Miguel Soniel & Jan Edgar Tupas Grp 13 **CS171 BM1 FA2**

Text Processing

Dataset includes 50 tweets, reduced to 43 after pre-processing. The inputs shown below are the first 20 tweets in the dataset

1. **Pre-process name: Lower Casing Text**
   The input:

['In 1995, I wrote about "the internet tidal wave." I even mentioned how the CD-ROM business would be dramatically impacted by the Internet. It's fun to look back and think about how far technology has come. https://t.co/IxntEFc6bh', 'The world has made incredible progress against polio and other infectious diseases. I am thankful for partners like @UNICEF_Nigeria for their work to ensure all children receive the vaccines they need. https://t.co/1ogLudBllI', 'Vaccines work to #EndPolio and protect communities from dozens of other diseases. I am heartened to see that the #BigCatchUp is underway. https://t.co/oaLSgsD6Wv', '.@bjornlomborg and I agree—the Global Goals are a phenomenal idea. This is the perfect time to assess our progress, recognize what's working, admit what isn't, and refine our approach so we can do the most good for the people most in need. https://t.co/H62VbgRkyl', 'Here's what I'm listening to this summer. Hopefully, you'll think it's not too bad for a granddad. https://t.co/Li0dybFd9J', 'I hope you have plenty of content to enjoy this summer. If you're looking to add to your list, I've got a few recommendations: https://t.co/KJo4B0BDP9 https://t.co/iXhEWv7sXm', 'I'm excited by the positive momentum around the Global Health Emergency Corps at last week's #WHA76. It's critical that countries don't let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats. https://t.co/qraK78Dz3p', 'For the past decade, I've recommended books to read each summer. This year, I'm mixing it up. https://t.co/Y4gdaGYxzU', 'It was an honor to meet Navajo Nation President @BuuVanNygren and First Lady @NygrenJasmine. https://t.co/COYtIsvGTP', 'Really good news: Renewables now account for about one third of electricity generation globally, and the share is increasing. https://t.co/U3zsdUjbUm', 'I often say Richard Feynman is the best teacher I never had. If you haven't already, check out his work. Beyond his immense talent as an educator, he's also quite the character. https://t.co/jC6iP7uaKT', 'My message to the class of 2023. https://t.co/Zduz6ldbN5', 'Ric was employee number 2 at Microsoft and a dear friend. He was also a tireless philanthropist, an advocate for LGBTQ+ rights and HIV/AIDS research, and an inspiration to me and so many others. His story is worth a watch: https://t.co/ASS0hp3Eug', 'Building a clean energy future will require a lot of new wires—the U.S. will need to grow the grid by almost two-thirds its current size in the next decade. https://t.co/XFkonneN3X', 'It's possible to build nuclear reactors without enormous costs—to our environment, our economy, or our health. This article details how it can be done to increase clean energy infrastructure. https://t.co/r01GHEg4Iq', 'I had the best day in Kemmerer, Wyoming. Over a coal plant tour, a site visit of the future Natrium plant, burgers, ice cream, and coffee, I got excited about the clean energy transition and the promise of next-generation nuclear: https://t.co/V8RO00hXKU https://t.co/kOInWbsJBE', 'Expanding access to single-dose HPV vaccines could save 45 million lives worldwide over the next 100 years. This is a great example of why I'm optimistic about the future of global health. https://t.co/DjaMRXOQle', 'I'm very encouraged by the E-MOTIVE Trial results released today at #IMNHC2023 to treat postpartum hemorrhage, the #1 cause of maternal mortality. This new approach led by @unibirmingham and @WHO could drastically improve women's chances of surviving childbirth globally.…

https://t.co/HP7fclWQV8', 'By prioritizing accessibility, affordability, and economic mobility, @NAU is redefining—and increasing—the value of a college degree. I can't wait to visit the school this weekend and see the incredible work they're doing up close. \nhttps://t.co/LwkO0KQUwl', 'Eradicating smallpox was hard, but it was also one of the most important achievements in human history. That makes it worth celebrating—and learning from. https://t.co/OeblsUug3w']

The code/process:

```
#In analyzing the dataset, the presence of uppercase characters showed no significance. Lower casing allows us to easily process text and handle sparsity.
df['Lower_Text'] = df['Text'].str.lower() #all strings under the text column are converted to lower case. They are stored in a new column named Lower_Text.

#Shows the text that are now in Lower Case. Values from df Lower_Text column replace the previous values in the list dataset.
dataset = df.loc[:,'Lower_Text'].tolist()
print(dataset)
```

The output:

['in 1995, i wrote about "the internet tidal wave." i even mentioned how the cd-rom business would be dramatically impacted by the internet. it's fun to look back and think about how far technology has come. https://t.co/ixntefc6bh', 'the world has made incredible progress against polio and other infectious diseases. i am thankful for partners like @unicef_nigeria for their work to ensure all children receive the vaccines they need. https://t.co/1ogludblli', 'vaccines work to #endpolio and protect communities from dozens of other diseases. i am heartened to see that the #bigcatchup is underway. https://t.co/oalsgsd6wv', '.@bjornlomborg and i agree—the global goals are a phenomenal idea. this is the perfect time to assess our progress, recognize what's working, admit what isn't, and refine our approach so we can do the most good for the people most in need. https://t.co/h62vbgrkyl', 'here's what i'm listening to this summer. hopefully, you'll think it's not too bad for a granddad. https://t.co/li0dybfd9j', 'i hope you have plenty of content to enjoy this summer. if you're looking to add to your list, i've got a few recommendations: https://t.co/kjo4b0bdp9 https://t.co/ixhewv7sxm', 'i'm excited by the positive momentum around the global health emergency corps at last week's #wha76. it's critical that countries don't let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats. https://t.co/qrak78dz3p', 'for the past decade, i've recommended books to read each summer. this year, i'm mixing it up. https://t.co/y4gdagyxzu', 'it was an honor to meet navajo nation president @buuvannygren and first lady @nygrenjasmine. https://t.co/coytisvgtp', 'really good news: renewables now account for about one third of electricity generation globally, and the share is increasing. https://t.co/u3zsdujbum', 'i often say richard feynman is the best teacher i never had. if you haven't already, check out his work. beyond his immense talent as an educator, he's also quite the character. https://t.co/jc6ip7uakt', 'my message to the class of 2023. https://t.co/zduz6ldbn5', 'ric was employee number 2 at microsoft and a dear friend. he was also a tireless philanthropist, an advocate for lgbtq+ rights and hiv/aids research, and an inspiration to me and so many others. his story is worth a watch: https://t.co/ass0hp3eug', 'building a clean energy future will require a lot of new wires—the u.s. will need to grow the grid by almost two-thirds its current size in the next decade. https://t.co/xfkonnen3x', 'it's possible to build nuclear reactors without enormous costs—to our environment, our economy, or our health. this article details how it can be done to increase clean energy

infrastructure. https://t.co/r01gheg4iq', 'i had the best day in kemmerer, wyoming. over a coal plant tour, a site visit of the future natrium plant, burgers, ice cream, and coffee, i got excited about the clean energy transition and the promise of next-generation nuclear: https://t.co/v8ro00hxku https://t.co/koinwbsjbe', 'expanding access to single-dose hpv vaccines could save 45 million lives worldwide over the next 100 years. this is a great example of why i'm optimistic about the future of global health. https://t.co/djamrxoqle', 'i'm very encouraged by the e-motive trial results released today at #imnhc2023 to treat postpartum hemorrhage, the #1 cause of maternal mortality. this new approach led by @unibirmingham and @who could drastically improve women's chances of surviving childbirth globally.… https://t.co/hp7fclwqv8', 'by prioritizing accessibility, affordability, and economic mobility, @nau is redefining—and increasing—the value of a college degree. i can't wait to visit the school this weekend and see the incredible work they're doing up close. \nhttps://t.co/lwko0kquwl', 'eradicating smallpox was hard, but it was also one of the most important achievements in human history. that makes it worth celebrating—and learning from. https://t.co/oeblsuug3w']

