Team Profile	
Category	**Content**
Name of Group	Data Beninging
Names of Group Members	BIO, Justine Anne D. ESPINO, Chloe Irish T. PEREZ, Jester
Text Preprocessing Techniques	
Category	**Content**
Description of the two (2) text sets	1. The first data set was downloaded from Kaggle. There are exactly 558,837 words and 4 columns including the "make", "cleaned_text", "tokenized_text, and "tagged_text". 2. The Second data set was downloaded from The CC19 - Datasets folder provided. There are exactly 3,325,313 words and 2 columns which are the "text" and "label" columns.
Analysis of Text Set One	1. After the text cleaning phase, the text became more uniform with each other because every letter was lowercase. The text didn't have other characters except for letters so the other text-cleaning techniques were not utilized. 2. The text tokenization phase didn't affect the text much because it only consisted of a single word (Car Brand). Also, it didn't have any other characters except letters. 3. It didn't have any effects at all because it only consisted of a single word. However, all the words were tagged. 4. I've used the "Text Inspector" Tagger, the results are the same. **5. Preprocessed Texts Unprocessed Texts**
Analysis of Text Set Two	1. After the text was cleaned, the texts were more readable compared to before it was cleaned. The emojis were removed, as well as unrelated characters. It all retained its meaning but the characters all did become lowercase for easier readability and

consistency.

2. After the tokenized text was finished, the words were all separated into individual brackets. This broke down the texts further for easier understanding of each word. Some texts needed to be tokenized correctly. These texts were mainly words with typos that had non-existent spaces between them.
3. Most text data were successfully tagged during the text tagging phase. Elements like linking/transition words (coordinating conjunctions) were not properly tagged in the text data.
4. We've used the "Parts-of-speech.Info" Tagger, and the results are the same.
5. **Sample screenshots of Unprocessed and Preprocessed text data**

index	text	cleaned_text
0	stop scaring people wearing masks firsteval im sure right subreddit post im still posting this point even think covid real care wearing mask still wear mask many people believe corona virus afraid persons like people wear masks wear mask make life harder hard wear mask sometimes forget im wearing mask walk even to example would be serial killer escaped prison media warns everyone stay house make lights off serial killer know someone believe it friend roommate believe saying wants show truth decides switchs lights on plays loud music stuff like this scared hell serial killer coming house turns roommate right serial killer exist still fucking horror him know good example someone ever read it wanted say that much easier wear mask stress hope good day stay safe sorry spelling mistakes native language english	stop scaring people wearing masks firsteval im sure right subreddit post im still posting this point even think covid real care wearing mask still wear mask many people believe corona virus afraid persons like people wear masks wear mask make life harder hard wear mask sometimes forget im wearing mask walk even to example would be serial killer escaped prison media warns everyone stay house make lights off serial killer know someone believe it friend roommate believe saying wants show truth decides switchs lights on plays loud music stuff like this scared hell serial killer coming house turns roommate right serial killer exist still fucking horror him know good example someone ever read it wanted say that much easier wear mask stress hope good day stay safe sorry spelling mistakes native language english
1	#ë°□ë´□ì □ë°□ë□´ë²□ë ¤ë□□ì□¸ì□□Depression is not a joke...instead of supporting her...netizens are dragging her...what a shame...Act like a human	ëëì ëëë²ë ëìdepression is not a jokeinstead of supporting hernetizens are dragging herwhat a shameact like a human
2	saw description movie tcm let run like peter ustinov maggie smith delightfully surprised find really liked movie found quite exceptional course seriously dated period piece well worth watching subtle humour insight life lifestyle almost forty years ago problem trying find dvd watch often also quite taken performances smith ustinov leads karl malden bob newhart cameo appearances robert morley cesar romero	saw description movie tcm let run like peter ustinov maggie smith delightfully surprised find really liked movie found quite exceptional course seriously dated period piece well worth watching subtle humour insight life lifestyle almost forty years ago problem trying find dvd watch often also quite taken performances smith ustinov leads karl malden bob newhart cameo appearances robert morley cesar romero
3	somebody please play yugioh duel links me id bored friends dont play yugioh duel links	somebody please play yugioh duel links me id bored friends dont play yugioh duel links
4	watching old video of dance team and such make me miss it	watching old video of dance team and such make me miss it
5	Hello world! Listen the Radio Dabas online: www.radiodabas.hu left corner above - PRï¿½BA SZERENCSE	hello world listen the radio dabas online wwwradiodabashu left corner above prï½ba szerencse
6	maybe ill start upvting every comment like ill cus mabe someone like somethin idk	maybe ill start upvting every comment like ill cus mabe someone like somethin idk
7	@berutt sorry for not voting earlier...I was still in bed	berutt sorry for not voting earlieri was still in bed
8	Goodnight my lovelys	goodnight my lovelys

