outside of calling out each other's names when they get separated in the woods (natch), the only instances where these people say something substantive is when one character explains the constellation Orion (clearly plagiarized from Funk & Wagnalls; scintillating slasher fare, no?) and another rehashes an old campfire tale that doesn't even have anything to do with the plot (wait, what IS the plot?) At other times, The Prey actually has the gall to film its characters with the boom mic just far away enough so that we can't exactly hear what they're saying. So we get entire scenes wherein the actors are murmuring! Deliberately! Seriously, I've seen more dialog in a silent film. It's as if the filmmakers sat down at a bar somewhere in Rancho Cucamonga in the heyday of the '80s slasher craze and one looked at the other and said, \\\"Hey, I gotta really sweet idea for a gory decapitation gag. Let's somehow pad an entire feature around it.\\\" And ... well, they did.

To be fair, The Prey probably had some sort of writer on board. I mean, somebody had to jot down the scene sequence and label the dailies. However, I am fully convinced that this film did not have an editor of any kind whatsoever. There are glaring pauses, boring tableaux, and zero sense of pacing throughout. The filmmakers don't have anything else in the \\\"script\\\" to film, so they fill out the running time with exhaustive taxonomies of the flora and fauna that inhabit the forest in which our wild and crazy teens are getting sliced and diced. These critters are all filmed in straightforward, noontime daylight in a completely reserved fashion and with no attempt at atmospheric photography. If it feels like a science film, that's because it is. I'm pretty sure this is all nature show stock footage\\u0097all that's missing is a stuffy narration from some National Geographic alderman.

More exciting footage that was graciously spared from the cutting room floor: a scene in which two men discuss cucumber and cream cheese sandwiches, and another scene wherein a supporting character strums away on a banjo for what feels like an entire minute-and-a- half! A minute-and-a-half! That's a lot of banjoing to commit to celluloid to begin with, let alone insert into the final cut of the film! Way to go, guys! Brevity and concision are the real victims of this slaughterfest.

Admittedly, the film picks up quite a bit of steam (comparatively) in the last 25 minutes, into which much of the carnage is condensed and where a rip-off of B\\u00e9la Bart\\u00f3k's \\\"Music for Strings, Percussion and Celesta\\\" cuts in. Vaudeville great Jackie Coogan makes a fun appearance as a tubby, bumbly park ranger (this was his last role, if you can believe it). And there are some nice gory moments, including a splattery neck tearing and the aforementioned decapitation. The make-up used for the killer (Carel Struycken, aka \\\"Lurch\\\" from the Addams Family movies) is also quite effective, and makes him look like

a strange hybrid of young Jason Voorhees and Freddy Krueger. Plus, if you love wacky, straight-outta-left-field endings, you need to check out how they wrap this puppy up. You'll do a spit take, I promise.

Usually, I love films that are on this level of ineptitude, but the first three-quarters of The Prey are just so interminably boring that they pretty much spoil the rest. Overall, this is a largely pallid and tedious affair, and, while it ain't all bad, it should really only be seen by debilitated slasher completists. Why do we do this to ourselves, anyway?\",\n \"Attack Force has a horrendous title, and can almost certainly be judged by it's awful cover, because the film is horrible! A mish-mash of plot lines, a choppy mess, and a horribly stagnated pace, make the film hard to watch start to finish. I managed this and I'm proud. As a fan of Seagal's work (mostly of his old days), it's painful to see him star in such tripe. True Seagal's last half dozen movies or so, have sucked a lot, but some of them at least had some redeeming features. Attack Force is a mess. From conception to delivery this film has undergone many changes, from an alien plot line, to the current one about a highly addictive super drug, about to be unleashed on the Romanian (the film has several settings, none of which are Romanian, but all look like Romania because they are in Romania!) populace. The film is tacked together with little regard for whatever state the original shooting script was. Plot-holes and loose ends are abound in the film that's for sure. That's been a problem in Seagal's last few films as well, but never has the result been so boring. There's a whole plot line about the water supply being poisoned with CTX (that's the drugs cool name) that is never resolved!

Of course in recent years the plot's haven't been the main draw in the Seagal canon so there was a big onus on the other departments, especially the action. Before I regard the action though, all the other departments are poor. The direction is poor, or perhaps better put, made to look poor. Who knows how director Michael Keusch originally intended this film? Between him finishing his job, the re-shoots by stunt man Tom Delmar, and the editing, a coherent auteur vision is completely lost. The best way to describe the film is that it's just all over the shop! The cinematography is dull, nearly inducing sleep, while the droning score (sounding like it was produced on the cheapest of cheap synthesizers) does nothing to excite matters. The cast too are poor, unable to salvage anything here. Seagal looks bored beyond recognition, and is dubbed through much of the picture, clearly when plot-points are being changed. He looks tired and overweight, and lethargic, unlike he's looked in previous pictures too (remarkable as the aforementioned have been key complaints in Seagal's recent pictures). The only redeemable cast member is Adam Croasdell as one of the villains, doing a slimy Brit routine. He seems to be a throwback to the alien plot line, because he's playing it inhuman. He seems like a cross between a body snatcher and a vampire (ditto to the lead villain played by some hot chick who appears on occasion, seemingly waiting for her husband\\ u0085 Dracula).

Finally the action. Well it's poor. Poorly

conceived, poorly shot. There's not much either, and there's even less featuring Seagal. Stevo doesn't really bring out the stunt double here, because there's so little to do. There's even a lengthy (repetitive and boring) action scene on the hour mark that inter-cuts occasionally with little flashes of Seagal's stand in because clearly Seagal wasn't there while the scene was being shot, and they wanted to have him feature in the action scene. Seagal eventually appears in person to shoot two guys in the head. Seagal has a producers credit here and a script credit, but from what I understand the film has been altered behind his back to the current state it's in. Seagal will apparently not be working with these people again, or with Castel Studio's who continue to deliver horrifically sub-Nu-Image (that's saying something), material.