2. **Pre-process name: Punctuation, Target Mentions, and Hashtag Removal**
   The input:

['in 1995, i wrote about "the internet tidal wave." i even mentioned how the cd-rom business would be dramatically impacted by the internet. it's fun to look back and think about how far technology has come. https://t.co/ixntefc6bh', 'the world has made incredible progress against polio and other infectious diseases. i am thankful for partners like @unicef_nigeria for their work to ensure all children receive the vaccines they need. https://t.co/1ogludblli', 'vaccines work to #endpolio and protect communities from dozens of other diseases. i am heartened to see that the #bigcatchup is underway. https://t.co/oalsgsd6wv', '.@bjornlomborg and i agree—the global goals are a phenomenal idea. this is the perfect time to assess our progress, recognize what's working, admit what isn't, and refine our approach so we can do the most good for the people most in need. https://t.co/h62vbgrkyl', 'here's what i'm listening to this summer. hopefully, you'll think it's not too bad for a granddad. https://t.co/li0dybfd9j', 'i hope you have plenty of content to enjoy this summer. if you're looking to add to your list, i've got a few recommendations: https://t.co/kjo4b0bdp9 https://t.co/ixhewv7sxm', 'i'm excited by the positive momentum around the global health emergency corps at last week's #wha76. it's critical that countries don't let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats. https://t.co/qrak78dz3p', 'for the past decade, i've recommended books to read each summer. this year, i'm mixing it up. https://t.co/y4gdagyxzu', 'it was an honor to meet navajo nation president @buuvannygren and first lady @nygrenjasmine. https://t.co/coytisvgtp', 'really good news: renewables now account for about one third of electricity generation globally, and the share is increasing. https://t.co/u3zsdujbum', 'i often say richard feynman is the best teacher i never had. if you haven't already, check out his work. beyond his immense talent as an educator, he's also quite the character. https://t.co/jc6ip7uakt', 'my message to the class of 2023. https://t.co/zduz6ldbn5', 'ric was employee number 2 at microsoft and a dear friend. he was also a tireless philanthropist, an advocate for lgbtq+ rights and hiv/aids research, and an inspiration to me and so many others. his story is worth a watch: https://t.co/ass0hp3eug', 'building a clean energy future will require a lot of new wires—the u.s. will need to grow the grid by almost two-thirds its current size in the next decade. https://t.co/xfkonnen3x', 'it's possible to build nuclear reactors without enormous costs—to our environment, our economy, or our health. this article details how it can be done to increase clean energy infrastructure. https://t.co/r01gheg4iq', 'i had the best day in kemmerer, wyoming. over a coal plant tour, a site visit of the future natrium plant, burgers, ice cream,

and coffee, i got excited about the clean energy transition and the promise of next-generation nuclear: https://t.co/v8ro00hxku https://t.co/koinwbsjbe', 'expanding access to single-dose hpv vaccines could save 45 million lives worldwide over the next 100 years. this is a great example of why i'm optimistic about the future of global health. https://t.co/djamrxoqle', 'i'm very encouraged by the e-motive trial results released today at #imnhc2023 to treat postpartum hemorrhage, the #1 cause of maternal mortality. this new approach led by @unibirmingham and @who could drastically improve women's chances of surviving childbirth globally.… https://t.co/hp7fclwqv8', 'by prioritizing accessibility, affordability, and economic mobility, @nau is redefining—and increasing—the value of a college degree. i can't wait to visit the school this weekend and see the incredible work they're doing up close. \nhttps://t.co/lwko0kquwl', 'eradicating smallpox was hard, but it was also one of the most important achievements in human history. that makes it worth celebrating—and learning from. https://t.co/oeblsuug3w']

The code/process:

```
#Removing punctuation marks, target mentions and hashtags lessens noise in the dataset. We use regular expression and str.replace method to match and replace punctuation marks, mentions, and hashtags, removing them from text.
df['Clean_Text_A'] = df['Lower_Text'].str.replace ('@[A-Za-z0-9_]+','') #Removes all words that begin with @. @ signifies a target mention. The texts are stored in a new column named Clean_Text_A.
df['Clean_Text_A'] = df['Clean_Text_A'].str.replace ('#[A-Za-z0-9_]+','') #Removes all words that begin with #. # signifies a hashtage
df['Clean_Text_A'] = df['Clean_Text_A'].str.replace ('[^\w\s]','') #Removes all characters that do not match Unicode word characters.

#Shows the text without punctuation marks, target mentions, and hashtags. Values from df Clean_Text_A column replace the previous values in the list dataset.
dataset = df.loc[:,'Clean_Text_A'].tolist()
print(dataset)
```

The output:

['in 1995 i wrote about the internet tidal wave i even mentioned how the cdrom business would be dramatically impacted by the internet its fun to look back and think about how far technology has come httpstcoixntefc6bh', 'the world has made incredible progress against polio and other infectious diseases i am thankful for partners like for their work to ensure all children receive the vaccines they need httpstco1ogludblli', 'vaccines work to and protect communities from dozens of other diseases i am heartened to see that the is underway httpstcooalsgsd6wv', ' and i agreethe global goals are a phenomenal idea this is the perfect time to assess our progress recognize whats working admit what isnt and refine our approach so we can do the most good for the people most in need httpstcoh62vbgrkyl', 'heres what im listening to this summer hopefully youll think its not too bad for a granddad httpstcoli0dybfd9j', 'i hope you have plenty of content to enjoy this summer if youre looking to add to your list ive got a few recommendations httpstcokjo4b0bdp9 httpstcoixhewv7sxm', 'im excited by the positive momentum around the global health emergency corps at last weeks its critical that countries dont let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats httpstcoqrak78dz3p', 'for the past decade ive recommended books to read each summer this year im mixing it up httpstcoy4gdagyxzu', 'it was an honor to meet navajo nation president and first lady httpstcocoytisvgtp', 'really good news

renewables now account for about one third of electricity generation globally and the share is increasing httpstcou3zsdujbum', 'i often say richard feynman is the best teacher i never had if you havent already check out his work beyond his immense talent as an educator hes also quite the character httpstcojc6ip7uakt', 'my message to the class of 2023 httpstcozduz6ldbn5', 'ric was employee number 2 at microsoft and a dear friend he was also a tireless philanthropist an advocate for lgbtq rights and hivaids research and an inspiration to me and so many others his story is worth a watch httpstcoass0hp3eug', 'building a clean energy future will require a lot of new wiresthe us will need to grow the grid by almost twothirds its current size in the next decade httpstcoxfkonnen3x', 'its possible to build nuclear reactors without enormous coststo our environment our economy or our health this article details how it can be done to increase clean energy infrastructure httpstcor01gheg4iq', 'i had the best day in kemmerer wyoming over a coal plant tour a site visit of the future natrium plant burgers ice cream and coffee i got excited about the clean energy transition and the promise of nextgeneration nuclear httpstcov8ro00hxku httpstcokoinwbsjbe', 'expanding access to singledose hpv vaccines could save 45 million lives worldwide over the next 100 years this is a great example of why im optimistic about the future of global health httpstcodjamrxoqle', 'im very encouraged by the emotive trial results released today at to treat postpartum hemorrhage the cause of maternal mortality this new approach led by and could drastically improve womens chances of surviving childbirth globally httpstcohp7fclwqv8', 'by prioritizing accessibility affordability and economic mobility is redefiningand increasingthe value of a college degree i cant wait to visit the school this weekend and see the incredible work theyre doing up close \nhttpstcolwko0kquwl', 'eradicating smallpox was hard but it was also one of the most important achievements in human history that makes it worth celebratingand learning from httpstcooeblsuug3w']

3. **Pre-process name: URL, Newline, Re-Tweets Removal**
   The input:

['in 1995 i wrote about the internet tidal wave i even mentioned how the cdrom business would be dramatically impacted by the internet its fun to look back and think about how far technology has come httpstcoixntefc6bh', 'the world has made incredible progress against polio and other infectious diseases i am thankful for partners like for their work to ensure all children receive the vaccines they need httpstco1ogludblli', 'vaccines work to and protect communities from dozens of other diseases i am heartened to see that the is underway httpstcooalsgsd6wv', ' and i agreethe global goals are a phenomenal idea this is the perfect time to assess our progress recognize whats working admit what isnt and refine our approach so we can do the most good for the people most in need httpstcoh62vbgrkyl', 'heres what im listening to this summer hopefully youll think its not too bad for a granddad httpstcoli0dybfd9j', 'i hope you have plenty of content to enjoy this summer if youre looking to add to your list ive got a few recommendations httpstcokjo4b0bdp9 httpstcoixhewv7sxm', 'im excited by the positive momentum around the global health emergency corps at last weeks its critical that countries dont let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats httpstcoqrak78dz3p', 'for the past decade ive recommended books to read each summer this year im mixing it up httpstcoy4gdagyxzu', 'it was an honor to meet navajo nation president and first lady httpstcocoytisvgtp', 'really good news renewables now account for about one third of electricity generation globally and the share is increasing httpstcou3zsdujbum', 'i often say richard feynman is the best teacher i never had if you havent already check out his work beyond his immense talent as an educator hes also quite the character httpstcojc6ip7uakt', 'my message to the class of 2023 httpstcozduz6ldbn5', 'ric was employee number 2 at microsoft and a dear friend he was also a tireless philanthropist an advocate for lgbtq rights and hivaids research and an inspiration to me and so many others his story is worth a

```
watch httpstcoass0hp3eug', 'building a clean energy future will require a lot of new wiresthe us will need to grow the grid by almost twothirds its current size in the next decade httpstcoxfkonnen3x', 'its possible to build nuclear reactors without enormous coststo our environment our economy or our health this article details how it can be done to increase clean energy infrastructure httpstcor01gheg4iq', 'i had the best day in kemmerer wyoming over a coal plant tour a site visit of the future natrium plant burgers ice cream and coffee i got excited about the clean energy transition and the promise of nextgeneration nuclear httpstcov8ro00hxku httpstcokoinwbsjbe', 'expanding access to singledose hpv vaccines could save 45 million lives worldwide over the next 100 years this is a great example of why im optimistic about the future of global health httpstcodjamrxoqle', 'im very encouraged by the emotive trial results released today at  to treat postpartum hemorrhage the cause of maternal mortality this new approach led by  and  could drastically improve womens chances of surviving childbirth globally httpstcohp7fclwqv8', 'by prioritizing accessibility affordability and economic mobility  is redefiningand increasingthe value of a college degree i cant wait to visit the school this weekend and see the incredible work theyre doing up close \nhttpstcolwko0kquwl', 'eradicating smallpox was hard but it was also one of the most important achievements in human history that makes it worth celebratingand learning from httpstcooeblsuug3w']
```

The code/process:

```
#Removing url, newlines and re-tweets further lessens noise in the dataset. Re-tweets are not original text posted by the user which causes sparsity. We use regular expression and str.replace method to match and replace words that are urls, newline character \n, and the entire row of re-tweets, removing them from text.
df['Clean_Text_B'] = df['Clean_Text_A'].str.replace('https[A-Za-z0-9_]+','') #Removes all words that begin with https. https signifies a url. The texts are stored in a new column named Clean_Text_B
df['Clean_Text_B'] = df['Clean_Text_B'].str.replace('\n','') #Removes all \n characters. \n signifies newline.
df.drop(df[df['Clean_Text_B'].str.startswith('rt')].index, inplace = True)
#Identifies all texts that begin with rt and removes the entire row from the data frame df. rt signifies a re-tweet.
df.reset_index(inplace = True) #Reset the index because of the rows Clean_Text_B.

#Shows the text without urls, newlines, and are not re-tweets. Values from df Clean_Text_B column replace the previous values in the list dataset.
dataset = df.loc[:,'Clean_Text_B'].tolist()
print(dataset)
```

The output:

```
['in 1995 i wrote about the internet tidal wave i even mentioned how the cdrom business would be dramatically impacted by the internet its fun to look back and think about how far technology has come ', 'the world has made incredible progress against polio and other infectious diseases i am thankful for partners like  for their work to ensure all children receive the vaccines they need ', 'vaccines work to and protect communities from dozens of other diseases i am heartened to see that the is underway ', ' and i agreethe global goals are a phenomenal idea this is the perfect time to assess our progress recognize whats working admit what isnt and refine our approach so we can do the most good for the people most in need ', 'heres what im listening to this summer hopefully youll think its not too bad for a granddad ', 'i hope you have plenty of content to enjoy this summer if youre looking to add to
```

your list ive got a few recommendations ', 'im excited by the positive momentum around the global health emergency corps at last weeks its critical that countries dont let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats ', 'for the past decade ive recommended books to read each summer this year im mixing it up ', 'it was an honor to meet navajo nation president and first lady ', 'really good news renewables now account for about one third of electricity generation globally and the share is increasing ', 'i often say richard feynman is the best teacher i never had if you havent already check out his work beyond his immense talent as an educator hes also quite the character ', 'my message to the class of 2023 ', 'ric was employee number 2 at microsoft and a dear friend he was also a tireless philanthropist an advocate for lgbtq rights and hivaids research and an inspiration to me and so many others his story is worth a watch ', 'building a clean energy future will require a lot of new wiresthe us will need to grow the grid by almost twothirds its current size in the next decade ', 'its possible to build nuclear reactors without enormous coststo our environment our economy or our health this article details how it can be done to increase clean energy infrastructure ', 'i had the best day in kemmerer wyoming over a coal plant tour a site visit of the future natrium plant burgers ice cream and coffee i got excited about the clean energy transition and the promise of nextgeneration nuclear ', 'expanding access to singledose hpv vaccines could save 45 million lives worldwide over the next 100 years this is a great example of why im optimistic about the future of global health ', 'im very encouraged by the emotive trial results released today at to treat postpartum hemorrhage the cause of maternal mortality this new approach led by and could drastically improve womens chances of surviving childbirth globally ', 'by prioritizing accessibility affordability and economic mobility is redefiningand increasingthe value of a college degree i cant wait to visit the school this weekend and see the incredible work theyre doing up close ', 'eradicating smallpox was hard but it was also one of the most important achievements in human history that makes it worth celebratingand learning from ']

4. **Pre-process name: Stop Words Removal**
   The input:

['in 1995 i wrote about the internet tidal wave i even mentioned how the cdrom business would be dramatically impacted by the internet its fun to look back and think about how far technology has come ', 'the world has made incredible progress against polio and other infectious diseases i am thankful for partners like for their work to ensure all children receive the vaccines they need ', 'vaccines work to and protect communities from dozens of other diseases i am heartened to see that the is underway ', ' and i agreethe global goals are a phenomenal idea this is the perfect time to assess our progress recognize whats working admit what isnt and refine our approach so we can do the most good for the people most in need ', 'heres what im listening to this summer hopefully youll think its not too bad for a granddad ', 'i hope you have plenty of content to enjoy this summer if youre looking to add to your list ive got a few recommendations ', 'im excited by the positive momentum around the global health emergency corps at last weeks its critical that countries dont let their guard down and join the effort to build a strong global network to help prevent and contain pandemic threats ', 'for the past decade ive recommended books to read each summer this year im mixing it up ', 'it was an honor to meet navajo nation president and first lady ', 'really good news renewables now account for about one third of electricity generation globally and the share is increasing ', 'i often say richard feynman is the best teacher i never had if you havent already check out his work beyond his immense talent as an educator hes also quite the character ', 'my message to the class of 2023 ', 'ric was employee number 2 at microsoft and a dear friend he was also a tireless philanthropist an advocate for

lgbtq rights and hivaids research and an inspiration to me and so many others his story is worth a watch ', 'building a clean energy future will require a lot of new wiresthe us will need to grow the grid by almost twothirds its current size in the next decade ', 'its possible to build nuclear reactors without enormous coststo our environment our economy or our health this article details how it can be done to increase clean energy infrastructure ', 'i had the best day in kemmerer wyoming over a coal plant tour a site visit of the future natrium plant burgers ice cream and coffee i got excited about the clean energy transition and the promise of nextgeneration nuclear ', 'expanding access to singledose hpv vaccines could save 45 million lives worldwide over the next 100 years this is a great example of why im optimistic about the future of global health ', 'im very encouraged by the emotive trial results released today at to treat postpartum hemorrhage the cause of maternal mortality this new approach led by and could drastically improve womens chances of surviving childbirth globally ', 'by prioritizing accessibility affordability and economic mobility is redefiningand increasingthe value of a college degree i cant wait to visit the school this weekend and see the incredible work theyre doing up close ', 'eradicating smallpox was hard but it was also one of the most important achievements in human history that makes it worth celebratingand learning from ']

