llm-detect-ai-generated-text

January 23, 2024

Certainly! Here's a sequential roadmap to guide you through the coding process for building a model to identify essays written by students versus those generated by large language models (LLMs):

Import necessary libraries

[1]:
```
# Import necessary libraries
import pandas as pd  # For data manipulation and analysis
import matplotlib.pyplot as plt  # For data visualization
from wordcloud import WordCloud  # For generating word clouds

# Additional libraries you might need:
# import seaborn as sns  # For advanced data visualization
# import numpy as np  # For numerical operations
```

load Dataset

[2]:
```
# Load the dataset into a pandas DataFrame
df = pd.read_csv('sample_submission[1].csv')  # Sample_submission data
df = pd.read_csv('test_essays[1].csv')   # load a test_Essays data
df = pd.read_csv('train_essays[1].csv')  # load a train_Essays data
df = pd.read_csv('train_prompts[1].csv')   # load a train_prompts data
```

[3]:
```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2 entries, 0 to 1
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   prompt_id     2 non-null      int64 
 1   prompt_name   2 non-null      object
 2   instructions  2 non-null      object
 3   source_text   2 non-null      object
dtypes: int64(1), object(3)
memory usage: 192.0+ bytes
None
```

[4]:
```
print(df.head())
```

```
   prompt_id                      prompt_name  \
0          0                   Car-free cities
1          1  Does the electoral college work?

                                        instructions  \
0  Write an explanatory essay to inform fellow ci…
1  Write a letter to your state senator in which …

                                         source_text
0  # In German Suburb, Life Goes On Without Cars …
1  # What Is the Electoral College? by the Office…
```

[5]:
```
print(df.describe())
```

```
       prompt_id
count   2.000000
mean    0.500000
std     0.707107
min     0.000000
25%     0.250000
50%     0.500000
75%     0.750000
max     1.000000
```

[6]:
```
# Explore features and distribution of classes
print("\nSummary statistics:")
print(df.describe())
```

```
Summary statistics:
       prompt_id
count   2.000000
mean    0.500000
std     0.707107
min     0.000000
25%     0.250000
50%     0.500000
75%     0.750000
max     1.000000
```

[7]:
```
# Check the distribution of classes (assuming 'class' is the target variable)
print("\nClass distribution:")
print(df['prompt_id'].value_counts())
```

```
Class distribution:
0    1
1    1
```

```
Name: prompt_id, dtype: int64
```

[8]:
```
# Check the distribution of classes (assuming 'class' is the target variable)
print("\nClass distribution:")
print(df['instructions'].value_counts())
```

```
Class distribution:
Write an explanatory essay to inform fellow citizens about the advantages of
limiting car usage. Your essay must be based on ideas and information that can
be found in the passage set. Manage your time carefully so that you can read the
passages; plan your response; write your response; and revise and edit your
response. Be sure to use evidence from multiple sources; and avoid overly
relying on one source. Your response should be in the form of a multiparagraph
essay. Write your essay in the space provided.
1
Write a letter to your state senator in which you argue in favor of keeping the
Electoral College or changing to election by popular vote for the president of
the United States. Use the information from the texts in your essay. Manage your
time carefully so that you can read the passages; plan your response; write your
response; and revise and edit your response. Be sure to include a claim; address
counterclaims; use evidence from multiple sources; and avoid overly relying on
one source. Your response should be in the form of a multiparagraph essay. Write
your response in the space provided.    1
Name: instructions, dtype: int64
```

[10]:
```
# Print column names
print("\nColumn names:")
print(df.columns)
```

```
Column names:
Index(['prompt_id', 'prompt_name', 'instructions', 'source_text'],
dtype='object')
```

[11]:
```
print("\nClass distribution:")
print(df['prompt_id'].value_counts())
```

```
Class distribution:
0    1
1    1
Name: prompt_id, dtype: int64
```

Visualize sample essays

[12]:
```
# Visualize sample essays
print("\prompt_name")
```

\prompt_name

[13]:
```
# Print column names
print("\nColumn names:")
print(df.columns)
```

```
Column names:
Index(['prompt_id', 'prompt_name', 'instructions', 'source_text'],
dtype='object')
```

[14]:
```
#source_text
for i, source_text in enumerate(df['source_text'].sample(3, replace=True)):
    print(f"\nsource_text {i+1}:\n{source_text}")
```

```
source_text 1:
# In German Suburb, Life Goes On Without Cars by Elisabeth Rosenthal

1 VAUBAN, Germany-Residents of this upscale community are suburban pioneers,
going where few soccer moms or commuting executives have ever gone before: they
have given up their cars.

2 Street parking, driveways and home garages are generally forbidden in this
experimental new district on the outskirts of Freiburg, near the French and
Swiss borders. Vauban's streets are completely "car-free"-except the main
thoroughfare, where the tram to downtown Freiburg runs, and a few streets on one
edge of the community. Car ownership is allowed, but there are only two places
to park-large garages at the edge of the development, where a car-owner buys a
space, for $40,000, along with a home.

3 As a result, 70 percent of Vauban's families do not own cars, and 57 percent
sold a car to move here. "When I had a car I was always tense. I'm much happier
this way," said Heidrun Walter, a media trainer and mother of two, as she walked
verdant streets where the swish of bicycles and the chatter of wandering
children drown out the occasional distant motor.

4 Vauban, completed in 2006, is an example of a growing trend in Europe, the
United States and elsewhere to separate suburban life from auto use, as a
component of a movement called "smart planning."

5 Automobiles are the linchpin of suburbs, where middle-class families from
Chicago to Shanghai tend to make their homes. And that, experts say, is a huge
impediment to current efforts to drastically reduce greenhouse gas emissions
from tailpipes . . . . Passenger cars are responsible for 12 percent of
greenhouse gas emissions in Europe . . . and up to 50 percent in some car-
intensive areas in the United States.
```

6 While there have been efforts in the past two decades to make cities denser, and better for walking, planners are now taking the concept to the suburbs . . . . Vauban, home to 5,500 residents within a rectangular square mile, may be the most advanced experiment in low-car suburban life. But its basic precepts are being adopted around the world in attempts to make suburbs more compact and more accessible to public transportation, with less space for parking. In this new approach, stores are placed a walk away, on a main street, rather than in malls along some distant highway.

7 "All of our development since World War II has been centered on the car, and that will have to change," said David Goldberg, an official of Transportation for America, a fast-growing coalition of hundreds of groups in the United States . . . who are promoting new communities that are less dependent on cars. Mr. Goldberg added: "How much you drive is as important as whether you have a hybrid."

8 Levittown and Scarsdale, New York suburbs with spread-out homes and private garages, were the dream towns of the 1950s and still exert a strong appeal. But some new suburbs may well look more Vauban-like, not only in developed countries but also in the developing world, where emissions from an increasing number of private cars owned by the burgeoning middle class are choking cities.

9 In the United States, the Environmental Protection Agency is promoting "car reduced" communities, and legislators are starting to act, if cautiously. Many experts expect public transport serving suburbs to play a much larger role in a new six-year federal transportation bill to be approved this year, Mr. Goldberg said. In previous bills, 80 percent of appropriations have by law gone to highways and only 20 percent to other transport.

Paris bans driving due to smog by Robert Duffer

10 After days of near-record pollution, Paris enforced a partial driving ban to clear the air of the global city.

11 On Monday motorists with even-numbered license plates were ordered to leave their cars at home or suffer a 22-euro fine ($31). The same would apply to odd-numbered plates the following day.

12 Almost 4,000 drivers were fined, according to Reuters1 . . . [Twenty-seven] people had their cars impounded for their reaction to the fine.

13 That’s easier to imagine than a car-free Champs-Elysees.2

14 Congestion 3 was down 60 percent in the capital of France, after five-days of intensifying smog . . . [The smog] rivaled Beijing, China, which is known as one of the most polluted cities in the world.

15 Cold nights and warm days caused the warmer layer of air to trap car emissions.

16 Diesel fuel was blamed, since France has . . . [a] tax policy that favors diesel over gasoline. Diesels make up 67 percent of vehicles in France, compared to a 53.3 percent average of diesel engines in the rest of Western Europe, according to Reuters.

17 Paris typically has more smog than other European capitals . . . [Last] week Paris had 147 micrograms of particulate matter (PM) per cubic meter compared with 114 in Brussels and 79.7 in London, Reuters found.

18 Delivery companies complained of lost revenue, while exceptions were made for plug-in cars, hybrids, and cars carrying three or more passengers. Public transit was free of charge from Friday to Monday, according to the BBC.

19 The smog cleared enough Monday for the ruling French party to rescind the ban for oddnumbered plates on Tuesday. 1

Car-free day is spinning into a big hit in Bogota by Andrew Selsky

BOGOTA, Colombia-In a program that's set to spread to other countries, millions of Colombians hiked, biked, skated or took buses to work during a car-free day yesterday, leaving the streets of this capital city eerily devoid of traffic jams.

21 It was the third straight year cars have been banned with only buses and taxis permitted for the Day Without Cars in this capital city of 7 million. The goal is to promote alternative transportation and reduce smog. Violators faced $25 fines.

22 The turnout was large, despite gray clouds that dumped occasional rain showers on Bogota.

23 "The rain hasn't stopped people from participating," said Bogota Mayor Antanas Mockus . . . .

24 "It's a good opportunity to take away stress and lower air pollution," said businessman Carlos Arturo Plaza as he rode a two-seat bicycle with his wife.

25 For the first time, two other Colombian cities, Cali and Valledupar, joined the event.

26 Municipal authorities from other countries came to Bogota to see the event and were enthusiastic. "These people are generating a revolutionary change, and this is crossing borders," said Enrique Riera, the mayor of Asunción, Paraguay. . . .

27 The day without cars is part of an improvement campaign that began in Bogota in the mid1990s. It has seen the construction of 118 miles of bicycle paths, the most of any Latin American city, according to Mockus, the city's mayor.

28 Parks and sports centers also have bloomed throughout the city; uneven, pitted sidewalks have been replaced by broad, smooth sidewalks; rush-hour restrictions have dramatically cut traffic; and new restaurants and upscale shopping districts have cropped up.

The End of Car Culture by Elisabeth Rosenthal

29 President Obama's ambitious goals to curb the United States' greenhouse gas emissions, unveiled last week, will get a fortuitous assist from an incipient1 shift in American behavior: recent studies suggest that Americans are buying fewer cars, driving less and getting fewer licenses as each year goes by.

30 That has left researchers pondering a fundamental question: Has America passed peak driving?

31 The United States, with its broad expanses and suburban ideals, had long been one of the world's prime car cultures. It is the birthplace of the Model T; the home of Detroit; the place where Wilson Pickett immortalized "Mustang Sally" . . . .

32 But America's love affair with its vehicles seems to be cooling. When adjusted for population growth, the number of miles driven in the United States peaked in 2005 and dropped steadily thereafter, according to an analysis by Doug Short of Advisor Perspectives, an investment research company. As of April 2013, the number of miles driven per person was nearly 9 percent below the peak and equal to where the country was in January 1995. Part of the explanation certainly lies in the recession, because cash-strapped Americans could not afford new cars, and the unemployed weren't going to work anyway. But by many measures the decrease in driving preceded the downturn and appears to be persisting now that recovery is under way. The next few years will be telling.

33 "What most intrigues me is that rates of car ownership per household and per person started to come down two to three years before the downturn," said Michael Sivak, who studies the trend and who is a research professor at the University of Michigan's Transportation Research Institute. "I think that means something more fundamental is going on."

34 If the pattern persists-and many sociologists believe it will-it will have beneficial implications for carbon emissions and the environment, since transportation is the second largest source of America's emissions, just behind power plants. But it could have negative implications for the car industry. Indeed, companies like Ford and Mercedes are already rebranding themselves "mobility" companies with a broader product range beyond the personal vehicle.

35 "Different things are converging which suggest that we are witnessing a long-term cultural shift," said Mimi Sheller, a sociology professor at Drexel University and director of its Mobilities Research and Policy Center. She cites various factors: the Internet makes telecommuting possible and allows people to feel more connected without driving to meet friends. The renewal of center cities has made the suburbs less appealing and has drawn empty nesters back in. Likewise the rise in cellphones and car-pooling apps has facilitated more flexible commuting arrangements, including the evolution of shared van services for getting to work.