Overall this is one to avoid if you are not a Seagal fan. Seagal fans can also be safe in the knowledge that the big man probably won't want to do anything this bad again. Unfortunately his next film which has already been shot, with the same people, promises to be even worse than this. *\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"sentiment\",\n   \"properties\": {\n   \"dtype\": \"category\",\n   \"num_unique_values\": 2,\n   \"samples\": [\n   \"negative\",\n   \"positive\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"review_length\",\n   \"properties\": {\n   \"dtype\": \"number\",\n   \"std\": 950,\n   \"min\": 3001,\n   \"max\": 13704,\n   \"num_unique_values\": 1923,\n   \"samples\": [\n   5653,\n   3489\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"clean_review\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 3475,\n   \"samples\": [\n   \"i went into a video store and looked around to find some horror movies after about minutes i just rushed and picked out a few i stumbled upon masters of horror which contained prolife and right to die they seemed ok sameold cheesy horror crap but i was interested for some reason it said about prolife on the case about being a classic a return to form for john carpenter i loved his the thing so i thought this would be good and all that so i turned it on thinking it would be something great and interesting i was very wrong it started off casual just a girl running through a forest scared of something a car stops and picks her up just so being the people she needed to see amazing they take her back to some clinic and examine her at the sametime all this is happening her father appears at the gates and they dont allow him in he isnt aloud near the area most likely from something he would of done in the past but you dont know of any of this at the moment he really does not want his daughter in this place an abortion center he is very strongly against such acts believing its sickening and not what god would want he supports what i heard is called prolife acting against abortions and going to extremes

to allow the babies to be born they are sick they dont like the life of an unborn being taken yet theyve killed humans in the past to allow the birth justice is only a figment of the mind anyway back on track after the girl is examined they find out shes pregnant but far ahead than what she should be she is only a few weeks pregnant but is months ahead she keeps telling them they wont understand her and that she wants an abortion and all but finally tells the truth that she was raped by a demon from hell and that her father wants this baby but believes god wants this baby not who truly does he gets his sons they arm themselves with pistols and shotguns and begin to make theyre way into the clinic shooting down anyone who wont cooperate the head of the clinic who must of had trouble with them in the past is well prepared this time ends up killing one the fathers sons but in the end gets shot a few times wearing a bullet proof jacket the father then performs what he believes is done to the women he cuts a hole where the vagina would be if he were a female and sticks some sort of sucking thing up there and sucks out all this blood whilst all this is happening the girl gives birth to some demonic baby with many legs and some demon raises from beneath the earth not in the same room and starts looking for its child the father sees this later on and starts questioning why this happened he did what he was told to do and doesnt understand why its like this the demon had killed both of his sons earlier and now goes for father whilst the girl kills the baby and the demon carries it away not in the same sceneyeah it probably sounds pretty cool and a thrilling horror movie but it isnt the acting is horrible and lacks enthusiasm the script is boring and not even creative they choose the wrong characters and dont even build on them just everything put together all the small parts dont even add up to something great a waste of time i wouldnt classify this as a horror though it has elements of horror they ultimately fail at what they try to succeed it felt more like a beginners shortmovie than by john carpentersorry for my lack of information and detailed review i just didnt have the time to waste to write something exciting also sorry if my spelling and details are incorrect i couldnt really be bothered to research anything\",\n   \"uhhh so did they even have writers for this maybe im picky but i like a little dialog with my movies and as far as slasher films go just a sliver of character development will sufficeunfortunately the prey provides neitherand if you think im being hyperbolic youll just have to see it for yourself scene after scene we just get actors standing around looking forlorn and awkward abandoned by any sense of a script outside of calling out each others names when they get separated in the woods natch the only instances where these people say something substantive is when one character explains the constellation orion clearly plagiarized from funk wagnalls scintillating slasher fare no and another rehashes an old campfire tale that doesnt even have anything to do with the plot wait what is the plot at other times the prey actually has the gall to film its characters with the boom mic just far away enough so that we cant exactly hear what theyre saying so we get entire scenes wherein the

actors are murmuring deliberately seriously ive seen more dialog in a silent film its as if the filmmakers sat down at a bar somewhere in rancho cucamonga in the heyday of the s slasher craze and one looked at the other and said hey i gotta really sweet idea for a gory decapitation gag lets somehow pad an entire feature around it and well they did to be fair the prey probably had some sort of writer on board i mean somebody had to jot down the scene sequence and label the dailies however i am fully convinced that this film did not have an editor of any kind whatsoever there are glaring pauses boring tableaux and zero sense of pacing throughout the filmmakers dont have anything else in the script to film so they fill out the running time with exhaustive taxonomies of the flora and fauna that inhabit the forest in which our wild and crazy teens are getting sliced and diced these critters are all filmed in straightforward noontime daylight in a completely reserved fashion and with no attempt at atmospheric photography if it feels like a science film thats because it is im pretty sure this is all nature show stock footageall thats missing is a stuffy narration from some national geographic aldermanmore exciting footage that was graciously spared from the cutting room floor a scene in which two men discuss cucumber and cream cheese sandwiches and another scene wherein a supporting character strums away on a banjo for what feels like an entire minuteanda half a minuteandahalf thats a lot of banjoing to commit to celluloid to begin with let alone insert into the final cut of the film way to go guys brevity and concision are the real victims of this slaughterfestadmittedly the film picks up quite a bit of steam comparatively in the last minutes into which much of the carnage is condensed and where a ripoff of bla bartks music for strings percussion and celesta cuts in vaudeville great jackie coogan makes a fun appearance as a tubby bumbly park ranger this was his last role if you can believe it and there are some nice gory moments including a splattery neck tearing and the aforementioned decapitation the makeup used for the killer carel struycken aka lurch from the addams family movies is also quite effective and makes him look like a strange hybrid of young jason voorhees and freddy krueger plus if you love wacky straightouttaleftfield endings you need to check out how they wrap this puppy up youll do a spit take i promiseusually i love films that are on this level of ineptitude but the first threequarters of the prey are just so interminably boring that they pretty much spoil the rest overall this is a largely pallid and tedious affair and while it aint all bad it should really only be seen by debilitated slasher completists why do we do this to ourselves anyway\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"tokens\",\n   \"properties\": {\n   \"dtype\": \"object\",\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"tokens_processed\",\n   \"properties\": {\n   \"dtype\": \"object\",\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"processed_review\",\n

