Mental health in college students during the COVID-19 pandemic: a text mining approach

Arantxa Gomez
Faculty of Engineering
Universidad del Pacifico
Lima, Peru
aa.gomezc@alum.up.edu.pe

Mariana Moyano
Faculty of Engineering
Universidad del Pacifico
Lima, Peru
md.moyanom@alum.up.edu.pe

Aramis Palacios
Faculty of Engineering
Universidad del Pacifico
Lima, Peru
aa.palaciosa@alum.up.edu.pe

ABSTRACT

The great amount of posts made by Peruvian college students in social media regarding the deterioration of their mental health drew attention to what could be done to solve this issue. It was unclear why, as reflected in social media, the mental health of Peruvian college students didn't seem to improve despite the constant work of student counseling services at universities. Thus, the objective of this study was to analyze posts from college students and classify the emotions they express regarding topics like mental health, the COVID-19 pandemic, political and economic crisis. 60,790 posts from Facebook and Twitter from Peruvian college students were extracted and preprocessed, ending as 33,103. Applying the K-Means method, nine clusters were obtained. With the assistance of two psychologists from the Universidad del Pacifico, these clusters were tagged corresponding to their main theme. The main obtained insight is that college students from Peru expressed worry about the pandemic and its effects on the population's mental health, disconformity about the Peruvian political crisis, concern about Peru's future, the importance of a social media channel to express anonymously their ideas and the growing need of psychological counseling to deal with mental health issues. In conclusion, this study provides a full view on how college students express their feelings about how they live everyday with stressful situations like the COVID-19 pandemic and political crisis in Peru while dealing with academic and personal worries, all at the same time.

Keywords: COVID-19, political crisis, mental health, college students, k means, text mining, peru, social media.

I. INTRODUCTION

Mental health, compared to other areas of health, has been neglected in terms of research and acknowledgement by many societies around the world. However, since the past decade, the number of research projects related to mental health have increased enormously, with a high preponderance of these works being developed in countries from North America or Europe. Nevertheless, as mental health care has gained more importance over the years, the number of cases of people suffering from some kind of mental health issue has also increased. Indeed, 10.7% of the world population, around 792 million people, was experiencing some mental health disorder by 2017[1]. Moreover, with the arrival of the COVID-19 pandemic, this situation has become widespread. Adding to this lockdowns in almost every country, the cases of people experiencing symptoms related to mental disorders have increased. This has also been influenced by the fact that 93% of countries across the world have seen a disruption of their mental health services while only 2% of national health budgets is destined for mental health [2]. That's why it is relevant to pay attention to this subject matter in order to understand and characterize its impact on citizens' lives. Among the citizens as a whole, one of the most affected groups is the group of college students, which this investigation is focused on.

Ever since the pandemic started, in Peru in 2020, the social media posts about deteriorated mental health in Peruvian college students has increased in anonymous pages meant to entertain the college communities (i.e. "Confesiones" facebook pages for each university where people communicate anonymously). It remains unclear why, as reflected in social media, the mental health of Peruvian college students doesn't seem to have improved despite the constant work of student counseling services at the universities. It's a concern to know what kind of counseling or treatment services could be adapted into the necessities that groups like college students express in social media. Thus, the purpose of this paper is to analyze posts from college students and classify the emotions they express regarding topics like mental health, the COVID-19 pandemic, political and economic issues in the country. Also, by identifying the most recurrent emotions that are expressed on social networks such as Twitter and Facebook, the results of this exploration may help universities to offer better counseling and accompaniment services.

The questions that we intend to answer with this research are the following:

1) What are the most used words to describe the feelings of college students during the COVID-19 pandemic?
2) What might colleges' student services do to improve their efficiency on treating student individual mental health issues based on the results of this research?

The structure of this research is as it follows: a literature review in which we discuss the latest research about similar explorations in other countries and in different contexts; the methodology, where we explain the process of data collection and techniques that we use to analyze the data; a section of results in which we show the outputs of the techniques applied; a discussion part, where we evaluate the outcomes and state the implications of those findings; a conclusion section that summarizes the research and gives

advice for future research; and, finally, the references that we used to support our explanations.

II. Literature review

A. How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds

As a result of the stay-at-home directives and social distancing measures that many countries around the world established to contain the spread of the COVID-19 disease, there has been a notorious rise in the amount of people experiencing loneliness. Thus, [3] examine the expression of loneliness on Twitter during the COVID-19 pandemic.

The data used for this research was extracted from Twitter feeds of individual users that contained the words "loneliness" and "COVID-19", considering a total of 19,095 feeds from May 1st 2020 to July 1st 2020.

The techniques that were applied to this dataset consist of machine learning and hierarchical modeling. The first one encompasses topic modeling in order to identify key topics from posts and, the second one, for identifying overarching themes, which are: (1) community impact of loneliness during COVID-19, (2) social distancing during COVID-19 and its effects on loneliness and, (3) mental health effects of loneliness during COVID-19 [3]. After identifying and labeling these themes, the authors perform an analysis in R using the packages: "rtweet", "spacyr", "quanteda", "stm" and, "ClustOfVar".