The code/process:

```
#Stop words are common, low information words and should be removed to further reduce noise and sparity. We use all nltk pre-defined stop words to match and remove them from text.
stop_words = stopwords.words('english') #all pre-defined stopwords are stored in list stop_words

#remove_stop function runs through all words in text and joins only words that are not in list stop_words.
def remove_stop(x):
  return ' '.join([word for word in str(x).split() if word not in stop_words])

#remove_stop function is applied in the values of column Clean_Text_B. Results are stored in a new column called Clean_Text_C.
df['Clean_Text_C'] = df['Clean_Text_B'].apply(lambda x: remove_stop(x))

#Shows the text without stop words. Values from df Clean_Text_C column replace the previous values in the list dataset.
dataset = df.loc[:,'Clean_Text_C'].tolist()
print(dataset)
```

The output:

['1995 wrote internet tidal wave even mentioned cdrom business would dramatically impacted internet fun look back think far technology come', 'world made incredible progress polio infectious diseases thankful partners like work ensure children receive vaccines need', 'vaccines work protect communities dozens diseases heartened see underway', 'agreethe global goals phenomenal idea perfect time assess progress recognize whats working admit isnt refine approach good people need', 'heres im listening summer hopefully youll think bad granddad', 'hope plenty content enjoy summer youre looking add list ive got recommendations', 'im excited positive momentum around global health emergency corps last weeks critical countries dont let guard join effort build strong global network help prevent contain pandemic threats', 'past

decade ive recommended books read summer year im mixing', 'honor meet navajo nation president first lady', 'really good news renewables account one third electricity generation globally share increasing', 'often say richard feynman best teacher never havent already check work beyond immense talent educator hes also quite character', 'message class 2023', 'ric employee number 2 microsoft dear friend also tireless philanthropist advocate lgbtq rights hivaids research inspiration many others story worth watch', 'building clean energy future require lot new wiresthe us need grow grid almost twothirds current size next decade', 'possible build nuclear reactors without enormous coststo environment economy health article details done increase clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit future natrium plant burgers ice cream coffee got excited clean energy transition promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save 45 million lives worldwide next 100 years great example im optimistic future global health', 'im encouraged emotive trial results released today treat postpartum hemorrhage cause maternal mortality new approach led could drastically improve womens chances surviving childbirth globally', 'prioritizing accessibility affordability economic mobility redefiningand increasingthe value college degree cant wait visit school weekend see incredible work theyre close', 'eradicating smallpox hard also one important achievements human history makes worth celebratingand learning']

5. **Pre-process name: Numerical Data Removal**
   The input:

['1995 wrote internet tidal wave even mentioned cdrom business would dramatically impacted internet fun look back think far technology come', 'world made incredible progress polio infectious diseases thankful partners like work ensure children receive vaccines need', 'vaccines work protect communities dozens diseases heartened see underway', 'agreethe global goals phenomenal idea perfect time assess progress recognize whats working admit isnt refine approach good people need', 'heres im listening summer hopefully youll think bad granddad', 'hope plenty content enjoy summer youre looking add list ive got recommendations', 'im excited positive momentum around global health emergency corps last weeks critical countries dont let guard join effort build strong global network help prevent contain pandemic threats', 'past decade ive recommended books read summer year im mixing', 'honor meet navajo nation president first lady', 'really good news renewables account one third electricity generation globally share increasing', 'often say richard feynman best teacher never havent already check work beyond immense talent educator hes also quite character', 'message class 2023', 'ric employee number 2 microsoft dear friend also tireless philanthropist advocate lgbtq rights hivaids research inspiration many others story worth watch', 'building clean energy future require lot new wiresthe us need grow grid almost twothirds current size next decade', 'possible build nuclear reactors without enormous coststo environment economy health article details done increase clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit future natrium plant burgers ice cream coffee got excited clean energy transition promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save 45 million lives worldwide next 100 years great example im optimistic future global health', 'im encouraged emotive trial results released today treat postpartum hemorrhage cause maternal mortality new approach led could drastically improve womens chances surviving childbirth globally', 'prioritizing accessibility affordability economic mobility redefiningand increasingthe value college degree cant wait visit school weekend see incredible work theyre close', 'eradicating smallpox hard also one important achievements human history makes worth celebratingand learning']

The code/process:

```
#In the chosen dataset, numerical data do not signify keywords necessary for nlp
resulting in further noise and sparity. We use regular expression to match and
replace numerical data, removing them from text.
df['No_Num_Text'] = df['Clean_Text_C'].str.replace('[0-9]','') #Removes all numbers.
The texts are stored in a new column named No_Num_Text

#Shows the text without numerical data. Values from df No_Num_Text column replace the
previous values in the list dataset.
dataset = df.loc[:,'No_Num_Text'].tolist()
print(dataset)
```

The output:

```
[' wrote internet tidal wave even mentioned cdrom business would dramatically
impacted internet fun look back think far technology come', 'world made incredible
progress polio infectious diseases thankful partners like work ensure children
receive vaccines need', 'vaccines work protect communities dozens diseases heartened
see underway', 'agreethe global goals phenomenal idea perfect time assess progress
recognize whats working admit isnt refine approach good people need', 'heres im
listening summer hopefully youll think bad granddad', 'hope plenty content enjoy
summer youre looking add list ive got recommendations', 'im excited positive momentum
around global health emergency corps last weeks critical countries dont let guard
join effort build strong global network help prevent contain pandemic threats', 'past
decade ive recommended books read summer year im mixing', 'honor meet navajo nation
president first lady', 'really good news renewables account one third electricity
generation globally share increasing', 'often say richard feynman best teacher never
havent already check work beyond immense talent educator hes also quite character',
'message class ', 'ric employee number  microsoft dear friend also tireless
philanthropist advocate lgbtq rights hivaids research inspiration many others story
worth watch', 'building clean energy future require lot new wiresthe us need grow
grid almost twothirds current size next decade', 'possible build nuclear reactors
without enormous coststo environment economy health article details done increase
clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit
future natrium plant burgers ice cream coffee got excited clean energy transition
promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save
million lives worldwide next  years great example im optimistic future global
health', 'im encouraged emotive trial results released today treat postpartum
hemorrhage cause maternal mortality new approach led could drastically improve womens
chances surviving childbirth globally', 'prioritizing accessibility affordability
economic mobility redefiningand increasingthe value college degree cant wait visit
school weekend see incredible work theyre close', 'eradicating smallpox hard also one
important achievements human history makes worth celebratingand learning']
```

6. **Pre-process name: White Space Removal**
   The input:

```
[' wrote internet tidal wave even mentioned cdrom business would dramatically
impacted internet fun look back think far technology come', 'world made incredible
progress polio infectious diseases thankful partners like work ensure children
receive vaccines need', 'vaccines work protect communities dozens diseases heartened
see underway', 'agreethe global goals phenomenal idea perfect time assess progress
recognize whats working admit isnt refine approach good people need', 'heres im
listening summer hopefully youll think bad granddad', 'hope plenty content enjoy
summer youre looking add list ive got recommendations', 'im excited positive momentum
around global health emergency corps last weeks critical countries dont let guard
```

```
join effort build strong global network help prevent contain pandemic threats', 'past decade ive recommended books read summer year im mixing', 'honor meet navajo nation president first lady', 'really good news renewables account one third electricity generation globally share increasing', 'often say richard feynman best teacher never havent already check work beyond immense talent educator hes also quite character', 'message class ', 'ric employee number  microsoft dear friend also tireless philanthropist advocate lgbtq rights hivaids research inspiration many others story worth watch', 'building clean energy future require lot new wiresthe us need grow grid almost twothirds current size next decade', 'possible build nuclear reactors without enormous coststo environment economy health article details done increase clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit future natrium plant burgers ice cream coffee got excited clean energy transition promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save million lives worldwide next  years great example im optimistic future global health', 'im encouraged emotive trial results released today treat postpartum hemorrhage cause maternal mortality new approach led could drastically improve womens chances surviving childbirth globally', 'prioritizing accessibility affordability economic mobility redefiningand increasingthe value college degree cant wait visit school weekend see incredible work theyre close', 'eradicating smallpox hard also one important achievements human history makes worth celebratingand learning']
```