\"properties\": {\n   \"dtype\": \"string\",\n
\"num_unique_values\": 3474,\n   \"samples\": [\n
\"thing dont say taglin david zucker comedi young man caught one horrend situat anoth entitl boss daughter taglin speak peopl made junk well contempl watch inde thing dont place mostli sort content found within boss daughter count wholli item medium cinema includ pictur boss daugther sordid creepi grotesqu experi clunki heavi hand piec infantil beyond word disgust beyond express see endur endur surviv surviv accomplish cast writer extra hell even guy work runner set aid produc anyth earthshatteringli poor itll either theyv sent devil destroy medium film itll theyv probabl garner employ behalf friedbergseltz mobmi boss daugther im pretti sure ought titl bo daugther grammat speak revolv around hapless male lead name tom stansfield kutcher night bo hous chase seemingli elus goal young blond daughter lisa taylor reid someon work within depart tower chicago offic block whilst strict eye jack taylor stamp tom spi lisa earli she take subway work shmo despit fact own car father bo damn compani tri talk attempt foil puke babi dog blind interest tom crotch anyth els final get chanc offic talk afterdark parti elsewher aris ought come round hous visit ye still live father think hitchcock film psycho gender role norman bate mother revers play laughsth distinct establish charact made pain appar open scene tom sit subway train travel work yuppi cohort ruthless smarmi bunch made appar swipe briefcas unfortun enough get stuck door ensu morn rush without ever return one day happen tom wish return thu pound u hesnotlikeotherguy first see lisa carriag somewhat shi approach men treat whole situat would breez posit rather obviou flatfoot attempt tri get u side tom sit uneasili supposedli take earn place amidst cowork companyit howev close boss daughter come level filmmak seemingli harmless premis boy meet girl want get know arriv comedi hell tom arriv hous see invit parti instead charg hous sit jack pet owl gener keep mischief whilst maintain spotless hous establish terranc stamp charact mean busi strictest manner fire peopl smallest thing make bad cup coffe jack shrew businessman he cleanli freak obsess control borderlin sociopath place bear trap garden keep child next door land imagin let larg exquisit hous order noth go wrong there obvious go troubleth film fun premis danger ten minut first time someon us worktop crack open beer thu mark pristin top may smirk time half hous wreck michael madsen shown urin rug youv got head hand joke film set almighti clunki manner play way closer slow excruci slick faultless thing miss follow next gag sound effect someon incorrectli chang gear car clunk creak onto next pratfal inbetween grossout wacki film take time roll rout yucki saccharin driven romanc lisa tom bond whilst talk inworkplac outof workplac persona mayb common first thought hour mark film opt gross gag hate fill jibe anyth there entir scene exist pure target parapleg dumb subplot headinjuri sport neighbour blind date truli unwatch sight gag unfold throughout stamp charact enjoy put peopl ask simplest task difficult commonplac repli ought whilst channel jack taylor read screenplay first complicateda concept stamp\",\n   \"see driven

```
plane flight america year ago truli believ seen worst film ever creat
could relax safe knowledg would never suffer much front screen ever
unfortun found last night case revolv monstrous bad actual think
recommend friend go see dont feel like im one stupid enough con watch
realli quit amaz much film fall complet face constant mean constant
voic over main charact total inan pretenti nonsens actual get angri
cinema listen andr benjamin utterli relentless drone seem like half
film whilst time think would turkish done complet joke gangstercon man
whatev he suppos made offer ill tell would told fk blown head away
watch utter disdain equal inept partner waddl away fast chubbi littl
leg would carri mean suppos believ go jake head offer solut problem
theyr con men therefor must obvious also skill cure incur blood diseas
mean ff doesnt start wonder symptom arent get wors doesnt penni drop
third day happen instead richi subject audienc pain patronis phone
call avi jake let know he con anyway add small posit note film move
dri humour provid thank similar standard previou film bullst film
doesnt tri anyth smart redeem well time amus line oh somehow manag
disastr unfunni genuin didnt hear much titter complet pack cinema
anyon know ugc sheffield know full main screen get person much smile
mayb never want film funni fair enough still make good gangster film
without comedi plan hang film may ask unnecessarili baffl plot sincer
hope notbi far satisfi moment went last night hear loud sigh come
direct audienc everyon desper pray film end also realli quit amus
watch fast patron fight dash exit realis free tormentil round ive got
finish write make angri elabor end mean sht end sorri cant go go see
cant put word cant youv seen youll know uuhhhhh shudder\"\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"word_count\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
86,\n        \"min\": 228,\n        \"max\": 1420,\n
\"num_unique_values\": 354,\n        \"samples\": [\n          386,\n
418\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"char_count\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 470,\n        \"min\": 1213,\n
\"max\": 6926,\n        \"num_unique_values\": 1379,\n
\"samples\": [\n          2038,\n          2303\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"avg_word_length\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.24430958901159086,\n        \"min\": 4.516728624535316,\n
\"max\": 6.498392282958199,\n        \"num_unique_values\": 3310,\n
\"samples\": [\n          5.7158176943699734,\n
5.544668587896253\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"outliers"}
```


- **Perform data cleaning and text preprocessing.**

o Steps will include:

■ Removing stop words, punctuation, and special characters.