Finally, after processing these tweets, the resultant number of feeds included in the study were 4,492; where most of the posts were from North America or Europe. Theme 1 was present in approximately 41,3% of the Twitter feeds, while theme 2 in 30,9% and theme 3 in 27,8% of the feeds. For theme 1, some related keywords were death, home, work and hope; for theme 2, isolation, distance, feel and lockdown; and, for theme 3, health, anxiety, depression and support. Nevertheless, these themes demonstrated temporal variations as the COVID-19 pandemic evolved [3].

B. Detecting Community Depression Dynamics Due to COVID-19 pandemic in Australia

The long-term social activity restriction policies adopted during the pandemic period may further amplify the mental issues of people [4]. For that reason, the authors center their attention to examine the community depression dynamics due to the COVID-19 pandemic in Australia. They introduce a new approach based on multimodal features from tweets and term frequency-inverse document frequency [4].

The dataset consists of tweets from Twitter users who live in different local government areas of New South Wales in Australia, from the period between January 1st 2020 and May 22nd 2020. The main tool used for this study was Python, in which with the library "Tweepy" collected 94,707,264 tweets with an average of 739,901 tweets for each local area [4]. Additionally, datasets concerning COVID-tests and confirmed cases in the state were collected from the Australian government open data web.

Thus, the techniques applied for this research consist of combining multimodal features with TF-IDF features. The first ones contain: (1) emotional features (use of positive or negative emojis/slang in each tweet), (2) topic-level features (extract topics from text) and, (3) domain specific features (criteria to describe depressive symptoms). TF refers to the occurrence frequency of a term in a tweet, while IDF does the opposite, it uses the frequency of the term in a collection giving a higher value to a rare word. TF-IDF is only useful as a lexical-level feature [4]. Then, for modeling depression in tweets, three mainstream classification methods are used: logistic regression, linear discriminant analysis and Gaussian Naive Bayes in order to compare their results.

After testifying their model, [4] ended up with 49 million tweets, where nearly two million of them were classified as depressed. Also, people's depression level remarkably escalated with a significant increase in confirmed cases of COVID-19.

C. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach

For people suffering from mental disorders, sharing their experiences, challenges, and coping mechanisms with online communities through social media platforms has become a way to emotionally support each other and exchange information. While various members of online health communities proclaim that online interactions have influenced their improvement from depression, anxiety, stress, and negative moods; it's also possible they have a negative impact on individuals through interaction itself, increasing anxiety, anger, and negative emotions after reports of celebrity suicides. So far, there hasn't been research comparing discussion topics in online mental health communities. That's why [5] focuses on making that comparison in the online health communities from the platform Reddit about depression, anxiety and Post-Traumatic Stress Disorder (PTSD).

To extract the data needed, the authors use Python Reddit API Wrapper (Reddit's official API). From the three sub-communities or subreddits (r/Depression, r/Anxiety, and r/PTSD), they collect the title, author id, timestamp, post or comment id, parent id, number of direct replies, scores and content. 7410 posts and 132599 associated comments from January of 2011 to December of 2015 are downloaded, going from 'top' rated posts, 'hot' posts, and 'new' posts to collect larger datasets.

To organize the data, they use k-means clustering in order to identify the main discussion themes that will be qualitatively examined to understand their similarities and differences. Many visualization techniques, such as D3, Gephi, ForceAtlas2, are used to comprehend how each theme relates to another for the three mental health conditions.

Finally, [5] recognizes four common themes between the three conditions: sharing of positive emotion, gratitude for receiving emotional support, sleep and work-related issues. While the Anxiety and PTSD subreddits have more similar discussion topics, addressing treatment and medication related issues; the depression subreddit focuses on self-expressed aspects of it.

D. Text Mining Mental Health Reports for Issues Impacting Today's College Students: Qualitative Study

A growing number of college students are experiencing personal circumstances or encountering situations that feel overwhelming and negatively affect their academic studies and other aspects of life on campus. To meet this growing demand for counseling services, US colleges and universities are offering a growing variety of mental health services that provide support and services to students in distress [6].

To address this problem, the researchers explored 165 references in *The Chronicle of Higher Education, Insider Higher Education, Diverse, Ethnic Newswatch, and the Journal of Blacks in Higher Education* from the period 2010 - 2015 which are focused on mental health in college students all across the US.

The method applied is text mining using combinations of the keywords, "college students," "mental health," and "mental services". The main tool used is SAS Enterprise Miner with Text Miner 12.1. This tool produces four outputs: (1) a text parsing analysis that pulls words from the documents and determines its part of speech and importance, (2) a text filtering analysis that determines which words to discard or keep for further analysis, (3) a text topic analysis that groups articles into similar topics defined by words, and (4) a text cluster analysis that places articles into clusters based on common words that are shared between articles [6].