9	Atlantis & Hubble current location now is above Indonesia. Wave your hand guys, wave! #fb	atlantis amp hubble current location now is above indonesia wave your hand guys wave fb
10	deserve livei raised normal healthy im emotionally defective toxic social failure potential friend push away hope stop interacting me think im healthy around want ruin life unfortunate upbringingtraumas think explains end total fucked psychologically told much lies like im daughter high ranking military man lies im still cringing think day reason general hate feel towards although kind authentic person dislike something her god still eager friends me im like please leave im toxic reclusive talk months except initiate first im bad friend know friendship works leave cant everyones cup tea thats ok person know that think im able talk stop interacting me admit envy living normal life normal mind blessed likeable face sounds pretty dumb immature know well tried work myself spent money time effort get better turns damages cant repair talk anonymously online helps time im back suicidal idealation contemplating much could cause serious harm people completely lost self control hate admit im bit narcy sometimes cant help it like whats going inside head either want cause possible harm could someone want associate cant help it here strong potential commit something awful law im done life	deserve livei raised normal healthy im emotionally defective toxic social failure potential friend push away hope stop interacting me think im healthy around want ruin life unfortunate upbringingtraumas think explains end total fucked psychologically told much lies like im daughter high ranking military man lies im still cringing think day reason general hate feel towards although kind authentic person dislike something her god still eager friends me im like please leave im toxic reclusive talk months except initiate first im bad friend know friendship works leave cant everyones cup tea thats ok person know that think im able talk stop interacting me admit envy living normal life normal mind blessed likeable face sounds pretty dumb immature know well tried work myself spent money time effort get better turns damages cant repair talk anonymously online helps time im back suicidal idealation contemplating much could cause serious harm people completely lost self control hate admit im bit narcy sometimes cant help it like whats going inside head either want cause possible harm could someone want associate cant help it here strong potential commit something awful law im done life
11	im tonightim already tired nothing makes happy anymore one help me	im tonightim already tired nothing makes happy anymore one help me
12	@britneyfrancis the regular.. trying to find some food, nothing new	britneyfrancis the regular trying to find some food nothing new

13	takashi miike one favorite directors worried kids film would hate see depart films came love visitor q gozu izo ichi killer black socioty trilogy lately seems exploring new territory think hes succeeding still first films id seen take direction nervous coarse bought without seeing glad didbr br great yokai war perfect kids film adults like too whole film reminded much movies loved child neverending story labyrinth return oz etc enjoyed films treat kids like theyre stupid one either dark underlying morals there but also silly kids film be personally bothered cgi prosthetics feel like fit well think kids noticebr br if die hard takashi miike fan may like one but suggest giving shot proves miike diverse talented suspected is also continues make signature miike films outside ones reassuringbr br to people new takashi miike want something light hearted dramatic like one suggest miike films zebraman the happiness katakuris sabu the bird people china br br good job takashi miike stars	takashi miike one favorite directors worried kids film would hate see depart films came love visitor q gozu izo ichi killer black socioty trilogy lately seems exploring new territory think hes succeeding still first films id seen take direction nervous coarse bought without seeing glad didbr br great yokai war perfect kids film adults like too whole film reminded much movies loved child neverending story labyrinth return oz etc enjoyed films treat kids like theyre stupid one either dark underlying morals there but also silly kids film be personally bothered cgi prosthetics feel like fit well think kids noticebr br if die hard takashi miike fan may like one but suggest giving shot proves miike diverse talented suspected is also continues make signature miike films outside ones reassuringbr br to people new takashi miike want something light hearted dramatic like one suggest miike films zebraman the happiness katakuris sabu the bird people china br br good job takashi miike stars
14	answer question previous reviewer asked name us official mentioned lumumba name character mr carlucci frank carlucci reported time second secretary us embassy congo subsequently among assignments appointed us ambassador portugal deputy director central intelligence agency secretary defense chairman carlyle group hardly surprising carluccis biographical sketch wwwcarlylegroupcom web site fails credit service belgian congo name deliberately censored hbo version lumumba may avoid possibility hbos sued us courts carluccis name however clearly mentioned theatre version lumumba saw recently event expect would deny involvement lumumbas murderbr br others commented evenhandedness film lumumba treats parties concerned lumumbasupporters congolese even belgians somewhat sinister view emerges think bbc documentary entitled who killed lumumba based book the murder lumumba belgian historian ludo de witte examined closely films demonstrate fate lumumba history congo matter black white lumumbas murderers believe that	answer question previous reviewer asked name us official mentioned lumumba name character mr carlucci frank carlucci reported time second secretary us embassy congo subsequently among assignments appointed us ambassador portugal deputy director central intelligence agency secretary defense chairman carlyle group hardly surprising carluccis biographical sketch wwwcarlylegroupcom web site fails credit service belgian congo name deliberately censored hbo version lumumba may avoid possibility hbos sued us courts carluccis name however clearly mentioned theatre version lumumba saw recently event expect would deny involvement lumumbas murderbr br others commented evenhandedness film lumumba treats parties concerned lumumbasupporters congolese even belgians somewhat sinister view emerges think bbc documentary entitled who killed lumumba based book the murder lumumba belgian historian ludo de witte examined closely films demonstrate fate lumumba history congo matter black white lumumbas murderers believe that
15	i wa at the peak of my depression during lockdown level yoh	i wa at the peak of my depression during lockdown level yoh

