Machine Learning for Feeling Analysis in Twitter Communications: A Case Study in HEYDRU!, Perú

https://doi.org/10.3991/ijim.v16i24.35493

Rosa Alegre-Veliz[1], Pedro Gaspar-Ortiz[1], Javier Gamboa-Cruzado[2](✉),
Liset Rodríguez Baca[1], Waldy Grandez Pizarro[3], Rosa Menéndez Mueras[2],
Carlos Chávez Herrera[2]
[1] Universidad Autónoma del Perú, Lima, Perú
[2] Universidad Nacional Mayor de San Marcos, Lima, Perú
[3] Universidad de San Martín de Porres, Lima, Perú
jgamboac@unmsm.edu.pe

Abstract—Nowadays, feeling analysis has become a trend; above all, in digital product development companies, as it is essential for rapid and automatic analysis. Feeling analysis deals with emotions with the help of software, and it is playing an unavoidable role in workplaces. The constant growth of social networks, especially the Twitter social network, has made the ability to understand and comprehend users or clients take a greater scope regarding their needs; and therefore, increase the complexity of analysis of this social network, causing excessive expenses in time, personnel, and money. This work presents a solution through the application of Machine Learning (ML) for feeling analysis and thus improve analysis, execution time and customer satisfaction. The scope of this research is limited to using the Support Vector Machine (SVM), a supervised learning technique for the intended analysis. The model derives from the ML technique making use of cross validation. CRISP-ML(Q) is the applied Methodology. The results show that the use of ML allows efficient feeling analysis in Twitter communications.

Keywords—machine learning, feeling analysis, Twitter, algorithms, classification, CRISP-ML(Q), SVM

1 Introduction

Today the care and effective service to the user is being promoted as a new competitive value to be considered within companies; the position of the consumer as sovereign of this is evident. Therefore, the objective is to avoid the bad relationship between the group of users and customers with the company. This situation is not always clearly perceived by both parties, but this causes expensive effects. The exploration and analysis of the content of social networks has aroused the interest of both researchers and companies in general.

Following these guidelines, [4] mentions that, to obtain sentiment data through tweets, it is required to build learning models and obtain labeled data, which are usually difficult and expensive to obtain. For this reason, it promotes semi-supervised learning, which generates a vast amount of data to classify feeling analysis of tweets; likewise, [9] expresses that, due to the large amount of data on audiences, communicators should use micro-segmentation with data analysis, for which it makes use of the CRISP-DM methodology, relying on data analysis models that lead to clustering, prediction and/or classification of data in a massive way.

On the other hand, [19] deals with the frequency of conversations in his research about menthol cigarettes to get the tweet and the sentiment characteristics. Through SVM classifiers, a large amount of data was extracted to analyze, one of the results shows that 47% of the data analyzed was positive; nevertheless, it can express the idea that can lead to addiction, opposing a greater number of negative opinions. As [22] mentions, to support the creation of successful startups, a model that determines the analysis of tweets and a feeling analysis supported by the SVM algorithm is needed to know how positive or negative comments about a startup can predict its popularity [13]. Accordingly, a business plan can be generated, and continuous feeling analysis can be carried out to keep the market active regarding the startup.

Currently, Machine Learning techniques are increasingly relevant at the business level in terms of automation and restructuring, for which, indicators with adequate information should be taken as a base to guide an optimal change within the decision-making in the company. Systematic data analysis also should be performed to develop a predictive model; thus, the automatic detection of sentiments in tweets becomes a powerful and useful tool for analyzing social networks and many other applications.

Given this worrisome reality, i.e., the inefficiency of software solutions for sentiment analysis in Twitter communications using Machine Learning worldwide, the present research will allow to close this technological and business gap.

The main objective of this research is to study the application of Machine Learning through sentiment analysis in the technology company Heydru!, which will be used as a case study. It shows how Machine Learning performs efficiently against a sentiment analysis in the company to automate its process in its marketing and sales areas. The proposed approach is to use decentralized technology, by which it is intended to develop a system for issuing certificates.

2 Background and related works

Currently, Machine Learning is one of the most popular ways to examine emotional behaviors, which generates intelligent algorithms that can learn without relying on rule-based programming. The application of machine learning has been prioritized in various fields, with the business environment as the main environment [14][15]. In relation to what was said above, different agencies are being adapted in the application of machine learning for their different processes [8].

In recent years, feeling analysis has become a frequent research topic due to the great demand in the market and the need to analyze public opinion [10]. Through the time, new techniques have appeared, as well as libraries and tools to apply sentiment analysis processing [1]. With the positioning of social networks, users also have all kind of facilities to express their opinions on different topics of interest [4]. Being aware of the opinions regarding a brand or product and measuring its impact is currently of vital importance for all companies, since the image of the company is what is at stake [11].

There is also an intense application of technologies in analysis [2]. It is worth noting the recent prevalence of social networks, especially Twitter, which is characterized by being a social network that generates a large amount of data and messages; with the possibility that it can be linked to a live event anywhere in the world [5]. Given this diversity of audiences, they can be segmented and located in a geographical area or in a hashtag environment [9]. Regarding the ranking of the sentiment of the messages on Twitter, there is a history of studies applied to different themes and languages; among them, the inclusion of emoticons is considered as a relevant element to support the context and to increase the accuracy of the model [14].

On the other hand, the use of hashtags, where the label system is considered to build the classifier [12], the semantics-based approach suggests the removal of stop words. These without apparent load of meaning such as the articles “one”, “an”, “the” or “the” [2]. In addition, the use of Machine Learning technology allows the development of the Support Vector Machine (SVM) algorithm [3][15]. The application of Machine Learning technology has also allowed it to be applied in the field of customer service; as, somehow, the analysis provided by the algorithm alerts the service team to any new issues that need to be taken into account; and, in this way, the preparation of a plan or strategy is made possible [17].