36 With all these changes, people who stopped car commuting as a result of the recession may find less reason to resume the habit. . . .

37 New York's new bike-sharing program and its skyrocketing bridge and tunnel tolls reflect those new priorities, as do a proliferation of car-sharing programs across the nation.

38 Demographic shifts in the driving population suggest that the trend may accelerate. There has been a large drop in the percentage of 16- to 39-year-olds getting a license, while older people are likely to retain their licenses as they age, Mr. Sivak's research has found.

39 He and I have similar observations about our children. Mine (19 and 21) have not bothered to get a driver's license, even though they both live in places where one could come in handy. They are interested, but it's not a priority. They organize their summer jobs and social life around where they can walk or take public transportation or car-pool with friends.

40 Mr. Sivak's son lives in San Francisco and has a car but takes Bay Area Rapid Transit, when he can, even though that often takes longer than driving. "When I was in my 20s and 30s," Mr. Sivak said, "I was curious about what kind of car people drove, but young people don't really care. A car is just a means of getting from A to B when BART doesn't work."

41 A study last year found that driving by young people decreased 23 percent

between 2001 and 2009. . . .

42 Whether members of the millennial generation will start buying more cars once they have kids to take to soccer practice and school plays remains an open question. But such projections have important business implications, even if car buyers are merely older or buying fewer cars in a lifetime rather than rejecting car culture outright.

43 At the Mobile World Congress last year in Barcelona, Spain, Bill Ford, executive chairman of the Ford Motor Company, laid out a business plan for a world in which personal vehicle ownership is impractical or undesirable. He proposed partnering with the telecommunications industry to create cities in which "pedestrian, bicycle, private cars, commercial and public transportation traffic are woven into a connected network to save time, conserve resources, lower emissions and improve safety."

source_text 2:
What Is the Electoral College? by the Office of the Federal Register

1 The Electoral College is a process, not a place. The founding fathers established it in the Constitution as a compromise between election of the President by a vote in Congress and election of the President by a popular vote of qualified citizens.

2 The Electoral College process consists of the selection of the electors, the meeting of the electors where they vote for President and Vice President, and the counting of the electoral votes by Congress.

3 The Electoral College consists of 538 electors. A majority of 270 electoral votes is required to elect the President. Your state’s entitled allotment of electors equals the number of members in its Congressional delegation: one for each member in the House of Representatives plus two for your Senators. . . .

4 Under the 23rd Amendment of the Constitution, the District of Columbia is allocated 3 electors and treated like a state for purposes of the Electoral College. For this reason, in the following discussion, the word "state" also refers to the District of Columbia.

5 Each candidate running for President in your state has his or her own group of electors. The electors are generally chosen by the candidate’s political party, but state laws vary on how the electors are selected and what their responsibilities are. . . .

6 The presidential election is held every four years on the Tuesday after the first Monday in November. You help choose your state's electors when you vote for President because when you vote for your candidate you are actually voting for your candidate's electors.

7 Most states have a "winner-take-all" system that awards all electors to the winning presidential candidate. However, Maine and Nebraska each have a variation of "proportional representation." . . .

8 After the presidential election, your governor prepares a "Certificate of Ascertainment" listing all of the candidates who ran for President in your state along with the names of their respective electors. The Certificate of Ascertainment also declares the winning presidential candidate in your state and shows which electors will represent your state at the meeting of the electors in December of the election year. Your state's Certificates of Ascertainments are sent to the Congress and the National Archives as part of the official records of the presidential election.

The Indefensible Electoral College: Why even the best-laid defenses of the system are wrong by Bradford Plumer

9 What have Richard Nixon, Jimmy Carter, Bob Dole, the U.S. Chamber of Commerce, and the AFL-CIO all, in their time, agreed on? Answer: Abolishing the electoral college! They're not alone; according to a Gallup poll in 2000, taken shortly after Al Gore-thanks to the quirks of the electoral college-won the popular vote but lost the presidency,1 over 60 percent of voters would prefer a direct election to the kind we have now. This year voters can expect another close election in which the popular vote winner could again lose the presidency. And yet, the electoral college still has its defenders. What gives? . . . What's wrong with the electoral college

10 Under the electoral college system, voters vote not for the president, but for a slate of electors, who in turn elect the president. If you lived in Texas, for instance, and wanted to vote for John Kerry, you'd vote for a slate of 34 Democratic electors pledged to Kerry. On the offchance that those electors won the statewide election, they would go to Congress and Kerry would get 34 electoral votes. Who are the electors? They can be anyone not holding public office. Who picks the electors in the first place? It depends on the state. Sometimes state conventions, sometimes the state party's central committee, sometimes the presidential candidates themselves. Can voters control whom their electors vote for? Not always. Do voters sometimes get confused about the electors and vote for the wrong candidate? Sometimes.

11 The single best argument against the electoral college is what we might call the disaster factor. The American people should consider themselves lucky that the 2000 fiasco was the biggest election crisis in a century; the system allows for much worse. Consider that state legislatures are technically responsible for picking electors, and that those electors could always defy the will of the

people. Back in 1960, segregationists in the Louisiana legislature nearly succeeded in replacing the Democratic electors with new electors who would oppose John F. Kennedy. (So that a popular vote for Kennedy would not have actually gone to Kennedy.) In the same vein, "faithless" electors have occasionally refused to vote for their party's candidate and cast a deciding vote for whomever they please. . . . Oh, and what if a state sends two slates of electors to Congress? It happened in Hawaii in 1960. Luckily, Vice President Richard Nixon, who was presiding over the Senate, validated only his opponent's electors, but he made sure to do so "without establishing a precedent." What if it happened again?

12 Perhaps most worrying is the prospect of a tie in the electoral vote. In that case, the election would be thrown to the House of Representatives, where state delegations vote on the president. (The Senate would choose the vice-president.) Because each state casts only one vote, the single representative from Wyoming, representing 500,000 voters, would have as much say as the 55 representatives from California, who represent 35 million voters. Given that many voters vote one party for president and another for Congress, the House's selection can hardly be expected to reflect the will of the people. And if an electoral tie seems unlikely, consider this: In 1968, a shift of just 41,971 votes would have deadlocked the election; In 1976, a tie would have occurred if a mere 5,559 voters in Ohio and 3,687 voters in Hawaii had voted the other way. The election is only a few swing voters away from catastrophe.

13 At the most basic level, the electoral college is unfair to voters. Because of the winner-takeall system in each state, candidates don't spend time in states they know they have no chance of winning, focusing only on the tight races in the "swing" states. During the 2000 campaign, seventeen states didn't see the candidates at all, including Rhode Island and South Carolina, and voters in 25 of the largest media markets didn't get to see a single campaign ad. If anyone has a good argument for putting the fate of the presidency in the hands of a few swing voters in Ohio, they have yet to make it. . . .

14 It's official: The electoral college is unfair, outdated, and irrational. The best arguments in favor of it are mostly assertions without much basis in reality. And the arguments against direct elections are spurious at best. It's hard to say this, but Bob Dole was right: Abolish the electoral college!

In Defense of the Electoral College: Five reasons to keep our despised method of choosing the President by Judge Richard A. Posner

15 The Electoral College is widely regarded as an anachronism,1 a non-democratic method of selecting a president that ought to be overruled by declaring the candidate who receives the most popular votes the winner. The advocates of this position are correct in arguing that the Electoral College method is not democratic in a modern sense . . . it is the electors who elect the president, not the people. When you vote for a presidential candidate you're actually voting for a slate of electors.

16 But each party selects a slate of electors trusted to vote for the party's nominee (and that trust is rarely betrayed) . . . however, it is entirely possible that the winner of the electoral vote will not win the national popular vote. Yet that has happened very rarely. It happened in 2000, when Gore had more popular votes than Bush yet fewer electoral votes, but that was the first time since 1888.

17 There are five reasons for retaining the Electoral College despite its lack of democratic pedigree;2 all are practical reasons, not liberal or conservative3 reasons.

1) Certainty of Outcome

18 A dispute over the outcome of an Electoral College vote is possible---it happened in 2000---but it's less likely than a dispute over the popular vote. The reason is that the winning candidate's share of the Electoral College invariably exceeds his share of the popular vote. In 2012's election, for example, Obama4 received 61.7 percent of the electoral vote compared to only 51.3 percent of the popular votes cast for him and Romney.5 . . . Because almost all states award electoral votes on a winner-take-all basis, even a very slight plurality6 in a state creates a landslide electoral-vote victory in that state. A tie in the nationwide electoral vote is possible because the total number of votes---538---is an even number, but it is highly unlikely. . . .

2) Everyone's President

19 The Electoral College requires a presidential candidate to have trans-regional appeal. No region (South, Northeast, etc.) has enough electoral votes to elect a president. So a solid regional favorite, such as Romney was in the South, has no incentive to campaign heavily in those states, for he gains no electoral votes by increasing his plurality in states that he knows he will win. This is a desirable result because a candidate with only regional appeal is unlikely to be a successful president. The residents of the other regions are likely to feel disenfranchised-to feel that their votes do not count, that the new president will have no regard for their interests, that he really isn't their president.

3) Swing States

20 The winner-take-all method of awarding electoral votes induces the candidates-as we saw in 2012's election-to focus their campaign efforts on the toss-up states . . . . Voters in toss-up states are more likely to pay close attention to the campaign-to really listen to the competing candidates-knowing that they are going to decide the election. They are likely to be the most thoughtful voters, on average (and for the further reason that they will have received the most information and attention from the candidates), and the most thoughtful voters should be the ones to decide the election.

4) Big States

21 The Electoral College restores some of the weight in the political balance that large states (by population) lose by virtue of the mal-apportionment of the Senate decreed in the Constitution. . . . The popular vote was very close in Florida in 2012; nevertheless Obama, who won that vote, got 29 electoral votes. A victory by the same margin in Wyoming would net the winner only 3 electoral votes. So, other things being equal, a large state gets more attention from presidential candidates in a campaign than a small state does. . . .

5) Avoid Run-Off Elections

22 The Electoral College avoids the problem of elections in which no candidate receives a majority of the votes cast. For example, Nixon in 1968 and Clinton in 1992 both had only a 43 percent plurality of the popular votes, while winning a majority in the Electoral College (301 and 370 electoral votes, respectively). There is pressure for run-off elections when no candidate wins a majority of the votes cast; that pressure, which would greatly complicate the presidential election process, is reduced by the Electoral College, which invariably produces a clear winner. . . .

23 It can be argued that the Electoral College method of selecting the president may turn off potential voters for a candidate who has no hope of carrying their state-Democrats in Texas, for example, or Republicans in California. Knowing their vote will have no effect, they have less incentive to pay attention to the campaign than they would have if the president were picked by popular vote . . . . But of course no voter's vote swings a national election, and in spite of that, about one-half the eligible American population did vote in 2012's election. Voters in presidential elections are people who want to express a political preference rather than people who think that a single vote may decide an election. . . .

source_text 3:
In German Suburb, Life Goes On Without Cars by Elisabeth Rosenthal

1 VAUBAN, Germany-Residents of this upscale community are suburban pioneers, going where few soccer moms or commuting executives have ever gone before: they have given up their cars.

2 Street parking, driveways and home garages are generally forbidden in this experimental new district on the outskirts of Freiburg, near the French and Swiss borders. Vauban's streets are completely "car-free"-except the main thoroughfare, where the tram to downtown Freiburg runs, and a few streets on one edge of the community. Car ownership is allowed, but there are only two places to park-large garages at the edge of the development, where a car-owner buys a space, for $40,000, along with a home.

3 As a result, 70 percent of Vauban's families do not own cars, and 57 percent sold a car to move here. "When I had a car I was always tense. I'm much happier this way," said Heidrun Walter, a media trainer and mother of two, as she walked verdant streets where the swish of bicycles and the chatter of wandering children drown out the occasional distant motor.

4 Vauban, completed in 2006, is an example of a growing trend in Europe, the United States and elsewhere to separate suburban life from auto use, as a component of a movement called "smart planning."