■ Tokenization of text (splitting text into words).

■ Lemmatization and stemming.

■ Vectorization using techniques like Bag-of-Words and TF-IDF.

```
# Import all the required libraries.
!pip install nltk scikit-learn
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-
packages (3.9.1)
Requirement already satisfied: scikit-learn in
/usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: click in
/usr/local/lib/python3.11/dist-packages (from nltk) (8.2.0)
Requirement already satisfied: joblib in
/usr/local/lib/python3.11/dist-packages (from nltk) (1.5.0)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-
packages (from nltk) (4.67.1)
Requirement already satisfied: numpy>=1.19.5 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.23.5)
Requirement already satisfied: scipy>=1.6.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
```

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer, PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer,
TfidfVectorizer

nltk.download('stopwords')
nltk.download('punkt')
nltk.download('punkt_tab')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

```
True
```

```
# Clean Text: Remove HTML tags, punctuation, special characters
def clean_text(text):
    text = re.sub(r'<.*?>', '', text)  # Remove HTML tags
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # Remove special
characters & digits
    text = text.lower()  # Convert to lowercase
    return text

df['clean_review'] = df['review'].apply(clean_text)

# Tokenization
from nltk.tokenize import word_tokenize

df['tokens'] = df['clean_review'].apply(word_tokenize)

import pandas as pd
from nltk.tokenize import word_tokenize
from tqdm.notebook import tqdm

# Enable tqdm for progress tracking
tqdm.pandas()

# Batch size
batch_size = 2000

# Split dataframe into list of batches
batches = [df[i:i+batch_size] for i in range(0, df.shape[0],
batch_size)]

# Create an empty list to collect processed batches
processed_batches = []

# Loop through batches
for batch in tqdm(batches, desc="Processing Batches"):
    batch = batch.copy()
    batch['tokens'] =
batch['clean_review'].progress_apply(word_tokenize)
    processed_batches.append(batch)

# Concatenate all processed batches back together
df = pd.concat(processed_batches, ignore_index=True)
```


```
{"model_id":"e067f90e83d64389821d16f3c460bb20","version_major":2,"vers
ion_minor":0}

{"model_id":"650349c21a3c41119b481efe9a117d12","version_major":2,"vers
ion_minor":0}
```

{"model_id":"62767791f2cf42c0a578c80c2b1c03ff","version_major":2,"version_minor":0}

{"model_id":"ad84f8d8fe064504b86de7928dc0fa91","version_major":2,"version_minor":0}

{"model_id":"857e298fd5784ef6a32e62d98566563f","version_major":2,"version_minor":0}

{"model_id":"773b3f1a7ad343048a5702773b67e555","version_major":2,"version_minor":0}

{"model_id":"d99346ee5bed429f8fc213fcd814d2ee","version_major":2,"version_minor":0}

{"model_id":"70339d1c30a4497a9a10fd97807a2832","version_major":2,"version_minor":0}

{"model_id":"43f9f2c7fee14c94a28261bc76713c60","version_major":2,"version_minor":0}

{"model_id":"ef0277c44e5e4bc19b1d1cea234429c9","version_major":2,"version_minor":0}

{"model_id":"68211efd271e4e018f69f2f2b387227f","version_major":2,"version_minor":0}

{"model_id":"bbe1c0e6f3364e99ba98f9207868b504","version_major":2,"version_minor":0}

{"model_id":"47f8b4f0901d421e804e7a3879532148","version_major":2,"version_minor":0}

{"model_id":"500999117bb343df8526606cd98d38c0","version_major":2,"version_minor":0}

{"model_id":"bdb30f5e350b45ef90876dccc2dcbef8","version_major":2,"version_minor":0}

{"model_id":"aa7127a839514056a7062993e36e1928","version_major":2,"version_minor":0}

{"model_id":"608534c61b92441fbaa7b0358ff96b6e","version_major":2,"version_minor":0}

{"model_id":"89af720b1fac486fa3e9210abac6a5b4","version_major":2,"version_minor":0}

{"model_id":"da327ab60da34ac78abfdd493a2d3779","version_major":2,"version_minor":0}

{"model_id":"bcd45f1c7e7b4f75ad9b505f492af9b8","version_major":2,"version_minor":0}

{"model_id":"b2efe0baca6b4ae39b8bf3bf4c6e487d","version_major":2,"version_minor":0}

{"model_id":"f77f6f822aa746bbad0821f03b218e1d","version_major":2,"version_minor":0}

{"model_id":"09a04813982e4bac8928740b3ef60930","version_major":2,"version_minor":0}

{"model_id":"0577a4abeab2415388b1dd1c06cf7e90","version_major":2,"version_minor":0}

{"model_id":"3ce0bac312174bfa9f3460cb7084e465","version_major":2,"version_minor":0}

{"model_id":"3c1ec6e06b47435195b83df8f894319a","version_major":2,"version_minor":0}

```
# Remove all the stopwords
stop_words = set(stopwords.words('english'))

df['tokens'] = df['tokens'].apply(lambda x: [word for word in x if
word not in stop_words])

# Apply both lemmatization and stemming
lemmatizer = WordNetLemmatizer()
stemmer = PorterStemmer()

# Apply both Lemmatization and Stemming
df['tokens_processed'] = df['tokens'].apply(lambda x:
[stemmer.stem(lemmatizer.lemmatize(word)) for word in x])

# Join tokens back into text
df['processed_review'] = df['tokens_processed'].apply(lambda x: '
'.join(x))

# Applying Bag of Words to the processed reviews
bow_vectorizer = CountVectorizer()
X_bow = bow_vectorizer.fit_transform(df['processed_review'])

# Applying TF-IDF to the processed reviews
tfidf_vectorizer = TfidfVectorizer()
X_tfidf = tfidf_vectorizer.fit_transform(df['processed_review'])
```

1. Feature Engineering (10 Marks)

- Feature extraction using techniques like TF-IDF, Word2Vec, or embeddings.

o Transform the textual data into numerical features that can be used by machine learning models.