The results from the clustering analysis delivered six main themes, which include age, race, crime, student services, aftermath and victim. From these six themes, two of them had more incidences in the 165 references analyzed: student services and race. Therefore, it represents the increasing demand of student services provided by campuses and the increasing mental health risks faced by ethnic minorities [6].

III. METHODOLOGY

For this study, we applied the KDD methodology. In Fig.1 we describe the followed steps:

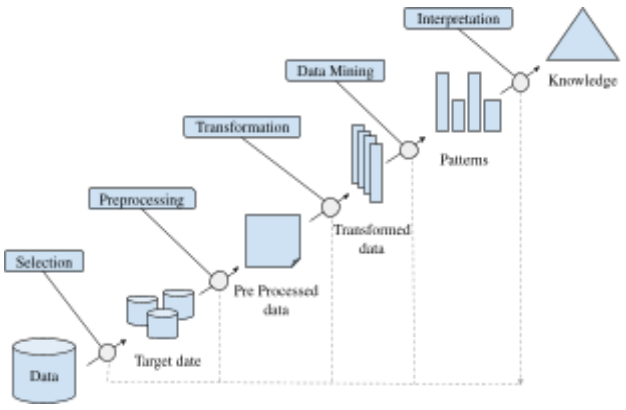

Fig. 1 Phases in KDD Methodology

A. Selection

In this study, to analyze how peruvian college students express their negative sentiments related to mental health, we collected tweets from Twitter users who live in Peru and posts from Facebook pages about peruvian colleges.

To search tweets, we set up two lists as shown in Fig. 2 (See Appendix): one with contextual words ("Contextos") such as "universidad", "coronavirus", "peru", "merino", etc.; another with feeling words ("Sentimientos") such as "desanimada", "triste", "ansioso", "preocupacion", etc. We filtered tweets made from March 5, 2020 to November 19, 2021 and excluded retweets. However, due to library restrictions we couldn't get tweets that far.

The time span of the collected Facebook posts was from March 5, 2020, to November 14, 2021, which covers important events related to COVID 19, political and economical crisis, election period and the transfer to virtual classes in Peru. Thus, in order to identify the college students´ opinions, we resorted to popular "confessions pages", that students usually use to express their feelings or concerns in their community (See Table 1 in Appendix for the complete list of pages).

Using the Tweepy library connected to the Twitter API, we extracted 53,265 tweets that mentioned combinations of the two lists of words. This way, the contents of each tweet included one contextual word and one feeling word. We also extracted the username, date of creation and location since these were the data to be used. Extracting this quantity of tweets took approximately 5hrs 25min. The downside of using this library was that it extracted only the most recent tweets. In addition, we tried to extract tweets from previous dates with Twint and Twarc, however, some module restrictions made by Twitter imposibilitated the smooth running of the first library that gave us access to older tweets. On the contrary, Twarc worked faster when extracting the data but only considered tweets from the most recent weeks.

On the other hand, the facebook-scraper library, in contrast to Tweepy, didn't require any credentials. With this tool we extracted 8,883 posts in total from the confession pages, which included the name of the publisher page, the post's url, the full text and the publication date. A limitation of this library was that it scraped a determined number of pages and didn't include a date parameter. Moreover, when trying to scrape a bigger quantity of posts, it didn't allow checking the exact date of the posts. Thus, we had to try with different numbers of pages until reaching the span time of our interest, however the limit of pages was 400 until getting blocked temporarily by Facebook. In addition, it is important to highlight that every single page was a special case, given the fact that some of them have more users activity than others. For example, for "Confesiones UP" we needed around 398 pages for scraping posts in the selected span of time, while for "Confesiones UDEP" we only required 110 pages. It is important to note these differences between each Facebook page by cause of the computational cost and running time that it involves, while for "Confesiones UP" the facebook-scraper get_posts function took around 33min 10s, for "Confesiones UDEP", 10min 43s.

B. *Preprocessing*

In the first place, it was necessary to make sure the data collected belonged to peruvian students. Due to restrictions of the Tweepy library, it was not possible to filter the location from the users who wrote the tweets collected. Meanwhile, the dataset obtained from facebook didn't present any of those issues, because the pages scraped belonged to peruvian universities. In order to reduce the bias of the data, some imputation and filtering techniques that increased the value of the data were applied. So, from the Wikipedia webpage, "List of populated places in Peru", we extracted a useful list of places to filter tweets that were written in this country. Moreover, we applied the "fillna" method from the pandas library to take advantage of the tweets that didn't show any location. Despite this last one being a questionable technique, it helped a lot to reach the minimum goal of rows that our dataset needed to be meaningful. After the filtering techniques, the resulting data frames from twitter and facebook were concatenated and set ready for the cleaning process.