The code/process:

```
#White spaces should be removed from text to improve and simplify text preprocessing. We use regular expression to match and replace all white spaces, removing them from text.
df['No_White_Text'] = df['No_Num_Text'].str.replace(' +',' ').str.strip() #Removes all white spaces. The texts are stored in a new column named No_White_Text

#Shows the text without white spaces. Values from df No_White_Text column replace the previous values in the list dataset.
dataset = df.loc[:,'No_White_Text'].tolist()
print(dataset[0:20])
```

The output:

```
['wrote internet tidal wave even mentioned cdrom business would dramatically impacted internet fun look back think far technology come', 'world made incredible progress polio infectious diseases thankful partners like work ensure children receive vaccines need', 'vaccines work protect communities dozens diseases heartened see underway', 'agreethe global goals phenomenal idea perfect time assess progress recognize whats working admit isnt refine approach good people need', 'heres im listening summer hopefully youll think bad granddad', 'hope plenty content enjoy summer youre looking add list ive got recommendations', 'im excited positive momentum around global health emergency corps last weeks critical countries dont let guard join effort build strong global network help prevent contain pandemic threats', 'past decade ive recommended books read summer year im mixing', 'honor meet navajo nation president first lady', 'really good news renewables account one third electricity generation globally share increasing', 'often say richard feynman best teacher never havent already check work beyond immense talent educator hes also quite character', 'message class', 'ric employee number microsoft dear friend also tireless philanthropist advocate lgbtq rights hivaids research inspiration many others story worth watch', 'building clean energy future require lot new wiresthe us need grow grid almost twothirds current size next decade', 'possible build nuclear reactors without enormous coststo environment economy health article details done increase
```

clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit future natrium plant burgers ice cream coffee got excited clean energy transition promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save million lives worldwide next years great example im optimistic future global health', 'im encouraged emotive trial results released today treat postpartum hemorrhage cause maternal mortality new approach led could drastically improve womens chances surviving childbirth globally', 'prioritizing accessibility affordability economic mobility redefiningand increasingthe value college degree cant wait visit school weekend see incredible work theyre close', 'eradicating smallpox hard also one important achievements human history makes worth celebratingand learning']

7. **Pre-process name: Tokenization**
   The input:

['wrote internet tidal wave even mentioned cdrom business would dramatically impacted internet fun look back think far technology come', 'world made incredible progress polio infectious diseases thankful partners like work ensure children receive vaccines need', 'vaccines work protect communities dozens diseases heartened see underway', 'agreethe global goals phenomenal idea perfect time assess progress recognize whats working admit isnt refine approach good people need', 'heres im listening summer hopefully youll think bad granddad', 'hope plenty content enjoy summer youre looking add list ive got recommendations', 'im excited positive momentum around global health emergency corps last weeks critical countries dont let guard join effort build strong global network help prevent contain pandemic threats', 'past decade ive recommended books read summer year im mixing', 'honor meet navajo nation president first lady', 'really good news renewables account one third electricity generation globally share increasing', 'often say richard feynman best teacher never havent already check work beyond immense talent educator hes also quite character', 'message class', 'ric employee number microsoft dear friend also tireless philanthropist advocate lgbtq rights hivaids research inspiration many others story worth watch', 'building clean energy future require lot new wiresthe us need grow grid almost twothirds current size next decade', 'possible build nuclear reactors without enormous coststo environment economy health article details done increase clean energy infrastructure', 'best day kemmerer wyoming coal plant tour site visit future natrium plant burgers ice cream coffee got excited clean energy transition promise nextgeneration nuclear', 'expanding access singledose hpv vaccines could save million lives worldwide next years great example im optimistic future global health', 'im encouraged emotive trial results released today treat postpartum hemorrhage cause maternal mortality new approach led could drastically improve womens chances surviving childbirth globally', 'prioritizing accessibility affordability economic mobility redefiningand increasingthe value college degree cant wait visit school weekend see incredible work theyre close', 'eradicating smallpox hard also one important achievements human history makes worth celebratingand learning']

The code/process:

```
#Tokenization splits text into tokens which are individual words to easily assign meaning when used for nlp. We use re.split method and regular expression to tokenize the text into words.

#We apply re.split method to df column No_White_Text to tokenize its values. The results are stored in a new column named Word_Token
df['Word_Token'] = df['No_White_Text'].apply(lambda x: re.split('\W+',x))

#Shows the text that has been tokenized. Values from df Word_Token column replace the previous values in the list dataset.
```

```
dataset = df.loc[:,'Word_Token'].tolist()
dataset = [item for sublist in dataset for item in sublist]
print(dataset)
```

The output:

```
[['wrote', 'internet', 'tidal', 'wave', 'even', 'mentioned', 'cdrom', 'business',
'would', 'dramatically', 'impacted', 'internet', 'fun', 'look', 'back', 'think',
'far', 'technology', 'come'], ['world', 'made', 'incredible', 'progress', 'polio',
'infectious', 'diseases', 'thankful', 'partners', 'like', 'work', 'ensure',
'children', 'receive', 'vaccines', 'need'], ['vaccines', 'work', 'protect',
'communities', 'dozens', 'diseases', 'heartened', 'see', 'underway'], ['agreethe',
'global', 'goals', 'phenomenal', 'idea', 'perfect', 'time', 'assess', 'progress',
'recognize', 'whats', 'working', 'admit', 'isnt', 'refine', 'approach', 'good',
'people', 'need'], ['heres', 'im', 'listening', 'summer', 'hopefully', 'youll',
'think', 'bad', 'granddad'], ['hope', 'plenty', 'content', 'enjoy', 'summer',
'youre', 'looking', 'add', 'list', 'ive', 'got', 'recommendations'], ['im',
'excited', 'positive', 'momentum', 'around', 'global', 'health', 'emergency',
'corps', 'last', 'weeks', 'critical', 'countries', 'dont', 'let', 'guard', 'join',
'effort', 'build', 'strong', 'global', 'network', 'help', 'prevent', 'contain',
'pandemic', 'threats'], ['past', 'decade', 'ive', 'recommended', 'books', 'read',
'summer', 'year', 'im', 'mixing'], ['honor', 'meet', 'navajo', 'nation', 'president',
'first', 'lady'], ['really', 'good', 'news', 'renewables', 'account', 'one', 'third',
'electricity', 'generation', 'globally', 'share', 'increasing'], ['often', 'say',
'richard', 'feynman', 'best', 'teacher', 'never', 'havent', 'already', 'check',
'work', 'beyond', 'immense', 'talent', 'educator', 'hes', 'also', 'quite',
'character'], ['message', 'class'], ['ric', 'employee', 'number', 'microsoft',
'dear', 'friend', 'also', 'tireless', 'philanthropist', 'advocate', 'lgbtq',
'rights', 'hivaids', 'research', 'inspiration', 'many', 'others', 'story', 'worth',
'watch'], ['building', 'clean', 'energy', 'future', 'require', 'lot', 'new',
'wiresthe', 'us', 'need', 'grow', 'grid', 'almost', 'twothirds', 'current', 'size',
'next', 'decade'], ['possible', 'build', 'nuclear', 'reactors', 'without',
'enormous', 'coststo', 'environment', 'economy', 'health', 'article', 'details',
'done', 'increase', 'clean', 'energy', 'infrastructure'], ['best', 'day', 'kemmerer',
'wyoming', 'coal', 'plant', 'tour', 'site', 'visit', 'future', 'natrium', 'plant',
'burgers', 'ice', 'cream', 'coffee', 'got', 'excited', 'clean', 'energy',
'transition', 'promise', 'nextgeneration', 'nuclear'], ['expanding', 'access',
'singledose', 'hpv', 'vaccines', 'could', 'save', 'million', 'lives', 'worldwide',
'next', 'years', 'great', 'example', 'im', 'optimistic', 'future', 'global',
'health'], ['im', 'encouraged', 'emotive', 'trial', 'results', 'released', 'today',
'treat', 'postpartum', 'hemorrhage', 'cause', 'maternal', 'mortality', 'new',
'approach', 'led', 'could', 'drastically', 'improve', 'womens', 'chances',
'surviving', 'childbirth', 'globally'], ['prioritizing', 'accessibility',
'affordability', 'economic', 'mobility', 'redefiningand', 'increasingthe', 'value',
'college', 'degree', 'cant', 'wait', 'visit', 'school', 'weekend', 'see',
'incredible', 'work', 'theyre', 'close'], ['eradicating', 'smallpox', 'hard', 'also',
'one', 'important', 'achievements', 'human', 'history', 'makes', 'worth',
'celebratingand', 'learning']]
```

