Research Type (Original Article)

Text Mining and Unsupervised Learning Techniques for Tweet Classification in the Peruvian Social Context

Victor Guevara-Ponce[1*], Roque Paredes Ofelia[2], Orlando Iparraguirre-Villanueva[3]
[1,2] Instituto de Datos e Inteligencia Artificial, Universidad Ricardo Palma, Lima, Perú
[3] Ho Nexus, Nuevo Chimbote, Perú

Article info

Keywords:
Text mining
unsupervised learning
tweet classification
social upheaval
Peru
quit

Abstract

Background: Currently, there has been an exponential growth in the volume of unstructured data, especially with the use of social networks. Technological progress has allowed the adoption of processes, techniques, and methods to obtain information from these data.
Objective: This work aims to analyze and classify Tweets in the context of Peru's social upheaval, using text mining (TM) and unsupervised learning (UL) techniques.
Methods: More than 268k tweets were collected and processed, with the hashtag of the trends occurred in the first two weeks of February 2023: #ParoNacional, #RenunciaYa and #Renuncia. Within a radius of 1000km from the city of Lima. Data cleaning and feature selection techniques were used. Then, UL techniques, such as clustering and sentiment analysis, were applied to classify the Tweets generated in social networks using the Loss Distribution Approach (LDA) model.
Results: The result of the analysis of Tweets related to the social upheaval in Peru shows a polarization in the opinions of users, with one group supporting the protests and another criticizing them. Recurring themes have been identified as the resignation of the president, Lima centralism, congressional shutdown, corruption, and social inequality. The sentiment analysis shows a mix of positive and negative emotions, with words grouping negative sentiment being the most recurrent. Sentiment was classified into three polarity categories: positive, negative, and neutral, corresponding to 37%, 58% and 5%, respectively. There is also criticism of the government for its lack of action and police violence during demonstrations.
Conclusions: Finally, it is concluded that UL techniques have been effective in classifying Tweets according to their polarity.

1. Introduction

The growing popularity of social networks has led to a huge amount of information generated by users, particularly on the Twitter platform [1], this social space facilitates people to communicate virtually with family, friends and even create new ones. In addition, this tool allows users to share content, interact with each other and create communities of people with similar interests: work, reading; it also facilitates political relations and business contacts [2]. Peru currently has a total population of 33.52 million users, 78.7% of whom live in urban areas; the number of users with Internet access is 21.89 million, representing 65.3% of all residents who have access to this space, and 28.1 million are active in social networks. Twitter has 2.2 million users connected to the Peruvian population [3], [4]. Peru is currently experiencing a period of political crisis unprecedented since the restoration of democracy. Various demonstrations have been held in different parts of the country, specifically in southern Peru. The demonstrations began in December after the arrest of the then president of the Republic Pedro Castillo. Various politicians, researchers, students, and people in general use social networks to express their disagreement or support to the current president of the Republic Dina Boluarte. On Twitter, a user makes an average of 6 comments on all types of tweets for 30 days [5].

Due to technological advancement and computers, most documents are digitally unstructured. As the term indicates, TM involves the search, retrieval, and analysis of unstructured natural language text that is not structured in any way. It is for the purpose of observing how text data embedded in these documents can be used for various purposes [6]. This huge volume of unstructured data provides a unique opportunity to apply data science techniques to extract valuable information [7]. TM and UL techniques are fundamental tools in this field and have been used for data classification in various contexts [8]. In addition, TM provides methodologies integrating issues related to artificial intelligence, knowledge discovery, data mining, among others. Also, the use of algorithms to analyze their own data [9]. These tasks include preprocessing, text

classification, information retrieval and search, document clustering, and document information retrieval [10].

This paper studies the application of TM and UL techniques to classify tweets in the context of social unrest in Peru. The objective is to identify patterns of behavior and related topics in tweets generated by Peruvian users related to social and political events in the country. Sentiment analysis, clustering, and word frequency analysis techniques will be used to explore the data and classify tweets into different categories. For this purpose, comments have been extracted, using the hashtags of the trends that occurred in the first two weeks of February 2023: #ParoNacional, #RenunciaYa #Renuncia. This study is expected to be of significant importance in understanding public opinion on social and political events in Peru. In addition, they are expected to identify opportunities for improvement in online communication and marketing. In addition, it is expected that this research will provide a solid foundation for future research on TM and UL as it relates to social uprisings in Peru.

2. Literature Review

The last decade has seen a steady evolution of TM and sentiment analysis in social networks. Real-time information analysis has gained popularity, given the growing interest in the meaning of data. In this section, we present research related to tweet classification using TM and UL.

2.1. Sentiment analysis on Twitter

In recent years, Twitter data classification has become a popular research topic. Sentiment analysis is a TM technique used to determine the emotional side of a text. Within the realm of Twitter, sentiment analysis is utilized to establish the positivity, negativity, or neutrality of a tweet. For example, in [11] conducted research focused on Twitter sentiment analysis, for which they employed UL techniques. They used aggregation models to organize the tweets into classes such as: positive, neutral, and negative. The results achieved a classification accuracy of 83%. Also, in [12] conducted work with UL and TM techniques to analyze tweets posted on Twitter. The distribution of the results of the analysis was organized into three types of polarity classes: positive, negative, and neutral, with the following results: 22%, 4% and 74%, respectively. Similarly, in [13] researchers analyzed the feelings and emotions of people from 10 countries during the COVID-19 pandemic. Their results allow concluding that all countries tweeted positive feelings about COVID19; from the analysis of word clouds from different countries, people tweet words such as epidemic, COVID, coronavirus, clinics, health conditions, struggle, stay, subsist, protected, assistance, emergency, death, and Masks with different emotions, mostly positive. Likewise, in [14] a method based on nested tweet representations is presented to analyze Twitter data related to Peruvian politicians' comments and evaluate the underlying sentiment polarity of such messages based on neural networks to predict politicians' approval ratings. It concludes with an accuracy level of 91%.