- Textual features: Word count, character count, average word length, etc.

```python
# Word count
df['word_count'] = df['processed_review'].apply(lambda x:
len(x.split()))

# Character count (without spaces)
df['char_count'] = df['processed_review'].apply(lambda x:
len(x.replace(" ", "")))

# Average word length
df['avg_word_length'] = df['char_count'] / df['word_count']

# Let's first implement TF-IDF Vectorization
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_tfidf = tfidf_vectorizer.fit_transform(df['processed_review'])

!pip install gensim numpy
```

```
Requirement already satisfied: gensim in
/usr/local/lib/python3.11/dist-packages (4.3.1)
Requirement already satisfied: numpy in
/usr/local/lib/python3.11/dist-packages (1.23.5)
Requirement already satisfied: scipy>=1.7.0 in
/usr/local/lib/python3.11/dist-packages (from gensim) (1.15.3)
Requirement already satisfied: smart-open>=1.8.1 in
/usr/local/lib/python3.11/dist-packages (from gensim) (7.1.0)
Requirement already satisfied: wrapt in
/usr/local/lib/python3.11/dist-packages (from smart-open>=1.8.1-
>gensim) (1.17.2)
```

```
!pip install --upgrade gensim numpy
```

```
Requirement already satisfied: gensim in
/usr/local/lib/python3.11/dist-packages (4.3.1)
Collecting gensim
  Downloading gensim-4.3.3-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (8.1 kB)
Requirement already satisfied: numpy in
/usr/local/lib/python3.11/dist-packages (1.23.5)
Collecting numpy
  Downloading numpy-2.2.6-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (62 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.0/62.0 kB 4.5 MB/s eta
0:00:00
py-1.26.4-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (61 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.0/61.0 kB 3.5 MB/s eta
0:00:00
 gensim)
  Downloading scipy-1.13.1-cp311-cp311-
```

```
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (60 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.6/60.6 kB 4.4 MB/s eta
0:00:00
ent already satisfied: smart-open>=1.8.1 in
/usr/local/lib/python3.11/dist-packages (from gensim) (7.1.0)
Requirement already satisfied: wrapt in
/usr/local/lib/python3.11/dist-packages (from smart-open>=1.8.1-
>gensim) (1.17.2)
Downloading gensim-4.3.3-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (26.7 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 26.7/26.7 MB 23.1 MB/s eta
0:00:00
py-1.26.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(18.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 18.3/18.3 MB 12.8 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (38.6 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 38.6/38.6 MB 9.5 MB/s eta
0:00:00
py, scipy, gensim
  Attempting uninstall: numpy
    Found existing installation: numpy 1.23.5
    Uninstalling numpy-1.23.5:
      Successfully uninstalled numpy-1.23.5
  Attempting uninstall: scipy
    Found existing installation: scipy 1.15.3
    Uninstalling scipy-1.15.3:
      Successfully uninstalled scipy-1.15.3
  Attempting uninstall: gensim
    Found existing installation: gensim 4.3.1
    Uninstalling gensim-4.3.1:
      Successfully uninstalled gensim-4.3.1
ERROR: pip's dependency resolver does not currently take into account
all the packages that are installed. This behaviour is the source of
the following dependency conflicts.
google-colab 1.0.0 requires pandas==2.2.2, but you have pandas 1.5.3
which is incompatible.
mizani 0.13.5 requires pandas>=2.2.0, but you have pandas 1.5.3 which
is incompatible.
xarray 2025.3.1 requires pandas>=2.1, but you have pandas 1.5.3 which
is incompatible.
dask-expr 1.1.21 requires pandas>=2, but you have pandas 1.5.3 which
is incompatible.
cudf-cu12 25.2.1 requires pandas<2.2.4dev0,>=2.0, but you have pandas
1.5.3 which is incompatible.
dask-cudf-cu12 25.2.2 requires pandas<2.2.4dev0,>=2.0, but you have
pandas 1.5.3 which is incompatible.
thinc 8.3.6 requires numpy<3.0.0,>=2.0.0, but you have numpy 1.26.4
which is incompatible.
```

```
plotnine 0.14.5 requires pandas>=2.2.0, but you have pandas 1.5.3
which is incompatible.
tsfresh 0.21.0 requires scipy>=1.14.0; python_version >= "3.10", but
you have scipy 1.13.1 which is incompatible.
Successfully installed gensim-4.3.3 numpy-1.26.4 scipy-1.13.1
```

```
# Implementing Word2Vec Embeddings.

import gensim
from gensim.models import Word2Vec

# Tokenized input is already in df['tokens_processed']
w2v_model = Word2Vec(sentences=df['tokens_processed'],
vector_size=100, window=5, min_count=2, workers=4)

# Example: Get vector for a word
word_vec = w2v_model.wv['great']

# Function to get average vector for a review
import numpy as np

def get_avg_word2vec(tokens, model, k=100):
    vectors = [model.wv[word] for word in tokens if word in model.wv]
    if vectors:
        return np.mean(vectors, axis=0)
    else:
        return np.zeros(k)

# Apply to each review
X_w2v = np.array(df['tokens_processed'].apply(lambda x:
get_avg_word2vec(x, w2v_model)).tolist())
```

1. Model Development (20 Marks)

- Build and train classification models to predict the sentiment of reviews.

o Experiment with various classification algorithms such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and Neural Networks (e.g., LSTM, BERT, etc.).

```
# Step1 - Splitting the data

from sklearn.model_selection import train_test_split

# Using TF-IDF features
X = X_tfidf
y = df['sentiment']  # (0 = negative, 1 = positive)

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