In order to clean the text mined, we used a series of libraries, such as re, string, NLTK and unicodedata. "re" library is used for deleting mentions and links from text through regular expressions. "string" library is used for removing punctuation. "NLTK" library is used for eliminating stopwords and lemmatizing words. "unicodedata" is used for normalizing the text, which means it clears accents and changes special letters, such as "ñ", to their root, such as "n". The final size of the dataset was 33,103 rows.

After the cleaning function was defined, we added a new column to each data frame which will show the preprocessed text. We lowered the strings in each raw text and stored its cleaned version in the corresponding column. The final columns that were selected were "FechaCreacion", "Texto" and "limpio". Finally, we saved the data frame into a csv file and read them in another notebook to get to the next step, model training.

C. *Transformation*

In this phase, we started by identifying the 10, 50, 100 and 200 top words in the column "limpio" of the 33,013 rows. Different sizes for this vector were tried in order to identify which one offers a better clustering for the data.

We then defined a term frequency (tf) function which returns a column of vectors that represents each one of the texts in the "limpio" column according to the number of top words selected. Every vector obtained depends on the size of the top words list, which means there exist 4 different columns of term frequency for the dataset. So, for every top word list a different data frame has been created which contains the "limpio" column and a new one showing the term frequencies for each row.

After a term frequency column has been created for each 4 data frames, we added a column that contains the standardized form of the vectors from the last column. This method was used in order to scale each vector to values that show a known behaviour. As Scikit scientists explain, "if a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected" [7].

Once this last column is created for each data frame, it is possible to start the machine learning process and apply the clusterization techniques for the data.

D. *Data Mining*

In this phase, we applied the unsupervised learning method of K-Means in order to group all the observations according to their similarities. We applied the KMeans function from the "sklearn.cluster" library. To determine the optimal value of k, meaning the optimal number of clusters we should make, we calculated the inertia for k as a value ranging from 1 to 30. We did this for each new data frame created taking as the input variable for the model the scaled TF vector. With the help of the pyplot function from the "matplotlib" library, four graphics were obtained that let us see the values of k and their respective inertias through the Elbow Method.