8. **Pre-process name: Lemmatization**

   The input:

```
[['wrote', 'internet', 'tidal', 'wave', 'even', 'mentioned', 'cdrom', 'business',
'would', 'dramatically', 'impacted', 'internet', 'fun', 'look', 'back', 'think',
'far', 'technology', 'come'], ['world', 'made', 'incredible', 'progress', 'polio',
```

```
'infectious', 'diseases', 'thankful', 'partners', 'like', 'work', 'ensure',
'children', 'receive', 'vaccines', 'need'], ['vaccines', 'work', 'protect',
'communities', 'dozens', 'diseases', 'heartened', 'see', 'underway'], ['agreethe',
'global', 'goals', 'phenomenal', 'idea', 'perfect', 'time', 'assess', 'progress',
'recognize', 'whats', 'working', 'admit', 'isnt', 'refine', 'approach', 'good',
'people', 'need'], ['heres', 'im', 'listening', 'summer', 'hopefully', 'youll',
'think', 'bad', 'granddad'], ['hope', 'plenty', 'content', 'enjoy', 'summer',
'youre', 'looking', 'add', 'list', 'ive', 'got', 'recommendations'], ['im',
'excited', 'positive', 'momentum', 'around', 'global', 'health', 'emergency',
'corps', 'last', 'weeks', 'critical', 'countries', 'dont', 'let', 'guard', 'join',
'effort', 'build', 'strong', 'global', 'network', 'help', 'prevent', 'contain',
'pandemic', 'threats'], ['past', 'decade', 'ive', 'recommended', 'books', 'read',
'summer', 'year', 'im', 'mixing'], ['honor', 'meet', 'navajo', 'nation', 'president',
'first', 'lady'], ['really', 'good', 'news', 'renewables', 'account', 'one', 'third',
'electricity', 'generation', 'globally', 'share', 'increasing'], ['often', 'say',
'richard', 'feynman', 'best', 'teacher', 'never', 'havent', 'already', 'check',
'work', 'beyond', 'immense', 'talent', 'educator', 'hes', 'also', 'quite',
'character'], ['message', 'class'], ['ric', 'employee', 'number', 'microsoft',
'dear', 'friend', 'also', 'tireless', 'philanthropist', 'advocate', 'lgbtq',
'rights', 'hivaids', 'research', 'inspiration', 'many', 'others', 'story', 'worth',
'watch'], ['building', 'clean', 'energy', 'future', 'require', 'lot', 'new',
'wiresthe', 'us', 'need', 'grow', 'grid', 'almost', 'twothirds', 'current', 'size',
'next', 'decade'], ['possible', 'build', 'nuclear', 'reactors', 'without',
'enormous', 'coststo', 'environment', 'economy', 'health', 'article', 'details',
'done', 'increase', 'clean', 'energy', 'infrastructure'], ['best', 'day', 'kemmerer',
'wyoming', 'coal', 'plant', 'tour', 'site', 'visit', 'future', 'natrium', 'plant',
'burgers', 'ice', 'cream', 'coffee', 'got', 'excited', 'clean', 'energy',
'transition', 'promise', 'nextgeneration', 'nuclear'], ['expanding', 'access',
'singledose', 'hpv', 'vaccines', 'could', 'save', 'million', 'lives', 'worldwide',
'next', 'years', 'great', 'example', 'im', 'optimistic', 'future', 'global',
'health'], ['im', 'encouraged', 'emotive', 'trial', 'results', 'released', 'today',
'treat', 'postpartum', 'hemorrhage', 'cause', 'maternal', 'mortality', 'new',
'approach', 'led', 'could', 'drastically', 'improve', 'womens', 'chances',
'surviving', 'childbirth', 'globally'], ['prioritizing', 'accessibility',
'affordability', 'economic', 'mobility', 'redefiningand', 'increasingthe', 'value',
'college', 'degree', 'cant', 'wait', 'visit', 'school', 'weekend', 'see',
'incredible', 'work', 'theyre', 'close'], ['eradicating', 'smallpox', 'hard', 'also',
'one', 'important', 'achievements', 'human', 'history', 'makes', 'worth',
'celebratingand', 'learning']]
```

The code/process:

```
#Lemmatization reduces a word to its root form and meaning to improve the process of
nlp. We use WordNetLemmatizer method from nltk to lemmatize the words if possible.

#lemmatize_word function uses WordNetLemmatizer method to lemmatize the word and
stores it into list lemmatized_word. When the function is called, the list
lemmatized_word is returned.
def lemmatize_word(txt):
  lemmatized_word = [WordNetLemmatizer().lemmatize(word) for word in txt]
  return lemmatized_word

#We apply lemmatize_word function to the values of df column Word_Token. The results
are stored in a new column named Word_Lemmatize
df['Word_Lemmatize'] = df['Word_Token'].apply(lambda x: lemmatize_word(x))
```

```
#Shows the text that has been lemmatized. Values from df Word_Lemmatize column
replace the previous values in the list dataset.
dataset = df.loc[:,'Word_Lemmatize'].tolist()
print(dataset)
```