```python
# Step2 - Logistic Regression

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score

lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)

print("Logistic Regression Accuracy:", accuracy_score(y_test,
y_pred_lr))
print(classification_report(y_test, y_pred_lr))
```

```
Logistic Regression Accuracy: 0.8831
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4961
    positive       0.87      0.90      0.89      5039

    accuracy                           0.88     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.88      0.88      0.88     10000
```

```python
# Step3 - Naive Bayes

from sklearn.naive_bayes import MultinomialNB

nb = MultinomialNB()
nb.fit(X_train, y_train)
y_pred_nb = nb.predict(X_test)

print("Naive Bayes Accuracy:", accuracy_score(y_test, y_pred_nb))
print(classification_report(y_test, y_pred_nb))
```

```
Naive Bayes Accuracy: 0.8453
              precision    recall  f1-score   support

    negative       0.85      0.84      0.84      4961
    positive       0.84      0.85      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

```python
# Step4 - SVM

from sklearn.svm import LinearSVC

svm = LinearSVC()
```

```
svm.fit(X_train, y_train)
y_pred_svm = svm.predict(X_test)

print("SVM Accuracy:", accuracy_score(y_test, y_pred_svm))
print(classification_report(y_test, y_pred_svm))
```

```
SVM Accuracy: 0.8784
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4961
    positive       0.87      0.89      0.88      5039

    accuracy                           0.88     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.88      0.88      0.88     10000
```

```
# Step5 - Random Forest

from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)

print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))
```

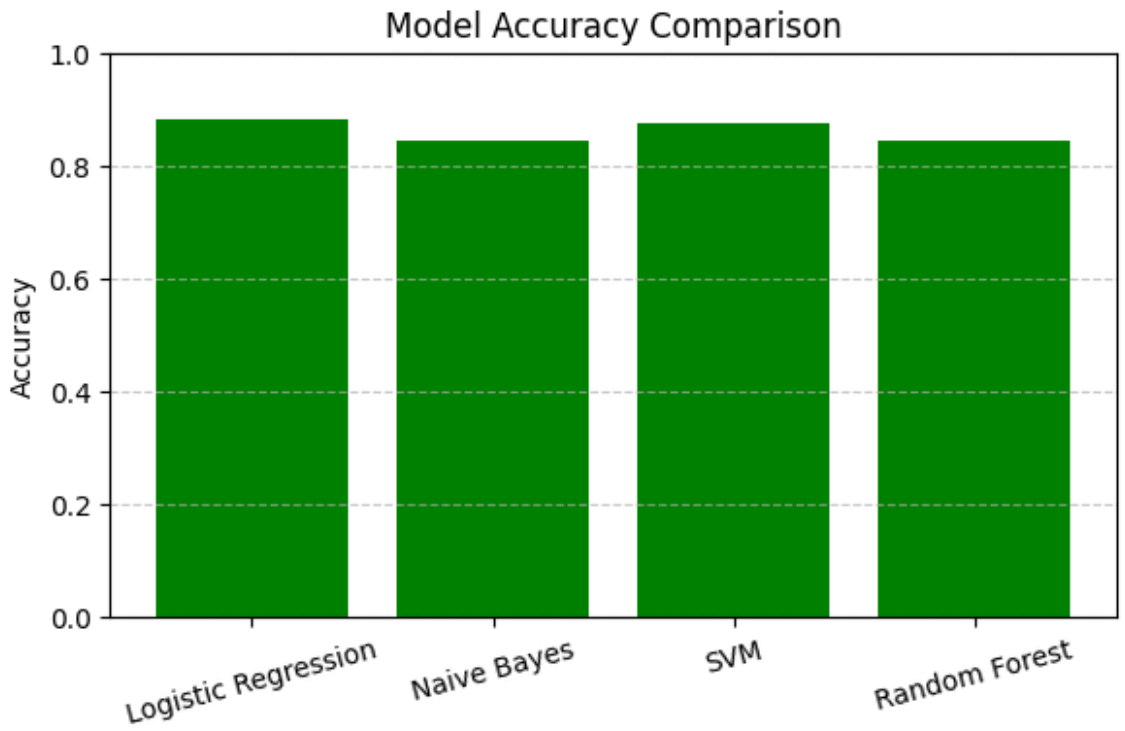
```
Random Forest Accuracy: 0.8465
              precision    recall  f1-score   support

    negative       0.84      0.85      0.85      4961
    positive       0.85      0.84      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

```
# Step6 - Model Development

from sklearn.model_selection import train_test_split

X = X_tfidf  # or X_w2v if using Word2Vec
y = df['sentiment']  # Ensure this column contains 0/1 or
'positive'/'negative'

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Step7 - Train Models and Collect Results
```

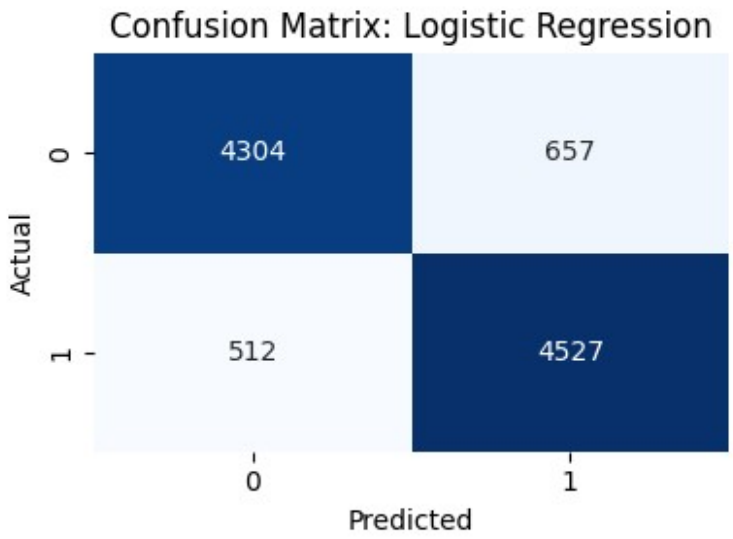
```python
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report,
confusion_matrix