16	reason miss school want fall asleep bus way home wake eyes feeling cold sleepy hoodie keeping comfy getting bus barely able walk way home tired much bones hurt day school get home let fall onto bed fall asleep grandma wakes lunch	reason miss school want fall asleep bus way home wake eyes feeling cold sleepy hoodie keeping comfy getting bus barely able walk way home tired much bones hurt day school get home let fall onto bed fall asleep grandma wakes lunch
17	hurt knowing dad wanted athletic hard working kid liked different stuff instead got hurts	hurt knowing dad wanted athletic hard working kid liked different stuff instead got hurts
18	Sitting on the deck with Robert Holden, author of Be Happy. A beautiful man and gifted coach. I'm happy	sitting on the deck with robert holden author of be happy a beautiful man and gifted coach im happy
19	@carobuckman Have fun!!!	carobuckman have fun
20	would u describe perfect girl filler	would u describe perfect girl filler
21	lets goooooooo every single one singles placement matches rocket league im platinum for know rocket league car soccer free epic games store wanna play really good game thanks listening celebration stay safe goodbye	lets goooooooo every single one singles placement matches rocket league im platinum for know rocket league car soccer free epic games store wanna play really good game thanks listening celebration stay safe goodbye
22	Something to Talk About - Badly Drawn Boy * RB @jackryan - nice 2 know â™« http://blip.fm/~8b3dr	something to talk about badly drawn boy rb jackryan nice 2 know â httpblipfm8b3dr
23	i have been caught in a cycle of depression v being hyper active it is a struggle to stay focused in either of those state of being doe anyone have a track or a song or even something you made yourself to help get out of this hole i love music and i need a new sound please help	i have been caught in a cycle of depression v being hyper active it is a struggle to stay focused in either of those state of being doe anyone have a track or a song or even something you made yourself to help get out of this hole i love music and i need a new sound please help
24	just posted a new blog @ www.awakendesigncompany.com/blog - check it out for news about a big wedding coming up -- and sales!! GO GO GO!	just posted a new blog wwwawakendesigncompanycomblog check it out for news about a big wedding coming up and sales go go go

Tokenized and Tagged text data

	tokenized_text	tagged_text
0	[stop, scaring, people, wearing, masks, firste...	[(stop, VB), (scaring, VBG), (people, NNS), (w...
1	[ëëì, ëëë²ë, ëìdepression, is, not, a, jokein...	[(ëëì, JJ), (ëëë²ë, NNP), (ëìdepression, NN),...
2	[saw, description, movie, tcm, let, run, like,...	[(saw, JJ), (description, NN), (movie, NN), (t...
3	[somebody, please, play, yugioh, duel, links, ...	[(somebody, NN), (please, VB), (play, VB), (yu...
4	[watching, old, video, of, dance, team, and, s...	[(watching, VBG), (old, JJ), (video, NN), (of,...
...	...	...
9200	[anyone, else, miss, magici, feel, like, last,...	[(anyone, NN), (else, RB), (miss, JJ), (magici...
9201	[posting, daily, longer, single, day, playing,...	[(posting, VBG), (daily, RB), (longer, RBR), (...
9202	[maynaseric, beep, beep]	[(maynaseric, JJ), (beep, NN), (beep, NN)]
9203	[vitamin, b12, deficiency, can, cause, memory,...	[(vitamin, NN), (b12, NN), (deficiency, NN), (...
9204	[want, outi, want, quick, easy, escape, perman...	[(want, JJ), (outi, NN), (want, VBP), (quick, ...

Comparison of the Effects of Text Preprocessing on Both Sets of Text

Compare the two sets of text data, their results, and how they were affected by the different text preprocessing techniques:

1. Which text set had more errors or issues during the text preprocessing steps? Why do you think that specific text sets encountered more issues or problems with the text preprocessing techniques? Was it an issue with the data itself or the techniques? If they both performed similarly, why do you think both text sets behaved similarly when they were preprocessed?

	- After analyzing both sets, Text Set Two encountered more issues during the preprocessing steps compared to Text Set One. This is evident from the need to remove emojis and unrelated characters, as well as the challenge of tokenizing words correctly, especially those with typos and non-existent spaces. - Text Set Two also had some difficulties in properly tagging certain elements like coordinating conjunctions. - The issues in Text Set Two seem to stem from the nature of the data itself rather than the techniques used. Emojis, typos, and non-existent spaces pose challenges for text preprocessing, especially tokenization and tagging. 2. Based on the analysis done, do you think certain types of text data perform better or worse when they are run through text preprocessing techniques? Discuss what type of text data works well with text preprocessing techniques. Discuss what type of text data works poorly with text preprocessing techniques. - After analyzing both text process outcomes, certain types of text data may perform better or worse during preprocessing depending on their characteristics. Text data with consistent formatting, correct spellings, and minimal special characters tend to perform better with text preprocessing techniques. These texts are easier to tokenize, tag, and clean, resulting in more accurate and reliable preprocessing outcomes. - On the other hand, text data with irregular formatting, misspellings, emojis, and special characters may encounter more challenges during preprocessing. These texts may require additional cleaning steps and manual interventions to ensure accurate tokenization, tagging, and cleaning.