This generates contributions in various fields in addition to the business field; for example, it is a resource for presidential elections [2], to obtain the classification of the candidates according to positive and negative indicators to know who the winner will be [6].

3 Research methodology

3.1 Machine learning application development methodology

For the development of the solution in this research, the CRISP-ML(Q) methodology was used; it has recently been proposed by the machine learning community to ensure the quality of the result of the project. Careful maintenance of each phase is kept in mind to reduce the risk of performance degradation over time. It has six phases (See Figure 1): business and data understanding, data engineering (data preparation), Machine Learning Model Engineering, quality assurance for machine learning applications, deployment and finally, monitoring and maintenance [31].

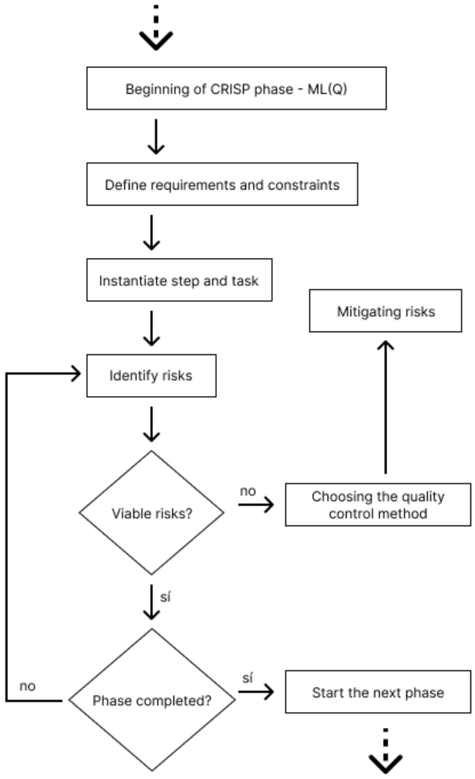

Beginning of CRISP phase - ML(Q)
Define requirements and constraints
Instantiate step and task
Mitigating risks
Identify risks
Viable risks?
no
Choosing the quality control method
sí
no
Phase completed?
sí
Start the next phase

Fig. 1. CRISP-ML(Q) methodology workflow

3.2 Applied research method

Operationalization of the variables: the indicators and their details, considered in the research, are shown in Table 1.

Table 1. Operationalization of the dependent variable

Indicator	Index	Unit of Measurement
Analysis time	[1-12]	Days
Analysis cost	[5400 – 108000]	Nuevo sol
Number of people involved	[1-11]	People
Acceptance level	Totally disagree, In disagreement, Neither agree nor disagree, In agreement, Totally agree	Likert scale

Research Design. This research presents a pure experimental design, which applies the Post Test method for the Experimental Group (Ge) and for the Control Group (Gc).

The elements of the sample are chosen randomly (R), to which a stimulus or experimental condition (X) is applied. When the stimulus (Machine Learning solution) is applied to the Ge, values for the indicators are obtained (O1) and when the stimulus is not applied, values for the Gc are obtained (O2).

RGe X O1
RGc -- O2

Universe and sample. It was established as a universe to all the processes of analysis of sentiments of communications by Twitter in marketing agencies in Peru.

In case of the sample, the process of sentiment analysis was taken in the communications by Twitter. With n = 30 transactions.

Data collection procedures. The direct observation technique was applied in the research; the tool, observation sheet, was used for collecting the data for each study indicator. This technique was applied from the beginning.

Statement of hypotheses. H_1: If you use a Machine Learning Application using CRISP-ML(Q) then the time to Analyze Feelings in Twitter Communications is reduced.

H_2: If a Machine Learning Application is used using CRISP-ML(Q) then the cost to Analyze Feelings in Twitter Communications is reduced.

H_3: If you use a Machine Learning Application using CRISP-ML(Q) then the number of people involved in Analyzing Feelings in Twitter Communications is reduced.

H_4: If a Machine Learning Application is used using CRISP-ML(Q) then the level of acceptance of the end user to Analyze Feelings in Twitter Communications is improved.

For the hypothesis test, with the purpose of contrasting each one of them, the following was proposed:

μ_1 = Population Mean (H_1, H_2, H_4) for G_c PosTest
μ_2 = Population Mean (H_1, H_2, H_4) for G_e PosTest
Where: H_o: $\mu_1 \leq \mu_2$ and H_a: $\mu_1 > \mu_2$
Also:
μ_1 = Population Mean (H_3) for G_c PosTest
μ_2 = Population Mean (H_3) for G_e PosTest
Where: H_o: $\mu_1 \geq \mu_2$ and H_a $\mu_1 < \mu_2$

To test the hypotheses, two statistical tests were applied: Student's t test (quantitative values) and the Mann-Whitney U test (qualitative values) using the Minitab software.

4 Case study

The research was carried out for a specific case in the organization called Heydru! This company's main activity is to develop software in Peru. The architecture of the solution is shown in Figure 2.

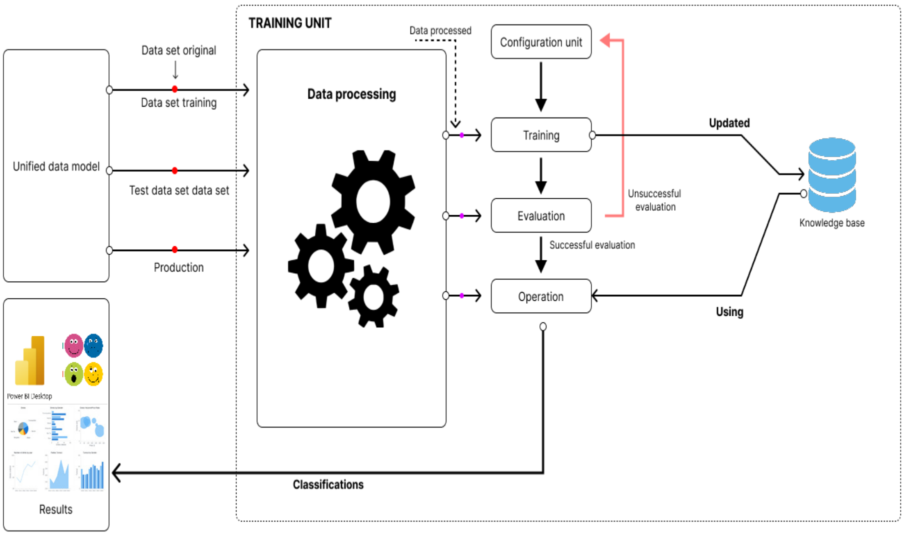


Fig. 2. Twitter data analysis workflow

Next, the case of the company Heydru! is described in detail and rigorously, following step by step the life cycle of the CRISP-ML(Q) methodology that has been shown in Figure 1.

4.1 Business and data understanding

Three success criteria are reflected for the validation of the first phase. Regarding the business criterion, it seeks to reduce the time of investigation, the time of creating reports and decision making. In the case of machine learning criteria, criteria such as the performance evaluated by the F1, and the soft measures are established. They assess robustness, the possibility of explanation, scalability, complexity, and resource demand. In addition, for the feasibility of the data, the availability of data, resources and regulatory constraints are established.

4.2 Data engineering (data preparation)

It is divided into four sub-phases, the first phase in which the following data are selected: creation_date, id, text, source, reply_to_tweet, reply_to_user_id, reply_to_user_name; the second is related to data cleaning, in which the cleaning of links, mentions, hashtag, multiple_spaces, lowercase, special characters and elimination of duplicates is developed; in the third subphase, feature engineering, there are no new features to add to the data or to be derived from existing ones. In the last subphase, the development of data standardization is contemplated, in which the definition of functions is established that allows guaranteeing the reproducibility of the application;

therefore, the functions they fulfill are store_tweet, divifitData, cross_validation and plot_matConfusion.

4.3 Engineering machine learning models

First, the quality measure is defined, and a quality and validity check of the model to be used is carried out. F1 is used. This metric is the combination of the accuracy and recall metrics:

$$F1 = 2 * (precision * recall) / (precision + recall) \tag{1}$$

And for the development of the product, the SVM (support vector machine) algorithm is used since a classification model is going to be made, and this is the optimal way to do it, so from a database, logical construction diagrams. The greatest difficulty, produced in the test of the data with the application of the model, was the data cleaning since the expected automation was not contemplated, so a function for the adequate and uniform cleaning of the data had to be created. In this way, a set of data suitable for the application of the model was achieved.

4.4 Quality assurance for machine learning applications

For this phase, it was obtained the approval and verification from an expert, who analyzed the data and the respective model with the expected results. For this, the CEO of the company Heydru! was introduced to an expert who validates the technologies to be implemented in the company; additionally, it was presented to the person in charge of the marketing area to validate the operation of the project's product.

4.5 Solution deployment

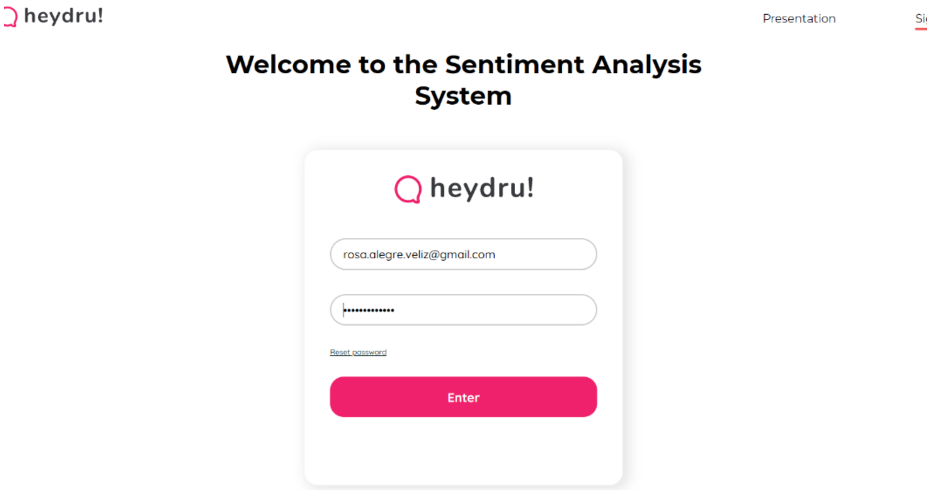
The hardware requirements to deploy the project are as follows:

- Laptop: Core i7 9th Gen 3GHz, 8GB RAM 1600MHz, 500GB SSD, 2GB Integrated Graphics, Windows 10 Pro OS.
- Computer: Core i7 5th Generation 2.5GHz Processor, 12GB RAM 1800MHz, 500GB SSD, NVIDIA 4GB, Windows 10 Pro OS.

On the other hand, a contingency plan was established: a Script 2 must be generated, the queries are kept as a base, and the list of variables is established beforehand. It is verified that the input variables can only be manageable.

For the implementation and deployment, online tests are carried out and one of them is the A/B test.

To make use of the model, a web system is developed so that the end user can analyze the feeling of his users (See Figure 3, Figure 4, Figure 5 and Figure 6)).

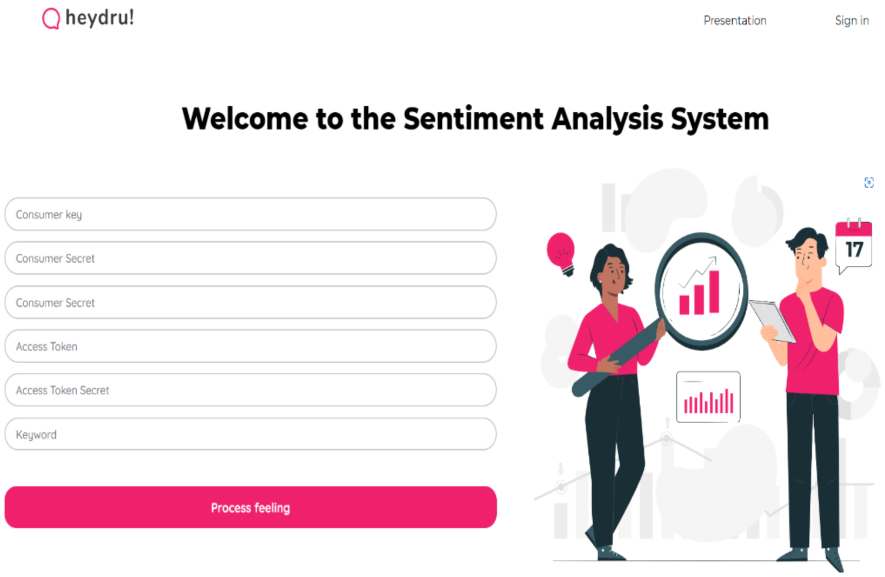


Welcome to the Sentiment Analysis System

rosa.alegre.veliz@gmail.com

••••••••••••••

Fig. 3. Login of the sentiment analysis system



Welcome to the Sentiment Analysis System

Fig. 4. Inputs section and sentiment analysis process

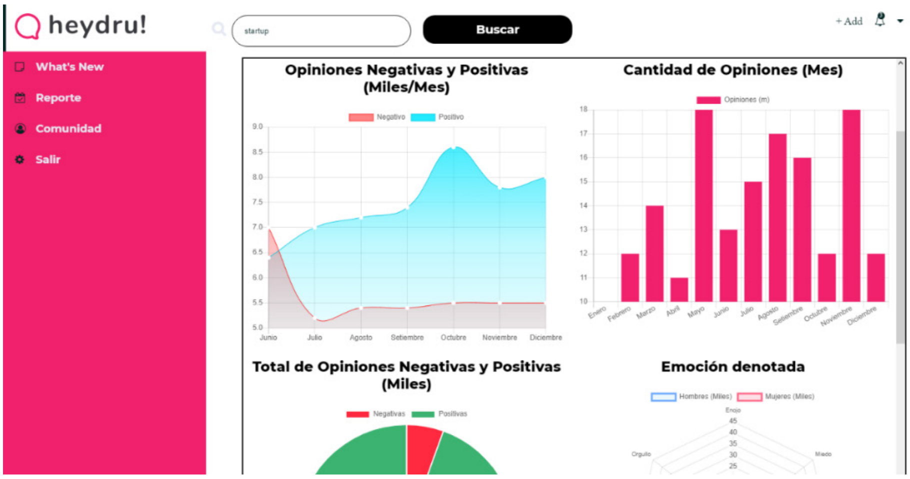


Fig. 5. Presentation of the sentiment analysis interface based on the word startup

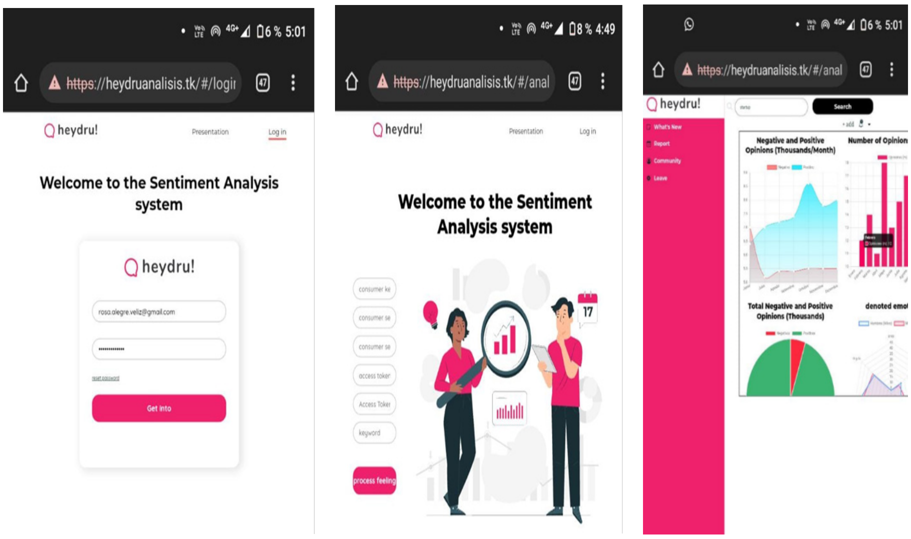

https://heydruanalisis.tk/#/login
heydru!
Presentation
Log in
Welcome to the Sentiment Analysis system
heydru!
Get into
https://heydruanalisis.tk/#/anal
Welcome to the Sentiment Analysis system
consumer ke
consumer se
consumer se
access toker
Access Toker
keyword
process feeling
https://heydruanalisis.tk/#/anal
Search
What's New
Report
Community
Leave
Negative and Positive Opinions (Thousands/Month)
Number of Opinion
Total Negative and Positive Opinions (Thousands)
denoted emot

Fig. 6. Sentiment analysis interface in responsive format

4.6 Monitoring and maintenance

In this phase, practices are established to avoid a drop in the model's performance, which consists of carrying out constant supervision so that it is evaluated, and it is decided when it is necessary to train the model again.

The ML model is updated. In addition to monitoring and retraining, reflect on the business use and ML task, it is valuable to adjust the ML process.

5 Experiments, results and discussions

5.1 Results: Reduction / Increase of indicators I1, I2, I3, I4

100% of the data obtained from the Control Group (Gc) and the Experimental Group (Ge), for each research indicator, were recorded with an observation sheet. It is obtained that 50.43% of the times, to analyze feelings, is less than the average time. Regarding the cost to develop the sentiment analysis, 73.61% is lower than the average. On the other hand, the number of personnel involved represents 50.0% less than the average with respect to the given implementation. Finally, in the level of acceptance of the end user, an increase of 69.0% was noted.

The direct observation technique was done using the stopwatch as a measuring instrument, which was very useful to understand the current state of the Twitter Sentiment Analysis process. In addition, the Minitab software was applied to perform the statistical calculations to provide information as evidence for the results obtained.

Normality test. Next (See Figure 7), the test is performed to determine which indicators have data with normal behavior. This serves to determine the statistical test to be applied.

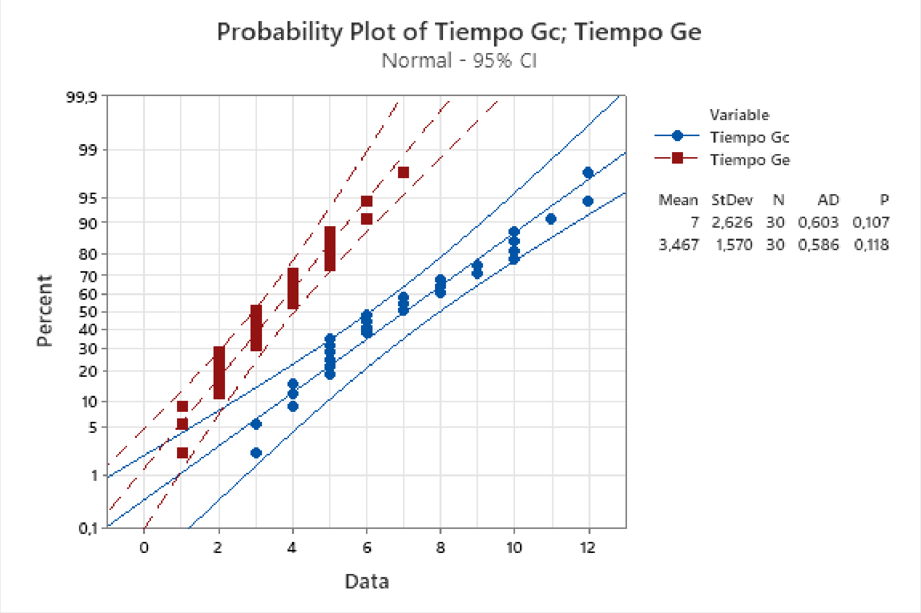


Fig. 7. Average sentiment analysis time in Twitter startup data

It is observed that in the Gc Post Test and the Ge Post Test the p-value (0.107 and 0.118) > α (0.05). Therefore, the values of the Time to analyze indicator behave normally. For the I2: Cost of feeling analysis in the Twitter data, in the Gc Post-Test and in the Ge Post-Test the p-value (0.271 and 0.243) > α (0.05). Therefore, the values of the indicator Cost to analyze have a normal behavior.

Similarly, it was done for Indicator 3 (Number of people involved). Its data are also shown to have normal behavior.

5.2 Discussion: Effect on sentiment analysis in Twitter communications

Descriptive statistics. Table 2 shows the results obtained and the application of descriptive statistics for each indicator.

Table 2. Operationalization of the dependent variable

Sample	N	95% confidence intervals for the mean	Kurtosis	Asymetry	Q3
I1: PosTest (Ge)	30	2.8805 - 4.0529	-0.504850	0.295076	5.0000
I2: PosTest (Ge)	30	12022 – 14000	-0.606303	0.381776	17000
I3: PosTest (Ge)	30	3.0000 - 4.0000	-0.668777	0.182359	5.0000

In summary, for each indicator in Table 2 it shows that, around 95% of the values, they are within 2 standard deviations for the comparison of the average. La Kurtosis indicates that there are values with peaks that are too low; similarity, the asymmetry indicates that most of the values are presented as low, the 3rd quarter indicates that the 75% of the values are less than or equal to this value. For indicator I1, the results are like those of [12] that in their research about the perceptions of Twitter users on menthol cigarettes: analysis of content and sentiment, expressed that the average time to analyze sentiments in communications via Twitter is estimated at Ge (4 days), was significantly shorter at Gc (10 days). There are similar results in [3] which, in a survey and a comparative study of the analysis of the sentiment of the tweets through the semi-supervised learning, estimating that the Ge (2.12 days) was significantly smaller than the Gc (8 days). Similarly, [20] is found with the investigation based on the analysis of sentiment on customer satisfaction with digital payment in Indonesia.

A comparative study, using KNN and Naive Bayes, highlighted in Ge (5.30 days) being significantly shorter at Gc (8 days). In addition to this, in a research on the detection of indicators for the success of an emerging company, the analysis of sentiments using text data mining estimates that Ge (1.12 days) was significantly lower than Gc (4 days) [15].

For indicator I2, it should be noted that the results for this indicator, in relation to the average cost, have been like those of [4], referring to the exploration of customer reviews online for the development of new products: the case of identifying reinforcers in the cosmetic industry, it is estimated in Ge (S/. 18 789.00), it was significantly lower for Gc (S/. 55 432.00).

In a similar way, the similar results in the study [1] with its empirical investigation on the prediction of customer abandonment behavior using the Twitter mining approach, it is estimated that the Ge (S/. 9 780.00) was significantly lower than Gc (S/27987.30). Thus, the study [9] on the analysis of sentiment of multimodal data from Twitter estimates that Ge (S/. 15 400.00) was significantly lower than Gc (S/. 60

780.50). On the other hand, in relation to a comparison of the performance of the supervised automatic learning models for the analysis of the opinion of the Covid-19 tweets, it is estimated that the Ge (S/. 10 950.23) was significantly lower than the Gc (S/. 19,789.56) [14].

Finally, the results of the I3 indicator with the [19] show similarities referring to an efficient preprocessing method for the supervised feeling analysis through the conversion of sentences into numerical vectors: a case study of Twitter, estimates that in Ge (2 personas) was significantly smaller than Gc (5 personas). On the other hand, [5] in his investigation Analysis of social networks of Covid-19 feelings: application of Artificial Intelligence estimates that Ge (6 people) was significantly lower than Gc (10 people). The study [2] referring to a hybrid N-gram model using Naive Bayes to classify political feelings on Twitter estimates that Ge (1 person) was significantly lower than Gc (3 personas). Finally, in an investigation into the deep analysis of sentiment: a case study derived from Turkish Twitter; it is estimated that Ge (4 persons) was significantly lower than Gc (8 persons) [17].

As in developing countries similar to Peru, the Machine Learning solutions for the Analysis of Sentiments in Communications via Twitter are still under-developed. This poses a challenge for the development of these solutions aimed at commercial users from different business sectors in various remote areas.

Inferential statistics. In Tables 3 and 4, the values of the application of the statistical principles are shown for the contrast of the hypotheses.

Table 3. Operationalization of the dependent variable

Sample	n	Ho	t-value	p-value
I1: PosTest (Gc)	30	$\mu1 > \mu2$	6.40	0.000
I1: PosTest (Ge)				
I2: PosTest (Gc)	30	$\mu1 > \mu2$	17.23	0.000
I2: PosTest (Ge)				
I3: PosTest (Gc)	30	$\mu1 > \mu2$	10.04	0.000
I3: PosTest (Ge)				

Table 4. Operationalization of the dependent variable

Sample	n	Ho	w-value	p-value
I4: PosTest (Gc)	30	$\mu1> \mu2$	475.00	0.000
I4: PosTest (Ge)				

Since all p values are less than 0.05, the results provide enough evidence to reject the null hypotheses (Ho), and the alternate hypotheses were correct. The tests will turn out to be significant.

Research implications. Other applications of the solution were in the political sphere, which leads to evaluating the comments on social networks, in this case Twitter, and as a result, the candidate's level of acceptance. Regarding the application, it started with the understanding of business and data, in which the scope is defined, we evaluate success criteria. In this case, it corresponds to the publications of the candidates and the comments done by the readers. It can be also the media, natural persons, and other politicians' interest. In the following phase, data engineering, a function for data cleaning and development of data for the model is planned and developed. During the model engineering phase, the application of the model is used with the data collected. However, the model is evaluated with the use of other noisy or incorrect input data to obtain the validity of the model. Finally, the deployment is carried out with the implementation of the model programmed on an existing software system.

The Machine Learning solution that has allowed to significantly optimize the values of Analysis cost, Number of people involved and Acceptance level can be perfectly applied in a wide variety of business sectors, in different geographical regions, worldwide, today and in the future.

6 Conclusions and future research

The Machine Learning models used for the analysis of feelings in social networks are effectively integrated in different areas, highlighting in the companies of Marketing that optimize their work and reduce costs related to time and dedicated personnel to carry out the analysis of feelings in social networks, verifying the growth of startups in which they invest and obtain good investment results. Most applications carried out for the analysis of feelings using predictive models with high precision, such as convolutional neural networks (CNN) and Support Vector Machine (SVM), generate an automation of processes for data management, interpretation of predictive analysis to carry out and customize the tasks of the assigned personnel. In the present investigation, an application was used for the analysis of feelings for the management of trends in startups, which is of great importance for the marketing area of the company Heydru! The solution was developed using the CRISP-LM(Q) methodology, which adapts to the use of massively changing information, ensuring the quality of the model in its deployment as the information changes or increases. As a result, it was able to improve the indicators of the process in the company Heydru!, related to the time of an analysis, the cost to carry out an analysis, the amount of staff involved and the level of satisfaction; solving these limitations and finding the solutions one person would need.

For future investigations, it is proposed to improve the algorithm used to increase the level of precision, adding neural networks of different types [15][16][17], as well as using new techniques of Deep Learning. The platform for the analysis of feelings can be optimized to increase the amount of information to be handled, as well as to increase the different statistical graphs for the report that should be presented in the respective area, optimizing the cost of the personal and the time dedicated to the analysis [13]. It is also necessary to implement the CRISP-LM(Q) methodology in a continuous way for this type of analysis, which requires the use of information in constant change and increase, since this methodology has recently been implemented, it is necessary to follow up and study the new changes and updates.

It is necessary to expand the use of the feeling analysis applications to different areas that need to carry out monitoring of the performance of a company, as well as the development of cloud platforms [29][30] with adaptive use to carry out a feeling analysis, supporting this is the way for SME that wants to implement the quality analysis process.

Some limitations have been identified in some algorithms for sentiment analysis on Twitter, which limits the efficiency of the results among decision makers; however, this has not impaired the interpretation of the results. Future implementations of Machine Learning solutions should consider the use of more efficient algorithms to achieve better results and thus eliminate bias in decision making.

7 Acknowledgment

We would like to thank the Universidad Autónoma del Perú for its support to our work carried out, specifically the career of de Systems Engineering. Also, to the Heydru company! for allowing us to apply our research in its institution.

8 References

[1] L. Almuqren, F. S. Alrayes, and A. I. Cristea, "An empirical study on customer churn behaviours prediction using arabic twitter mining approach," Futur. Internet, vol. 13, no. 7, pp. 1–19, 2021, https://doi.org/10.3390/fi13070175

[2] J. Awwalu, A. Bakar Abu, and M. Yaakub Ridzwan, "Hybrid N-gram model using Naïve Bayes for classification of political sentiments on Twitter," Neural Comput. Appl., vol. 31, no. 12, pp. 9207–9220, 2019, https://doi.org/10.1007/s00521-019-04248-z

[3] X. M. Cuzcano and V. H. Ayma, "A comparison of classification models to detect cyberbullying in the Peruvian Spanish language on twitter," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 10, pp. 132–138, 2020, https://doi.org/10.14569/IJACSA.2020.0111018

[4] N. F. F. Da Silva, L. F. S. Coletta, and E. R. Hruschka, "A survey and comparative study of tweet sentiment analysis via semi-supervised learning," ACM Comput. Surv., vol. 49, no. 1, pp. 1–26, 2016, https://doi.org/10.1145/2932708

[5] M. Denegri Coria, J. C. Morales Arevalo, J. L. Hilario-Rivas, R. Hilario-Cárdenas, and J. I. Prado-Juscamaita, "Supervised Sentiment Analysis Algorithms," vol. 12, no. 14, pp. 2000–2012, 2021, https://turcomat.org/index.php/turkbilmat/article/view/10547

[6] M. Haddara, J. Hsieh, A. Fagerstrøm, N. Eriksson, and V. Sigurðsson, "Exploring cus-tomer online reviews for new product development: The case of identifying reinforcers in the cosmetic industry," Manag. Decis. Econ., vol. 41, no. 2, pp. 250–273, 2020, https://doi.org/10.1002/mde.3078

[7] M. Hung et al., "Social network analysis of COVID-19 sentiments: Application of artifi-cial intelligence," J. Med. Internet Res., vol. 22, no. 8, pp. 1–13, 2020, https://doi.org/10.2196/22590

[8] J. Ji, H. Wang, S. Song, and J. Pi, "Sentiment analysis of comments of wooden furniture based on naive Bayesian model," Prog. Artif. Intell., vol. 10, no. 1, pp. 23–35, 2021, https://doi.org/10.1007/s13748-020-00221-3

[9] M. Kapatamoyo, "Data analytics in mass communication: New methods for an old craft," RISTI - Rev. Iber. Sist. e Tecnol. Inf., vol. 2019, no. E20, pp. 504–515, 2019.

[10] J. D. Kinyua, C. Mutigwe, D. J. Cushing, and M. Poggi, "An analysis of the impact of President Trump's tweets on the DJIA and S&P 500 using machine learning and sentiment analysis," J. Behav. Exp. Financ., vol. 29, p. 100447, 2021, https://doi.org/10.1016/j.jbef.2020.100447

[11] A. Kumar and G. Garg, "Sentiment analysis of multimodal twitter data," Multimed. Tools Appl., no. February, 2019, https://doi.org/10.1007/s11042-019-7390-1

[12] S. Kurniawan, W. Gata, D. A. Puspitawati, I. K. S. Parthama, H. Setiawan, and S. Hartini, "Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis," IOP Conf. Ser. Mater. Sci. Eng., vol. 835, no. 1, 2020, https://doi.org/10.1088/1757-899X/835/1/012057

[13] J. Murga et al., "A Sentiment Analysis Software Framework for the support of Business information architecture in the tourist sector," vol. 12390, 2020, https://doi.org/10.1007/978-3-662-62308-4_8

[14] H. L. Nguyen and J. E. Jung, "Statistical approach for figurative sentiment analysis on Social Networking Services: a case study on Twitter," Multimed. Tools Appl., vol. 76, no. 6, pp. 8901–8914, 2017, https://doi.org/10.1007/s11042-016-3525-9

[15] J. Ochoa-Luna and D. Ari, "Deep Neural Network Approaches for Spanish Sentiment Analysis of Short Texts," vol. 1, pp. 206–216, 2018, https://doi.org/10.1007/978-3-030-03928-8

[16] J. Ochoa-Luna and D. Ari, "Word embeddings and deep learning for spanish twitter sentiment analysis," Commun. Comput. Inf. Sci., vol. 898, pp. 19–31, 2019, https://doi.org/10.1007/978-3-030-11680-4_4

[17] D. Palomino and J. Ochoa-Luna, "Advanced Transfer Learning Approach for Improving Spanish Sentiment Analysis," Adv. Soft Comput., vol. 11835 LNAI, no. November, pp. 112–123, 2019, https://doi.org/10.1007/978-3-030-33749-0_10

[18] G. A. Pierina, P. J. Guzman Ramos, L. A. Chipana Vila, C. A. Trigoso Valeriano, and J. Fabian Arteaga, "Bag of embedding words for sentiment analysis of tweets," Computers, vol. 14, no. 3, pp. 223–231, 2019, https://doi.org/10.17706/jcp.14.3.223-231

[19] S. W. Rose, C. L. Jo, S. Binns, M. Buenger, S. Emery, and K. M. Ribisl, "Perceptions of menthol cigarettes among twitter users: Content and sentiment analysis," J. Med. Internet Res., vol. 19, no. 2, pp. 1–16, 2017, https://doi.org/10.2196/jmir.5694

[20] J. K. Rout, K. K. Raymond Choo, A. Kumar Dash, S. Bakshi, S. Kumar Jena, and K. L. Williams, "A model for sentiment and emotion analysis of unstructured social media text," Electron. Commer. Res., vol. 18, no. 1, pp. 181–199, 2018, https://doi.org/10.1007/s10660-017-9257-8

[21] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. Sang Choi, "A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis," PLoS One, vol. 16, no. 2, pp. 1–23, 2021, https://doi.org/10.1371/journal.pone.0245909

[22] J. R. Saura, P. Palos-Sanchez, and A. Grilo, "Detecting indicators for startup business success: Sentiment analysis using text data mining," Sustain., vol. 11, no. 3, pp. 1–14, 2019, https://doi.org/10.3390/su11030917

[23] P. Sharma and A. K. Sharma, "Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms," Mater. Today Proc., 2020, https://doi.org/10.1016/j.matpr.2020.09.351

[24] H. A. Shehu et al., "Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data," IEEE Access, vol. 9, pp. 56836–56854, 2021, https://doi.org/10.1109/ACCESS.2021.3071393

[25] K. Sigit, A. P. Dewi, G. Windu, Nurmalasari, T. Muhamad, and N. Kadinar, "Comparison of Classification Methods on Sentiment Analysis of Political Figure Electability Based on

Public Comments on Online News Media Sites," IOP Conf. Ser. Mater. Sci. Eng., vol. 662, no. 4, 2019, https://doi.org/10.1088/1757-899X/662/4/042003

[26] M. K. Sohrabi and F. Hemmatian, "An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study," Multimed. Tools Appl., 2019, https://doi.org/10.1007/s11042-019-7586-4

[27] G. Vizcarra, A. Mauricio, and L. Mauricio, "A deep learning approach for sentiment analysis in Spanish Tweets," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11141 LNCS, pp. 622–629, 2018, https://doi.org/10.1007/978-3-030-01424-7_61

[28] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes," J. Phys. Conf. Ser., vol. 1444, no. 1, 2020, https://doi.org/10.1088/1742-6596/1444/1/012034

[29] G. Zapata, J. Murga, C. Raymundo, J. Alvarez, and F. Dominguez, "Predictive model based on sentiment analysis for peruvian smes in the sustainable tourist sector," IC3K 2017 - Proc. 9th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag., vol. 3, no. Kmis, pp. 232–240, 2017, https://doi.org/10.5220/0006583302320240

[30] G. Zapata, J. Murga, C. Raymundo, F. Dominguez, J. M. Moguerza, and J. M. Alvarez, "Business information architecture for successful project implementation based on sentiment analysis in the tourist sector," J. Intell. Inf. Syst., vol. 53, no. 3, pp. 563–585, 2019, https://doi.org/10.1007/s10844-019-00564-x

[31] S. Studer et al., "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," Machine Learning and Knowledge Extraction, vol. 3, no. 2, pp. 392–413, Apr. 2021, https://doi.org/10.3390/make3020020

9 Authors

Rosa Alegre-Veliz, student at the Faculty of Engineering and Architecture at the Universidad Autónoma del Perú, with extensive experience in database modeling and software development (email: rale-grev@autonoma.edu.pe).

Pedro Gaspar-Ortiz, student at the Faculty of Engineering and Architecture at the Universidad Autónoma del Perú, with extensive experience in database management and development of mobile applications (email: pgas-par@autonoma.edu.pe).

Javier Gamboa-Cruzado, Systems Engineer, Doctor in Systems Engineering, Doctor in Administration. Professor-Researcher in the postgraduate programs at Systems Engineering Faculty at the Universidad Nacional Mayor de San Marcos, Peru (email: jgamboac@unmsm.edu.pe).

Liset Rodriguez Baca, Systems Engineer, graduated in Education, Master in Systems Engineering with Mention in Management and Management in Information Technology, Master in Strategic Business Management, Doctor in Education Sciences. Director of the Professional School of Systems Engineering at the Universidad Autónoma del Perú (email: liset.rodriguez@autonoma.pe).

Waldy Grandez Pizarro, Computing and Systems Engineer. Professor at the Faculty of Engineering and Architecture at the Universidad de San Martin de Porres, Peru (email: wgrandezp@usmp.pe).

Rosa Menéndez Mueras, Systems Engineer. Professor at the Facultad de Ingeniería de Sistemas e Informática at the Universidad Nacional Mayor de San Marcos, Peru (email: rmenendezm@unmsm.edu.pe).

Carlos Chávez Herrera, Systems Engineer. Professor at the Facultad de Ingeniería de Sistemas e Informática at the Universidad Nacional Mayor de San Marcos, Peru (email: cchavezh@unmsm.edu.pe).

Article submitted 2022-08-22. Resubmitted 2022-10-05. Final acceptance 2022-10-21. Final version published as submitted by the authors.