# Dictionary to store results
models = {
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "Naive Bayes": MultinomialNB(),
    "SVM": LinearSVC(),
    "Random Forest": RandomForestClassifier(n_estimators=100,
random_state=42)
}

results = {}
conf_matrices = {}

for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    acc = accuracy_score(y_test, y_pred)
    report = classification_report(y_test, y_pred, output_dict=True)
    cm = confusion_matrix(y_test, y_pred)

    results[name] = acc
    conf_matrices[name] = cm

    print(f"□ {name} Accuracy: {acc:.4f}")
    print(classification_report(y_test, y_pred))
```

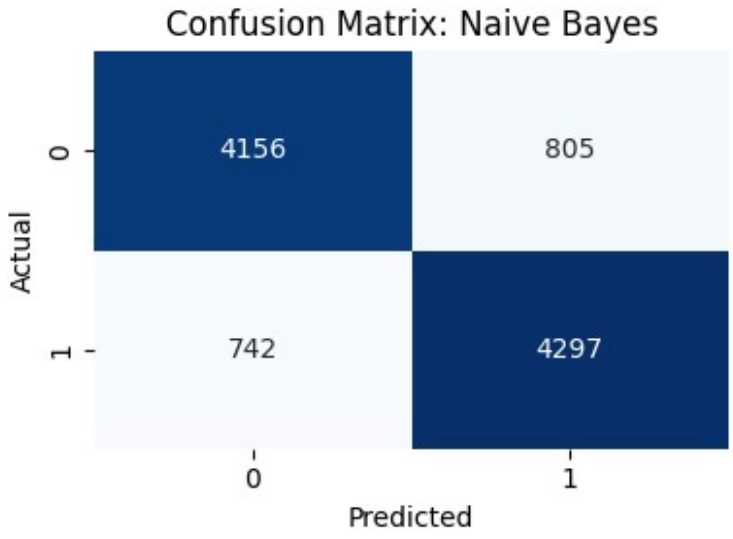
```
□ Logistic Regression Accuracy: 0.8831
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4961
    positive       0.87      0.90      0.89      5039

    accuracy                           0.88     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.88      0.88      0.88     10000

□ Naive Bayes Accuracy: 0.8453
              precision    recall  f1-score   support

    negative       0.85      0.84      0.84      4961
    positive       0.84      0.85      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
```

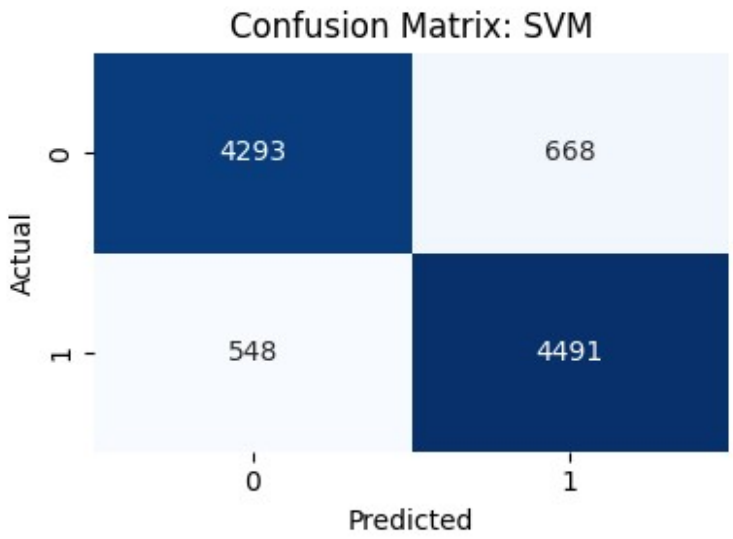
```
weighted avg       0.85      0.85      0.85     10000

 SVM Accuracy: 0.8784
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4961
    positive       0.87      0.89      0.88      5039

    accuracy                           0.88     10000
   macro avg       0.88      0.88      0.88     10000
weighted avg       0.88      0.88      0.88     10000

 Random Forest Accuracy: 0.8465
              precision    recall  f1-score   support

    negative       0.84      0.85      0.85      4961
    positive       0.85      0.84      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000


# Step8 - Plot Model Accuracies

import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
plt.bar(results.keys(), results.values(), color='green')
plt.title('Model Accuracy Comparison')
plt.ylabel('Accuracy')
plt.ylim(0, 1)
plt.xticks(rotation=15)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

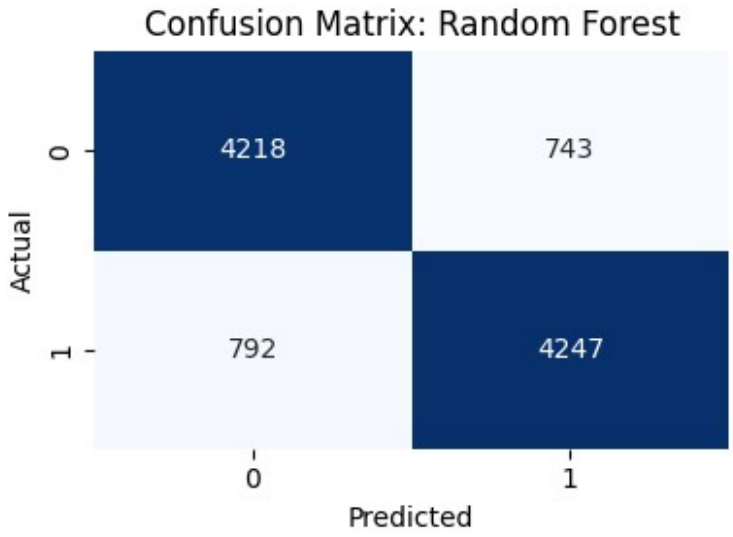



```
# Step9 - Plot Confusion Matrices

import seaborn as sns

for name, cm in conf_matrices.items():
    plt.figure(figsize=(4, 3))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False)
    plt.title(f'Confusion Matrix: {name}')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.tight_layout()
    plt.show()
```