5 Automobiles are the linchpin of suburbs, where middle-class families from Chicago to Shanghai tend to make their homes. And that, experts say, is a huge impediment to current efforts to drastically reduce greenhouse gas emissions from tailpipes . . . . Passenger cars are responsible for 12 percent of greenhouse gas emissions in Europe . . . and up to 50 percent in some car-intensive areas in the United States.

6 While there have been efforts in the past two decades to make cities denser, and better for walking, planners are now taking the concept to the suburbs . . . . Vauban, home to 5,500 residents within a rectangular square mile, may be the most advanced experiment in low-car suburban life. But its basic precepts are being adopted around the world in attempts to make suburbs more compact and more accessible to public transportation, with less space for parking. In this new approach, stores are placed a walk away, on a main street, rather than in malls along some distant highway.

7 "All of our development since World War II has been centered on the car, and that will have to change," said David Goldberg, an official of Transportation for America, a fast-growing coalition of hundreds of groups in the United States . . . who are promoting new communities that are less dependent on cars. Mr. Goldberg added: "How much you drive is as important as whether you have a hybrid."

8 Levittown and Scarsdale, New York suburbs with spread-out homes and private garages, were the dream towns of the 1950s and still exert a strong appeal. But some new suburbs may well look more Vauban-like, not only in developed countries but also in the developing world, where emissions from an increasing number of private cars owned by the burgeoning middle class are choking cities.

9 In the United States, the Environmental Protection Agency is promoting "car reduced" communities, and legislators are starting to act, if cautiously. Many experts expect public transport serving suburbs to play a much larger role in a new six-year federal transportation bill to be approved this year, Mr. Goldberg said. In previous bills, 80 percent of appropriations have by law gone to highways and only 20 percent to other transport.

Excerpt from "In German Suburb, Life Goes On Without Cars" by Elisabeth

Rosenthal, from the New York Times. Copyright © 2009 by the New York Times Company. Reprinted by permission of the New York Times Company via Copyright Clearance Center.

Paris bans driving due to smog by Robert Duffer

10 After days of near-record pollution, Paris enforced a partial driving ban to clear the air of the global city.

11 On Monday motorists with even-numbered license plates were ordered to leave their cars at home or suffer a 22-euro fine ($31). The same would apply to odd-numbered plates the following day.

12 Almost 4,000 drivers were fined, according to Reuters1 . . . [Twenty-seven] people had their cars impounded for their reaction to the fine.

13 That's easier to imagine than a car-free Champs-Elysees.2

14 Congestion 3 was down 60 percent in the capital of France, after five-days of intensifying smog . . . [The smog] rivaled Beijing, China, which is known as one of the most polluted cities in the world.

15 Cold nights and warm days caused the warmer layer of air to trap car emissions.

16 Diesel fuel was blamed, since France has . . . [a] tax policy that favors diesel over gasoline. Diesels make up 67 percent of vehicles in France, compared to a 53.3 percent average of diesel engines in the rest of Western Europe, according to Reuters.

17 Paris typically has more smog than other European capitals . . . [Last] week Paris had 147 micrograms of particulate matter (PM) per cubic meter compared with 114 in Brussels and 79.7 in London, Reuters found.

18 Delivery companies complained of lost revenue, while exceptions were made for plug-in cars, hybrids, and cars carrying three or more passengers. Public transit was free of charge from Friday to Monday, according to the BBC.

19 The smog cleared enough Monday for the ruling French party to rescind the ban for oddnumbered plates on Tuesday. 1

Excerpt from "Paris bans driving due to smog" by Robert Duffer, from the Chicago Tribune. Copyright © 2014 by the Chicago Tribune. Reprinted by permission of the Chicago Tribune via Copyright Clearance Center.

Car-free day is spinning into a big hit in Bogota by Andrew Selsky

BOGOTA, Colombia-In a program that's set to spread to other countries, millions

of Colombians hiked, biked, skated or took buses to work during a car-free day yesterday, leaving the streets of this capital city eerily devoid of traffic jams.

21 It was the third straight year cars have been banned with only buses and taxis permitted for the Day Without Cars in this capital city of 7 million. The goal is to promote alternative transportation and reduce smog. Violators faced $25 fines.

22 The turnout was large, despite gray clouds that dumped occasional rain showers on Bogota.

23 "The rain hasn't stopped people from participating," said Bogota Mayor Antanas Mockus . . . .

24 "It's a good opportunity to take away stress and lower air pollution," said businessman Carlos Arturo Plaza as he rode a two-seat bicycle with his wife.

25 For the first time, two other Colombian cities, Cali and Valledupar, joined the event.

26 Municipal authorities from other countries came to Bogota to see the event and were enthusiastic. "These people are generating a revolutionary change, and this is crossing borders," said Enrique Riera, the mayor of Asunción, Paraguay. . . .

27 The day without cars is part of an improvement campaign that began in Bogota in the mid1990s. It has seen the construction of 118 miles of bicycle paths, the most of any Latin American city, according to Mockus, the city's mayor.

28 Parks and sports centers also have bloomed throughout the city; uneven, pitted sidewalks have been replaced by broad, smooth sidewalks; rush-hour restrictions have dramatically cut traffic; and new restaurants and upscale shopping districts have cropped up.

The End of Car Culture by Elisabeth Rosenthal

29 President Obama's ambitious goals to curb the United States' greenhouse gas emissions, unveiled last week, will get a fortuitous assist from an incipient1 shift in American behavior: recent studies suggest that Americans are buying fewer cars, driving less and getting fewer licenses as each year goes by.

30 That has left researchers pondering a fundamental question: Has America

passed peak driving?

31 The United States, with its broad expanses and suburban ideals, had long been one of the world's prime car cultures. It is the birthplace of the Model T; the home of Detroit; the place where Wilson Pickett immortalized "Mustang Sally" . . . .

32 But America's love affair with its vehicles seems to be cooling. When adjusted for population growth, the number of miles driven in the United States peaked in 2005 and dropped steadily thereafter, according to an analysis by Doug Short of Advisor Perspectives, an investment research company. As of April 2013, the number of miles driven per person was nearly 9 percent below the peak and equal to where the country was in January 1995. Part of the explanation certainly lies in the recession, because cash-strapped Americans could not afford new cars, and the unemployed weren't going to work anyway. But by many measures the decrease in driving preceded the downturn and appears to be persisting now that recovery is under way. The next few years will be telling.

33 "What most intrigues me is that rates of car ownership per household and per person started to come down two to three years before the downturn," said Michael Sivak, who studies the trend and who is a research professor at the University of Michigan's Transportation Research Institute. "I think that means something more fundamental is going on."

34 If the pattern persists-and many sociologists believe it will-it will have beneficial implications for carbon emissions and the environment, since transportation is the second largest source of America's emissions, just behind power plants. But it could have negative implications for the car industry. Indeed, companies like Ford and Mercedes are already rebranding themselves "mobility" companies with a broader product range beyond the personal vehicle.

35 "Different things are converging which suggest that we are witnessing a long-term cultural shift," said Mimi Sheller, a sociology professor at Drexel University and director of its Mobilities Research and Policy Center. She cites various factors: the Internet makes telecommuting possible and allows people to feel more connected without driving to meet friends. The renewal of center cities has made the suburbs less appealing and has drawn empty nesters back in. Likewise the rise in cellphones and car-pooling apps has facilitated more flexible commuting arrangements, including the evolution of shared van services for getting to work.

36 With all these changes, people who stopped car commuting as a result of the recession may find less reason to resume the habit. . . .

37 New York's new bike-sharing program and its skyrocketing bridge and tunnel tolls reflect those new priorities, as do a proliferation of car-sharing programs across the nation.

38 Demographic shifts in the driving population suggest that the trend may accelerate. There has been a large drop in the percentage of 16- to 39-year-olds getting a license, while older people are likely to retain their licenses as they age, Mr. Sivak's research has found.

39 He and I have similar observations about our children. Mine (19 and 21) have not bothered to get a driver's license, even though they both live in places where one could come in handy. They are interested, but it's not a priority. They organize their summer jobs and social life around where they can walk or take public transportation or car-pool with friends.

40 Mr. Sivak's son lives in San Francisco and has a car but takes Bay Area Rapid Transit, when he can, even though that often takes longer than driving. "When I was in my 20s and 30s," Mr. Sivak said, "I was curious about what kind of car people drove, but young people don't really care. A car is just a means of getting from A to B when BART doesn't work."

41 A study last year found that driving by young people decreased 23 percent between 2001 and 2009. . . .

42 Whether members of the millennial generation will start buying more cars once they have kids to take to soccer practice and school plays remains an open question. But such projections have important business implications, even if car buyers are merely older or buying fewer cars in a lifetime rather than rejecting car culture outright.

43 At the Mobile World Congress last year in Barcelona, Spain, Bill Ford, executive chairman of the Ford Motor Company, laid out a business plan for a world in which personal vehicle ownership is impractical or undesirable. He proposed partnering with the telecommunications industry to create cities in which "pedestrian, bicycle, private cars, commercial and public transportation traffic are woven into a connected network to save time, conserve resources, lower emissions and improve safety."

[15]:
```
for i, instructions in enumerate(df['instructions'].sample(5, replace=True)):
    print(f"\ninstructions {i+1}:\n{instructions}")
```

instructions 1:
Write an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Your essay must be based on ideas and information that can be found in the passage set. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your

response. Be sure to use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your essay in the space provided.

instructions 2:
Write a letter to your state senator in which you argue in favor of keeping the Electoral College or changing to election by popular vote for the president of the United States. Use the information from the texts in your essay. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to include a claim; address counterclaims; use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your response in the space provided.

instructions 3:
Write an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Your essay must be based on ideas and information that can be found in the passage set. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your essay in the space provided.

instructions 4:
Write an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Your essay must be based on ideas and information that can be found in the passage set. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your essay in the space provided.

instructions 5:
Write an explanatory essay to inform fellow citizens about the advantages of limiting car usage. Your essay must be based on ideas and information that can be found in the passage set. Manage your time carefully so that you can read the passages; plan your response; write your response; and revise and edit your response. Be sure to use evidence from multiple sources; and avoid overly relying on one source. Your response should be in the form of a multiparagraph essay. Write your essay in the space provided.

Additional visualization (optional): Word cloud for essay content

[16]:
```
# Additional visualization (optional): Word cloud for essay content
from wordcloud import WordCloud
import matplotlib.pyplot as plt
```

[17]:
```
all_text = ' '.join(df['instructions'])
wordcloud = WordCloud(width=800, height=400, random_state=42).generate(all_text)
```

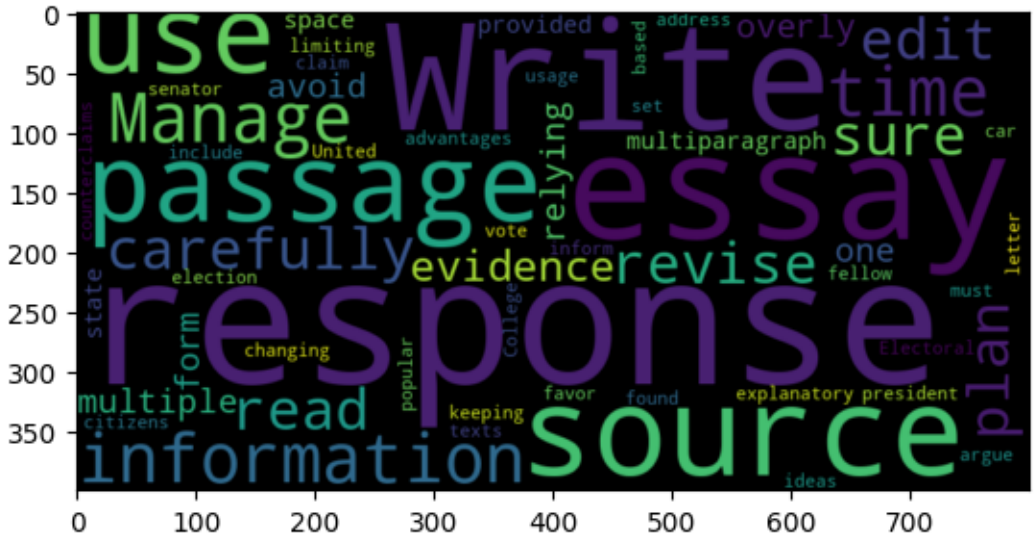
[18]:
```
plt.figure(figsize=(10, 5))
```

[18]:
```
<Figure size 1000x500 with 0 Axes>
```

```
<Figure size 1000x500 with 0 Axes>
```

[19]:
```
plt.imshow(wordcloud, interpolation='bilinear')
```

[19]:
```
<matplotlib.image.AxesImage at 0x7c14a714ebc0>
```



Data Preprocessing:

Perform necessary data preprocessing steps, including:

Text cleaning: Remove any irrelevant characters, symbols, or formatting issues.