2.2. Unsupervised learning techniques

UL techniques are widely used in TM for document classification and clustering. For example, in [15] developed a project in which they used the LDA model to classify topics in a corpus of texts and then identify the most relevant terms. In this processing they used more than 10k documents (curricula) related to data science. The work was able to cluster relevant topics with an accuracy of 87%, and they conclude that those interested in the discipline of data science should have certain skills, such as being technical in some line of technology. Also, in [16] a work was developed using the UL approach to classify tweets written in English, from customers of multiple telecommunication companies in Saudi Arabia. For the classification they used ML algorithms, neural networks, k- Nearest Neighbors, Naive Bayes and classified into the following categories: positive, neutral, and negative. And they concluded that KNN model with one variant achieved an accuracy of 80.1%.

2.3. Text Mining

TM is a research field that combines data mining, statistics, and computational language for the purpose of extracting useful information from texts. In [17], a work based on TM for the classification of Tweets is presented, where the processing of Twitter data, including their opinions and sentiments by processing the subjectivity of interactions is sought. This study concludes that using machine learning (ML) and TM techniques, it is possible to classify comments on the social network with an accuracy of over 60%. Similarly, in the work [18] they addressed a TM case, for which they analyzed approximately 600 000 Twitter posts published in a group (A) of users and compared it with more than 400 000 messages in another group (B) of Twitter users. The purpose of this research was to determine the preferences of users belonging

to group A, regarding aspects such as the use of varied terms, the expression of emotions and feelings, as well as the use of grammatical structures in their publications. To achieve this, techniques such as sentiment analysis, natural language processing and topic modeling were employed. As a result, it was concluded that users shared mostly positive emotions, used a wide range of emoticons, and had a medium level of lexical diversity. In addition, the topics they tended to address ranged from more mental health-related issues to more practical work- related topics. Similarly, in [19] data collected from the major media outlets in Uganda, both print and non-print, were analyzed using the Twitter platform and TM techniques, with the aim of obtaining valuable and relevant information. After analyzing the data, it was determined that sentiments associated with security, politics and economy were mostly negative, while those related to sports were positive. Also, in [20] they conducted a sentiment analysis work to collect people's comments about tourism in Oman through social media posts, the dataset consisted of tourists' comments about the country. The authors proposed a proprietary tourism ontology for Oman based on ConceptNet.

3. Methodology

This section describes the methods and materials used in the research for the classification of tweets using TM and UL techniques to analyze trends on the current political situation in Peru.

3.1. Machine learning / Unsupervised learning

as a branch of computer science is one of the most application-oriented fields of artificial intelligence. They are inspired by the ability to learn, the ability to acquire new or additional information, which is one of the important characteristics of a living being. ML as a science focuses on finding and developing algorithms that can mimic or simulate the mental abilities of living beings in terms of learning [21], [22]. Among the main tasks is supervised learning that allows solving the input and output case on available data sets [23]. In other words, training data consists of supervised learning of input-output pairs. For each training data input, we know the corresponding output, which can be used to drive the learning algorithm as a control signal [24]. UL on the other hand only solves problems where one has access to the input data by collecting the training data with the proper models of this class, as seen in Fig. 1. A good algorithm should be able to meet some criteria to group only similar inputs using information from all possible inputs, since two inputs are considered similar only if they are expected to have the same output label [25], [26].

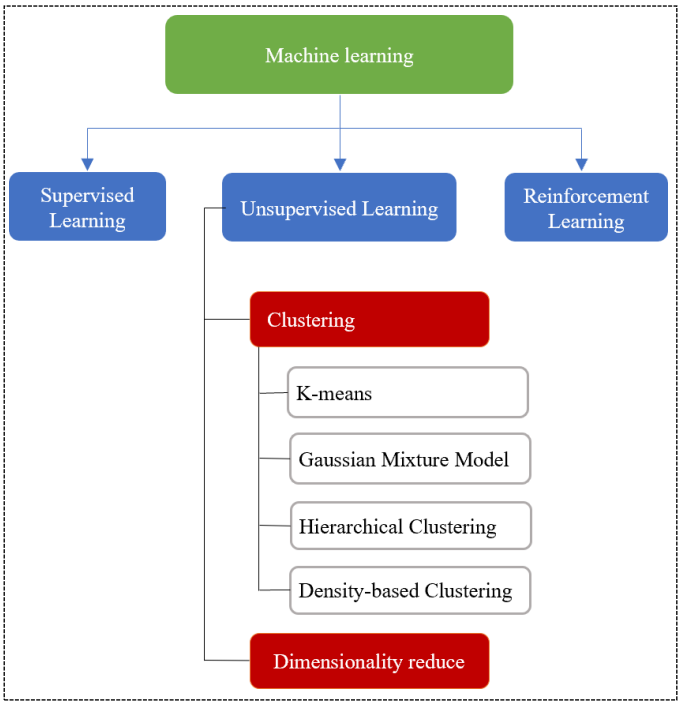

Fig. 1Unsupervised learning models

3.2. Text mining

It is an exciting field with new research techniques and software tools being used in a variety of settings, including academia, business, and government agencies [27]. Currently, researchers are using TM

in an ambitious project to forecast a wide range of events, from stock market fluctuations to the timing of political protests. In addition, TM is applied in many commercial fields, such as market research, and is also used in government and defense operations [28], [29]. TM seeks to discover valuable information within a dataset, and to do so, several distinct steps are carried out. These steps, summarized in 6 points, are based on the proposal presented in reference [30] and are detailed in Fig. 2.