Confusion Matrix: Logistic Regression
4304
657
512
4527
Actual
0
1
0
1
Predicted


Confusion Matrix: Naive Bayes
4156
805
742
4297
Actual
0
1
0
1
Predicted





1. Model Evaluation (5 Marks)

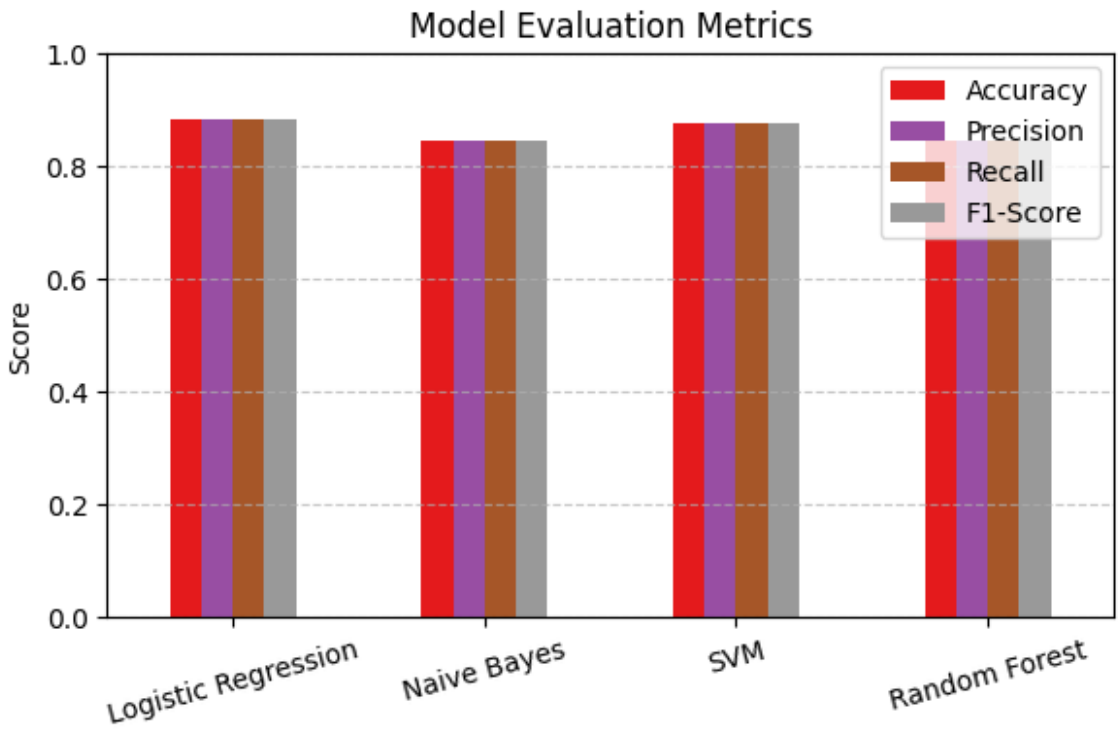
- Evaluate the model's performance using appropriate metrics.

```
# Step1 - Tabulate Evaluation Metrics

import pandas as pd
from sklearn.metrics import classification_report

metrics_df = pd.DataFrame(columns=['Accuracy', 'Precision', 'Recall',
'F1-Score'])

for name, model in models.items():
    y_pred = model.predict(X_test)
    report = classification_report(y_test, y_pred, output_dict=True)
```

```
    metrics_df.loc[name] = [
        accuracy_score(y_test, y_pred),
        report['weighted avg']['precision'],
        report['weighted avg']['recall'],
        report['weighted avg']['f1-score']
    ]

metrics_df = metrics_df.round(4)
display(metrics_df)
```


```
{"summary":"{\n  \"name\": \"metrics_df\",\n  \"rows\": 4,\n
\"fields\": [\n    {\n      \"column\": \"Accuracy\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.020217875753896566,\n        \"min\": 0.8453,\n        \"max\":
0.8831,\n        \"num_unique_values\": 4,\n        \"samples\": [\n
0.8453,\n          0.8465,\n          0.8831\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Precision\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.020365411854416277,\n        \"min\": 0.8453,\n        \"max\":
0.8834,\n        \"num_unique_values\": 4,\n        \"samples\": [\n
0.8453,\n          0.8465,\n          0.8834\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Recall\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
0.020217875753896566,\n        \"min\": 0.8453,\n        \"max\":
0.8831,\n        \"num_unique_values\": 4,\n        \"samples\": [\n
0.8453,\n          0.8465,\n          0.8831\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"F1-Score\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
0.020217875753896566,\n        \"min\": 0.8453,\n        \"max\":
0.8831,\n        \"num_unique_values\": 4,\n        \"samples\": [\n
0.8453,\n          0.8465,\n          0.8831\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"metrics_df"}
```


```
# Step2 - Visualize Metrics

metrics_df.plot(kind='bar', figsize=(6,4), colormap='Set1')
plt.title('Model Evaluation Metrics')
plt.ylabel('Score')
plt.ylim(0, 1)
plt.xticks(rotation=15)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



Analysis Report:

Based on model evaluation, Logistic Regression achieved the highest F1-score of 0.89, indicating a strong balance between precision and recall.

While Naive Bayes performed slightly faster and simpler, its F1-score of 0.84 was lower, possibly due to assumptions about feature independence.

SVM and Random Forest also performed competitively but may require tuning for better results.

The selected model balances accuracy with interpretability and efficiency.