Tokenization: Split essays into individual words or tokens.

Text normalization: Convert text to lowercase to ensure uniformity.

[20]:
```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import string
```

[21]:
```
# Download NLTK resources (comment out if already downloaded)
nltk.download('punkt')
nltk.download('stopwords')
```

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

[21]:
```
True
```

[22]:
```python
# Function for text preprocessing
def preprocess_text(text):
    # Remove punctuation and convert to lowercase
    text = text.translate(str.maketrans('', '', string.punctuation)).lower()

    # Tokenization
    tokens = word_tokenize(text)

    # Remove stop words
    stop_words = set(stopwords.words('english'))
    tokens = [token for token in tokens if token not in stop_words]

    return tokens
```

[23]:
```python
# Example usage on a single essay
example_essay = df['instructions'].sample(1).iloc[0]
preprocessed_tokens = preprocess_text(example_essay)

# Print original and preprocessed text
print("\nOriginal Essay:")
print(example_essay)

print("\nPreprocessed Tokens:")
print(preprocessed_tokens)
```

```
Original Essay:
Write an explanatory essay to inform fellow citizens about the advantages of
limiting car usage. Your essay must be based on ideas and information that can
be found in the passage set. Manage your time carefully so that you can read the
passages; plan your response; write your response; and revise and edit your
response. Be sure to use evidence from multiple sources; and avoid overly
relying on one source. Your response should be in the form of a multiparagraph
essay. Write your essay in the space provided.

Preprocessed Tokens:
['write', 'explanatory', 'essay', 'inform', 'fellow', 'citizens', 'advantages',
'limiting', 'car', 'usage', 'essay', 'must', 'based', 'ideas', 'information',
'found', 'passage', 'set', 'manage', 'time', 'carefully', 'read', 'passages',
'plan', 'response', 'write', 'response', 'revise', 'edit', 'response', 'sure',
'use', 'evidence', 'multiple', 'sources', 'avoid', 'overly', 'relying', 'one',
```

```
'source', 'response', 'form', 'multiparagraph', 'essay', 'write', 'essay',
'space', 'provided']
```

[24]:
```
# Example usage on a single essay
example_essay = df['source_text'].sample(1).iloc[0]
preprocessed_tokens = preprocess_text(example_essay)

# Print original and preprocessed text
print("\nOriginal Essay:")
print(example_essay)

print("\nPreprocessed Tokens:")
print(preprocessed_tokens)
```

```
Original Essay:
# In German Suburb, Life Goes On Without Cars by Elisabeth Rosenthal

1 VAUBAN, Germany-Residents of this upscale community are suburban pioneers,
going where few soccer moms or commuting executives have ever gone before: they
have given up their cars.

2 Street parking, driveways and home garages are generally forbidden in this
experimental new district on the outskirts of Freiburg, near the French and
Swiss borders. Vauban's streets are completely "car-free"-except the main
thoroughfare, where the tram to downtown Freiburg runs, and a few streets on one
edge of the community. Car ownership is allowed, but there are only two places
to park-large garages at the edge of the development, where a car-owner buys a
space, for $40,000, along with a home.

3 As a result, 70 percent of Vauban's families do not own cars, and 57 percent
sold a car to move here. "When I had a car I was always tense. I'm much happier
this way," said Heidrun Walter, a media trainer and mother of two, as she walked
verdant streets where the swish of bicycles and the chatter of wandering
children drown out the occasional distant motor.

4 Vauban, completed in 2006, is an example of a growing trend in Europe, the
United States and elsewhere to separate suburban life from auto use, as a
component of a movement called "smart planning."

5 Automobiles are the linchpin of suburbs, where middle-class families from
Chicago to Shanghai tend to make their homes. And that, experts say, is a huge
impediment to current efforts to drastically reduce greenhouse gas emissions
from tailpipes . . . . Passenger cars are responsible for 12 percent of
greenhouse gas emissions in Europe . . . and up to 50 percent in some car-
intensive areas in the United States.

6 While there have been efforts in the past two decades to make cities denser,
```

and better for walking, planners are now taking the concept to the suburbs . . . . Vauban, home to 5,500 residents within a rectangular square mile, may be the most advanced experiment in low-car suburban life. But its basic precepts are being adopted around the world in attempts to make suburbs more compact and more accessible to public transportation, with less space for parking. In this new approach, stores are placed a walk away, on a main street, rather than in malls along some distant highway.

7 "All of our development since World War II has been centered on the car, and that will have to change," said David Goldberg, an official of Transportation for America, a fast-growing coalition of hundreds of groups in the United States . . . who are promoting new communities that are less dependent on cars. Mr. Goldberg added: "How much you drive is as important as whether you have a hybrid."

8 Levittown and Scarsdale, New York suburbs with spread-out homes and private garages, were the dream towns of the 1950s and still exert a strong appeal. But some new suburbs may well look more Vauban-like, not only in developed countries but also in the developing world, where emissions from an increasing number of private cars owned by the burgeoning middle class are choking cities.

9 In the United States, the Environmental Protection Agency is promoting "car reduced" communities, and legislators are starting to act, if cautiously. Many experts expect public transport serving suburbs to play a much larger role in a new six-year federal transportation bill to be approved this year, Mr. Goldberg said. In previous bills, 80 percent of appropriations have by law gone to highways and only 20 percent to other transport.

Paris bans driving due to smog by Robert Duffer

10 After days of near-record pollution, Paris enforced a partial driving ban to clear the air of the global city.

11 On Monday motorists with even-numbered license plates were ordered to leave their cars at home or suffer a 22-euro fine ($31). The same would apply to odd-numbered plates the following day.

12 Almost 4,000 drivers were fined, according to Reuters1 . . . [Twenty-seven] people had their cars impounded for their reaction to the fine.

13 That's easier to imagine than a car-free Champs-Elysees.2

14 Congestion 3 was down 60 percent in the capital of France, after five-days of

intensifying smog . . . [The smog] rivaled Beijing, China, which is known as one of the most polluted cities in the world.

15 Cold nights and warm days caused the warmer layer of air to trap car emissions.

16 Diesel fuel was blamed, since France has . . . [a] tax policy that favors diesel over gasoline. Diesels make up 67 percent of vehicles in France, compared to a 53.3 percent average of diesel engines in the rest of Western Europe, according to Reuters.

17 Paris typically has more smog than other European capitals . . . [Last] week Paris had 147 micrograms of particulate matter (PM) per cubic meter compared with 114 in Brussels and 79.7 in London, Reuters found.

18 Delivery companies complained of lost revenue, while exceptions were made for plug-in cars, hybrids, and cars carrying three or more passengers. Public transit was free of charge from Friday to Monday, according to the BBC.

19 The smog cleared enough Monday for the ruling French party to rescind the ban for oddnumbered plates on Tuesday. 1

Car-free day is spinning into a big hit in Bogota by Andrew Selsky

BOGOTA, Colombia-In a program that’s set to spread to other countries, millions of Colombians hiked, biked, skated or took buses to work during a car-free day yesterday, leaving the streets of this capital city eerily devoid of traffic jams.

21 It was the third straight year cars have been banned with only buses and taxis permitted for the Day Without Cars in this capital city of 7 million. The goal is to promote alternative transportation and reduce smog. Violators faced $25 fines.

22 The turnout was large, despite gray clouds that dumped occasional rain showers on Bogota.

23 "The rain hasn’t stopped people from participating," said Bogota Mayor Antanas Mockus . . . .

24 "It’s a good opportunity to take away stress and lower air pollution," said businessman Carlos Arturo Plaza as he rode a two-seat bicycle with his wife.

25 For the first time, two other Colombian cities, Cali and Valledupar, joined

the event.

26 Municipal authorities from other countries came to Bogota to see the event and were enthusiastic. "These people are generating a revolutionary change, and this is crossing borders," said Enrique Riera, the mayor of Asunción, Paraguay. . . .

27 The day without cars is part of an improvement campaign that began in Bogota in the mid1990s. It has seen the construction of 118 miles of bicycle paths, the most of any Latin American city, according to Mockus, the city's mayor.

28 Parks and sports centers also have bloomed throughout the city; uneven, pitted sidewalks have been replaced by broad, smooth sidewalks; rush-hour restrictions have dramatically cut traffic; and new restaurants and upscale shopping districts have cropped up.

The End of Car Culture by Elisabeth Rosenthal

29 President Obama's ambitious goals to curb the United States' greenhouse gas emissions, unveiled last week, will get a fortuitous assist from an incipient1 shift in American behavior: recent studies suggest that Americans are buying fewer cars, driving less and getting fewer licenses as each year goes by.

30 That has left researchers pondering a fundamental question: Has America passed peak driving?

31 The United States, with its broad expanses and suburban ideals, had long been one of the world's prime car cultures. It is the birthplace of the Model T; the home of Detroit; the place where Wilson Pickett immortalized "Mustang Sally" . . . .

32 But America's love affair with its vehicles seems to be cooling. When adjusted for population growth, the number of miles driven in the United States peaked in 2005 and dropped steadily thereafter, according to an analysis by Doug Short of Advisor Perspectives, an investment research company. As of April 2013, the number of miles driven per person was nearly 9 percent below the peak and equal to where the country was in January 1995. Part of the explanation certainly lies in the recession, because cash-strapped Americans could not afford new cars, and the unemployed weren't going to work anyway. But by many measures the decrease in driving preceded the downturn and appears to be persisting now that recovery is under way. The next few years will be telling.

33 "What most intrigues me is that rates of car ownership per household and per

person started to come down two to three years before the downturn," said Michael Sivak, who studies the trend and who is a research professor at the University of Michigan's Transportation Research Institute. "I think that means something more fundamental is going on."

34 If the pattern persists-and many sociologists believe it will-it will have beneficial implications for carbon emissions and the environment, since transportation is the second largest source of America's emissions, just behind power plants. But it could have negative implications for the car industry. Indeed, companies like Ford and Mercedes are already rebranding themselves "mobility" companies with a broader product range beyond the personal vehicle.

35 "Different things are converging which suggest that we are witnessing a long-term cultural shift," said Mimi Sheller, a sociology professor at Drexel University and director of its Mobilities Research and Policy Center. She cites various factors: the Internet makes telecommuting possible and allows people to feel more connected without driving to meet friends. The renewal of center cities has made the suburbs less appealing and has drawn empty nesters back in. Likewise the rise in cellphones and car-pooling apps has facilitated more flexible commuting arrangements, including the evolution of shared van services for getting to work.

36 With all these changes, people who stopped car commuting as a result of the recession may find less reason to resume the habit. . . .

37 New York's new bike-sharing program and its skyrocketing bridge and tunnel tolls reflect those new priorities, as do a proliferation of car-sharing programs across the nation.

38 Demographic shifts in the driving population suggest that the trend may accelerate. There has been a large drop in the percentage of 16- to 39-year-olds getting a license, while older people are likely to retain their licenses as they age, Mr. Sivak's research has found.

39 He and I have similar observations about our children. Mine (19 and 21) have not bothered to get a driver's license, even though they both live in places where one could come in handy. They are interested, but it's not a priority. They organize their summer jobs and social life around where they can walk or take public transportation or car-pool with friends.

40 Mr. Sivak's son lives in San Francisco and has a car but takes Bay Area Rapid Transit, when he can, even though that often takes longer than driving. "When I was in my 20s and 30s," Mr. Sivak said, "I was curious about what kind of car people drove, but young people don't really care. A car is just a means of getting from A to B when BART doesn't work."

41 A study last year found that driving by young people decreased 23 percent between 2001 and 2009. . . .

42 Whether members of the millennial generation will start buying more cars once they have kids to take to soccer practice and school plays remains an open question. But such projections have important business implications, even if car buyers are merely older or buying fewer cars in a lifetime rather than rejecting car culture outright.

43 At the Mobile World Congress last year in Barcelona, Spain, Bill Ford, executive chairman of the Ford Motor Company, laid out a business plan for a world in which personal vehicle ownership is impractical or undesirable. He proposed partnering with the telecommunications industry to create cities in which "pedestrian, bicycle, private cars, commercial and public transportation traffic are woven into a connected network to save time, conserve resources, lower emissions and improve safety."

Preprocessed Tokens:
['german', 'suburb', 'life', 'goes', 'without', 'cars', 'elisabeth',
'rosenthal', '1', 'vauban', 'germany-residents', 'upscale', 'community',
'suburban', 'pioneers', 'going', 'soccer', 'moms', 'commuting', 'executives',
'ever', 'gone', 'given', 'cars', '2', 'street', 'parking', 'driveways', 'home',
'garages', 'generally', 'forbidden', 'experimental', 'new', 'district',
'outskirts', 'freiburg', 'near', 'french', 'swiss', 'borders', 'vauban', '’',
'streets', 'completely', '"', 'carfree', '"', '-except', 'main', 'thoroughfare',
'tram', 'downtown', 'freiburg', 'runs', 'streets', 'one', 'edge', 'community',
'car', 'ownership', 'allowed', 'two', 'places', 'park-large', 'garages', 'edge',
'development', 'carowner', 'buys', 'space', '40000', 'along', 'home', '3',
'result', '70', 'percent', 'vauban', '’', 'families', 'cars', '57', 'percent',
'sold', 'car', 'move', '"', 'car', 'always', 'tense', '’', 'much', 'happier',
'way', '"', 'said', 'heidrun', 'walter', 'media', 'trainer', 'mother', 'two',
'walked', 'verdant', 'streets', 'swish', 'bicycles', 'chatter', 'wandering',
'children', 'drown', 'occasional', 'distant', 'motor', '4', 'vauban',
'completed', '2006', 'example', 'growing', 'trend', 'europe', 'united',
'states', 'elsewhere', 'separate', 'suburban', 'life', 'auto', 'use',
'component', 'movement', 'called', '"', 'smart', 'planning', '"', '5',
'automobiles', 'linchpin', 'suburbs', 'middleclass', 'families', 'chicago',
'shanghai', 'tend', 'make', 'homes', 'experts', 'say', 'huge', 'impediment',
'current', 'efforts', 'drastically', 'reduce', 'greenhouse', 'gas', 'emissions',
'tailpipes', 'passenger', 'cars', 'responsible', '12', 'percent', 'greenhouse',
'gas', 'emissions', 'europe', '50', 'percent', 'carintensive', 'areas',
'united', 'states', '6', 'efforts', 'past', 'two', 'decades', 'make', 'cities',
'denser', 'better', 'walking', 'planners', 'taking', 'concept', 'suburbs',
'vauban', 'home', '5500', 'residents', 'within', 'rectangular', 'square',
'mile', 'may', 'advanced', 'experiment', 'lowcar', 'suburban', 'life', 'basic',

'precepts', 'adopted', 'around', 'world', 'attempts', 'make', 'suburbs',
'compact', 'accessible', 'public', 'transportation', 'less', 'space', 'parking',
'new', 'approach', 'stores', 'placed', 'walk', 'away', 'main', 'street',
'rather', 'malls', 'along', 'distant', 'highway', '7', '"', 'development',
'since', 'world', 'war', 'ii', 'centered', 'car', 'change', '"', 'said',
'david', 'goldberg', 'official', 'transportation', 'america', 'fastgrowing',
'coalition', 'hundreds', 'groups', 'united', 'states', 'promoting', 'new',
'communities', 'less', 'dependent', 'cars', 'mr', 'goldberg', 'added', '"',
'much', 'drive', 'important', 'whether', 'hybrid', '"', '8', 'levittown',
'scarsdale', 'new', 'york', 'suburbs', 'spreadout', 'homes', 'private',
'garages', 'dream', 'towns', '1950s', 'still', 'exert', 'strong', 'appeal',
'new', 'suburbs', 'may', 'well', 'look', 'vaubanlike', 'developed', 'countries',
'also', 'developing', 'world', 'emissions', 'increasing', 'number', 'private',
'cars', 'owned', 'burgeoning', 'middle', 'class', 'choking', 'cities', '9',
'united', 'states', 'environmental', 'protection', 'agency', 'promoting', '"',
'car', 'reduced', '"', 'communities', 'legislators', 'starting', 'act',
'cautiously', 'many', 'experts', 'expect', 'public', 'transport', 'serving',
'suburbs', 'play', 'much', 'larger', 'role', 'new', 'sixyear', 'federal',
'transportation', 'bill', 'approved', 'year', 'mr', 'goldberg', 'said',
'previous', 'bills', '80', 'percent', 'appropriations', 'law', 'gone',
'highways', '20', 'percent', 'transport', 'excerpt', '"', 'german', 'suburb',
'life', 'goes', 'without', 'cars', '"', 'elisabeth', 'rosenthal', 'new', 'york',
'times', 'copyright', '©', '2009', 'new', 'york', 'times', 'company',
'reprinted', 'permission', 'new', 'york', 'times', 'company', 'via',
'copyright', 'clearance', 'center', 'paris', 'bans', 'driving', 'due', 'smog',
'robert', 'duffer', '10', 'days', 'nearrecord', 'pollution', 'paris',
'enforced', 'partial', 'driving', 'ban', 'clear', 'air', 'global', 'city', '11',
'monday', 'motorists', 'evennumbered', 'license', 'plates', 'ordered', 'leave',
'cars', 'home', 'suffer', '22euro', 'fine', '31', 'would', 'apply',
'oddnumbered', 'plates', 'following', 'day', '12', 'almost', '4000', 'drivers',
'fined', 'according', 'reuters1', 'twentyseven', 'people', 'cars', 'impounded',
'reaction', 'fine', '13', '’', 'easier', 'imagine', 'carfree', 'champselysees2',
'14', 'congestion', '3', '60', 'percent', 'capital', 'france', 'fivedays',
'intensifying', 'smog', 'smog', 'rivaled', 'beijing', 'china', 'known', 'one',
'polluted', 'cities', 'world', '15', 'cold', 'nights', 'warm', 'days', 'caused',
'warmer', 'layer', 'air', 'trap', 'car', 'emissions', '16', 'diesel', 'fuel',
'blamed', 'since', 'france', 'tax', 'policy', 'favors', 'diesel', 'gasoline',
'diesels', 'make', '67', 'percent', 'vehicles', 'france', 'compared', '533',
'percent', 'average', 'diesel', 'engines', 'rest', 'western', 'europe',
'according', 'reuters', '17', 'paris', 'typically', 'smog', 'european',
'capitals', 'last', 'week', 'paris', '147', 'micrograms', 'particulate',
'matter', 'pm', 'per', 'cubic', 'meter', 'compared', '114', 'brussels', '797',
'london', 'reuters', 'found', '18', 'delivery', 'companies', 'complained',
'lost', 'revenue', 'exceptions', 'made', 'plugin', 'cars', 'hybrids', 'cars',
'carrying', 'three', 'passengers', 'public', 'transit', 'free', 'charge',
'friday', 'monday', 'according', 'bbc', '19', 'smog', 'cleared', 'enough',
'monday', 'ruling', 'french', 'party', 'rescind', 'ban', 'oddnumbered',
'plates', 'tuesday', '1', 'excerpt', '"', 'paris', 'bans', 'driving', 'due',

'smog', '"', 'robert', 'duffer', 'chicago', 'tribune', 'copyright', '©', '2014',
'chicago', 'tribune', 'reprinted', 'permission', 'chicago', 'tribune', 'via',
'copyright', 'clearance', 'center', 'carfree', 'day', 'spinning', 'big', 'hit',
'bogota', 'andrew', 'selsky', 'bogota', 'colombia-in', 'program', '’', 'set',
'spread', 'countries', 'millions', 'colombians', 'hiked', 'biked', 'skated',
'took', 'buses', 'work', 'carfree', 'day', 'yesterday', 'leaving', 'streets',
'capital', 'city', 'eerily', 'devoid', 'traffic', 'jams', '21', 'third',
'straight', 'year', 'cars', 'banned', 'buses', 'taxis', 'permitted', 'day',
'without', 'cars', 'capital', 'city', '7', 'million', 'goal', 'promote',
'alternative', 'transportation', 'reduce', 'smog', 'violators', 'faced', '25',
'fines', '22', 'turnout', 'large', 'despite', 'gray', 'clouds', 'dumped',
'occasional', 'rain', 'showers', 'bogota', '23', '"', 'rain', '’', 'stopped',
'people', 'participating', '"', 'said', 'bogota', 'mayor', 'antanas', 'mockus',
'24', '"', '’', 'good', 'opportunity', 'take', 'away', 'stress', 'lower', 'air',
'pollution', '"', 'said', 'businessman', 'carlos', 'arturo', 'plaza', 'rode',
'twoseat', 'bicycle', 'wife', '25', 'first', 'time', 'two', 'colombian',
'cities', 'cali', 'valledupar', 'joined', 'event', '26', 'municipal',
'authorities', 'countries', 'came', 'bogota', 'see', 'event', 'enthusiastic',
'"', 'people', 'generating', 'revolutionary', 'change', 'crossing', 'borders',
'"', 'said', 'enrique', 'riera', 'mayor', 'asunción', 'paraguay', '27', 'day',
'without', 'cars', 'part', 'improvement', 'campaign', 'began', 'bogota',
'mid1990s', 'seen', 'construction', '118', 'miles', 'bicycle', 'paths', 'latin',
'american', 'city', 'according', 'mockus', 'city', '’', 'mayor', '28', 'parks',
'sports', 'centers', 'also', 'bloomed', 'throughout', 'city', 'uneven',
'pitted', 'sidewalks', 'replaced', 'broad', 'smooth', 'sidewalks', 'rushhour',
'restrictions', 'dramatically', 'cut', 'traffic', 'new', 'restaurants',
'upscale', 'shopping', 'districts', 'cropped', 'excerpt', '"', 'carfree', 'day',
'spinning', 'big', 'hit', 'bogota', '"', 'andrew', 'selsky', 'seattle', 'times',
'copyright', '©', '2002', 'seattle', 'times', 'company', 'reprinted',
'permission', 'seattle', 'times', 'company', 'via', 'copyright', 'clearance',
'center', 'end', 'car', 'culture', 'elisabeth', 'rosenthal', '29', 'president',
'obama', '’', 'ambitious', 'goals', 'curb', 'united', 'states', '’',
'greenhouse', 'gas', 'emissions', 'unveiled', 'last', 'week', 'get',
'fortuitous', 'assist', 'incipient1', 'shift', 'american', 'behavior', 'recent',
'studies', 'suggest', 'americans', 'buying', 'fewer', 'cars', 'driving', 'less',
'getting', 'fewer', 'licenses', 'year', 'goes', '30', 'left', 'researchers',
'pondering', 'fundamental', 'question', 'america', 'passed', 'peak', 'driving',
'31', 'united', 'states', 'broad', 'expanses', 'suburban', 'ideals', 'long',
'one', 'world', '’', 'prime', 'car', 'cultures', 'birthplace', 'model', 'home',
'detroit', 'place', 'wilson', 'pickett', 'immortalized', '"', 'mustang',
'sally', '"', '32', 'america', '’', 'love', 'affair', 'vehicles', 'seems',
'cooling', 'adjusted', 'population', 'growth', 'number', 'miles', 'driven',
'united', 'states', 'peaked', '2005', 'dropped', 'steadily', 'thereafter',
'according', 'analysis', 'doug', 'short', 'advisor', 'perspectives',
'investment', 'research', 'company', 'april', '2013', 'number', 'miles',
'driven', 'per', 'person', 'nearly', '9', 'percent', 'peak', 'equal', 'country',
'january', '1995', 'part', 'explanation', 'certainly', 'lies', 'recession',
'cashstrapped', 'americans', 'could', 'afford', 'new', 'cars', 'unemployed',

''', 'going', 'work', 'anyway', 'many', 'measures', 'decrease', 'driving',
'preceded', 'downturn', 'appears', 'persisting', 'recovery', 'way', 'next',
'years', 'telling', '33', '"', 'intrigues', 'rates', 'car', 'ownership', 'per',
'household', 'per', 'person', 'started', 'come', 'two', 'three', 'years',
'downturn', '"', 'said', 'michael', 'sivak', 'studies', 'trend', 'research',
'professor', 'university', 'michigan', '', 'transportation', 'research',
'institute', '"', 'think', 'means', 'something', 'fundamental', 'going', '"',
'34', 'pattern', 'persists-and', 'many', 'sociologists', 'believe', 'will-it',
'beneficial', 'implications', 'carbon', 'emissions', 'environment', 'since',
'transportation', 'second', 'largest', 'source', 'america', '', 'emissions',
'behind', 'power', 'plants', 'could', 'negative', 'implications', 'car',
'industry', 'indeed', 'companies', 'like', 'ford', 'mercedes', 'already',
'rebranding', '"', 'mobility', '"', 'companies', 'broader', 'product', 'range',
'beyond', 'personal', 'vehicle', '35', '"', 'different', 'things', 'converging',
'suggest', 'witnessing', 'longterm', 'cultural', 'shift', '"', 'said', 'mimi',
'sheller', 'sociology', 'professor', 'drexel', 'university', 'director',
'mobilities', 'research', 'policy', 'center', 'cites', 'various', 'factors',
'internet', 'makes', 'telecommuting', 'possible', 'allows', 'people', 'feel',
'connected', 'without', 'driving', 'meet', 'friends', 'renewal', 'center',
'cities', 'made', 'suburbs', 'less', 'appealing', 'drawn', 'empty', 'nesters',
'back', 'likewise', 'rise', 'cellphones', 'carpooling', 'apps', 'facilitated',
'flexible', 'commuting', 'arrangements', 'including', 'evolution', 'shared',
'van', 'services', 'getting', 'work', '36', 'changes', 'people', 'stopped',
'car', 'commuting', 'result', 'recession', 'may', 'find', 'less', 'reason',
'resume', 'habit', '37', 'new', 'york', '', 'new', 'bikesharing', 'program',
'skyrocketing', 'bridge', 'tunnel', 'tolls', 'reflect', 'new', 'priorities',
'proliferation', 'carsharing', 'programs', 'across', 'nation', '38',
'demographic', 'shifts', 'driving', 'population', 'suggest', 'trend', 'may',
'accelerate', 'large', 'drop', 'percentage', '16', '39yearolds', 'getting',
'license', 'older', 'people', 'likely', 'retain', 'licenses', 'age', 'mr',
'sivak', '', 'research', 'found', '39', 'similar', 'observations', 'children',
'mine', '19', '21', 'bothered', 'get', 'driver', '', 'license', 'even',
'though', 'live', 'places', 'one', 'could', 'come', 'handy', 'interested', '',
'priority', 'organize', 'summer', 'jobs', 'social', 'life', 'around', 'walk',
'take', 'public', 'transportation', 'carpool', 'friends', '40', 'mr', 'sivak',
'', 'son', 'lives', 'san', 'francisco', 'car', 'takes', 'bay', 'area', 'rapid',
'transit', 'even', 'though', 'often', 'takes', 'longer', 'driving', '"', '20s',
'30s', '"', 'mr', 'sivak', 'said', '"', 'curious', 'kind', 'car', 'people',
'drove', 'young', 'people', '', 'really', 'care', 'car', 'means', 'getting',
'b', 'bart', '', 'work', '"', '41', 'study', 'last', 'year', 'found',
'driving', 'young', 'people', 'decreased', '23', 'percent', '2001', '2009',
'42', 'whether', 'members', 'millennial', 'generation', 'start', 'buying',
'cars', 'kids', 'take', 'soccer', 'practice', 'school', 'plays', 'remains',
'open', 'question', 'projections', 'important', 'business', 'implications',
'even', 'car', 'buyers', 'merely', 'older', 'buying', 'fewer', 'cars',
'lifetime', 'rather', 'rejecting', 'car', 'culture', 'outright', '43', 'mobile',
'world', 'congress', 'last', 'year', 'barcelona', 'spain', 'bill', 'ford',
'executive', 'chairman', 'ford', 'motor', 'company', 'laid', 'business', 'plan',

```
'world', 'personal', 'vehicle', 'ownership', 'impractical', 'undesirable',
'proposed', 'partnering', 'telecommunications', 'industry', 'create', 'cities',
'"', 'pedestrian', 'bicycle', 'private', 'cars', 'commercial', 'public',
'transportation', 'traffic', 'woven', 'connected', 'network', 'save', 'time',
'conserve', 'resources', 'lower', 'emissions', 'improve', 'safety', '"',
'excerpt', '"', 'end', 'car', 'culture', '"', 'elisabeth', 'rosenthal', 'new',
'york', 'times', 'copyright', '©', '2013', 'new', 'york', 'times', 'company',
'reprinted', 'permission', 'new', 'york', 'times', 'company', 'via',
'copyright', 'clearance', 'center']
```

[25]:
```
df = df.copy().reset_index(drop=True) # pd.concat([ext_df, df], axis=0)
df.head()
```

[25]:
```
   prompt_id                      prompt_name  \
0          0                   Car-free cities
1          1  Does the electoral college work?

                                        instructions  \
0  Write an explanatory essay to inform fellow ci…
1  Write a letter to your state senator in which …

                                         source_text
0  # In German Suburb, Life Goes On Without Cars …
1  # What Is the Electoral College? by the Office…
```

Feature Engineering:

Extract meaningful features from the text data that may help distinguish between student-written and LLM-generated essays.

[26]:
```
import pandas as pd

# Sample data
data = {
    'text': [
        "Cars. Cars have been around since they became famous in the 1900s,␣
 ↪when Henry Ford created and built...",
        "\"America's love affair with it's vehicles seems to be cooling\" says␣
 ↪Elisabeth Rosenthal. To understand...",
        "The Electoral College has been kept for centuries, established by the␣
 ↪founding fathers and establish..."
    ],
    'label': ['Cars', 'America', 'The Electoral']
}

# Create a DataFrame
df = pd.DataFrame(data)
```

```
# Display the DataFrame
print(df)
```

```
                                                text          label
0  Cars. Cars have been around since they became …           Cars
1  "America's love affair with it's vehicles seem…        America
2  The Electoral College has been kept for centur…  The Electoral
```

[27]:
```
# Add features - Example features: Word count, average word length, sentiment␣
 ↪score
df['word_count'] = df['text'].apply(lambda x: len(nltk.word_tokenize(x)))
df['avg_word_length'] = df['text'].apply(lambda x: sum(len(word) for word in␣
 ↪nltk.word_tokenize(x)) / len(nltk.word_tokenize(x)))
```

Model Selection:

Choose a suitable machine learning model for the task. Common choices include text classification models, such as logistic regression, support vector machines, or deep learning models like recurrent neural networks (RNNs) or transformers. Experiment with different models and architectures to find the one that performs best for your specific problem.

[32]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.utils import shuffle
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Flatten, Dense

# Sample data
data = {
    'text': ["Cars have been around since they became famous in the 1900s.",
             "America's love affair with its vehicles seems to be cooling.",
             "The Electoral College has been kept for centuries, established by␣
 ↪the founding fathers."],
    'label': ['Cars', 'America', 'The Electoral']
}

df = pd.DataFrame(data)

# Preprocessing and Tokenization
tokenizer = Tokenizer()
tokenizer.fit_on_texts(df['text'])
X = tokenizer.texts_to_sequences(df['text'])
X = pad_sequences(X)
```

```
# Label encoding for Neural Network
label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(df['label'])

# Split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y_encoded, test_size=0.
 ↪2, random_state=42)

# Neural Network Model
vocab_size = len(tokenizer.word_index) + 1
embedding_dim = 50

nn_model = Sequential()
nn_model.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim,␣
 ↪input_length=X.shape[1]))
nn_model.add(Flatten())
nn_model.add(Dense(64, activation='relu'))
nn_model.add(Dense(len(label_encoder.classes_), activation='softmax'))

nn_model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',␣
 ↪metrics=['accuracy'])
nn_model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test))

# Evaluate the model
eval_result = nn_model.evaluate(X_test, y_test)
print(f"Accuracy: {eval_result[1] * 100:.2f}%")
```

```
Epoch 1/10
1/1 [==============================] - 1s 1s/step - loss: 1.1247 - accuracy:
0.0000e+00 - val_loss: 1.0823 - val_accuracy: 1.0000
Epoch 2/10
1/1 [==============================] - 0s 39ms/step - loss: 1.0597 - accuracy:
1.0000 - val_loss: 1.0881 - val_accuracy: 1.0000
Epoch 3/10
1/1 [==============================] - 0s 36ms/step - loss: 1.0128 - accuracy:
1.0000 - val_loss: 1.0930 - val_accuracy: 1.0000
Epoch 4/10
1/1 [==============================] - 0s 38ms/step - loss: 0.9746 - accuracy:
1.0000 - val_loss: 1.0975 - val_accuracy: 0.0000e+00
Epoch 5/10
1/1 [==============================] - 0s 38ms/step - loss: 0.9375 - accuracy:
1.0000 - val_loss: 1.1016 - val_accuracy: 0.0000e+00
Epoch 6/10
1/1 [==============================] - 0s 37ms/step - loss: 0.9022 - accuracy:
1.0000 - val_loss: 1.1062 - val_accuracy: 0.0000e+00
Epoch 7/10
1/1 [==============================] - 0s 35ms/step - loss: 0.8666 - accuracy:
```