The output:

```
[['wrote', 'internet', 'tidal', 'wave', 'even', 'mentioned', 'cdrom', 'business',
'would', 'dramatically', 'impacted', 'internet', 'fun', 'look', 'back', 'think',
'far', 'technology', 'come'], ['world', 'made', 'incredible', 'progress', 'polio',
'infectious', 'disease', 'thankful', 'partner', 'like', 'work', 'ensure', 'child',
'receive', 'vaccine', 'need'], ['vaccine', 'work', 'protect', 'community', 'dozen',
'disease', 'heartened', 'see', 'underway'], ['agreethe', 'global', 'goal',
'phenomenal', 'idea', 'perfect', 'time', 'ass', 'progress', 'recognize', 'whats',
'working', 'admit', 'isnt', 'refine', 'approach', 'good', 'people', 'need'], ['here',
'im', 'listening', 'summer', 'hopefully', 'youll', 'think', 'bad', 'granddad'],
['hope', 'plenty', 'content', 'enjoy', 'summer', 'youre', 'looking', 'add', 'list',
'ive', 'got', 'recommendation'], ['im', 'excited', 'positive', 'momentum', 'around',
'global', 'health', 'emergency', 'corp', 'last', 'week', 'critical', 'country',
'dont', 'let', 'guard', 'join', 'effort', 'build', 'strong', 'global', 'network',
'help', 'prevent', 'contain', 'pandemic', 'threat'], ['past', 'decade', 'ive',
'recommended', 'book', 'read', 'summer', 'year', 'im', 'mixing'], ['honor', 'meet',
'navajo', 'nation', 'president', 'first', 'lady'], ['really', 'good', 'news',
'renewables', 'account', 'one', 'third', 'electricity', 'generation', 'globally',
'share', 'increasing'], ['often', 'say', 'richard', 'feynman', 'best', 'teacher',
'never', 'havent', 'already', 'check', 'work', 'beyond', 'immense', 'talent',
'educator', 'he', 'also', 'quite', 'character'], ['message', 'class'], ['ric',
'employee', 'number', 'microsoft', 'dear', 'friend', 'also', 'tireless',
'philanthropist', 'advocate', 'lgbtq', 'right', 'hivaids', 'research', 'inspiration',
'many', 'others', 'story', 'worth', 'watch'], ['building', 'clean', 'energy',
'future', 'require', 'lot', 'new', 'wiresthe', 'u', 'need', 'grow', 'grid', 'almost',
'twothirds', 'current', 'size', 'next', 'decade'], ['possible', 'build', 'nuclear',
'reactor', 'without', 'enormous', 'coststo', 'environment', 'economy', 'health',
'article', 'detail', 'done', 'increase', 'clean', 'energy', 'infrastructure'],
['best', 'day', 'kemmerer', 'wyoming', 'coal', 'plant', 'tour', 'site', 'visit',
'future', 'natrium', 'plant', 'burger', 'ice', 'cream', 'coffee', 'got', 'excited',
'clean', 'energy', 'transition', 'promise', 'nextgeneration', 'nuclear'],
['expanding', 'access', 'singledose', 'hpv', 'vaccine', 'could', 'save', 'million',
'life', 'worldwide', 'next', 'year', 'great', 'example', 'im', 'optimistic',
'future', 'global', 'health'], ['im', 'encouraged', 'emotive', 'trial', 'result',
'released', 'today', 'treat', 'postpartum', 'hemorrhage', 'cause', 'maternal',
'mortality', 'new', 'approach', 'led', 'could', 'drastically', 'improve', 'woman',
'chance', 'surviving', 'childbirth', 'globally'], ['prioritizing', 'accessibility',
'affordability', 'economic', 'mobility', 'redefiningand', 'increasingthe', 'value',
'college', 'degree', 'cant', 'wait', 'visit', 'school', 'weekend', 'see',
'incredible', 'work', 'theyre', 'close'], ['eradicating', 'smallpox', 'hard', 'also',
'one', 'important', 'achievement', 'human', 'history', 'make', 'worth',
'celebratingand', 'learning']]
```

9. **Pre-process name: Stemming**
   The input:

```
[['wrote', 'internet', 'tidal', 'wave', 'even', 'mentioned', 'cdrom', 'business',
'would', 'dramatically', 'impacted', 'internet', 'fun', 'look', 'back', 'think',
'far', 'technology', 'come'], ['world', 'made', 'incredible', 'progress', 'polio',
'infectious', 'disease', 'thankful', 'partner', 'like', 'work', 'ensure', 'child',
'receive', 'vaccine', 'need'], ['vaccine', 'work', 'protect', 'community', 'dozen',
'disease', 'heartened', 'see', 'underway'], ['agreethe', 'global', 'goal',
'phenomenal', 'idea', 'perfect', 'time', 'ass', 'progress', 'recognize', 'whats',
'working', 'admit', 'isnt', 'refine', 'approach', 'good', 'people', 'need'], ['here',
'im', 'listening', 'summer', 'hopefully', 'youll', 'think', 'bad', 'granddad'],
['hope', 'plenty', 'content', 'enjoy', 'summer', 'youre', 'looking', 'add', 'list',
'ive', 'got', 'recommendation'], ['im', 'excited', 'positive', 'momentum', 'around',
'global', 'health', 'emergency', 'corp', 'last', 'week', 'critical', 'country',
'dont', 'let', 'guard', 'join', 'effort', 'build', 'strong', 'global', 'network',
'help', 'prevent', 'contain', 'pandemic', 'threat'], ['past', 'decade', 'ive',
'recommended', 'book', 'read', 'summer', 'year', 'im', 'mixing'], ['honor', 'meet',
'navajo', 'nation', 'president', 'first', 'lady'], ['really', 'good', 'news',
'renewables', 'account', 'one', 'third', 'electricity', 'generation', 'globally',
'share', 'increasing'], ['often', 'say', 'richard', 'feynman', 'best', 'teacher',
'never', 'havent', 'already', 'check', 'work', 'beyond', 'immense', 'talent',
'educator', 'he', 'also', 'quite', 'character'], ['message', 'class'], ['ric',
'employee', 'number', 'microsoft', 'dear', 'friend', 'also', 'tireless',
'philanthropist', 'advocate', 'lgbtq', 'right', 'hivaids', 'research', 'inspiration',
'many', 'others', 'story', 'worth', 'watch'], ['building', 'clean', 'energy',
'future', 'require', 'lot', 'new', 'wiresthe', 'u', 'need', 'grow', 'grid', 'almost',
'twothirds', 'current', 'size', 'next', 'decade'], ['possible', 'build', 'nuclear',
'reactor', 'without', 'enormous', 'coststo', 'environment', 'economy', 'health',
'article', 'detail', 'done', 'increase', 'clean', 'energy', 'infrastructure'],
['best', 'day', 'kemmerer', 'wyoming', 'coal', 'plant', 'tour', 'site', 'visit',
'future', 'natrium', 'plant', 'burger', 'ice', 'cream', 'coffee', 'got', 'excited',
'clean', 'energy', 'transition', 'promise', 'nextgeneration', 'nuclear'],
['expanding', 'access', 'singledose', 'hpv', 'vaccine', 'could', 'save', 'million',
'life', 'worldwide', 'next', 'year', 'great', 'example', 'im', 'optimistic',
'future', 'global', 'health'], ['im', 'encouraged', 'emotive', 'trial', 'result',
'released', 'today', 'treat', 'postpartum', 'hemorrhage', 'cause', 'maternal',
'mortality', 'new', 'approach', 'led', 'could', 'drastically', 'improve', 'woman',
'chance', 'surviving', 'childbirth', 'globally'], ['prioritizing', 'accessibility',
'affordability', 'economic', 'mobility', 'redefiningand', 'increasingthe', 'value',
'college', 'degree', 'cant', 'wait', 'visit', 'school', 'weekend', 'see',
'incredible', 'work', 'theyre', 'close'], ['eradicating', 'smallpox', 'hard', 'also',
'one', 'important', 'achievement', 'human', 'history', 'make', 'worth',
'celebratingand', 'learning']]
```

   The code/process:

```
#Stemming reduces the inflection of a word to their root form to support and improve
text pre-processing and nlp. We use PorterStemmer method from nltk to stem words if
possible.

#stem_word function used PorterStemmer method to stem the word and stores it into
list stemmed_word. When the function is called, the list stemmed_word is returned.
def stem_word(txt):
  stemmed_word = [PorterStemmer().stem(word) for word in txt]
```

```
  return stemmed_word

#We apply stem_wordfunction to the values of df column Word_Lemmatize. The results are stored in a new column named Word_Stem
df['Word_Stem'] = df['Word_Lemmatize'].apply(lambda x: stem_word(x))

#Shows the text that has been stemmed. Values from df Word_Stem column replace the previous values in the list dataset.
dataset = df.loc[:,'Word_Stem'].tolist()
print(dataset)
```

