Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm

Orlando Iparraguirre-Villanueva[1], Victor Guevara-Ponce[2], Fernando Sierra-Liñan[3]
Saul Beltozar-Clemente[4], Michael Cabanillas-Carbonell[5]
Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima, Perú[1]
Escuela de Posgrado, Universidad Ricardo Palma, Lima, Perú[2]
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú[3]
Universidad Científica del Sur, Lima, Perú[4]
Vicerrectorado de Investigación, Universidad Norbert Wiener, Lima, Perú[5]

***Abstract*—Today, web content such as images, text, speeches, and videos are user-generated, and social networks have become increasingly popular as a means for people to share their ideas and opinions. One of the most popular social media for expressing their feelings towards events that occur is Twitter. The main objective of this study is to classify and analyze the content of the affiliates of the Pension and Funds Administration (AFP) published on Twitter. This study incorporates machine learning techniques for data mining, cleaning, tokenization, exploratory analysis, classification, and sentiment analysis. To apply the study and examine the data, Twitter was used with the hashtag #afp, followed by descriptive and exploratory analysis, including metrics of the tweets. Finally, a content analysis was carried out, including word frequency calculation, lemmatization, and classification of words by sentiment, emotions, and word cloud. The study uses tweets published in the month of May 2022. Sentiment distribution was also performed in three polarity classes: positive, neutral, and negative, representing 22%, 4%, and 74% respectively. Supported by the unsupervised learning method and the K-Means algorithm, we were able to determine the number of clusters using the elbow method. Finally, the sentiment analysis and the clusters formed indicate that there is a very pronounced dispersion, the distances are not very similar, even though the data standardization work was carried out.**

Keywords*—Techniques; machine learning; classification; twitter

I. Introduction

The analysis of data derived from social media is of growing importance from various points of view, including academic and economic ones. Currently, there are 4.2 billion users on social networks worldwide, equivalent to 53.6% of the total population [1]. In Peru, there are 26.41 million social network users, equivalent to 81.4% of the population [2].

On Twitter, an average user makes six comments on all types of tweets, follows one user, and shares two tweets on average in 30 days; the number of likes and comments on social media tweets influences users' self-presentation [3]. Social networks are new spaces where groups of users with specific characteristics are centered, allowing them to share expressions, and opinions, these are considered the new means of communication and exchange of web content [4], [5], and the amount of data provided by users and feedback is enormous, however, there is no predefined method or tool to sort and classify the comments, since knowing the opinions of users is of utmost importance to generate knowledge and make better decisions in the different areas [6],[5],[7].

The AFPs are private institutions whose sole purpose is to manage pension funds in the form of personal accounts that provide retirement, disability, and survivor's pensions and funeral expenses [8]. The objective of the AFP is to protect the older population from the risk of poverty and allow citizens to save for their retirement. This study aims to classify and analyze the sentiments of Twitter users (affiliates) using machine learning models, for this case we will work with the hashtag #afp.

Comments are composed of positive opinions that create a contagion effect, while criticism or negative comments also influence users' decisions [9], by which, it is important to classify and analyze users' sentiments to know their reactions to a certain topic or tweet. Supported by association rule learning and machine learning algorithms, it is possible to discover the relationships between words that associate sentiments [10],[11].

The article is organized as follows. Section II describes the main works related to sentiment analysis. Section III presents the method and case implementation. Section IV describes and discusses the results and discussions. Finally, Section V presents the conclusions.

II. Related Work

Currently, the Internet and social networks are environments in which large amounts of user data are collected [12]. Comments can negatively affect the reputation of a person or company and be harmed socially and/or economically, it is for this reason, that [13] argues that to generate social capital, it is necessary to work on the feedback generated by users on social networks.

The authors in [14],[15],[16] implemented an analysis model using machine learning classifiers, to measure and predict user profile credibility, the implemented model was evaluated on two different datasets with term frequency and three inverse document frequency variables. Similarly, in [17],[18],[19] identified features that can be useful for predicting whether a tweet or comment is a rumor or

information, using a rule-based approach involving regular expressions to categorize sentences.

In addition, [20] analyzed the existence of malicious users spreading malware or phishing in tweet comments, for which six classifiers were used: random tree, random forest, Naïve Bayes, Kstar, decision table, and decision stump. Likewise, [21] proposed an approach to extract from Twitter the image content, classify it and verify the authenticity of digital images and discover manipulation. Also [22] recommended a topic analysis and sentiment polarity classification with machine learning techniques for emergency management, finding that the most suitable classification model is random forests.

Likewise, [23] developed a system that employs sentiment analysis for product reviews (e.g., the company shares a photo of a new product in a tweet and it receives a thousand new product comments), the system classifies the comments as positive or negative.

Machine learning employs two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future results. Unsupervised learning finds hidden patterns or intrinsic structures in the input data. In the same context, in [24],[25],[26] the authors used natural language processing methods to detect fake news in social media posts, and also created a semi-supervised learning model for early detection.

In addition, in [27],[28] a supervised learning model for rumor verification using contextual features based on contextual word embeddings, speech acts, and the writing style was presented. In [29] the authors collected unstructured data from social networks, using an application implementing different supervised and unsupervised learning classification algorithms. Similarly, in [30] a novel method for generating opinion summaries in social networks was proposed, using the unsupervised learning technique, in [31] content analysis of textual Twitter data was performed using a set of supervised and unsupervised machine learning methods to appropriately cluster and classify traffic-related events.

Although previous studies considered many aspects of sentiment analysis in social networks, according to the review conducted, there is a gap to be investigated in the use of clustering rules to classify and analyze user sentiments, considering that these results would be of great use to stakeholders, as it would allow them to make better decisions supported with ML.

III. Method and Implementation of the Case

This section presents the construction of the machine learning terminology, the method, and the detailed implementation of the proposed case study that seeks to predict members' sentiments regarding AFP of Peru.

Machine learning (ML) is a discipline of artificial intelligence that, through algorithms, teaches computers to build models from experience [32]. ML is divided into three types: supervised learning, unsupervised learning, and reinforcement learning, and these in turn are classified into different learning models, as shown in Fig. 1. The study uses unsupervised learning, and the k-means algorithm, for which we will address only this category.

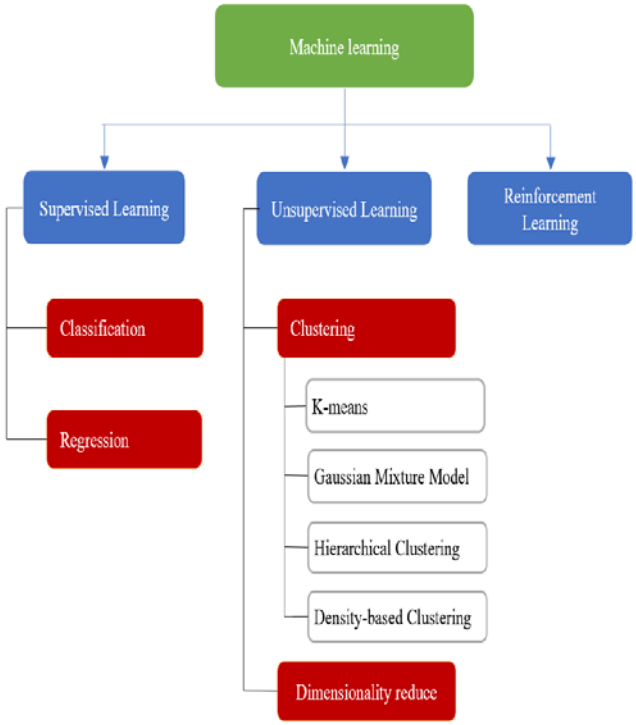

Fig. 1. Types of Learning.

Unsupervised Learning: It does not require labeled data and can be processed by clustering methods. Clustering is a typical example of unsupervised learning that finds visual classifications that match the hypothesis [14]. The goal of clustering is to find similarities, regardless of the kind of data. Therefore, a clustering algorithm needs to know how to calculate similarity, and then start running.

K-Means: It is a clustering algorithm that combines the "n" observations into "k" clusters that are aggregated together according to specific similarities [33], as shown in Fig. 2.

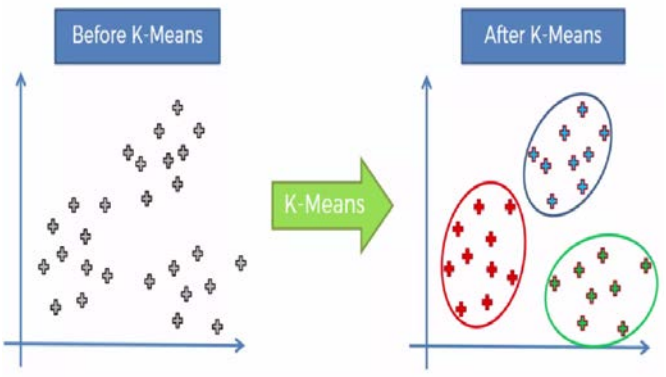

Fig. 2. Grouping of K-Means.

The K-Means algorithm follows the following steps:

- Initiation: The location of the centroids of the k clusters is chosen randomly.

- Assignment: Each datum is assigned to the nearest centroid.
- Updating: The centroid position is updated to the arithmetic mean of the data positions assigned to the cluster.

Steps 2 and 3 are followed iteratively until there are no more changes in the location of the centroids, or they move below a threshold distance from each step. The K-means algorithm seeks to solve an optimization problem, the function to be optimized being the sum of the quadratic distances from each object to the centroid of its cluster. The objects are represented by *d* dimensional real vectors (x_1, x_2, …, x_n) and the algorithm constructs *k* clusters where the sum of the distance of the objects, within each object s= {s_1, s_2, …, s_k}, to its centroid is minimized. The problem is formulated as follows:

$$\frac{min}{s} E(\mu_i) = \frac{min}{s} \sum_{i=1}^{k} \sum_{x_j \in s_i} ||x_j - \mu_i||^2 \quad (1)$$

Where S is the data set whose elements are the objects represented by vectors, where each of its elements represents an attribute. We have k cluster with its corresponding centroid μ_i. On each centroid update, from the mathematical point of view, we impose the necessary condition of extrema to the function $E(\mu_i)$ which, for the quadratic function (1) is.

$$\frac{\partial E}{\partial \mu_i} = 0 \Rightarrow \mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} + \sum_{x_j \in S_i^{(t)}} x_i$$

The k-means algorithm is simple and fast, however, it is necessary to decide the value of K and the final result will depend on the initialization of the centroids.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology was chosen for the development of the case. It is a model that divides the process into six main phases: business understanding, data compression, data preparation, modeling phase, evaluation, and implementation. CRISP-DM has a structured approach, establishing a set of tasks and activities for each phase.

A. Business Understanding

The AFP is a private institution in charge of managing people's provisional savings. Like any company, with the arrival of COVID-19, the economy of the region has been severely affected and, therefore, the annualized profitability of the AFP has not been favorable. This has led the Congress of the Republic to promote laws such as free disaffiliation, 100% of contributions for those over 55 years of age, among other authorizations for controlled withdrawals under the context of a pandemic for the years 2020 to 2022. Peruvians' perception of the AFP has never been positive, probably because retirement pensions do not meet expectations, among other factors. In this context, the sentiment analysis of the case study aims to analyze and classify the opinions of AFP members. This analysis can help the AFP to make better decisions in its offerings to the market. Fig. 3 shows the general process of sentiment analysis.

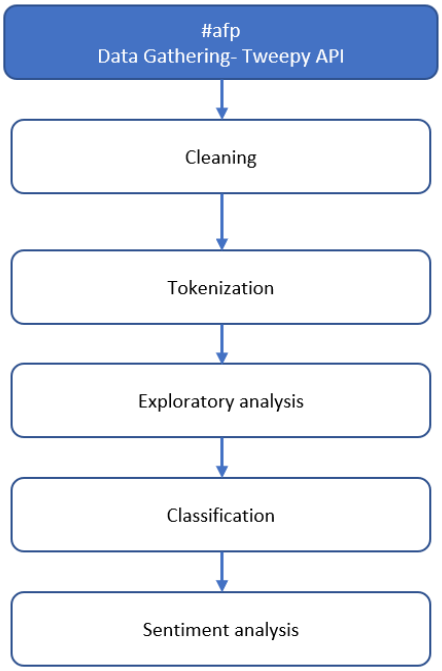

Fig. 3. General Sentiment Analysis Process.

Fig. 3 shows the general process of sentiment classification and analysis. It starts with extraction supported by libraries, then we move on to cleaning and tokenization, this process consists of removing from the text everything that does not provide information about the subject matter. For example, on Twitter, users can write in a way that they consider convenient, such as abbreviations, punctuation marks, web page URLs, single characters, numbers, etc.

Tokenization consists of dividing the text into its constituent units, in this case, words. Next, we perform the exploratory analysis, in this stage, we try to understand and study which words and how often they are used, as well as their meaning. In the Python or R language, one of the structures that facilitate the exploratory analysis is the DataFrame, which is where the information of the tweets is stored. The next step is to perform the classification, in this phase, to be able to apply classification algorithms to the text (tweets), a numerical representation of the text is created.

One of the most used ways is known as the Bag of words, this method consists of identifying the set formed by all the words (tokens) that arise in the corpus. Finally, we proceed with the sentiment analysis, for which it is necessary to have a dictionary in which a sentiment or sentiment level is associated with each word.

B. Data Comprehension

The dataset for the case study has been collected on recent Tweets from May 2022, related to the hashtag #afp, with coordinates -12.062152, -77.0361328 corresponding to the city of lima. The dataset is composed of 18k tweets. The data we collected contains relevant information about most of the tweets such as their attributes (user_id, status_id, created_at, screen_name, text, source, display_text_width,

reply_to_status_id, favorite_count, replay_to_user_id, replay_to_screen_name, is_quote, retweet_count1, quote_count, media_url, media_t. co, media_expanded_url, etc.) which, in sum, are 90 attributes with 18k tweets to be processed, as shown in Fig. 4.

```
tweets_afp <- search_tweets(
                q = "AFP",
                geocode = "-12.062152,-77.0361328,20mi",
                include_rts = FALSE,
                n = 18000
            )
```

Fig. 4. Tweet Search and Download Code.

1) DataSet preparation: This is an essential step before applying any machine learning model. The DataSet preparation phase starts with a crucial previous step, known as extraction, loading, and transformation, to load the extracted data to a source or data warehouse, in this case, it will be processed in a single computer with the necessary characteristics to support the processing of 18k tweets represented in a DataSet. The phases, as a whole, are described below:

2) Extraction, loading, and transformation: This process was implemented with the code illustrated in Fig. 4, the data were stored in a text file, then, supported with the *stringr* library, we performed the first preliminary cleaning, then the text file was converted to a csv file (write.csv ()), then we applied tokenization with the libraries tidyverse, tibble and a personal dictionary in Spanish loaded with the readxl library. And then we used lemmatization, that is, the root of a word that can be written in several forms and refer to the same thing. For example, if the words *pueblo, pueblito, pueblano* are found, the root of these three words is pueblo, thus avoiding redundancy in the analysis, this process is supported in a pre-trained model with the udpipe library, as shown in Table I.

In the next step, we performed an exploratory analysis of the data, using the bag-of-words(bow) model to obtain the most common words in the lexicon, obtaining a list of the most frequent words in the hashtag #afp. Fig. 5 shows a dense word cloud with some of the most used words in the generated corpus, also, it can be seen that several words are found in different positions.

TABLE I. LEMMATIZATION OF WORDS

sentence_id	sentence	token_id	token	lemma
1	chabelitacongre	1	chabelitacongre	chabelitacongre
1	ayudas	1	ayudas	ayudas
1	pueblo	1	pueblo	pueblo
1	quitando	1	quitando	quitando
1	jubilación	1	jubilación	jubilación
1	problema	1	problema	problema

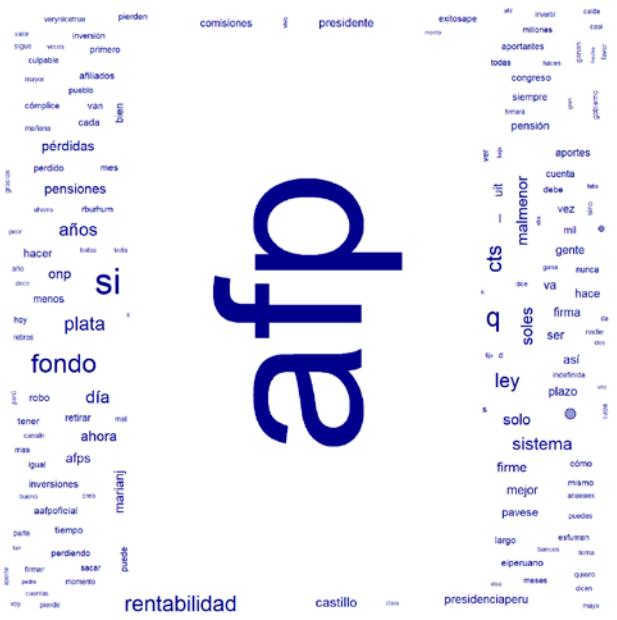

Fig. 5. Most Popular Words Related to #afp from our Corpus.

Next, the syuzhet dictionary package is used, this package works with four sentiment dictionaries, such as: Bing, Afinn, Stanford, and NRC, in our case, we will work with NRC, since it is the only one available in several languages, including Spanish. This dictionary has 14182 words with categories of feelings, positive, negative, and the emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, confidence, etc. This dictionary is used with the purpose of analyzing and grouping the words of the corpus and to know in which category the feelings are grouped. Fig. 6 shows that the negative sentiment of affiliates is most frequently repeated in the tweets, followed by positive sentiment, fear, confidence, anger, disgust, etc.

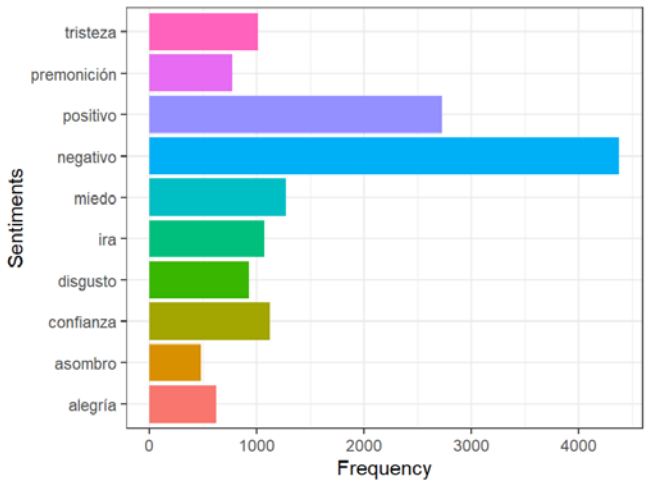

Fig. 6. Frequency of Terms in Sentiment Analysis.

Fig. 7 shows the words of the tweets grouped by sentiment, in it, we can analyze each of the words associated with a particular feeling. For example, the word money is the one that mostly generates the feeling of joy, followed by astonishment, confidence, anger, positive and premonition. In

the same line, the word congress is the one that mostly generates disgust, likewise, the word government is the one that mostly generates the feeling of fear.

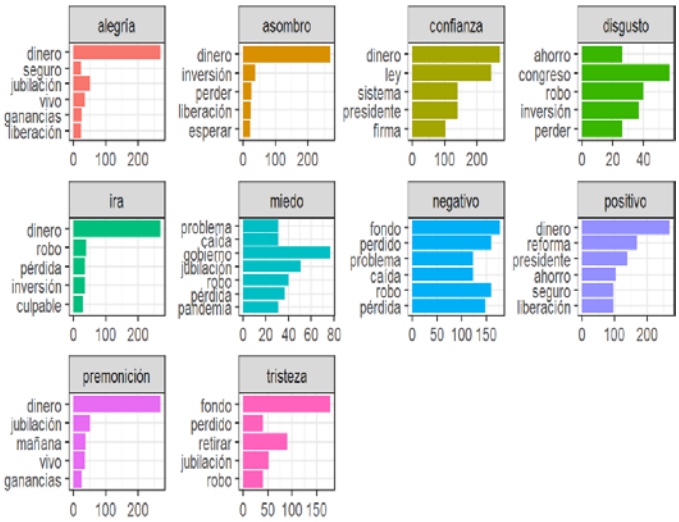

Fig. 7. Words for Sentiment.

C. Preparation and Clustering with K-means

At this stage, we have already extracted, cleaned the data and stored it in a text file. Next, we use the TM package, to call the functions VectorSource (), inspect () and removeSparseTerms (), the latter function to remove words that are not very frequent, and finally we convert the text into a word frequency matrix with the TermDocumentMatrix () function. The k-means clustering is a vector quantization method, which tends to find clusters of comparable spatial extent. Part of the code is shown in Fig. 8.

```
library(tm)
corpus <- Corpus(VectorSource(texto))
corpus
inspect(corpus)
# Printout of a corpus line

writeLines(as.character(corpus[[100]]))  # basic package
content(corpus[[100]])

d  <- tm_map(corpus, tolower)
inspect(d)
# Remove custom empty words
d <- tm_map(d, removeWords, sw.es)
inspect(d)

# Create Term Matrix with TermDocumentMatrix()
tdm <- TermDocumentMatrix(d)
tdm
```

Fig. 8. Use of Libraries in R.

D. Modeling Phase

In this context, it is very important to analyze the relationship between words with the findAssocs () function of the tm package. The calculation of the findAssocs () function is performed at the document level. Then, for each specific document containing the word in question, the other terms in those documents are associated, the other search terms are ignored. The function returns a list of all the other terms that meet or exceed the minimum threshold (findAssocs (tdm, "afp", 0.40)). The minimum correlation values are usually relatively low due to the diversity of words. The tweets of the hashtag #afp have been cleaned and organized with tweets_tdm. For the case study, the function findAssocs () searches for the association of terms and manipulates the results with list_vect2df () from the qdap package and then creates the diagram with the ggplot2 library. Fig. 9 shows the strong association represented by the thickness of the lines between the words. The word afp is the root for the association of the other words, such as afp-money, afp-withdrawal, afp-funds, afp-system, afp-years, afp-law, afp-now, funds-withdrawal, withdrawal-law, afp-castle, etc. This association process allows us to know the words most associated and expressed by AFP members.

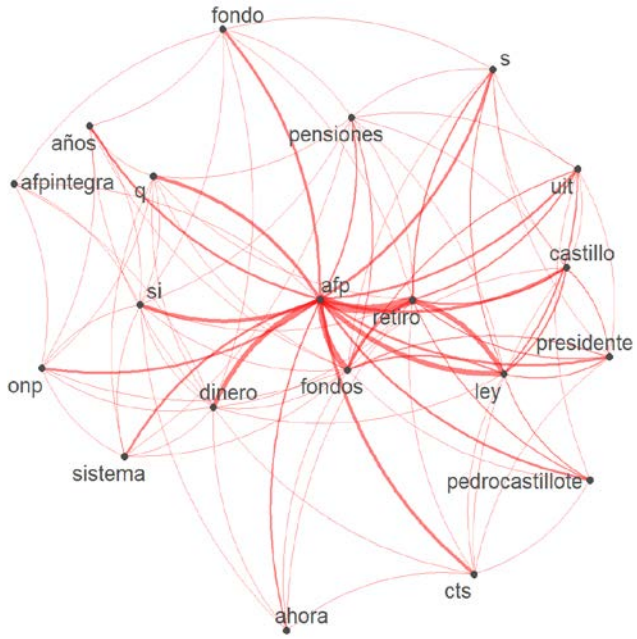

Fig. 9. Frequency Represented in Line Thickness.

In Fig. 10, the sentiment classification is presented with the overall percentage of each positive, negative and neutral tweet found in the dataset. It can be seen that the sentiment classes are unequal, as a large part of the affiliates express themselves negatively or neutral concerning the hashtag #afp.

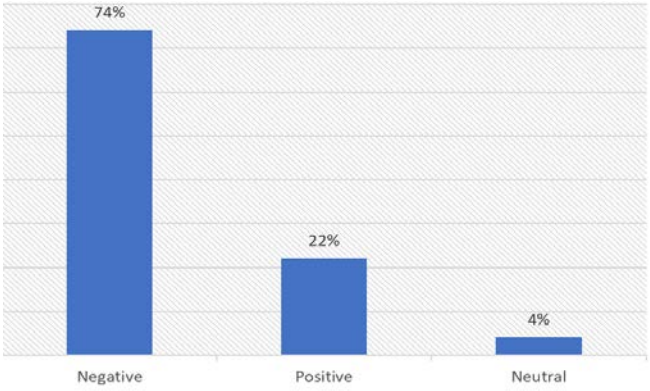

Fig. 10. Sentiment Distribution of the Three Polarity Classes along with the Percentage of Tweets Referring to the #afp Hashtag.

IV. Results and Discussion

In this section, we present the performance of the k-means algorithm to obtain the best number of clusters and the results of the case study. In general, most clustering algorithms, including k-means, employ different numbers of clusters, and are evaluated by multiple validity measures to determine the most meritorious clustering results. To measure clustering performance, we use a precision index (IP), where $IP=\sum_{k=1}^{c} n(c_k)/n$, where c_k is the number of data points that obtain the correct clustering for cluster k and n is the total number of data points.

The higher the accuracy index, the better the clustering performance. K-Means clustering is based on data partitioning, data that have the same features are grouped into one cluster, while data that have different features are grouped into other clusters. The steps of K-Means clustering are three: deciding the cluster number k, centroid initialization and assigning the data to the nearest cluster, the determination of the closeness of objects/data is determined based on the distance of objects/data, to calculate the distance of all data to each centroid point, the Euclidean distance theory is used, which is represented as follows.

$$D(\mathrm{i},j)=\sqrt{\left(x_{i1}-x_{J1}\right)^2+\left(x_{i2}-x_{J2}\right)^2+\cdots+\left(x_{ip}-x_{Jp}\right)^2}$$

Where:

D (i, j) = distance of the data to the cluster center j.

X_{ik} = i^{th} in the k^{th} date attribute.

X_{ij} = j^{th} center point in the k^{th} data attribute.

The centroid is the average of all data/objects within a particular cluster. Each data/object is reassigned using the new cluster centroid, the cluster does not change, then the clustering process is finished, otherwise, repeat the steps until there is no change for each cluster.

To determine the number of clusters in which the data are grouped, we use the elbow method, this method uses the mean of the observations to their centroid, i.e., it looks at the intra-cluster distances. The larger the number of clusters k, the intra-cluster variance tends to decrease. The smaller the intra-cluster distance the better.

The elbow method searches for the k value which satisfies that an increase of k does not substantially improve the mean intra-cluster distance. To find the number k, 15 clusters were assigned, in Fig. 11 it can be seen that the elbow(k) is formed at point 3, this means that the number of clusters that we will process will be k=3.

Once the number of clusters (k = 3) has been determined, we can start with the data processing. For this, the tweets had to be converted to numerical data, using one-hot encoding for the transformation, as shown in Table II.

It is advisable to scale values and not to introduce variables that are highly correlated or that are linear combinations of other variables, i.e., to avoid multicollinearity. The objective of scaling variables is to make them comparable, that is, to have a mean equal to zero and a standard deviation equal to one. The scaled method is achieved by using the following formula:

$$z(x)=\frac{x_i-centro\,(x)}{deviation(x)}$$

Where, center(x) is a measure of centrality with mean or median and variance (x) is a measure of dispersion such as standard deviation. The standardization of data is important as it makes the distances similar to that without scaling. Fig. 12 shows a representation by dimensions and the clusters formed. It is clearly observed that a cluster of comments or tweets are very homogeneous, with low values in cluster 1, we refer to the negative sentiment and high values in cluster 2, cluster 3 is presented in the central part with intermediate values in both cluster 1 and 2 corresponding to the positive and neutral sentiments.

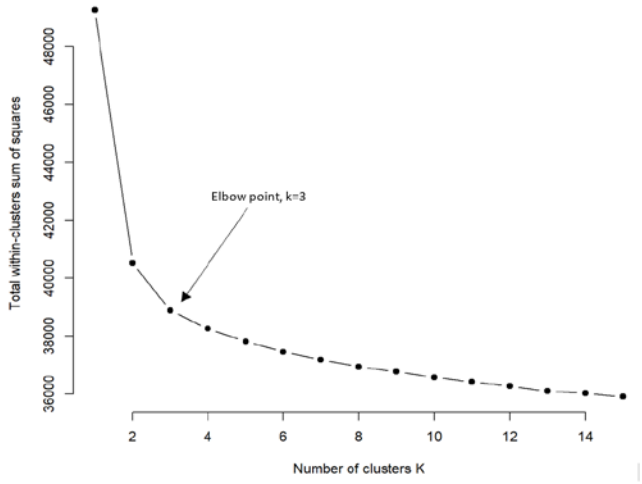

Fig. 11. Elbow Method - Number of Clusters.

TABLE II. Summary of One Hot Coding

Docs	afp	casa	hasta	los	que	seguro	son	tienes	todos	vida
1	0	0	0	0	0	0	0	0	0	0
10	1	0	0	2	2	0	0	0	0	1
2	1	0	0	3	1	0	1	0	1	0
3	0	2	1	1	0	0	1	0	1	0
4	1	0	0	0	1	0	0	0	0	0
5	1	0	0	0	0	0	0	1	0	0
6	1	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	0	0	0	0	0
8	1	0	0	0	0	3	1	1	0	3
9	1	0	1	1	3	0	0	1	0	0

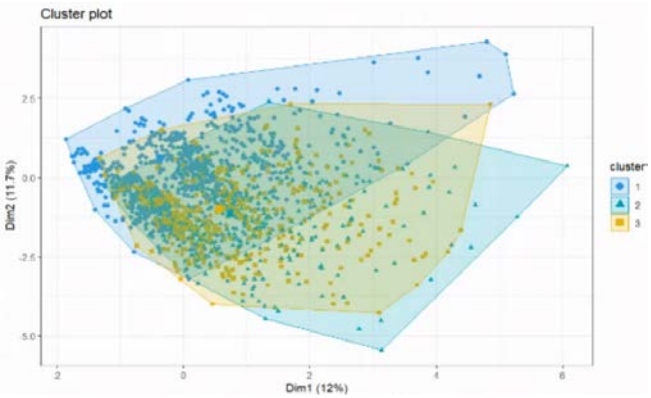

Fig. 12. Sentiment Analysis of #afp.

V. Conclusion

In recent years, several approaches have been developed for sentiment analysis on social network data, specifically on Twitter. The process of sentiment analysis is often complex due to the huge amount of data and the need to achieve a high level of accuracy. This study presents the sentiment analysis of Twitter hashtag #afp data, for which the k-means algorithm was used. This algorithm is based on data partitioning, data having the same features that are grouped into a cluster. In this study, eight types of feelings (sadness, foreboding, positive, negative, fear, anger, disgust, confidence, amazement and joy) were used. After applying unsupervised learning with K-Means, it is seen that the negative feeling is the most recurrent, followed by positive feeling, fear, confidence, etc. As shown in Fig. 6 out of a total of 18000 tweets related to the hashtag #afp, tweets with sentiments, positive, neutral and negative represent 22%, 4% and 74% respectively. This means, that machine learning applying the K-Means algorithm proves to be efficient and practical, and can be easily applied for sentiment classification on related topics, where we do not have input and output tags.

This study provides theoretical and practical scopes. Regarding the theoretical scope, the study applies a machine learning approach, with the unsupervised learning method for sentiment analysis of AFP members, using Twitter data with the hashtag #afp. Also, sentiment analysis with machine learning can be applied in different industries such as marketing, services and academia, etc. In terms of practical scope, this study recommends machine learning with the unsupervised method to be applied in cases similar to the study, allowing it to be adapted and improved to achieve a better level of accuracy, especially in complex situations when performing textual analysis.

In this study, from the classification to the analysis of feelings, there was evidence of certain feelings that were most repeated, with negative feelings being the predominant one, with which we can conclude that the members do not feel represented or reject the actions of the AFP, followed by positive feelings, fear, trust, anger, sadness, disgust, premonition, joy and amazement, in that order. This study is not without limitations. Attribute mapping was applied to the initial data set; within this, there may be a diversity of factors, combinations that may affect the classification results. In future work, machine learning with the unsupervised method and the K-Means algorithm can be optimized to improve the accuracy of sentiment detection, classification and analysis.

References

[1] P. Lara-Navarra, A. López-Borrull, J. Sánchez-Navarro, and P. Yànez, “Measuring the influence of users on social networks: SocialEngagement proposal,” p. 899, 2018, doi: 10.3145/epi.2018.jul.18.

[2] H. Xiang, K. Y. Chau, W. Iqbal, M. Irfan, and V. Dagar, “Determinants of Social Commerce Usage and Online Impulse Purchase: Implications for Business and Digital Revolution,” Frontiers in Psychology, vol. 13, Feb. 2022, doi: 10.3389/FPSYG.2022.837042.

[3] C. Marino, C. Lista, D. Solari, M. M. Spada, A. Vieno, and L. Finos, “Predicting comments on Facebook photos: Who posts might matter more than what type of photo is posted,” Addictive Behaviors Reports, vol. 15, p. 100417, Jun. 2022, doi: 10.1016/J.ABREP.2022.100417.

[4] O. R. Seryasat, I. Kor, H. G. Zadeh, and A. S. Taleghani, “Predicting the number of comments on facebook posts using an ensemble regression model,” International Journal of Nonlinear Analysis and Applications, vol. 12, no. Special Issue, pp. 49–62, Dec. 2021, doi: 10.22075/IJNAA.2021.4796.

[5] C. P. Chandrika and J. S. Kallimani, “Classification of Abusive Comments Using Various Machine Learning Algorithms,” Advances in Intelligent Systems and Computing, vol. 1040, pp. 255–262, 2020, doi: 10.1007/978-981-15-1451-7_28.

[6] J. Biswas, M. M. Rahman, A. A. Biswas, M. A. Z. Khan, A. Rajbongshi, and H. A. Niloy, “Sentiment Analysis on User Reaction for Online Food Delivery Services using BERT Model,” 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, pp. 1019–1023, Mar. 2021, doi: 10.1109/ICACCS51430.2021.9441669.

[7] H. Deng, D. Ergu, F. Liu, Y. Cai, and B. Ma, “Text sentiment analysis of fusion model based on attention mechanism,” Procedia Computer Science, vol. 199, pp. 741–748, 2021, doi: 10.1016/J.PROCS.2022.01.092.

[8] “Asociación de AFP.” https://www.asociacionafp.pe/ (accessed May 19, 2022).

[9] F. Erlandsson, A. Borg, H. Johnson, and P. Bródka, “Predicting user participation in social media,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9564, pp. 126–135, 2016, doi: 10.1007/978-3-319-28361-6_10.

[10] L. Parisi, R. Ma, N. RaviChandran, and M. Lanzillotta, “hyper-sinh: An accurate and reliable function from shallow to deep learning in TensorFlow and Keras,” Machine Learning with Applications, vol. 6, p. 100112, Dec. 2021, doi: 10.1016/J.MLWA.2021.100112.

[11] D. C. Gkikas, K. Tzafilkou, P. K. Theodoridis, A. Garmpis, and M. C. Gkikas, “How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook,” International Journal of Information Management Data Insights, vol. 2, no. 1, p. 100067, Apr. 2022, doi: 10.1016/J.JJIMEI.2022.100067.

[12] N. A. Awad and A. Mahmoud, “Analyzing customer reviews on social media via applying association rule,” Computers, Materials and Continua, vol. 68, no. 2, pp. 1519–1530, Apr. 2021, doi: 10.32604/CMC.2021.016974.

[13] A. Wang and K. Potika, “Cyberbullying Classification based on Social Network Analysis,” Proceedings - IEEE 7th International Conference on Big Data Computing Service and Applications, BigDataService 2021, pp. 87–95, 2021, doi: 10.1109/BIGDATASERVICE52369.2021.00016.

[14] E. A. Afify, A. S. Eldin, and A. E. Khedr, “Facebook Profile Credibility Detection using Machine and Deep Learning Techniques based on User’s Sentiment Response on Status Message,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 12, pp. 622–637, 2020, doi: 10.14569/IJACSA.2020.0111273.

[15] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, “GBSVM: Sentiment classification from unstructured reviews

using ensemble classifier," Applied Sciences (Switzerland), vol. 10, no. 8, Apr. 2020, doi: 10.3390/APP10082788.

[16] S. Adonai, H. Morales, M. Fabiola, M. Antayhua, and L. Andrade-Arenas, "Development of Predictions through Machine Learning for Sars-Cov-2 Forecasting in Peru," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 12, no. 11, p. 2021, doi: 10.14569/IJACSA.2021.0121188.

[17] D. Sharma and S. Singhal, "Detection of fake news on social media using classification data mining techniques," International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 3132–3138, Oct. 2019, doi: 10.35940/IJEAT.A1637.109119.

[18] R. Moin, Zahoor-ur-Rehman, K. Mahmood, M. E. Alzahrani, and M. Q. Saleem, "Framework for rumors detection in social media," International Journal of Advanced Computer Science and Applications, vol. 9, no. 5, pp. 439–444, 2018, doi: 10.14569/IJACSA.2018.090557.

[19] M. Cabanillas-Carbonell, R. Verdecia-Peña, E. Medina-Rafaile, J. Luis, H. Salazar, and O. Casazola-Cruz, "Data Mining to Determine Behavioral Patterns in Respiratory Disease in Pediatric Patients," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, p. 2021, doi: 10.14569/IJACSA.2021.0120749.

[20] N. Alias, C. F. M. Foozy, and S. N. Ramli, "Video spam comment features selection using machine learning techniques," Indonesian Journal of Electrical Engineering and Computer Science, vol. 15, no. 2, pp. 1046–1053, Aug. 2019, doi: 10.11591/IJEECS.V15.I2.PP1046-1053.

[21] N. M. AlShariah and A. K. Jilani Saudagar, "Detecting fake images on social media using machine learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 12, pp. 170–176, 2019, doi: 10.14569/IJACSA.2019.0101224.

[22] H. Park, H. Seo, K. J. Kim, and G. Moon, "Application of machine learning techniques to tweet polarity classification with news topic analysis," International Journal of Engineering and Technology(UAE), vol. 7, no. 4, pp. 40–41, 2018, doi: 10.14419/IJET.V7I4.4.19606.

[23] U. Suleymanov, B. K. Kalejahi, E. Amrahov, and R. Badirkhanli, "Text classification for azerbaijani language using machine learning," Computer Systems Science and Engineering, vol. 35, no. 6, pp. 467–475, 2020, doi: 10.32604/CSSE.2020.35.467.

[24] S. G. Taskin, E. U. Kucuksille, and K. Topal, "Detection of Turkish Fake News in Twitter with Machine Learning Algorithms," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2359–2379, Feb. 2022, doi: 10.1007/S13369-021-06223-0.

[25] P. M. Konkobo, R. Zhang, S. Huang, T. T. Minoungou, J. A. Ouedraogo, and L. Li, "A Deep Learning Model for Early Detection of Fake News on Social Media*," Proceedings of 2020 7th IEEE International Conference on Behavioural and Social Computing, BESC 2020, Nov. 2020, doi: 10.1109/BESC51023.2020.9348311.

[26] R. Shima, H. Yunan, O. Fukuda, H. Okumura, K. Arai, and N. Bu, "Object classification with deep convolutional neural network using spatial information," ICIIBMS 2017 - 2nd International Conference on Intelligent Informatics and Biomedical Sciences, vol. 2018-January, pp. 135–139, Feb. 2018, doi: 10.1109/ICIIBMS.2017.8279704.

[27] Z. Jahanbakhsh-Nagadeh, M. R. Feizi-Derakhshi, and A. Sharifi, "A semi-supervised model for Persian rumor verification based on content information," Multimedia Tools and Applications, vol. 80, no. 28–29, pp. 35267–35295, Nov. 2021, doi: 10.1007/S11042-020-10077-3.

[28] V. Gonzáles-Gutierrez et al., "Multitasking Behavior and Perceptions of Academic Performance in University Business Students in Mexico during the COVID-19 Pandemic," International Journal of Mental Health Promotion, vol. 24, no. 4, pp. 565–581, 2022, doi: 10.32604/ijmhp.2022.021176.

[29] S. K. Punia, M. Kumar, T. Stephan, G. G. Deverajan, and R. Patan, "Performance analysis of machine learning algorithms for big data classification: Ml and ai-based algorithms for big data analysis," International Journal of E-Health and Medical Communications, vol. 12, no. 4, pp. 60–75, Jul. 2021, doi: 10.4018/IJEHMC.20210701.OA4.

[30] A. Rao and K. Shah, "An Unsupervised Technique to Generate Summaries from Opinionated Review Documents," Advances in Intelligent Systems and Computing, vol. 1199, pp. 388–397, 2021, doi: 10.1007/978-981-15-6353-9_35.

[31] K. Kokkinos and E. Nathanail, "Exploring an Ensemble of Textual Machine Learning Methodologies for Traffic Event Detection and Classification," Transport and Telecommunication, vol. 21, no. 4, pp. 285–294, Dec. 2020, doi: 10.2478/TTJ-2020-0023.

[32] D. Solyali, "A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus," Sustainability (Switzerland), vol. 12, no. 9, May 2020, doi: 10.3390/SU12093612.

[33] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," Applied Statistics, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.