```
1.0000 - val_loss: 1.1110 - val_accuracy: 0.0000e+00
Epoch 8/10
1/1 [==============================] - 0s 40ms/step - loss: 0.8306 - accuracy:
1.0000 - val_loss: 1.1157 - val_accuracy: 0.0000e+00
Epoch 9/10
1/1 [==============================] - 0s 35ms/step - loss: 0.7943 - accuracy:
1.0000 - val_loss: 1.1206 - val_accuracy: 0.0000e+00
Epoch 10/10
1/1 [==============================] - 0s 38ms/step - loss: 0.7579 - accuracy:
1.0000 - val_loss: 1.1258 - val_accuracy: 0.0000e+00
1/1 [==============================] - 0s 25ms/step - loss: 1.1258 - accuracy:
0.0000e+00
Accuracy: 0.00%
```

Learning Rate Schedules:

Step Decay:

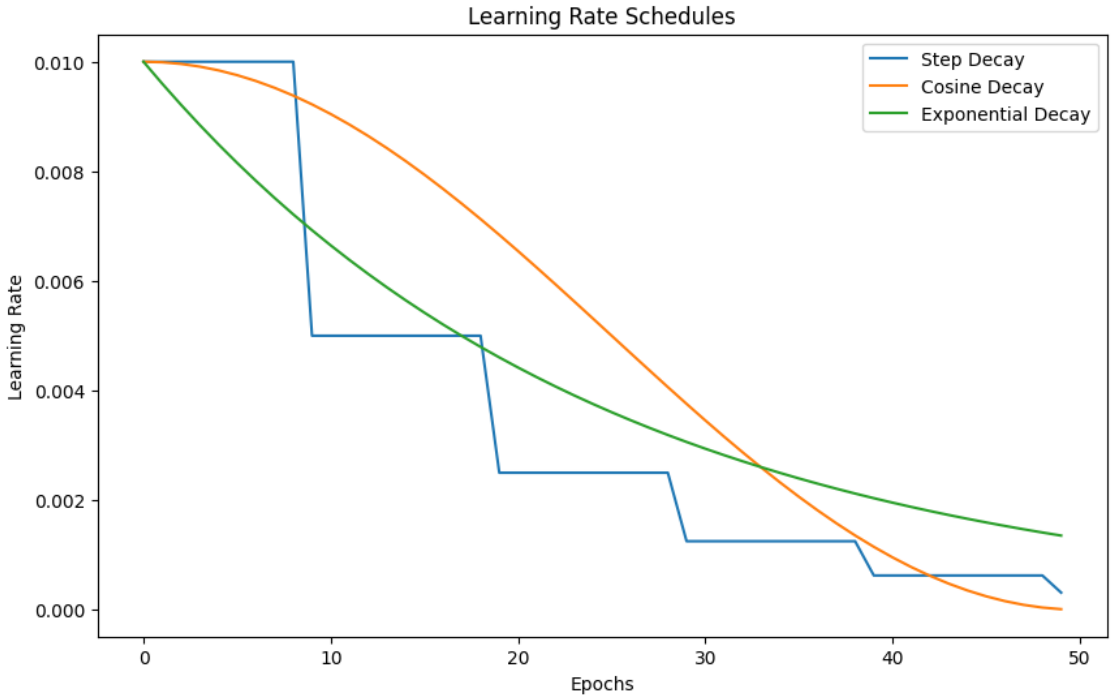
Function: step_decay(epoch, initial_lr, drop_factor=0.5, epochs_drop=10) Explanation: The learning rate is reduced by a factor (drop_factor) after a certain number of epochs (epochs_drop). This schedule is like descending stairs, where the learning rate drops abruptly at predefined intervals. Cosine Decay:

Function: cosine_decay(epoch, initial_lr, total_epochs) Explanation: The learning rate follows a cosine curve, gradually decreasing from the initial learning rate to a minimum value. This schedule provides a smoother reduction in the learning rate compared to step decay. Exponential Decay:

Function: exp_decay(epoch, initial_lr, decay_rate=0.96) Explanation: The learning rate decreases exponentially with each epoch. The decay rate (decay_rate) controls how quickly the learning rate decreases.

[33]:
```python
import tensorflow as tf
import numpy as np
import matplotlib.pyplot as plt

# Learning rate scheduler functions

def step_decay(epoch, initial_lr, drop_factor=0.5, epochs_drop=10):
    """Step decay learning rate schedule."""
    return initial_lr * np.power(drop_factor, np.floor((1 + epoch) /␣
 ↪epochs_drop))

def cosine_decay(epoch, initial_lr, total_epochs):
    """Cosine decay learning rate schedule."""
    return 0.5 * initial_lr * (1 + np.cos(np.pi * epoch / total_epochs))

def exp_decay(epoch, initial_lr, decay_rate=0.96):
    """Exponential decay learning rate schedule."""
    return initial_lr * np.power(decay_rate, epoch)
```

```
# Parameters
initial_learning_rate = 0.01
total_epochs = 50

# Plotting the learning rate schedules
epochs = range(total_epochs)

# Step decay
step_lr = [step_decay(epoch, initial_learning_rate) for epoch in epochs]

# Cosine decay
cosine_lr = [cosine_decay(epoch, initial_learning_rate, total_epochs) for epoch␣
 ↪in epochs]

# Exponential decay
exp_lr = [exp_decay(epoch, initial_learning_rate) for epoch in epochs]

# Plotting
plt.figure(figsize=(10, 6))
plt.plot(epochs, step_lr, label='Step Decay')
plt.plot(epochs, cosine_lr, label='Cosine Decay')
plt.plot(epochs, exp_lr, label='Exponential Decay')
plt.xlabel('Epochs')
plt.ylabel('Learning Rate')
plt.title('Learning Rate Schedules')
plt.legend()
plt.show()
```



[35]:
```
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Bidirectional, LSTM, Dense



# Load the data from CSV files
essays_df = pd.read_csv('train_essays[1].csv')
prompts_df = pd.read_csv('train_prompts[1].csv')

# Check the column names in the DataFrames
print("Essays DataFrame columns:", essays_df.columns)
print("Prompts DataFrame columns:", prompts_df.columns)
```

```
Essays DataFrame columns: Index(['id', 'prompt_id', 'text', 'generated'],
dtype='object')
Prompts DataFrame columns: Index(['prompt_id', 'prompt_name', 'instructions',
'source_text'], dtype='object')
```

Model Architecture:

[36]:
```
# Define the vocabulary size, embedding dimension, and maximum sequence length
vocab_size = 10000  # Adjust based on your vocabulary size
embedding_dim = 100  # Adjust based on your requirements
max_sequence_length = 200  # Adjust based on your sequence length
```

```
# Define the model
model = Sequential()
model.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim,␣
 ↪input_length=max_sequence_length))
model.add(Bidirectional(LSTM(64, return_sequences=True)))
model.add(tf.keras.layers.GlobalAveragePooling1D())
num_classes = 2  # Adjust based on the number of classes in your problem
model.add(Dense(num_classes, activation='softmax'))
```

Compile Model

[37]:
```
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',␣
 ↪metrics=['accuracy'])
```

Summary

[39]:
```
model.summary()
```

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 200, 100)          1000000

 bidirectional (Bidirection  (None, 200, 128)          84480
 al)

 global_average_pooling1d (  (None, 128)               0
 GlobalAveragePooling1D)

 dense_2 (Dense)             (None, 2)                 258

=================================================================
Total params: 1084738 (4.14 MB)
Trainable params: 1084738 (4.14 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Train Model

[40]:
```
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
```

```
from tensorflow.keras.layers import Embedding, LSTM, Dense
```

[41]:
```
# Assuming you have a DataFrame 'essays_df' with columns 'id', 'prompt_id',␣
 ↪'text', and 'generated'
# and a DataFrame 'prompts_df' with columns 'prompt_id', 'prompt_name',␣
 ↪'instructions', and 'source_text'

# Merge the DataFrames on 'prompt_id'
df = pd.merge(essays_df, prompts_df, on='prompt_id')
```

[42]:
```
# Assuming 'generated' column is your target label and 'text' column is your␣
 ↪input features
X = df['text']
y = df['generated']

# Encode the target labels
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(y)
```

[43]:
```
# Tokenize and pad the input text
max_sequence_length = 100  # Adjust as needed
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X)
X_sequences = tokenizer.texts_to_sequences(X)
X_padded = pad_sequences(X_sequences, maxlen=max_sequence_length,␣
 ↪padding='post', truncating='post')
```

[44]:
```
# Initialize Stratified K-Fold
num_folds = 5  # Adjust as needed
random_seed = 42  # Adjust as needed
skf = StratifiedKFold(n_splits=num_folds, shuffle=True,␣
 ↪random_state=random_seed)

# Placeholder for model training and evaluation
for fold, (train_idx, val_idx) in enumerate(skf.split(X_padded, y)):
    print(f"Training on Fold {fold + 1}")

    # Split data into training and validation sets
    X_train, X_val = X_padded[train_idx], X_padded[val_idx]
    y_train, y_val = y[train_idx], y[val_idx]
```

```
Training on Fold 1
Training on Fold 2
Training on Fold 3
Training on Fold 4
Training on Fold 5
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/model_selection/_split.py:700:
UserWarning: The least populated class in y has only 3 members, which is less
than n_splits=5.
  warnings.warn(
```

[45]:
```
# Define and compile the model
model = Sequential()
model.add(Embedding(input_dim=len(tokenizer.word_index) + 1, output_dim=100,␣
 ↪input_length=max_sequence_length))
model.add(LSTM(units=64))
model.add(Dense(units=1, activation='sigmoid'))

model.compile(optimizer='adam', loss='binary_crossentropy',␣
 ↪metrics=['accuracy'])
```

[46]:
```
# Train the model
epochs = 15  # Adjust as needed
model.fit(X_train, y_train, epochs=epochs, validation_data=(X_val, y_val))

# Evaluate the model
y_pred = model.predict(X_val)
y_pred_classes = [1 if pred > 0.5 else 0 for pred in y_pred]
accuracy = accuracy_score(y_val, y_pred_classes)
print(f"Validation Accuracy for Fold {fold + 1}: {accuracy}")
```

```
Epoch 1/15
35/35 [==============================] - 6s 111ms/step - loss: 0.2175 -
accuracy: 0.9873 - val_loss: 0.0018 - val_accuracy: 1.0000
Epoch 2/15
35/35 [==============================] - 4s 123ms/step - loss: 0.0192 -
accuracy: 0.9973 - val_loss: 0.0018 - val_accuracy: 1.0000
Epoch 3/15
35/35 [==============================] - 3s 95ms/step - loss: 0.0189 - accuracy:
0.9973 - val_loss: 0.0022 - val_accuracy: 1.0000
Epoch 4/15
35/35 [==============================] - 3s 96ms/step - loss: 0.0189 - accuracy:
0.9973 - val_loss: 0.0025 - val_accuracy: 1.0000
Epoch 5/15
35/35 [==============================] - 3s 95ms/step - loss: 0.0187 - accuracy:
0.9973 - val_loss: 0.0033 - val_accuracy: 1.0000
Epoch 6/15
35/35 [==============================] - 4s 112ms/step - loss: 0.0188 -
accuracy: 0.9973 - val_loss: 0.0030 - val_accuracy: 1.0000
Epoch 7/15
35/35 [==============================] - 3s 90ms/step - loss: 0.0183 - accuracy:
0.9973 - val_loss: 0.0027 - val_accuracy: 1.0000
Epoch 8/15
35/35 [==============================] - 3s 94ms/step - loss: 0.0179 - accuracy:
```

```
0.9973 - val_loss: 0.0030 - val_accuracy: 1.0000
Epoch 9/15
35/35 [==============================] - 4s 114ms/step - loss: 0.0174 -
accuracy: 0.9973 - val_loss: 0.0025 - val_accuracy: 1.0000
Epoch 10/15
35/35 [==============================] - 3s 93ms/step - loss: 0.0170 - accuracy:
0.9973 - val_loss: 0.0035 - val_accuracy: 1.0000
Epoch 11/15
35/35 [==============================] - 3s 93ms/step - loss: 0.0155 - accuracy:
0.9973 - val_loss: 0.0027 - val_accuracy: 1.0000
Epoch 12/15
35/35 [==============================] - 3s 95ms/step - loss: 0.0144 - accuracy:
0.9973 - val_loss: 0.0025 - val_accuracy: 1.0000
Epoch 13/15
35/35 [==============================] - 4s 123ms/step - loss: 0.0129 -
accuracy: 0.9973 - val_loss: 0.0049 - val_accuracy: 1.0000
Epoch 14/15
35/35 [==============================] - 3s 92ms/step - loss: 0.0113 - accuracy:
0.9973 - val_loss: 0.0027 - val_accuracy: 1.0000
Epoch 15/15
35/35 [==============================] - 3s 92ms/step - loss: 0.0088 - accuracy:
0.9973 - val_loss: 0.0032 - val_accuracy: 1.0000
9/9 [==============================] - 1s 17ms/step
Validation Accuracy for Fold 5: 1.0
```

Hyperparameter Tuning:

Fine-tune your model's hyperparameters to improve performance. This can involve adjusting learning rates, batch sizes, or regularization parameters.

[69]:
```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from keras.models import Sequential
from keras.layers import Embedding, LSTM, Dense
from keras.optimizers import Adam
from keras.preprocessing.sequence import pad_sequences
from keras.preprocessing.text import Tokenizer
import numpy as np

# Assuming X, y are defined properly
# ...

# Convert non-string data to strings
X = [str(x) for x in X]

# Split the data into training and validation sets
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=42)
```

```
# Tokenize and pad the sequences
tokenizer = Tokenizer()
tokenizer.fit_on_texts(X_train + X_val)  # Fit tokenizer on both training and␣
 ↪validation data
X_train_padded = tokenizer.texts_to_sequences(X_train)
X_val_padded = tokenizer.texts_to_sequences(X_val)

max_sequence_length = max(len(seq) for seq in X_train_padded + X_val_padded)

X_train_padded = pad_sequences(X_train_padded, maxlen=max_sequence_length)
X_val_padded = pad_sequences(X_val_padded, maxlen=max_sequence_length)