The output:

```
[['wrote', 'internet', 'tidal', 'wave', 'even', 'mention', 'cdrom', 'busi', 'would',
'dramat', 'impact', 'internet', 'fun', 'look', 'back', 'think', 'far', 'technolog',
'come'], ['world', 'made', 'incred', 'progress', 'polio', 'infecti', 'diseas',
'thank', 'partner', 'like', 'work', 'ensur', 'child', 'receiv', 'vaccin', 'need'],
['vaccin', 'work', 'protect', 'commun', 'dozen', 'diseas', 'hearten', 'see',
'underway'], ['agreeth', 'global', 'goal', 'phenomen', 'idea', 'perfect', 'time',
'ass', 'progress', 'recogn', 'what', 'work', 'admit', 'isnt', 'refin', 'approach',
'good', 'peopl', 'need'], ['here', 'im', 'listen', 'summer', 'hope', 'youll',
'think', 'bad', 'granddad'], ['hope', 'plenti', 'content', 'enjoy', 'summer', 'your',
'look', 'add', 'list', 'ive', 'got', 'recommend'], ['im', 'excit', 'posit',
'momentum', 'around', 'global', 'health', 'emerg', 'corp', 'last', 'week', 'critic',
'countri', 'dont', 'let', 'guard', 'join', 'effort', 'build', 'strong', 'global',
'network', 'help', 'prevent', 'contain', 'pandem', 'threat'], ['past', 'decad',
'ive', 'recommend', 'book', 'read', 'summer', 'year', 'im', 'mix'], ['honor', 'meet',
'navajo', 'nation', 'presid', 'first', 'ladi'], ['realli', 'good', 'news', 'renew',
'account', 'one', 'third', 'electr', 'gener', 'global', 'share', 'increas'],
['often', 'say', 'richard', 'feynman', 'best', 'teacher', 'never', 'havent',
'alreadi', 'check', 'work', 'beyond', 'immens', 'talent', 'educ', 'he', 'also',
'quit', 'charact'], ['messag', 'class'], ['ric', 'employe', 'number', 'microsoft',
'dear', 'friend', 'also', 'tireless', 'philanthropist', 'advoc', 'lgbtq', 'right',
'hivaid', 'research', 'inspir', 'mani', 'other', 'stori', 'worth', 'watch'],
['build', 'clean', 'energi', 'futur', 'requir', 'lot', 'new', 'wiresth', 'u', 'need',
'grow', 'grid', 'almost', 'twothird', 'current', 'size', 'next', 'decad'],
['possibl', 'build', 'nuclear', 'reactor', 'without', 'enorm', 'coststo', 'environ',
'economi', 'health', 'articl', 'detail', 'done', 'increas', 'clean', 'energi',
'infrastructur'], ['best', 'day', 'kemmer', 'wyom', 'coal', 'plant', 'tour', 'site',
'visit', 'futur', 'natrium', 'plant', 'burger', 'ice', 'cream', 'coffe', 'got',
'excit', 'clean', 'energi', 'transit', 'promis', 'nextgener', 'nuclear'], ['expand',
'access', 'singledos', 'hpv', 'vaccin', 'could', 'save', 'million', 'life',
'worldwid', 'next', 'year', 'great', 'exampl', 'im', 'optimist', 'futur', 'global',
'health'], ['im', 'encourag', 'emot', 'trial', 'result', 'releas', 'today', 'treat',
'postpartum', 'hemorrhag', 'caus', 'matern', 'mortal', 'new', 'approach', 'led',
'could', 'drastic', 'improv', 'woman', 'chanc', 'surviv', 'childbirth', 'global'],
['priorit', 'access', 'afford', 'econom', 'mobil', 'redefiningand', 'increasingth',
'valu', 'colleg', 'degre', 'cant', 'wait', 'visit', 'school', 'weekend', 'see',
'incred', 'work', 'theyr', 'close'], ['erad', 'smallpox', 'hard', 'also', 'one',
'import', 'achiev', 'human', 'histori', 'make', 'worth', 'celebratingand', 'learn']]
```

Sentiment Scores of the 43 Tweets in the Dataset

RoBERTa-base Model

LABEL_0 means NEGATIVE, LABEL_1 means NEUTRAL, LABEL_2 means POSITIVE


```
[{'label': 'LABEL_2', 'score': 0.6453714370727539},
{'label': 'LABEL_2', 'score': 0.8911813497543335},
{'label': 'LABEL_1', 'score': 0.49357324838638306},
{'label': 'LABEL_2', 'score': 0.7566262483596802},
{'label': 'LABEL_1', 'score': 0.5774592757225037},
{'label': 'LABEL_2', 'score': 0.956655740737915},
{'label': 'LABEL_2', 'score': 0.8767237067222595},
{'label': 'LABEL_1', 'score': 0.7798121571540833},
{'label': 'LABEL_1', 'score': 0.6674532294273376},
{'label': 'LABEL_2', 'score': 0.955295979976654},
{'label': 'LABEL_2', 'score': 0.7240356802940369},
{'label': 'LABEL_1', 'score': 0.6101581454277039},
{'label': 'LABEL_2', 'score': 0.7944564819335938},
{'label': 'LABEL_1', 'score': 0.5714358687400818},
{'label': 'LABEL_1', 'score': 0.5473719835281372},
{'label': 'LABEL_2', 'score': 0.9616201519966125},
{'label': 'LABEL_2', 'score': 0.9469780325889587},
{'label': 'LABEL_2', 'score': 0.8445103764533997},
{'label': 'LABEL_2', 'score': 0.9702820181846619},
{'label': 'LABEL_2', 'score': 0.785840630531311},
{'label': 'LABEL_2', 'score': 0.9252800345420837},
{'label': 'LABEL_1', 'score': 0.792206883430481},
{'label': 'LABEL_1', 'score': 0.7244055867195129},
{'label': 'LABEL_1', 'score': 0.8023517727851868},
{'label': 'LABEL_2', 'score': 0.9030808806419373},
{'label': 'LABEL_2', 'score': 0.4986259639263153},
{'label': 'LABEL_2', 'score': 0.7839193344116211},
{'label': 'LABEL_2', 'score': 0.48219624161720276},
{'label': 'LABEL_1', 'score': 0.5723499059677124},
{'label': 'LABEL_2', 'score': 0.6040951013565063},
{'label': 'LABEL_2', 'score': 0.6617953181266785},
{'label': 'LABEL_2', 'score': 0.6807810068130493},
{'label': 'LABEL_2', 'score': 0.9781891703605652},
{'label': 'LABEL_1', 'score': 0.6223629713058472},
{'label': 'LABEL_1', 'score': 0.8490437269210815},
{'label': 'LABEL_2', 'score': 0.6170380711555481},
{'label': 'LABEL_1', 'score': 0.868795216083526 6},
{'label': 'LABEL_2', 'score': 0.5615702867507935},
{'label': 'LABEL_2', 'score': 0.9616062641143799},
{'label': 'LABEL_2', 'score': 0.9365656971931458},
{'label': 'LABEL_2', 'score': 0.5762225985527039},
{'label': 'LABEL_2', 'score': 0.8900753855705261},
{'label': 'LABEL_2', 'score': 0.9728915095329285}]
```


TextBlob

score < 0 is NEGATIVE, score == 0 is NEUTRAL, score > 0 is POSITIVE


```
0 - 0.400000 - Positive
1 - 0.900000 - Positive
```

2 - 0.000000 - Neutral
3 - 0.525000 - Positive
4 - 0.666667 - Positive
5 - 0.500000 - Positive
6 - 0.413636 - Positive
7 - 0.250000 - Positive
8 - 0.333333 - Positive
9 - 0.200000 - Positive
10 - 0.650000 - Positive
11 - 0.000000 - Neutral
12 - 0.300000 - Positive
13 - 0.335909 - Positive
14 - 0.866667 - Positive
15 - 0.468750 - Positive
16 - 0.218750 - Positive
17 - 0.227273 - Positive
18 - 0.550000 - Positive
19 - 0.435417 - Positive
20 - 0.495238 - Positive
21 - 0.650000 - Positive
22 - 0.200000 - Positive
23 - 0.000000 - Neutral
24 - 0.295960 - Positive
25 - 0.700000 - Positive
26 - 0.150000 - Positive
27 - 0.312500 - Positive
28 - 0.412626 - Positive
29 - 0.251515 - Positive
30 - 0.533333 - Positive
31 - 0.600000 - Positive
32 - 0.552273 - Positive
33 - 0.200000 - Positive
34 - 0.650000 - Positive
35 - 0.375000 - Positive
36 - 0.125000 - Positive
37 - 0.276136 - Positive
38 - 0.583333 - Positive
39 - 0.133333 - Positive
40 - 0.535417 - Positive
41 - 0.000000 - Neutral
42 - 0.375000 - Positive