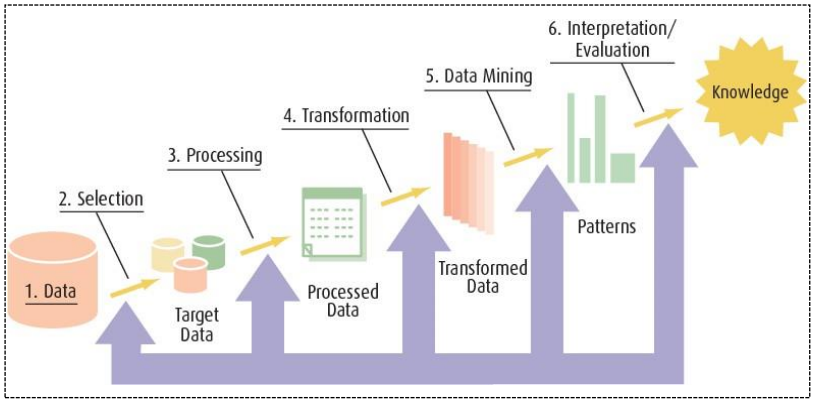

Fig. 2 Text mining process

3.3. Understanding the data set

In this section, we seek to understand the dataset, tweet classification involves understanding the nature and structure of the data used in the classification process. In this case, the data are tweets, which are short messages posted on Twitter. This process follows the following steps.

The first step consisted of collecting the data, for which the Twitter API was used to obtain the tweets published with the hashtag of the trends that occurred in the first two weeks of February 2023: #ParoNacional, #RenunciaYa #Renuncia. A total of 268,238 tweets and retweets were selected. In addition, the reference point was considered the city of Lima geocode = "-12.043180 -77.028240" and a radius of 1000 km. The dataset is composed of repetitive terms such as: "renuncia"," comunista"," lima", "boluarte", "dina", "Perú", "congreso", "tomar", "exigir" and asumir", which is frequently used during this stage of social upheaval. As a next point, it is important to define the characteristics that characterize the texts well and that are appropriate for the task at hand. Features are usually based on the content of the documents. A very simple approach, bag-of-words with binary attribute weighting, takes each word as a Boolean feature. Its value indicates whether the word is in a document or not. For the study, it uses word cloud and relations between words, as shown in Fig. 3. Previously, the data was preprocessed through text cleaning and tokenization.

Fig. 3 Most frequent words

Data preprocessing involves preparing and cleaning the text for classification, as online texts can contain a lot of noise and little useful content [31]. Keeping all words can increase the complexity of the problem, making classification difficult by considering each word as one dimension. It is believed that proper data preprocessing will reduce text noise, improve classifier performance, and facilitate real-time opinion analysis. The complete process includes several stages, such as the removal of unnecessary characters in the text, the elimination of blank spaces, the conversion of abbreviations to their full forms

and the reduction of words to their basic lexical form. The collected tweets were preprocessed to remove noise and redundancies. Data cleaning techniques, such as the removal of special characters, punctuation marks, and URLs, were performed. In addition, duplicate tweets were removed, and a tokenization and lemmatization process were performed to restore words to their base forms. As can be seen in Table 1.

Table 1Most likely terms within the probabilities

#	Topic 1	Topic 2	Topic 3	Topic 4	Item5
1	los	que	elecciones	del	las
2	elecciones	asi	como	son	quieren
3	izquierda	debe	año	les	eso
4	una	elecciones	presidenta	elecciones	este
5	mismo	estan	mas	con	ellos
6	que	quiere	congreso	candidatos	sin
7	pero	adelanto	que	hace	mismo
8	ello	hay	congreso	hoy	como
9	adelanto	votos	queremos	pueblo	tener
10	nos	como	estas	todo	que
11	Congreso	ellos	todos	renuncia	nuevas
12	Sus	pais	boluarte	esta	generales
13	democracia	para	nos	van	presidente
14	son	congreso	reformas	ser	van
15	peru	una	ellos	mas	elleccione s
16	proximas	ahora	las	quiere	adelanto
17	quieren	hacer	peru	proximas	asi
18	Congreso	que	por	congreso	los
19	como	presidenta	adelanto	asamblea	mas
20	las	elecciones	asi	mas	esta

Table 1 shows the trends in the data. For example, the main topics include elections, congress, and candidates, suggesting that the political situation is a significant concern in this context. The word "left" also appears on the list, suggesting political positions are discussed. Other words such as "resignation", "democracy", "reforms" and "president" suggest that political issues in Peru are significant to Twitter users.

During this phase of the process, the goal is to discover patterns in the data that can solve the processing. To achieve this, it is necessary to carefully select a specific and appropriate model for the type of problem being addressed. Once selected, the properties and parameters are defined and applied to the data to solve the problem. It is essential to remember that each model has different characteristics that can affect both the process and the results of the techniques used. It is crucial to correctly fit the model to avoid overfitting, which usually requires testing to find the right structure and optimal parameters. Furthermore, it is imperative to keep in mind that the model fit depends on the size of the data set. This means that the same model may perform well on one data set and be less accurate on another. Therefore, it is critical to select an appropriate model and make sure that the parameters fit well to the specific data set. In the case of sentiment analysis, this is an unstructured text classification problem, so it is essential to take these characteristics into account when selecting the appropriate model and methods to solve the problem.

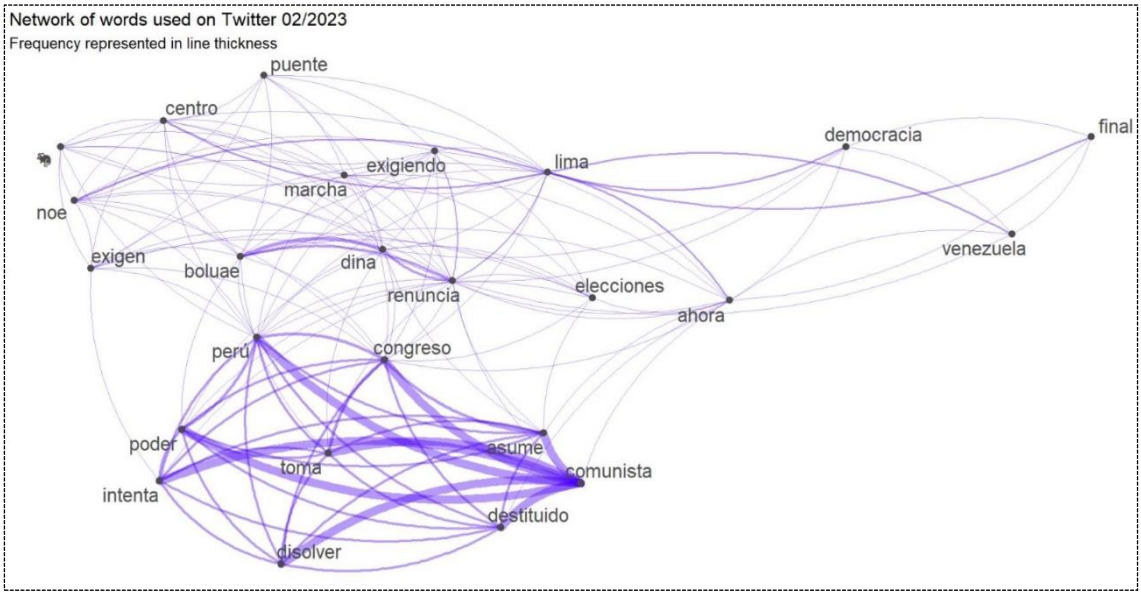

Fig. 4 Number of words most used on Twitter during the month of February 2023

In fig. 4, you can see the most used words on Twitter during the social upheaval in Peru after former president Pedro Castillo's self-coup. From this list, it is possible to identify some key trends and themes that emerged during that period. First, the words "Congress" and "communist" are the most frequently used words, suggesting that Congress's role in Peruvian politics was a significant issue at the time. The word 'elections' appears in Fig. 4, indicating that electoral politics was a highly contested topic for Twitter users. Similarly, the words "resignation," "impeached," and "try" suggest that President Castillo's power and authority were in question. In addition, the word 'power' appears in Fig. 4, suggesting that the struggle for power and authority was a significant issue at the time. In addition, the word "democracy" also appears on the list, indicating that concerns about democracy and governance were present during this period. Other words such as "march," "take," "dissolve," "demand," and "Lima" suggest that protest and social mobilization were significant themes during Peru's social upheaval. Overall, this set of words provided us with an overview of the issues and concerns that emerged during the social upheaval in Peru. This was following former president Pedro Castillo's self-coup and President Dina Bouarte's inauguration. It is imperative to note that this set of words is only a sample of comments on Twitter. It does not necessarily represent all Peruvians' opinions.

In this context, it is important to emphasize a thematic model such as LDA. This model helps to identify hidden concepts and salient features by reducing the dimension of the data set. Dimension reduction is performed by decomposing the original matrix into a factor matrix using probabilistic or non-probabilistic models [32]. LDA is a probabilistic generative model designed to work with text datasets. Its fundamental concept is that documents are composed of a random combination of potential topics, and each topic is defined by its distribution of words [33].

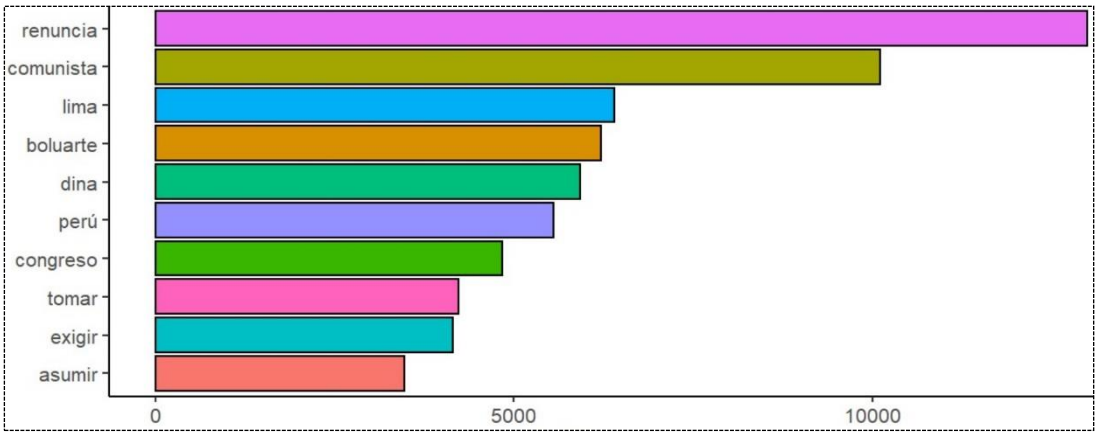

Fig. 5 Exploratory analysis of the top 10 most used words

In Fig. 5 we can appreciate the 10 most used words in the dataset which are: "Renunciation", "communist", "lima", "boluarte", "dina", "Peru", "congress", "take", "demand" and "assume". We can observe that the word "resignation" is the most used word with a frequency of more than 12000 times, suggesting that the demand for the resignation of the president was a central topic in the conversation on Twitter in the context of the social upheaval in Peru. The words "communist", "lima", "boluarte", "dina" and "congress" also appear with a significant frequency, suggesting that these topics have also been relevant in the conversation. Also, we can observe that the words "resignation" and "Peru" have been very recurrent themes, which suggests that the population was asking for the resignation of the president and is related to the political situation in Peru in general. In this work we conducted the exploratory analysis of the 10 most used words on Twitter during the period of social upheaval in Peru, we managed to obtain valuable information about the topics and concerns that are most relevant in the conversation of users. The request for the resignation of the The request for the president's resignation and political issues are clearly priorities, and a certain correlation between geographic and political issues in the conversation can also be observed.

4. Results and discussion

This work aims to use TM and UL techniques to classify tweets related to the social upheaval in Peru during February 2023. TM is used to extract useful information and knowledge from text data, while UL is used to find patterns and relationships in the data without the need for prior labels. Classifying tweets in this context can be valuable to better understand Twitter users' sentiments and opinions about the situation in Peru.

The data was collected using the Twitter API and filtered to include Spanish-only tweets containing keywords related to the social upheaval in Peru after the self-coup. The 268k+ data were preprocessed using text cleaning techniques, such as removing punctuation marks, numbers, and irrelevant words. Then, a word vectorization technique was applied, using the Word2Vec model, to represent each tweet as a feature vector.

The analysis revealed that tweets related to the social upheaval in Peru after the self-coup can be classified into five main categories: protests and demonstrations, violence and repression, politics and government, support and solidarity, news, and media. Many tweets fell into the protests and demonstrations category, followed by violence and repression.

The politics and government category had the least number of tweets. For data analysis and using sentiment analysis, this is classified into feelings (negative, positive, fear, trust, anger, sadness, disgust, foreboding, joy, and amazement), as presented in Fig. 6, the top 10 most frequent types of feelings

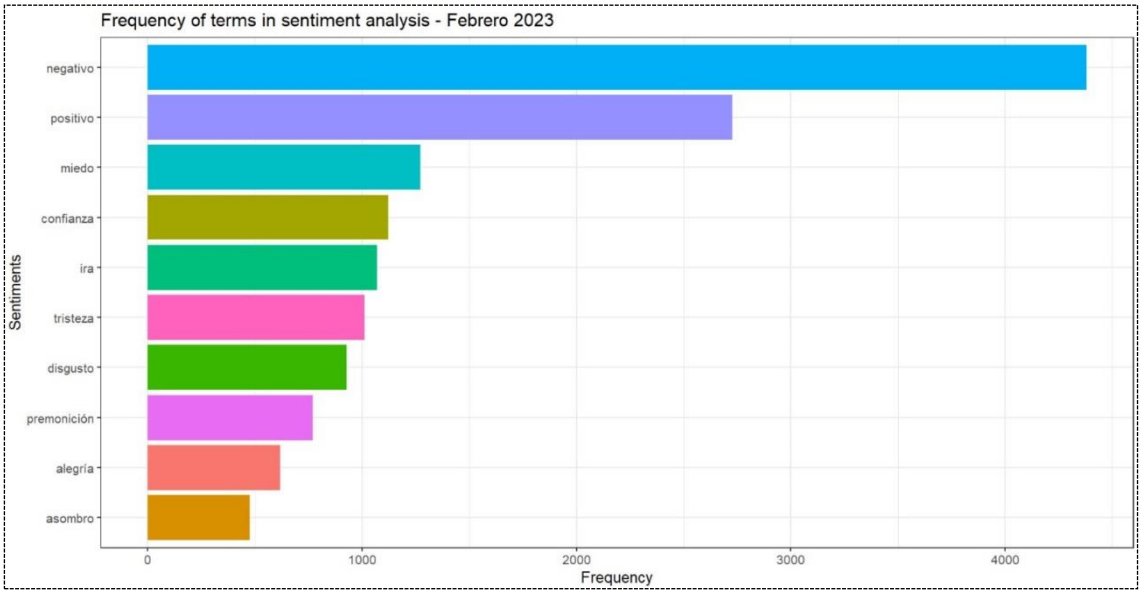

Fig. 6 Representation of feelings

In Fig. 6, natural language processing and ML techniques have been used to classify each tweet into one of these sentiment categories. Each sentiment is then analyzed:

- Negative sentiment: Negative sentiment was found to have the highest number of tweets in the context of the social upheaval in Peru in February 2023. Tweets with this sentiment expressed a wide range of negative emotions, such as frustration, anger, sadness, fear, worry and despair.
- Positive sentiment: Although negative sentiment was the most common, it was also found that many tweets expressed positive sentiments, such as hope, support, solidarity, and optimism. These tweets were less common than negative tweets but were still significant in number.
- Fear: The feeling of fear was one of the most common sentiments found in tweets related to the social upheaval in Peru in February 2023. Tweets with this sentiment expressed fear for the security and stability of the country, as well as for the situation of citizens.
- Trust: Tweets expressing trust were less common than negative or fear, but still significant in number.
- Anger: Tweets expressing anger were common and reflected frustration and anger at the current situation in Peru. These tweets also expressed outrage at the violence and repression that has been seen in some parts of the country.
- Sadness: The feeling of sadness was common in tweets related to the social upheaval in Peru. These tweets expressed sadness about the current situation, loss of life and violence in the country.
- Disgust: Tweets expressing disgust were less common than negative tweets, but still significant in number. These tweets expressed displeasure with the current situation and the lack of action by leaders and institutions.
- Premonition: The feeling of foreboding was less common than other feelings, but some tweets expressed a sense of unease and worry about what might happen in the future.
- Alegria: Tweets expressing joy were less common than negative ones, but still significant in number. These tweets expressed joy at the positive news and changes seen in some parts of the country.
- Amazement: The feeling of amazement was less common than other feelings, but some tweets expressed a sense of surprise at the current situation and the way events have unfolded.

It is important to discriminate which terms are considered in these ten sentiment categories. To do this, the words in the tweets are divided by type of sentiment, which allows us to analyze each of the words associated (5 words) to a particular sentiment. As shown in Fig. 7.

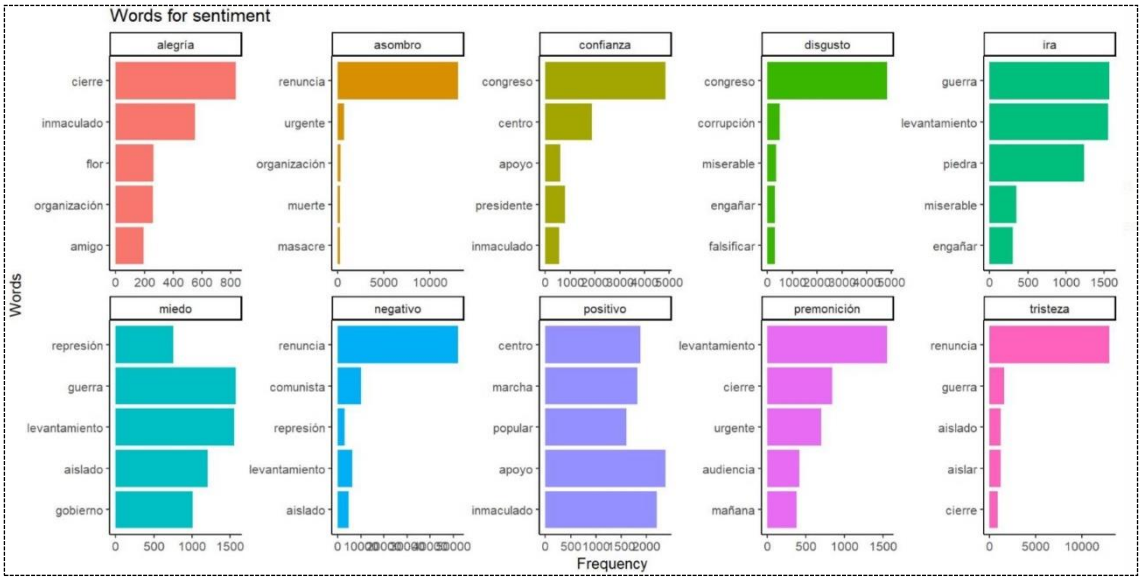

Fig. 7 Top 5 words associated with a type of feeling

Fig. 7 shows the ten categories of feelings: joy, amazement, confidence, disgust, anger, fear, negative, positive, premonition and sadness. Each category groups five words associated with this feeling. For example, the feeling joy, groups the words: close, immaculate, flower, organization, friend. Also, the word astonishment, groups the words: resignation, urgent, organization, death, and massacre. And so on, with the other feelings. At this point in the process, it is critical to implement the strategy to assess the model. Since it is about analyzing comments or ideas shared by users on Twitter, it is necessary to use an unsupervised ML method in order to obtain a larger amount of data and, thus, generate new knowledge. One of the most suitable models for data exploration is topic modeling is LDA. The main objective of topic modeling is to determine the proportional composition of a fixed number of topics in the documents of a collection. The first step to start with this technique is the number of topics "K", which is a fundamental parameter that must be defined beforehand. For this purpose, two metrics "Griffiths2004" and "Arun2010"

are proposed in this work. The metric "Griffiths2004" is used to evaluate the semantic coherence of the groups. The idea is that groups should be internally consistent and externally distinct. This metric is commonly used in topic analysis and topic modeling in texts. On the other hand, the "Arun2010" metric is based on mutual information and is used to evaluate the quality of clusters in terms of their ability to reveal relevant and distinctive information. Fig. 8 shows the results of the two metrics. The two metrics are approximately at K=20 topics, which is why this number of clusters was chosen.

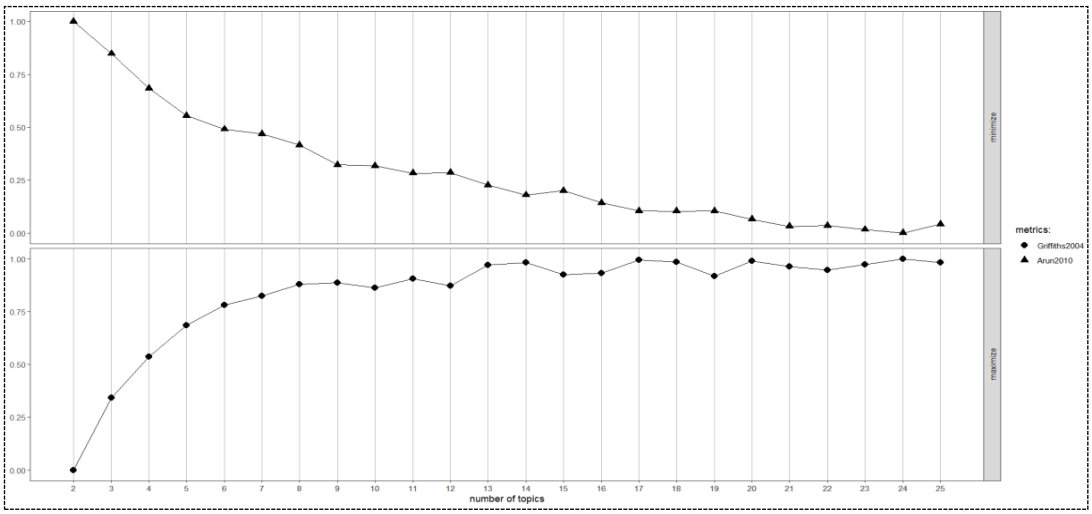

Fig. 8 Metrics for model evaluation

The distribution of topics within the document according to the model and used is shown in Fig. 9. All three documents show at least a small percentage of each topic. However, two or three topics dominate each document. For example, "elections", "advance" and "congress", are the most repeated words in each group, although, the term "elections" is the one that predominates in all three groups. The analysis tries to obtain a more meaningful order of the main terms according to the topic by reclassifying them with a specific score. The purpose of the reclassification of terms is like the idea of term frequency - inverse document frequency. The more frequently a term appears relative to its probability at higher levels, the less meaningful it is in describing a topic. Therefore, the extended evaluation prefers terms to describe the topic. The results can be seen in Fig. 9.

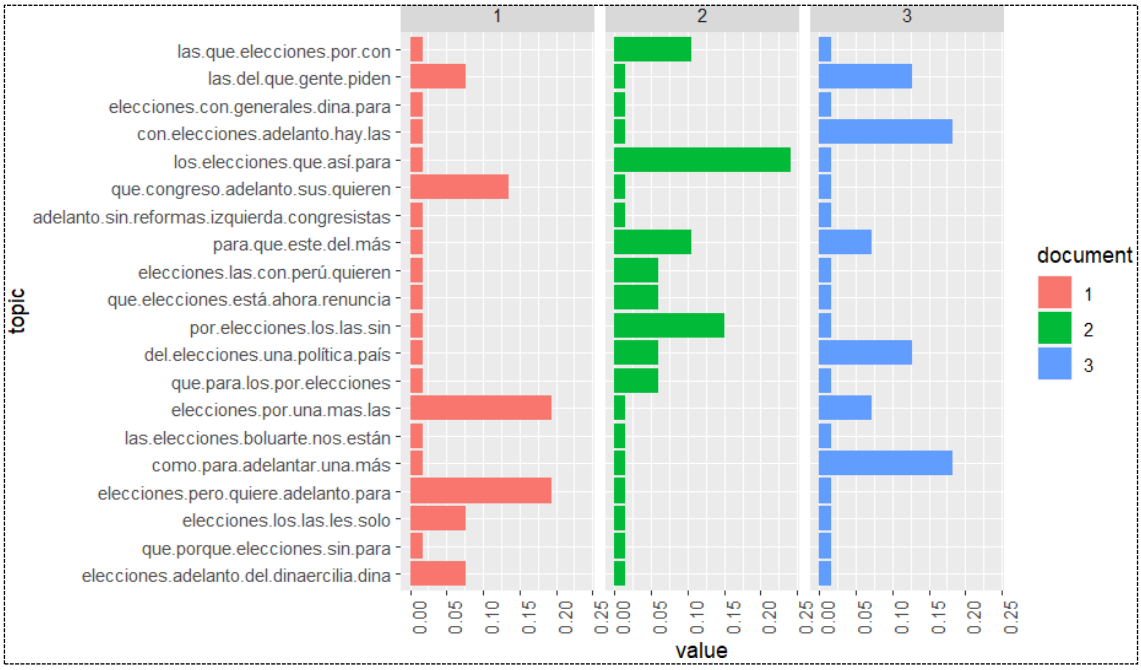

Fig. 9 Classification of topics

Finally, it can be said that the study conducted using TM and UL techniques to classify tweets related to the social upheaval in Peru in February 2023 has been very useful to understand the feelings and opinions of Twitter users in this context. In the preprocessing of the data, which included the removal of punctuation marks, numbers and irrelevant words, and the word vectorization technique, using the Word2Vec model, allowed us to represent each tweet as a feature vector. In this way, UL techniques could be applied to discover patterns and relationships in the data without the need for prior labels. The results revealed that tweets related to social upheaval in Peru can be classified into five main categories: protests and demonstrations, violence and repression, politics and government, support and solidarity, news, and media. Most tweets were classified in the protests and demonstrations category, followed by violence and repression. The first results synthesize that the most used words (top 10) correspond to: resignation, communist, Dina, Boluarte, Peru, Congress, Take, demand and assume. In Fig. 5 and Fig. 4, the most used words are represented for a better understanding. In addition, a sentiment analysis was applied that allowed classifying each tweet into one of the sentiment categories, including negative, positive, fear, trust, anger, sadness, disgust, foreboding, joy and amazement. It was found that negative sentiment (Resignation, communist, repression, uprising, and isolated) was the most common sentiment in the tweets, followed by the sentiment of manifestations fear and anger. However, tweets expressing positive sentiments such as hope, support, solidarity and optimism were also found. This result is closely related to the work [9], where they used UL and data mining techniques to classify and analyze the content posted on Twitter. The results of the analysis were distributed in three polarity categories: positive, neutral, and negative. The following percentages were obtained: 22% for positive polarity, 4% for neutral polarity, and 74% for negative polarity. Similarly, with the work [14], the results obtained in this work are linked in the political context, since in the research [14] they analyzed Twitter data related to the comments of Peruvian politicians and evaluated the polarity of the underlying sentiment of these messages based on neural networks to predict the approval ratings of politicians. Also, the results of topic modelling posit that 20 major themes or topics can be generated, among the words that stand out are: "congress advance elections, elections by one more and elections want advance for topic 1", "elections that so, by elections and that elections for topic 2", and "as to advance one more, with elections advance there are and general elections for topic 3", these results, are related to the work [15], in which they used LDA to identify the themes present in a set of texts and then classified the words that were relevant. The results of the work showed an accuracy of 87% in clustering important topics.

Sentiment analysis indicates that most of the comments have a negative connotation (over 60%). The words most associated with this sentiment are resignation, communist, uprising, isolated and repression (top 5). On the other hand, positive sentiment is less frequent, and the most related words are support, immaculate, center, march and popular (top 5). This analysis, also, has some similarity with the results obtained in the work [16], where they analyzed Twitter data from employees of communications companies, for which they used different ML models. Finally, they concluded that the KNN model with a variant reached an accuracy of 80.1%.