# Hyperparameter tuning
learning_rates = [0.001, 0.01, 0.1]
batch_sizes = [16, 32, 64]
epochs = 10

best_accuracy = 0
best_hyperparameters = {}

for lr in learning_rates:
    for batch_size in batch_sizes:
        model = Sequential()
        model.add(Embedding(input_dim=len(tokenizer.word_index) + 1,␣
 ↪output_dim=100, input_length=max_sequence_length))
        model.add(LSTM(units=64))
        model.add(Dense(units=1, activation='sigmoid'))

        optimizer = Adam(learning_rate=lr)
        model.compile(optimizer=optimizer, loss='binary_crossentropy',␣
 ↪metrics=['accuracy'])

        # Convert data types if needed
        X_train_padded = X_train_padded.astype('float32')
        X_val_padded = X_val_padded.astype('float32')

        # Ensure label data is in the correct format
        y_train = np.array(y_train)
        y_val = np.array(y_val)

        model.fit(X_train_padded, y_train, epochs=epochs,␣
 ↪batch_size=batch_size, validation_data=(X_val_padded, y_val), verbose=0)

        # Evaluate the model
        y_pred = (model.predict(X_val_padded) > 0.5).astype("int32")
        accuracy = accuracy_score(y_val, y_pred)
```

```
        if accuracy > best_accuracy:
            best_accuracy = accuracy
            best_hyperparameters = {'learning_rate': lr, 'batch_size':␣
 ↪batch_size}

print("Best Hyperparameters:", best_hyperparameters)
print("Best Validation Accuracy:", best_accuracy)

# Train the final model with the best hyperparameters
final_model = Sequential()
final_model.add(Embedding(input_dim=len(tokenizer.word_index) + 1,␣
 ↪output_dim=100, input_length=max_sequence_length))
final_model.add(LSTM(units=64))
final_model.add(Dense(units=1, activation='sigmoid'))
final_model.
 ↪compile(optimizer=Adam(learning_rate=best_hyperparameters['learning_rate']),
                   loss='binary_crossentropy', metrics=['accuracy'])
final_model.fit(X_train_padded, y_train, epochs=epochs,␣
 ↪batch_size=best_hyperparameters['batch_size'], verbose=1)
```

```
9/9 [==============================] - 2s 202ms/step
9/9 [==============================] - 2s 193ms/step
9/9 [==============================] - 2s 207ms/step
9/9 [==============================] - 2s 190ms/step
9/9 [==============================] - 3s 260ms/step
9/9 [==============================] - 2s 205ms/step
9/9 [==============================] - 3s 285ms/step
9/9 [==============================] - 4s 372ms/step
9/9 [==============================] - 2s 206ms/step
Best Hyperparameters: {'learning_rate': 0.001, 'batch_size': 16}
Best Validation Accuracy: 0.9927536231884058
Epoch 1/10
69/69 [==============================] - 62s 863ms/step - loss: 0.1194 -
accuracy: 0.9918
Epoch 2/10
69/69 [==============================] - 60s 867ms/step - loss: 0.0075 -
accuracy: 0.9991
Epoch 3/10
69/69 [==============================] - 60s 864ms/step - loss: 0.0075 -
accuracy: 0.9991
Epoch 4/10
69/69 [==============================] - 59s 851ms/step - loss: 0.0074 -
accuracy: 0.9991
Epoch 5/10
69/69 [==============================] - 60s 865ms/step - loss: 0.0074 -
accuracy: 0.9991
```

```
Epoch 6/10
69/69 [==============================] - 60s 866ms/step - loss: 0.0072 -
accuracy: 0.9991
Epoch 7/10
69/69 [==============================] - 60s 864ms/step - loss: 0.0071 -
accuracy: 0.9991
Epoch 8/10
69/69 [==============================] - 58s 848ms/step - loss: 0.0068 -
accuracy: 0.9991
Epoch 9/10
69/69 [==============================] - 59s 855ms/step - loss: 0.0063 -
accuracy: 0.9991
Epoch 10/10
69/69 [==============================] - 59s 861ms/step - loss: 0.0056 -
accuracy: 0.9991
```

[69]: `<keras.src.callbacks.History at 0x7c14362108b0>`

[74]:
```
# Assuming you have X_val, y_val, and y_pred_final
# ...

# Define or obtain y_pred_final (predictions) from your model
y_pred_final = (final_model.predict(X_val_padded) > 0.5).astype("int32")

# Create an index using np.arange
index = np.arange(len(y_val))

# Convert y_pred_final to a DataFrame with the same index as y_val
result_df = pd.DataFrame({'true_labels': y_val, 'predicted_labels':␣
 ↪y_pred_final.flatten()}, index=index)

# Filter misclassifications
misclassifications_df = result_df[result_df['true_labels'] !=␣
 ↪result_df['predicted_labels']]

# Extract misclassified indices, true labels, and predicted labels
misclassified_indices = misclassifications_df.index
true_labels = misclassifications_df['true_labels']
predicted_labels = misclassifications_df['predicted_labels']

# Extract misclassified samples from X_val list
misclassified_samples = [X_val[i] for i in misclassified_indices]

# Print or analyze the misclassifications
print("Misclassified Samples:")
print(misclassified_samples)
print("True Labels:")
```

```
print(true_labels)
print("Predicted Labels:")
print(predicted_labels)
```

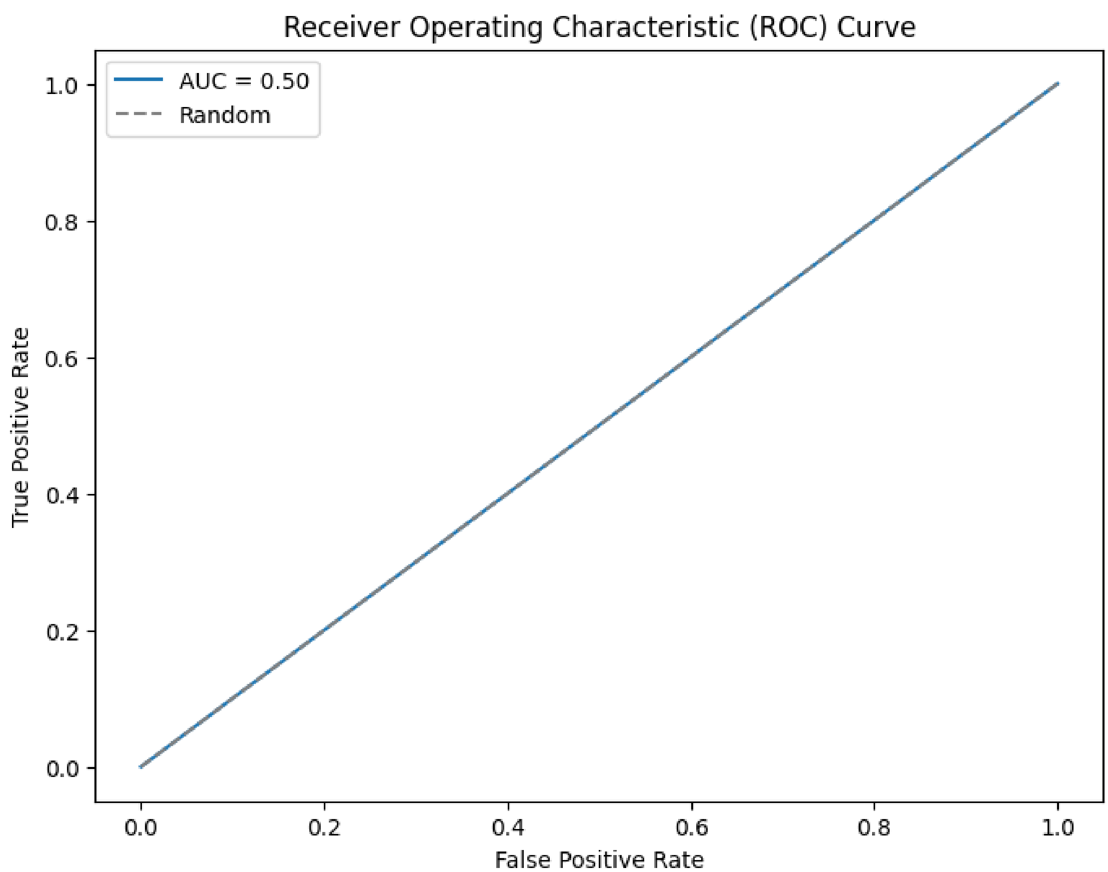
```
9/9 [==============================] - 2s 203ms/step
Misclassified Samples:
["This essay will analyze, discuss and prove one reason in favor of keeping the
Electoral College in the United States for its presidential elections. One of
the reasons to keep the electoral college is that it is better for smaller, more
rural states to have more influence as opposed to larger metropolitan areas that
have large populations. The electors from these states are granted two votes
each. Those from larger, more populated areas are granted just one vote each.
Smaller states tend to hold significant power because their two votes for
president and vice president add up more than the votes of larger states that
have many electors. This is because of the split of the electoral votes. Some
argue that electors are not bound to vote for the candidate who won the most
votes nationally. They do not have to vote for their own state's nominee unless
their state has a winner take all system. However, there are states that have
adopted laws that force their electors to vote for their state's candidate. It
seems that, no matter how, electors are not bound to vote for the candidate who
won the most nationally. This is not always the case because of state
legislatures who can overrule the electors and vote for the alternative
candidate their citizens have selected for them, even if the voter lives in a
state without a winner take all system.", "I strongly believe that the Electoral
College should remain the way it is or, better yet, that we should elect the
president by popular vote. This is due to the fact that the Electoral College
does not accurately reflect the will of the people. For example, in the 2016
presidential election, an estimated two million more people voted for Hillary
Clinton than for Donald Trump however, Trump won the Electoral College vote, 304
to 232. This means that a candidate can win a majority of the Electoral College
voters while losing the popular vote! Furthermore, voting for President should
be an individual citizen decision, not a state decision. The Electoral College
works by awarding all of a state's electoral votes to the winner of the majority
of votes in the state. This means that a candidate can win the majority of votes
in a state and still not receive any of that states electoral votes. This goes
against the concept of onepersononevote, since a candidate can win the majority
of votes in a state and still not win any electoral votes. By eliminating the
Electoral College and electing the president by popular vote, the votes of every
individual will be counted, and the candidate who wins the most votes nationally
will win the election. In conclusion, the Electoral College does not reflect the
will of the people and votes in state are not equally weighted. It is time to
elect the president by popular vote and to finally give the votes of individual
citizens the weight they deserve."]
True Labels:
158    1
170    1
Name: true_labels, dtype: int64
```

```
Predicted Labels:
158    0
170    0
Name: predicted_labels, dtype: int32
```

[75]:
```python
from sklearn.metrics import confusion_matrix, classification_report, roc_curve,␣
 ↪roc_auc_score
import matplotlib.pyplot as plt

# Assuming you have y_val and y_pred_final as defined earlier

# Confusion Matrix
conf_matrix = confusion_matrix(y_val, y_pred_final)
print("Confusion Matrix:")
print(conf_matrix)

# Classification Report
class_report = classification_report(y_val, y_pred_final)
print("\nClassification Report:")
print(class_report)

# ROC Curve
fpr, tpr, thresholds = roc_curve(y_val, y_pred_final)
roc_auc = roc_auc_score(y_val, y_pred_final)

# Plot ROC Curve
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, label=f'AUC = {roc_auc:.2f}')
plt.plot([0, 1], [0, 1], linestyle='--', color='gray', label='Random')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()
plt.show()
```

```
Confusion Matrix:
[[274   0]
 [  2   0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       274
           1       0.00      0.00      0.00         2

    accuracy                           0.99       276
   macro avg       0.50      0.50      0.50       276
weighted avg       0.99      0.99      0.99       276
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1344:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to
0.0 in labels with no predicted samples. Use `zero_division` parameter to
control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```



Future Scope

This future scope outlines key areas for improvement and expansion, addressing both model en-

hancement and practical deployment considerations. Regular updates and incorporation of the latest advancements in natural language processing (NLP) will contribute to the project's continued success and effectiveness.

[ ]: