In [ ]:
```
import pandas as pd
import numpy as np
import re
from langdetect import detect  # You might need to install langdetect using: pip in
from bs4 import BeautifulSoup  # You might need to install beautifulsoup4 using: pi

# Load the CSV file into a pandas DataFrame
df = pd.read_csv('all_comments.csv')
df.head()
```

Out[ ]:

	cid	text	time	author	
0	UgxbMfXzzFg0PIi7Yh54AaABAg	2021 still rewatching AG vlog 😘	3 years ago	@rogervsumalbag3042	UCTibHKJ6L
1	UgwQV35hNOeyHcPgKv54AaABAg	This is what i love with Alex Travel Vlog, hin...	3 years ago	@deyowzah	UCSUdzrtc08Z
2	UgzrhwnuCddIDn6w5VJ4AaABAg	Sana mayroong "Adulting 101 with Toni Gonzaga"...	5 years ago	@abanta7657	UCjsrbET
3	UgwPo4OPXUxGcsa8-Fd4AaABAg	Watching this in 2022. Ate Alex you're one of ...	1 year ago	@kim_1088	UCWsSMgh6oD
4	Ugx5alsNeN9-E-g1Qwd4AaABAg	Grabe yung tears of joy ni Uncle Jojo, halatan...	5 years ago	@eloisajenespina2852	UCFTtwd8U

In [ ]:
```
# 1. Text formatting
df['text'] = df['text'].apply(lambda x: x.lower())  # Convert text to lowercase, fo
df.head()
```

Out[ ]:

	cid	text	time	author	
0	UgxbMfXzzFg0Pli7Yh54AaABAg	2021 still rewatching ag vlog 😘	3 years ago	@rogervsumalbag3042	UCTibHKJ6L
1	UgwQV35hNOeyHcPgKv54AaABAg	this is what i love with alex travel vlog, hin...	3 years ago	@deyowzah	UCSUdzrtc08Z
2	UgzrhwnuCddIDn6w5VJ4AaABAg	sana mayroong "adulting 101 with toni gonzaga"...	5 years ago	@abanta7657	UCjsrbETj
3	UgwPo4OPXUxGcsa8-Fd4AaABAg	watching this in 2022. ate alex you're one of ...	1 year ago	@kim_1088	UCWsSMgh6oD
4	Ugx5alsNeN9-E-g1Qwd4AaABAg	grabe yung tears of joy ni uncle jojo, halatan...	5 years ago	@eloisajenespina2852	UCFTtwd8U(

In [ ]:

```
# 2. Removing emojis from comments
def remove_emojis(text):
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # Emoticons
                           u"\U0001F300-\U0001F5FF"  # Symbols & Pictographs
                           u"\U0001F680-\U0001F6FF"  # Transport & Map Symbols
                           u"\U0001F700-\U0001F77F"  # Alphanumeric Supplement
                           u"\U0001F780-\U0001F7FF"  # Geometric Shapes Extended
                           u"\U0001F800-\U0001F8FF"  # Supplemental Arrows-C
                           u"\U0001F900-\U0001F9FF"  # Supplemental Symbols and Pic
                           u"\U0001FA00-\U0001FA6F"  # Chess Symbols
                           u"\U0001FA70-\U0001FAFF"  # Symbols and Pictographs Exte
                           u"\U00002702-\U000027B0"  # Dingbats
                           u"\U000024C2-\U0001F251"
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

df['text'] = df['text'].apply(remove_emojis)
df.head()
```

Out[ ]:

	cid	text	time	author	
0	UgxbMfXzzFg0PIi7Yh54AaABAg	2021 still rewatching ag vlog	3 years ago	@rogervsumalbag3042	UCTibHKJ6L
1	UgwQV35hNOeyHcPgKv54AaABAg	this is what i love with alex travel vlog, hin...	3 years ago	@deyowzah	UCSUdzrtc08Z
2	UgzrhwnuCddIDn6w5VJ4AaABAg	sana mayroong "adulting 101 with toni gonzaga"...	5 years ago	@abanta7657	UCjsrbETj
3	UgwPo4OPXUxGcsa8-Fd4AaABAg	watching this in 2022. ate alex you're one of ...	1 year ago	@kim_1088	UCWsSMgh6oD
4	Ugx5alsNeN9-E-g1Qwd4AaABAg	grabe yung tears of joy ni uncle jojo, halatan...	5 years ago	@eloisajenespina2852	UCFTtwd8U(

In [ ]:

```
# 3. Removing languages other than English
def filter_english(text):
    try:
        if detect(text) == 'en':
            return text
        else:
            return ''
    except:
        return ''

df['text'] = df['text'].apply(filter_english)
df.head()
```

Out[ ]:

	cid	text	time	author	
0	UgxbMfXzzFg0PIi7Yh54AaABAg		3 years ago	@rogervsumalbag3042	UCTibHKJ6L
1	UgwQV35hNOeyHcPgKv54AaABAg		3 years ago	@deyowzah	UCSUdzrtc08Z
2	UgzrhwnuCddIDn6w5VJ4AaABAg	sana mayroong "adulting 101 with toni gonzaga"...	5 years ago	@abanta7657	UCjsrbETj
3	UgwPo4OPXUxGcsa8-Fd4AaABAg	watching this in 2022. ate alex you're one of ...	1 year ago	@kim_1088	UCWsSMgh6oD
4	Ugx5alsNeN9-E-g1Qwd4AaABAg		5 years ago	@eloisajenespina2852	UCFTtwd8U(

In [ ]:
```
import pandas as pd
from bs4 import BeautifulSoup  # Make sure to import BeautifulSoup if not already d
import numpy as np

# Assuming df is your original DataFrame with 'text' and 'votes' columns
df['text'] = df['text'].apply(lambda x: BeautifulSoup(x, 'html.parser').get_text())

# Remove repeated comments
df.drop_duplicates(subset='text', inplace=True)

# Remove NaN values
df.replace('', np.nan, inplace=True)
df.dropna(subset=['text'], inplace=True)

# Save the cleaned data to a new CSV file
df[['text', 'votes']].to_csv('cleaned_youtube_comments.csv', index=False)
```

```
C:\Users\etash\AppData\Local\Temp\ipykernel_7656\3447952269.py:6: MarkupResemblesLoc
atorWarning: The input looks more like a filename than markup. You may want to open
this file and pass the filehandle into Beautiful Soup.
  df['text'] = df['text'].apply(lambda x: BeautifulSoup(x, 'html.parser').get_text
())
```

In [ ]:
```
new_df = pd.read_csv('cleaned_youtube_comments.csv')
new_df.head()
```

Out[ ]:

	text	votes
0	sana mayroong "adulting 101 with toni gonzaga"...	593
1	watching this in 2022. ate alex you're one of ...	7
2	uncle jojo and ate alex = perfect match by blood	381
3	thank youuuu for keeping your videos filled wi...	41
4	back to the days when u don't need to wear fac...	59

In [ ]:
```
##Inserting data into Mongodb##
from pymongo import MongoClient

# Connect to MongoDB (Make sure MongoDB is running on your machine or specify the c
client = MongoClient('localhost', 27017)
db = client['SMA']  # Replace 'your_database_name' with your desired database name
collection = db['Experiments']  # Replace 'your_collection_name' with your desired
```

In [ ]:
```
import pandas as pd

# Load the cleaned data CSV
cleaned_df = pd.read_csv('cleaned_youtube_comments.csv')

# Convert the DataFrame to a list of dictionaries (each row becomes a dictionary)
cleaned_data_list = cleaned_df.to_dict(orient='records')

# Insert the data into MongoDB
collection.insert_many(cleaned_data_list)
```

In [ ]:
```
# Query data from MongoDB (limit to 20 records)
cursor = collection.find().limit(20)

# Display the results
for document in cursor:
    print(document)

# Close the MongoDB connection
client.close()
```

{'_id': ObjectId('65ddb4cc203ee0aa218d638f'), 'text': 'sana mayroong "adulting 101 with toni gonzaga" she\'s full of wisdom that needs to be heard by many '}
{'_id': ObjectId('65ddb4cc203ee0aa218d6390'), 'text': "watching this in 2022. ate alex you're one of my stress reliever po thank you ur humor is everything "}
{'_id': ObjectId('65ddb4cc203ee0aa218d6391'), 'text': 'uncle jojo and ate alex = perfect match by blood '}
{'_id': ObjectId('65ddb4cc203ee0aa218d6392'), 'text': 'thank youuuu for keeping your videos filled with high and joyous spirit!!! we owe you a good recognition. you make us proud ate alexxxxx '}
{'_id': ObjectId('65ddb4cc203ee0aa218d6393'), 'text': "back to the days when u don't need to wear face mask"}
{'_id': ObjectId('65ddb4cc203ee0aa218d6394'), 'text': 'laughtrip ako kay uncle jojo talaga pati kay alex. more vlog with uncle jojo pa po hahaha! i love mummy pinty din '}
{'_id': ObjectId('65ddb4cc203ee0aa218d6395'), 'text': 'okay this inspired me to travel the world. thanks ate alex!!! the best!!!'}
{'_id': ObjectId('65ddb4cc203ee0aa218d6396'), 'text': "2022 & still rewatching ag's travel vlogs "}
{'_id': ObjectId('65ddb4cc203ee0aa218d6397'), 'text': 'hi alex i’m a big big fan of yours. i love your vlog stress reliever .continue sharing your blessings making people happy god bless you and your family.'}
{'_id': ObjectId('65ddb4cc203ee0aa218d6398'), 'text': 'love to bits your uncle jojo! he’s just so cool and i know he’s fun to be with! warmest hi from '}
{'_id': ObjectId('65ddb4cc203ee0aa218d6399'), 'text': 'cant say no with ate toni ! like if you agree'}
{'_id': ObjectId('65ddb4cc203ee0aa218d639a'), 'text': 'i really love her travel vlogs especially her trips with the fam.'}
{'_id': ObjectId('65ddb4cc203ee0aa218d639b'), 'text': 'ann custodio 4th impact songs'}
{'_id': ObjectId('65ddb4cc203ee0aa218d639c'), 'text': "wow you're blessed alex ,for having such a wonderful family"}
{'_id': ObjectId('65ddb4cc203ee0aa218d639d'), 'text': "just three things:\n1. alex gonzaga is the coolest vlogger i have ever encountered on youtube. everything is raw when it comes to her vlogs. it's all natural.\n2. 2m followers? just wow! she deserves it! :)\n3. i will never get tired of watching her vlogs. doing so keeps me from a lot of things in the real world. when i am here on youtube and watching her vlogs i have seen some of them many times), i still get the same feeling of happiness."}
{'_id': ObjectId('65ddb4cc203ee0aa218d639e'), 'text': "i super loveeeeeee your vlogs ate alex keep that attitude of your's you are really a stress reliever.. girl pride ng pilipinas keep inspiring and god bless.."}
{'_id': ObjectId('65ddb4cc203ee0aa218d639f'), 'text': '2022 still rewatching your vlogs!!! love you ag! '}
{'_id': ObjectId('65ddb4cc203ee0aa218d63a0'), 'text': 'still watching 2019. nakakawala ng stress si alex and mommy pinty :*godbless always more vlog'}
{'_id': ObjectId('65ddb4cc203ee0aa218d63a1'), 'text': 'i need alex in my life. she never loses energy! '}
{'_id': ObjectId('65ddb4cc203ee0aa218d63a2'), 'text': 'winner tong vlog na to kasama lahat ng favorites ko from mommy pinty to uncle jojo '}