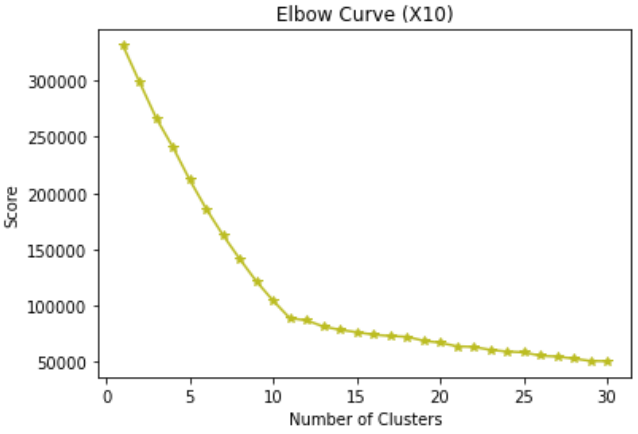

Fig. 3. Elbow curve made from the Scaled Vector size 10

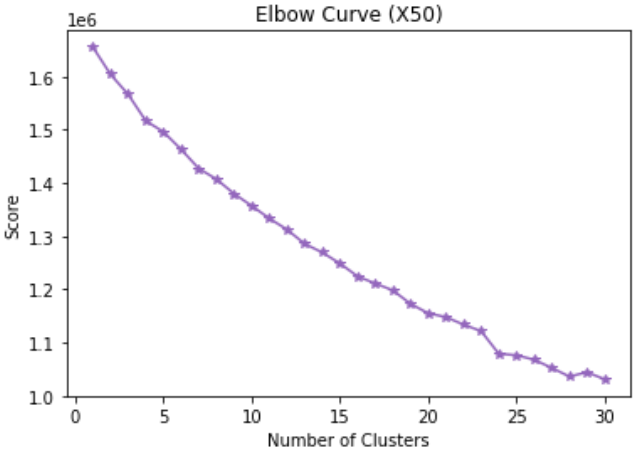

Fig. 4. Elbow curve made from the Scaled Vector size 50

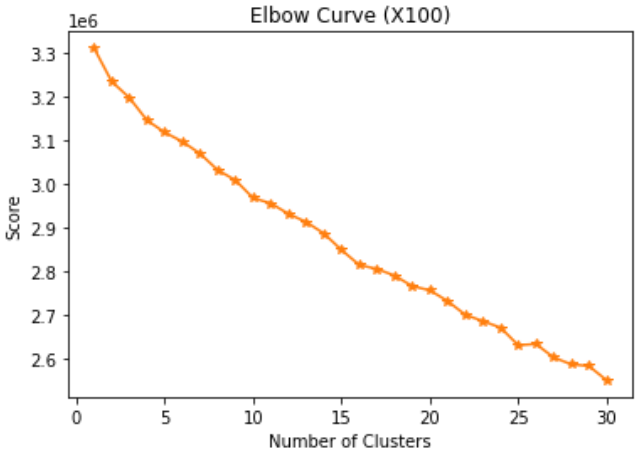

Fig. 5. Elbow curve made from the Scaled Vector size 100

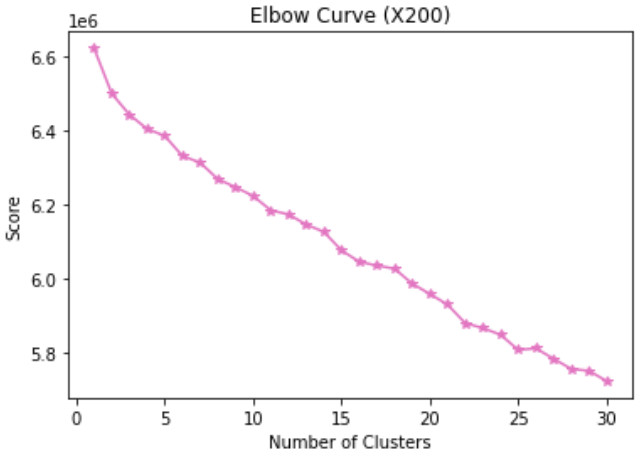

Fig. 6. Elbow curve made from the Scaled Vector size 200

E. Interpretation

For this phase, we had the advice from two psychologists at Universidad del Pacifico: Magaly Rubina, Director of Extracurricular Training and Fiorella Otiniano, Research and Development Coordinator. An interview was held on November 23rd, 2021, where we showed them the resultant clusters in word clouds and asked them to analyze them and give us their interpretation and tag for each. For access to the complete interview you can click the following link: Validation of clusters (2021-11-23).

IV. Results

After performing the preprocessing we had a total of 33103 records, then, we applied k-means and developed experiments with different top words. In that sense, we obtained these combinations of sets conformed by top N words and their theoretical optimal number of clusters:

Table 2. Number of clusters by top words

Top N words	10	50	100	200
# of clusters according the Elbow Method	11	4	2	9

We decided to do the validation with the combination conformed by the top 200 of words and 9 clusters. This was because the top 50 and 100 were really difficult to interpret given that the clusters grouped many themes together. Then, we only had left the top 10 and 200 and, as the top 200 had less clusters, we decided to do the validation with this case. In Fig. 6, the amount of words grouped in each cluster is shown.

	Etiqueta
0	743
1	2260
2	2144
3	23633
4	434
5	748
6	351
7	1691
8	1099

Fig. 7. Number of words per cluster

In that sense, we continued to show the respective word clouds to the psychologists mentioned above. Thus, they gave us their interpretation and what each cluster transmitted to them, tagged them and finally evaluated if there were clusters that needed to be splitted or joined.

The first obtained cluster (cluster[0]) had the words “solo”, “quiero”, “hace”, “hola”, “siento”, “hacer” as the most mentioned. However, this cluster wasn’t relevant to the psychologists as it didn’t provide any specific information on a determined topic; therefore, they recommended that we don’t consider this cluster. It was tagged as “No information” cluster. Also, they pointed out that “solo”, in english, probably referred more to an “only” than being associated with a loneliness feeling.

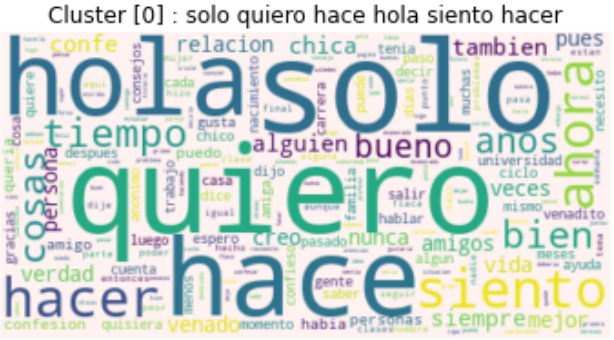

Fig. 8. “No information” word cloud

The second cluster (cluster[1]) had the words “hola”, “confe”, “confieso”, “alguien”, “confesiones”, “ciclo” as the most mentioned ones. This cluster was tagged as “Confesiones” to emphasize the role of these facebook pages for university students as an anonymous online space to vent. Additionally, a comment that came off with this cluster was about the need for anonymity that all of these pages reflect.

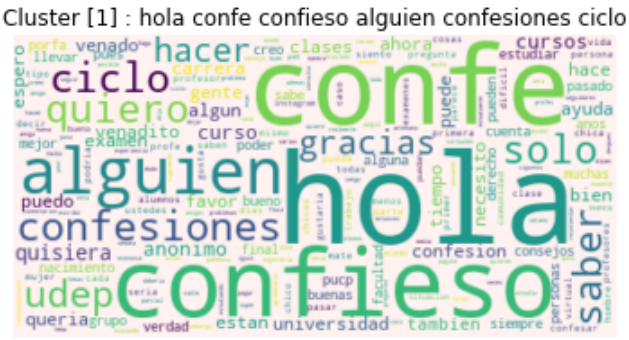

Fig. 9. “Confesiones” word cloud

The third cluster (cluster[2]) has the words “covid”, “solo”, “vacuna”, “miedo”, “contagio”, “pandemia” as the most mentioned words. Thus, it was tagged as a “Covid pandemic” cluster. The psychologists interpreted the relations of these words as a reflection of the fear of the pandemic and of catching the virus. Also, given the context, they inferred that this time “solo” did refer to the feeling of loneliness since the levels of anxiety and fear had increased in the general population during the covid pandemic.

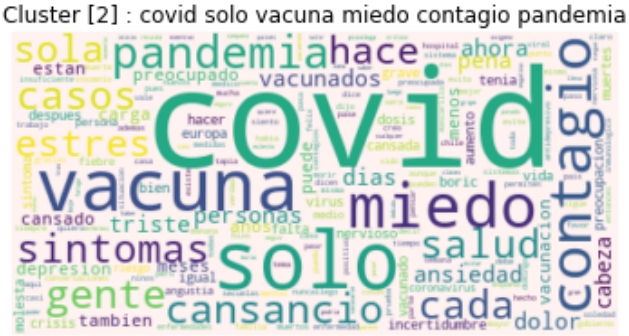

Fig. 10. “Covid pandemic” word cloud

The fourth cluster (cluster[3]) has the words “solo”, “peru”, “miedo”, “pena”, “castillo” and “congreso” as the most mentioned words. Thus, this cluster illustrated how the political crisis affects college students emotionally. “Grief and fear about what may happen in the country”. The psychologists also highlighted how this cluster in specific had more negative than positive emotions.

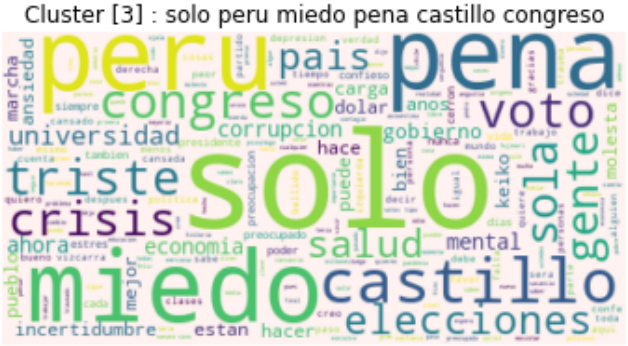

Fig. 11. “Political crisis during Castillo’s government” word cloud

The fifth cluster (cluster[4]) has the words “psicologo”, “salud”, “mental”, “crisis”, “universidad” and “ahora” as the most mentioned ones. For the psychologists, this was a clear example of how the mental health services have gained more relevance over the years. Before, these services were more related to seeking help because of personal problems, but now is also for getting support with shared activities like going to college. Additionally, in relation to the appearance of the word “trauma”, they pointed out that this is a word that can be used by students in a colloquial way; however, it could also mean how the tension situations that happened throughout the year marked the students.

Fig. 12. “Mental health services and support” word cloud

The sixth cluster (cluster[5]) has the words “nervios”, “problemas”, “salud”, “mental”, “crisis”and “dolor” as the most repeated ones. This cluster reflected the problems and concerns during these two years of pandemic. Words like “problema”, “salud”, “nervios”, “necesidades” could be related to familiar problem, while, in the other hand “elecciones”, “economia”, “corrupcion” to the concerns that arise with the political and economical context in Peru. Nevertheless, some words that caught up the attention of the psychologists were “sanados”, “rezar”, “jesus”, that translate into “cured”, “pray” respectively, which gave them the impression that this cluster in specific showed some signs of hope in the face of a difficult situation.

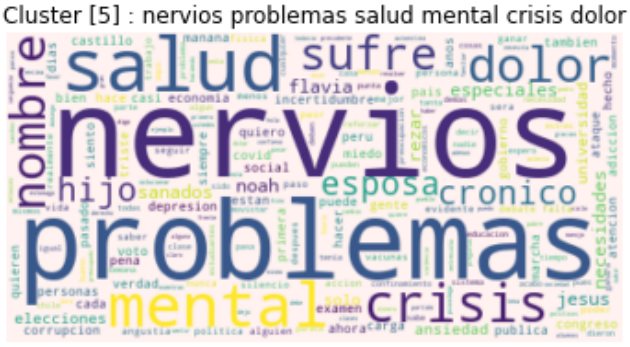

Fig. 13. “Worries” word cloud

The seventh cluster (cluster[6]) has the words “habla”, “peru”, “gente”, “solo”, “sigue”, “verdad” as the most mentioned words. This cluster summarized the feelings of fed up for the Peruvian political class. Also, the presence of words such as “risa”, “laugh”, in the cluster context, gave the impression that there was also a sarcastic feeling between the communities with different opinions. In this

part, the psychologists mentioned how there were many debates around the existence of the coronavirus and the political crisis in Peru.

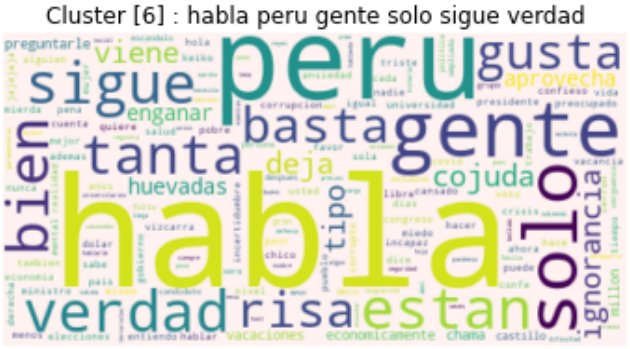

Fig. 14. “Fed up with the political crisis” word cloud

The eight cluster (cluster[7]) has the words “vacunas”, “impotencia”, “solo”, “covid”, “gente”and “carga”. This cluster translated into the topic of vaccination. However, the Peruvian context summed up to this cluster with the scandals related to the purchase of vaccines, that’s why there were many feelings associated with impotence and sadness.

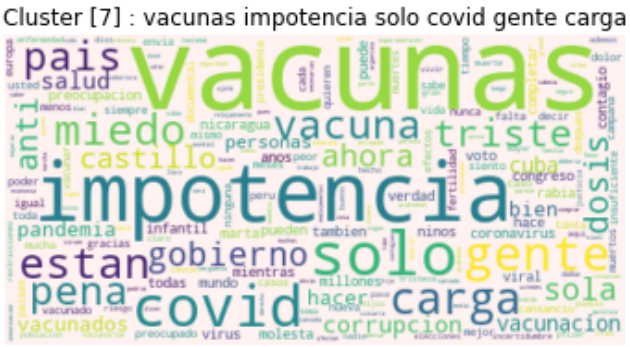

Fig. 15. “Vaccination” word cloud

The last cluster (cluster[8]) has the words “nadie”, “merino”, “solo”, “pena”, “gente” and “sola” as the most repeated ones. Again, the reaction of college students faced the political crisis and, through these comments, the psychologists saw the emotions felt in that period of time and the concerns of what could come next. In addition, the psychologists recommended joining this cluster with cluster[3]: “Political crisis during Castillo’s government”.

Fig. 16. “Political crisis during Merino’s government” word cloud

V. Conclusions and Recommendations

In conclusion, through data preprocessing, machine learning and visualization techniques it was possible to analyze the sentiments that Peruvian college students expressed in social media, such as Twitter and Facebook, during the COVID-19 pandemic. We were able to extract enough data from these environments to feed our techniques and consider the scope of the exploration as satisfactory**.** With this, it was possible to obtain the most frequent terms that these students used to describe their feelings on different topics, which vary from university-related to societal-political-related sentiments. The clusters obtained and represented in the form of word clouds allow us to conclude that college students make explicit their thoughts about their disconformities and fears around a series of topics. These topics include the pandemic and its consequences on the population’s mental health, the political crisis that Peru is passing through and the constant worry about the future of the nation, the importance of a channel to express anonymously their ideas and the growing need of psychological counseling to deal with mental health issues.

After consulting the psychologists contacted to tag the clusters, following their recommendations, the cluster “No information” could be excluded from the investigation since it doesn’t give any insights. Also, the clusters “Political crisis + Castillo” and “Political crisis + Merino” could be joined together to make one cluster about “Political crisis”. This way we would only have seven relevant clusters in the end.

In the making of this investigation, we understood that the way the psychologists work with the students is very personalized to their feelings. We can’t recommend specific strategies about their work since it’s not our field of study, but we hope the results of this paper add to the vision they have about college students and how they live everyday with stressful situations like the COVID19 pandemic and political crisis in Peru while dealing with academic and personal worries, all at the same time.

As said previously, this study could reach a satisfactory but not excellent scope due to some restrictions in the tools used. For example, the Tweepy and facebook-scraper libraries only offered few parameters that didn’t allow us to reach further dates in the past, specify localizations and obtain a certain amount of data from the websites. This happened mainly because we only disposed of standard accounts in the API’s used. Also, there might exist techniques to identify tweets that were written specifically by college students in a certain country, so that the imputation techniques affect the output as little as possible. Finally, further research could improve the search methods for the data extraction process. For example, future works might enrich the variety of search terms for scrapping. Moreover, when extracting data from confessions pages in social media, it would be recommendable to apply an extra cleaning process to those posts by removing words like the following: “confe”, “hola”, “tío”, etc., because as they are highly mentioned in the Facebook posts, it might be

skewing our TF vectorizer with words that are repeated many times but don't provide any additional information.

With all these recommendations together, researchers could reach new horizons in the mental health knowledge base and find new insights that help not only college counseling services, but the psychologists society as a whole to effectively fight this problem at a bigger scale.

References

[1] S. Dattani, H. Ritchie, and M. Roser. (2018, April). "Mental Health Worldwide," Global Burden of Disease. [Online] Available: https://ourworldindata.org/mental-health

[2] A. Brunier. (2020, October 5). COVID-19 disrupting mental health services in most countries, WHO survey. [Online] Available: https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

[3] J. X. Koh and T. M. Liew, "How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds," *Journal of Psychiatric Research*, 2020, doi: 10.1016/j.jpsychires.2020.11.015.

[4] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen, "Detecting Community Depression Dynamics Due to COVID-19 Pandemic in Australia," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 958–967, Aug. 2021, doi: 10.1109/TCSS.2020.3047604.

[5] A. Park, M. Conway, and A. T. Chen, "Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach," Comput. Human Behav., vol. 78, pp. 98–112, 2018, doi: 10.1016/j.chb.2017.09.001.

[6] Cobb, F., Yarger, L. K., & Pinter, A. T. (2018). Text Mining Mental Health Reports for Issues Impacting Today's College Students: Qualitative Study. JMIR Mental Health 2018;5(4):E10032 Https://Mental.Jmir.Org/2018/4/E10032, 5(4), e10032. https://doi.org/10.2196/10032

[7] Scikit-learn. (2011). sklearn.preprocessing.StandardScaler. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Appendix

1. TABLE 1. LIST OF CONSULTED FACEBOOK PAGES

N°	Page name	URL
1	Confesiones UP	https://www.facebook.com/ConfesionesUP
2	Confesiones ULima	https://www.facebook.com/Ulima-Confesiones-108361104374215/
3	Confesiones PUCP	https://www.facebook.com/confesionespucp1
4	Confesiones UNI	https://www.facebook.com/UNIconfesiones
5	Confesiones ESAN	https://www.facebook.com/ConfesionesEsanOficial
6	Confesiones San Luisanas	https://www.facebook.com/Confesiones-San-Luisanas-128712019139 9049/
7	Confesiones Molineras	https://www.facebook.com/ConfesionesAGRARIALAMOLINA
8	Confesiones AdUNI	https://www.facebook.com/ConfesionesAduni1
9	Confesiones San Juan Bautista	https://www.facebook.com/eltioeliasupsjb
10	Confesiones César Vallejo	https://www.facebook.com/Confesiones-Vallejo-Piura-101830881472 022
11	Confesiones USMP	https://www.facebook.com/USMPConfesiones
12	Confesiones PUCP	https://www.facebook.com/pucpconfesiones
13	Confesiones URP	https://www.facebook.com/Confesiones-URP-1566790033536399
14	Confesiones UPC	https://www.facebook.com/UPCConfesionesOficial
15	Confesiones UNSA Oficial	https://www.facebook.com/Confesiones-UNSA-Oficial-ツ -2199621426764072
17	Universidad de Piura	https://www.facebook.com/Confesionesudep

2. FIGURE 2. LISTS OF WORDS USED TO SEARCH TWEETS

```
Contextos = ['universidad', 'uni', 'covid', 'coronavirus', 'peru', 'castillo',
'elecciones', 'voto', 'economia', 'marcha', 'vizcarra', 'vacunas', 'salud mental'
'congreso', 'merino', 'corrupcion', 'dolar', 'vacunagate', 'remoto', 'bellido',
'cerron', 'virtualidad', 'crisis', 'clases virtuales', 'movistar', 'contagio',
'uci', 'oxigeno', 'keiko']

Sentimientos = ['triste', 'ansioso', 'ansiosa', 'ansiedad', 'depre', 'carga',
'depresión', 'desanimada', 'desanimado', 'desmotivado', 'desmotivada', 'sola',
'solo', 'soledad', 'molesta', 'molesto', 'insuficiente', 'incertidumbre','miedo',
'angustia', 'angustiada', 'angustiado', 'autoestima', 'cansancio', 'cansada',
'cansado', 'desaliento', 'desalentado', 'desalentada', 'estresada', 'estresado',
'estres', 'impotencia', 'impotente', 'insatisfaccion', 'insatisfecha',
'insatisfecho', 'insomnio', 'nervioso', 'nerviosa', 'nervios', 'nerviosismo',
'pena', 'apenada', 'apenado', 'psicologo', 'psicologa', 'recaida', 'adiccion',
'adicciones', 'adicto', 'adicta', 'trauma', 'traumada', 'traumado', 'afliccion',
'afligido', 'afligida', 'decaida', 'decaido', 'preocupada', 'preocupado',
'preocupacion']
```

Fig. 2. Lists of words related to context and feelings used to search tweets of interest.