6. Conclusions

After analyzing approximately 268k between tweets and retweets, with the hashtag of the trends occurred in the first two weeks of February 2023: #ParoNacional, #RenunciaYa #Renuncia, in a radius of 1000km from the city of Lima (geocode = "-12.043180 - 77.028240") and applying TM and UL techniques to classify the Tweets related to the social upheaval in Peru, it can be concluded with the following:

Firstly, it has been observed that most of the Tweets analyzed reflect a polarization in the opinions of users. On the one hand, there is a group of users who support the protests and demands for social change, with more than 60%, while, on the other hand, there is a group that criticizes the protests and demands to restore order. Also, a series of recurring themes have been identified in the Tweets analyzed. Among them, the resignation of the president, Lima's centralism, the closure of the congress, corruption and social inequality stand out as the main causes motivating the protests. Likewise, criticism of the government for its lack of action to solve these problems and for police violence during the demonstrations has also been observed.

In this work, ten types of feelings were used for the analysis: joy, amazement, confidence, disgust, anger, fear, negative, positive, premonition and sadness, as shown in Fig. 7. Likewise, it is evident that words such as: resignation, communist, repression and uprising that group the negative feeling are the most recurrent. Tweets with sentiments, positive, neutral, and negative represent 37%, 5% and 58% respectively. Also, UL techniques have been found to be effective in classifying Tweets according to their polarity,

allowing to identify those Tweets that reflect positive, negative, and neutral opinions. This suggests that by applying UL and the LDA model, it is possible to classify user sentiments.

Although TM and UL techniques are useful tools for analyzing large amounts of data, it is important to consider the limitations of these models and complement them with qualitative analysis to obtain a more complete understanding of the situation. It is very important to point out that this work was conducted in the Peruvian context during the social upheaval and in the Spanish language. Therefore, it may limit access to readers of other languages.

Finally, the results obtained in this work are very useful to understand public opinion on Twitter about the social upheaval that Peru experienced in February 2023. The findings can be used by governments, organizations, and society in general to make informed decisions and improve the current situation of the country. In addition, this study can serve as a basis for future research in TM and sentiment analysis.

7. References

[1] A. Urueña, A. Ferrari, D. Blanco and I. E. Valdecasa, "El Estudio Las Redes Sociales en Internet," Observatorio nacional de las telecomunicaciones y de la SI, 2011.

[2] C. Alvino "Estadísticas de la situación digital de Perú en el 2021-2022", in Digital 2022 Global Overview Report, Branch Agencia. 2022.

[3] DPL News, "Perú | Internet: más familias acceden al servicio dando paso a un país mejor conectado," dplnews, 14 noviembre, 2022

[4] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, "Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, pp. 571–578, 2022, doi: https://doi.org/10.14569/IJACSA.2022.0130669.

[5] S. Weiss, "Text mining : predictive methods for analyzing unstructured information," . New York, Ny: Springer.

[6] A. Alsaeedi and M. Zubair, "A Study on Sentiment Analysis Techniques of Twitter Data," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, pp. 10, 2019, doi: https://doi.org/10.14569/IJACSA.2019.0100248.

[7] A. Abdullah and K. Mohammad, "A Study on Sentiment Analysis Techniques of Twitter Data," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, 2019, doi: https://doi.org/10.14569/IJACSA.2019.0100248.

[8] P. Ficamos and Y. Liu, "A Topic based Approach for Sentiment Analysis on Twitter Data," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, 2019, doi: https://doi.org/10.14569/IJACSA.2019.0100248.

[9] MA. Kausar, A. Soosaimanickam and M. Nasar, "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak," International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, 2019, doi: https://doi.org/10.14569/IJACSA.2021.0120252.

[10] A. Mustafa, A. Alsuhibany and S. Ahmed, "Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies," International Journal of Advanced Computer Science and Applications, vol. 08, no. 1, 2017, doi: https://doi.org/10.14569/IJACSA.2017.080150.

[11] J. Yauri, E. Porras, M. Lagos, E. Tinoco and S. Solis, "Approval Rating of Peruvian Politicians and Policies using Sentiment Analysis on Twitter," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, doi: https://doi.org/10.14569/IJACSA.2022.0130696.

[12] G. Ignatow, and R. Mihalcea, "An introduction to text mining : research design, data collection, and analysis," Thousand Oaks, California: Sage Publications, Inc; 2018.

[13] Z. Zong, R. Xia and J. Zhang, "Text data mining," Singapore: Springer; 2021.

[14] M. Cebral-Loureda, A. Hernández-Baqueiro, and E. Tamés-Muñoz, "A text mining analysis of human flourishing on Twitter," Scientific Reports [Internet], 2023, 13(1):3403. Available from: https://www.nature.com/articles/s41598-023-30209-7.

[15] F. Namugera, R. Wesonga, and P. Jehopio, "Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda," Computational Social Networks. 2019 Apr 10;6(1).

[16] V. Ramanathan and T. Meyyappan, "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism," [Internet]. IEEE Xplore. 2019. p. 1–5. Available from: https://ieeexplore.ieee.org/document/8645596.

[17] J. Žižkaand, D. František and A. Svoboda, "A. Text Mining with Machine Learning," CRC Press; 2019.

[18] H. Jiang, "Machine Learning Fundamentals," Cambridge University Press; 2021.

[19] E. Haddi, X. Liu and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Computer Science. 2013;17:26–32.

[20] A. Alsaeedi and M. Zubair, "A Study on Sentiment Analysis Techniques of Twitter Data," International Journal of Advanced Computer Science and Applications", vol. 10, no. 2, pp. 10, 2019, doi: https://doi.org/10.14569/IJACSA.2019.0100248.

[21] T. Kwartler, "Text mining in practice with R". John Wiley & Sons. Copyright.

[22] ZG. Zhou, "Research on Sentiment Analysis Model of Short Text Based on Deep Learning," Liu J, editor. Scientific Programming. 2022 May 29;2022:1–7.

[23] P. Kherwa, "Topic Modeling: A Comprehensive Review". ICST Transactions on Scalable Information Systems. 2018 Jul 13;0(0):159623.

[24] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet Allocation. Journal of Machine Learning Research," [Internet]. 2003;3:993–1022. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf