

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

**Inteligencia artificial: Evaluación emocional y cognitiva enfocada
al análisis de tendencias y comportamientos del consumidor.
Casos de productos de consumo masivo.**

Camila Alejandra Revelo Rodríguez

Whitman Joel Marín Salazar

Jordy Vinicio Urquiza Solorzano

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 18 de mayo de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ
Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

**Inteligencia artificial: Evaluación emocional y cognitiva enfocada
al análisis de tendencias y comportamientos del consumidor.
Casos de productos de consumo masivo.**

Camila Alejandra Revelo Rodríguez
Whitman Joel Marín Salazar
Jordy Vinicio Urquizo Solorzano

Nombre del profesor, Título académico

Carlos Alberto Suárez Núñez, Ph.D.

Quito, 18 de mayo de 2021

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: Camila Alejandra Revelo Rodríguez

Código: 00136975

Cédula de identidad: 1717155368

Nombres y apellidos: Whitman Joel Marín Salazar

Código: 00138240

Cédula de identidad: 1718546128

Nombres y apellidos: Jordy Vinicio Urquizo Solorzano

Código: 00137952

Cédula de identidad: 1727354969

Lugar y fecha: Quito, 18 de mayo de 2021

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on <http://bit.ly/COPETHeses>.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en <http://bit.ly/COPETHeses>.

RESUMEN

La investigación de mercados es una herramienta que ayuda a las empresas o entidades a obtener información acerca de los consumidores y sus perspectivas. No obstante, en países en vías de desarrollo como Ecuador, muy pocas empresas realizan estos estudios, evitando que se puedan aprovechar oportunidades de crecimiento. Las redes sociales son una gran fuente de datos que pueden ser aprovechados para realizar investigación de mercados. Sin embargo, estos datos se presentan de manera no estructurada por lo que es difícil analizarlos. Actualmente, existen herramientas de procesamiento de lenguaje natural que permiten el manejo de estos datos y su análisis, implementando también algoritmos de inteligencia artificial. El objetivo de este estudio es desarrollar un modelo de análisis de sentimientos usando un algoritmo SVM (Support Vector Machine) y una red neuronal para predecir el sentimiento de los comentarios de las redes sociales de dos productos de consumo masivo en Ecuador. Además, se tiene como objetivo utilizar herramientas de NLP (Natural Language Processing) y análisis de texto para obtener información de las perspectivas de los consumidores sobre estos productos. Este estudio demuestra cómo se pueden utilizar los datos de las redes sociales para obtener información útil para llevar a cabo un análisis de contenido y poder evaluar estrategias de marketing y planificar próximas campañas. Adicionalmente se proponen recomendaciones para futuros estudios y aplicaciones de la información obtenida.

Palabras clave: redes neuronales, investigación de mercado, inteligencia artificial, análisis de sentimientos, análisis de contenido, NLP, redes sociales, machine learning.

ABSTRACT

Market research is a tool that helps companies or entities to obtain insights about consumers and their perspectives. Nevertheless, in developing countries such as Ecuador, few companies perform these studies. This prevents companies from taking advantage of growth opportunities. Nowadays, social media platforms are a powerful data source that can be harnessed to carry out market research. However, social media data is available as unstructured data, making it difficult to analyze. There are Natural Language Processing (NLP) tools that allow unstructured data management and analysis. The objective of this study is to develop a sentiment analysis model using a SVM (Support Vector Machine) algorithm and a Neural Network to predict the sentiment of comments available in two Ecuadorian massive consumption products' social media platforms. Another objective consists in using NLP tools and text analysis to obtain insights about consumers' perspectives towards these products. This study demonstrates how social media data can be useful information as an input for content analysis and evaluate marketing strategies and campaigns. Moreover, recommendations for future studies and applications with the collected information are proposed.

Key words: neural networks, market research, artificial intelligence, sentiment analysis, content analysis, NLP, social media, machine learning.

TABLE OF CONTENTS

Introduction.....	11
Development of the topic.....	14
1. Literature Review.....	14
1.1 Sentiment Analysis.....	14
1.2 Sentiment Analysis Techniques.....	16
1.3 Machine learning algorithms for sentiment analysis.....	18
2. Methodology.....	20
2.1 Business Understanding.....	22
2.2 Data Understanding.....	23
2.3 Data Preparation.....	24
2.3.1 Preprocessing Techniques.....	24
2.3.2 Feature Extraction.....	26
2.3.3 Dimensionality Reduction.....	26
2.4 Modeling.....	27
2.4.1 Support Vector Machine (SVM).....	27
2.4.2 Neural Network (NN).....	29
2.5 Evaluation.....	32
2.5.1 Results of SVM.....	32
2.5.2 Results of Neural Network.....	33
3. Data Presentation.....	35
3.1 Model Deployment.....	36
4. Applications.....	37
4.1 Mixed Methods Research.....	38
Challenges.....	40

Conclusions.....	41
Recommendations.....	42
Bibliographic References.....	44
Annex A: Most frequent unigrams for product A and product B.....	48
Annex B: Most frequent bigrams for product A and product B.....	49

INDEX OF TABLES

Table 1. Comparison between lexicon-based and machine learning approaches.....	17
Table 2. Algorithms for sentiment analysis.....	18
Table 3. Examples of emoji transformation.....	25
Table 4. Hyperparameters and model accuracy results of SVM.....	32
Table 5. Hyperparameters and model accuracy results of NN.....	34

INDEX OF FIGURES

Figure 1. CRISP-DM iterative process	21
Figure 2: SVM representation.....	28
Figure 3: Neural Network representation.....	30
Figure 4: Set of values for learning rate and batch size for Grid Search.....	31
Figure 5: Confusion Matrix of SVM.....	33
Figure 6: Confusion Matrix of NN.....	35
Figure 7: Word Clouds for Product A and Product B.....	36
Figure 8: Count of Comments by Sentiment for Product A and Product B.....	37
Figure 9: Convert Parallel Design.....	38
Figure 10: Exploratory Sequential Design.....	39

INTRODUCTION

Market research is the process by which entities gather and interpret information about individuals or organizations using statistical and analytical methods to gain insights or support decision making (ESOMAR, 2007). Today, a significant portion of market research is focused more on the customer than the firm (Pineiro, 2014). Traditional market research methods include focus groups, depth interviewing, and survey completion (Verma et al., 2018). However, the use of these methods has its drawbacks. Although researchers have a good control over the information gathered using traditional methods, the responses they get from their subjects is reactive. Hence, the information provided tends to only answer the question present in the interview. Respondents usually answer based on the interview guide or the questions that are being asked, which limits the information the subject might be able to provide and might bias research results. In addition, TMR (Traditional Market Research) methods can help companies visualize spoken needs, but there are other methods that aim to get information about consumers' unspoken latent needs. Moreover, TMR methods require long intervals of time to deploy and usually represent a high cost that small and medium companies may not be willing to accept (Price et al., 2015).

Inside market research, researchers can perform quantitative and qualitative analysis. One way to get information to perform quantitative and qualitative analysis is through content analysis (Bengtsson, 2016). Content analysis is defined as a research technique that makes inferences from texts or other meaningful content to the contexts of their use (Krippendorff, 2004). Therefore, texts and content generated by potential consumers can be exploited as input for content analysis. But why is this content useful for a brand or company?

A customer's perception of a brand or product is highly influenced by the relationship between both. The sentiments consumers develop about a brand reflect their opinion, attitude,

and the overall likelihood to support them. Nevertheless, customer's engagement depends on the intensity or passion of the sentiment they have. Consequently, if customers feel a strong positive sentiment, WoM (word of mouth) tends to be positive. Fetscerin et al. (2021) have developed a framework to understand how branding works in the customer's mind. This framework explains how customer's feelings about a brand influence their thinking, which then determine how customers act towards the brand. Customers can act based on their thinking in three dimensions: Share of Voice (Say), Share of Wallet (Choose), and Action Preparedness (Devote). Share of voice corresponds to how consumers actively communicate about their thoughts or feelings and can be displayed as feedback, complaints, brand defense, and positive or negative WoM.

As texts and content generated by customers and potential consumers are relevant to companies and brands, researchers need to find a way to retrieve this content and analyze it. Web 2.0 and the appearance of social media platforms have changed marketing because the brands' influence shifted in scalability and consumers changed the way in which they share, evaluate, and choose information (Smithee, 2011). On social media, consumers can choose whether to connect with others or not. Also, user-generated content in social media is perceived as unbiased and free of commercial influence, which promotes consumers to seek advice on these platforms (Voramontri & Klieb, 2018). In addition, information flow is open and the voice of the customer in social media is based on honesty and freedom (Patino et al., 2012). For those reasons and the limitations of TMR methods mentioned previously, social media has become an ideal alternative to TMR methods and can facilitate how companies obtain information about potential customers' perspective of the brand (Schlack, 2010).

Given that user-generated content is available, researchers need to find a tool to process textual and unstructured data from social media. Sentiment analysis is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written

language. One of the main fields in which sentiment analysis has been used is marketing research and consumer information. With sentiment analysis, large amounts of unstructured data in the form of text can be translated into useful business insights. As billions of sentences of user-generated content are being created on social media, these platforms have become a common place for sentiment analysis applications (Patino et al., 2012).

The purpose of this project is to use sentiment analysis to evaluate the functional performance and consumer's emotion response regarding two massive consumption products. Also, one of this study's objectives is to develop a tool that can be used by Ecuadorian companies to obtain an input to perform market research by applying content analysis to user generated content in social media platforms.

The relevance of this study in the Ecuadorian context is that it is difficult to find case studies about the use of sentiment analysis in the country, and the cases found do not address consumer research or product performance. In addition, as of 2017, only 1.7% of Ecuadorian companies had conducted market research, for commonly known methods require lengthy processes and high costs (Lazo et al., 2017). However, 78.7% of Ecuadorians with internet access use social media platforms (Alcázar, 2020). There is a great amount of data available in social media, yet it has not been harnessed by Ecuadorian companies.

DEVELOPMENT OF THE TOPIC

1. Literature Review

This section includes topics such as: the definition of sentiment analysis, techniques that are carried out for sentiment analysis and machine learning algorithms for sentiment analysis. Covering these topics to support the development of the proposed methodology for the main company.

1.1 Sentiment Analysis

Sentiment analysis is a computational study about people's attitudes, emotions and opinions in relation to a particular topic, which is usually covered by reviews made by interested users (Medhat et.al, 2014). This analysis comes from Natural Language Processing (NLP), which is basically the ability of a computer to process natural language. It is worth mentioning that some authors treat sentiment analysis as a synonym for opinion mining, while others such as Tsytsarau and Palpanas (2011) think that both terms have somewhat different notions. Sentiment analysis recognizes and analyzes a sentiment reflected in a text, while opinion mining is a field of study that is responsible for obtaining and analyzing different opinions from different people about a specific entity (Tsytsarau and Palpanas, 2011). In any case, the main objective of the sentiment analysis is to extract and analyze opinions, identify feelings and emotions, and then classify them based on their polarity (Medhat et.al, 2014).

It is important to detail what is a sentiment of an opinion, and for that Liu (2015) establishes that the sentiment is represented by three aspects:

- Type: Sentiment may be based on psychology (mental state of a person characterized by joy, anger, sadness or fear of a specific situation), linguistics (emotions generated by the form of speaking, related to sympathy or antipathy towards a person), or consumer research (evaluations from rational reasoning, tangible or intangible beliefs, utilitarian

attitudes and emotional responses to entities, such as giving value to the price of a product or being happy to buy it).

- Polarity: It constitutes the orientation of the sentiment, so it can be positive, negative or neutral.
- Intensity: It constitutes the levels of strength of the sentiment, in relation to its polarity, so it can be weak or strong.

For instance, to clarify these three aspects, if a person comments the following about an edible product: "It has a super delicious and different taste", the sentiment expressed in this comment is based on consumer research and it expresses a strongly positive sentiment, according to its polarity and intensity.

Additionally, sentiment analysis can be examined as a classification problem, so there are three levels of classification according to Medhat et.al (2014):

- Document-level: An entire document is considered the basic unit of information, so the level classifies it as a positive or negative opinion or sentiment.
- Sentence-level: A sentence is considered the basic unit of information, so the level classifies the opinion or sentiment, expressed in the sentence, as positive or negative.
- Aspect-level: Since users can give different opinions or express different sentiments for different aspects of the same topic, the level classifies the sentiments or opinions with respect to the aspects of the topics.

It is worth mentioning that, according to Liu (2012), there is no significant difference between document and sentence levels since sentences are considered as short documents. On the other hand, in terms of texts in the form of documents, sentences or aspects, it is important to consider the data sets that are used in sentiment analysis. Thus, the main sources of data are product reviews and comments from social media. Product reviews are very important for

companies as they can take decisions based on the analysis of user opinions about their products. Similarly, social media is important because people can freely share and give opinions about a specific product or topic (Medhat et.al, 2014).

One of the first steps in the sentiment analysis is the selection of the text features, and among the main ones according to Aggarwal and Zhai (2012), are the following:

- Term's presence and frequency: Words and their frequencies in the text, considering the binary weights of the words (for example 1 for the presence of the word and 0 otherwise) and the weights of the frequencies (as indicative of the importance of the features).
- Parts of speech (POS): Adjectives as markers of user opinions.
- Opinion words and phrases: Words and phrases used to reflect user opinions.
- Negations: Negative words that change the polarity of user opinions.

1.2 Sentiment Analysis Techniques

To classify the sentiments of an expression, two approaches can be used, such as lexical approaches and machine learning (Ulises and Torres, 2020). In lexical approaches, an expression is determined as positive if their semantic orientation (SO) is more related to “best” or it is considered as negative if it is more related to “poor” (Garg & Lal, 2018). Therefore, to calculate the result of the comment according to Garg and Lal (2018), it is required to take the average of the SO values of all the words that forms the comment. Then, it would be possible to have a positive or negative comment. In addition, it requires the use of lexical resources, such as SentiWordNet database (Jayasanka, et al., 2013). Its application appears in the model proposed by Xion et al. (2018) where the sentiments’ labels depend on the semantic orientation (SO) of the word as at the level of the comment.

On the other hand, machine learning approaches focus on supervised learning techniques, where basically the models are built by learning from a large number of training examples. Each one of these training examples has a label that indicates its output or classification (Zhi-Hua Zhou, 2018). Therefore, after the training phase, these models can be used to classify unlabeled data (Cunningham, Cord & Delany, 2008).

Further comparisons between lexicon-based and machine learning approaches are analyzed in Table 1.

Table 1. *Comparison between lexicon-based and machine learning approaches (Garg & Lal, 2018)*

Parameters	Machine Learning	Lexicon-Based										
Classification Approach	Supervised	Unsupervised										
Domain	Dependent	Independent										
Statistical Significance	More	Less (Small Dataset)										
Require prior training of Dataset	Yes	No										
Adaptive Learning	Yes	No										
Accuracy	High	Low, Depends on Lexical Resources. <table border="1" data-bbox="837 1220 1268 1411"> <thead> <tr> <th>Resource</th> <th>Coverage</th> </tr> </thead> <tbody> <tr> <td>SentiWordNet</td> <td>117,659</td> </tr> <tr> <td>WordNet Affect</td> <td>200</td> </tr> <tr> <td>SenticNet</td> <td>14,000</td> </tr> <tr> <td>MPQA</td> <td>8,222</td> </tr> </tbody> </table>	Resource	Coverage	SentiWordNet	117,659	WordNet Affect	200	SenticNet	14,000	MPQA	8,222
Resource	Coverage											
SentiWordNet	117,659											
WordNet Affect	200											
SenticNet	14,000											
MPQA	8,222											
Sensitive to quality and quantity of data	More	Little										
Time of Result Generation	Slow	Fast										
Maintenance	Not Required	Need maintenance of corpus										
Training Require	Yes	No										

Therefore, machine learning techniques were developed in this project due to its high accuracy, its sensitivity of quality and quantity of data, and its adaptive learning attribute. There are several supervised learning algorithms that are used to classify the content of the reviews to develop the sentiment analysis (Symeonidis et al., 2018).

1.3 Machine learning algorithms for sentiment analysis

The sentiment analysis in this project was based on a machine learning approach, in which its algorithms were used to solve sentiment analysis as a text classification problem. In this way, this type of problem is defined by "a set of training records, where each record is labeled as a class" (Medhat et.al, 2014). Consequently, a classification model (supervised learning) is used to predict a given class label. Table 2 shows the most used classification algorithms in sentiment analysis.

Table 2. *Algorithms for sentiment analysis*

Algorithm	Description	Advantages	Disadvantages	Reference
Naive Bayes (NB)	In function of the placement of words in the text, it uses Bayes' theorem to predict the probability that a certain number of BOWs (bag of words) characteristics belong to a given class label.	It is easy to understand and fast to implement. It is a simple and commonly used classifier in sentiment analysis.	Average accuracy decreases due to the gap between the negative classification accuracy and the positive classification accuracy of the sentiments.	(Kang et.al, 2012)
Bayesian Network (BN)	Acyclic graph where margins represent conditional dependencies and nodes represent random variables.	It is a good model for variables and their associations. Therefore, if all the variables are combined in the same sentiment classification task, the potential relationships between them can be exploited.	The computational complexity is very expensive, so it is not used regularly in sentiment analysis.	(Aggarwal and Zhai, 2012)

Maximum Entropy (ME)	Using encoding, it converts a set of labeled features into vectors, which are then used to calculate the weights of the mentioned features. These weights can get together to establish the most likely label for the feature set.	Allows to capture parallel sentences between any pair of languages that a text may have, to which sentiment analysis is applied.	The exact solution does not exist under certain circumstances, so the probability distribution resulting from the algorithm can lead to poor prediction accuracy for the sentiment analysis.	(Kaufmann, 2012)
Support Vector Machine (SVM)	Based on the limited nature of a text, few features are insignificant but at the same time correlate with each other. In this way, the main objective of the algorithm is to establish the linear separators that best separate the labeled classes.	It allows to explore opinions on different topics in real time, considering the credibility of the user and the subjectivity of the opinion as essential parts in the sentiment analysis.	The algorithm may fall into overfitting if the number of features is much greater than the number of samples in the training data set for sentiment analysis.	(Li and Li, 2013)
Neural Network (NN)	It is an algorithm made up of many neurons (nodes - basic unit) connected to each other. Neuron inputs are the frequencies of each word in a text. On the other hand, there is a set of weights related with each neuron that, in the end, is used to calculate a function of its inputs.	Based on a traditional BOWs (bag of words) model, it achieves better results in terms of classification accuracy levels of sentiments.	The learning and processing time for the prediction in relation to the polarity of sentiments is very high.	(Moraes et.al, 2013)
	Produces a hierarchical distinction of the	Very easy to interpret and resistant to noise	Trees can become very complex due to	

Decision Tree	training data, where a condition is used to divide the data, which is the absence or presence of words. The data division is recursive, so it ends until the leaf nodes have a minimum number of records, used for classification.	and lost words during the sentiment analysis.	possible duplication within the same subtree. This due to the similarity of the meaning of the words that express a desired sentiment.	(Hu and Li, 2011).
---------------	--	---	--	--------------------

Based on the description of each algorithm in Table 2, behavior patterns related to the predicted polarity of sentiment can be detected. In this way, machine learning approaches are ideal to carry out sentiment analysis (Medhat et.al, 2014).

2. Methodology

CRISP-DM or Cross Industry Standard Process for Data Mining is the most widely used methodology to address the analysis of data (Espinosa, 2020). Therefore, it is the methodology used across this project to analyze the sentiments of the comments.

CRISP-DM is a non-proprietary and open methodology that was conceived under the supervision of the European Union (Garcia-Osorio, 2019). It is an iterative process formed by six phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment (Guzman, 2018):

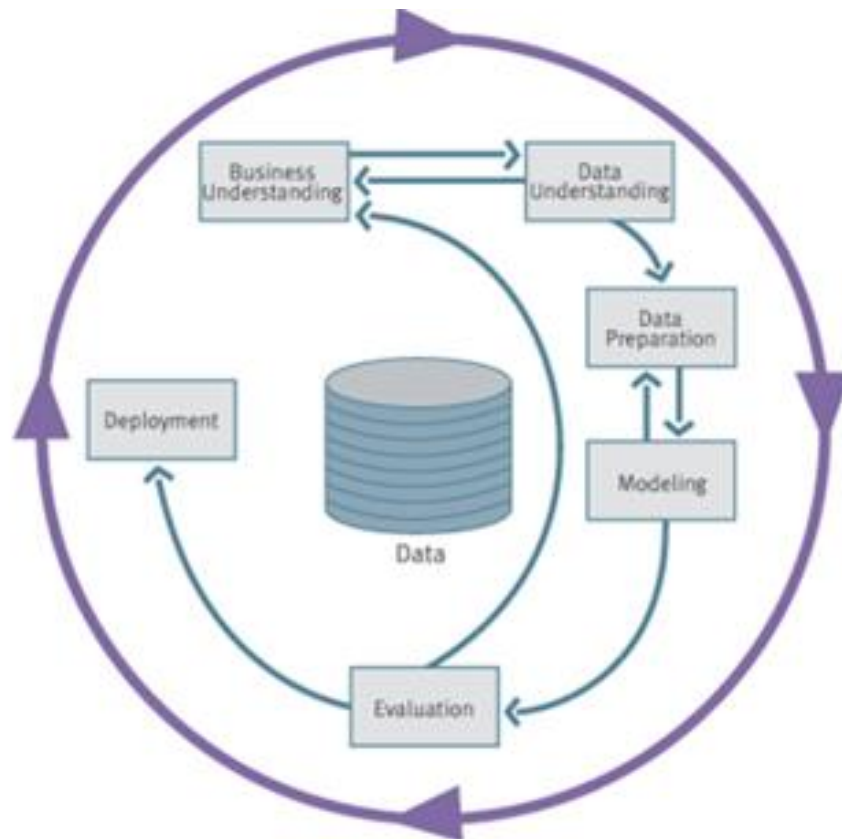


Figure 1: CRISP-DM iterative process (Guzman, 2018)

According to Bottero (2019), the first phase, Business Understanding, focuses on establishing the main business objectives, so they could be addressed under a data mining perspective to solve the problem. This requires an early evaluation of the situation of the project to visualize the path that the research will follow. The second phase, Data Understanding, gathers and explores the data to verify its quality and added value to the research. Third phase, Data Preparation, is used to clean and transform raw data into the format that is used by the machine learning algorithm. In the Modeling phase, the algorithm is chosen to satisfy the objectives of the project. Then, the Evaluation phase focuses on the evaluation of the results at addressing the requirements of the first phase. Finally, the Deployment phase reveal the implementation of the model in unseen data, which shapes the decision-making process of the project.

In the following sections, each of the phases of this methodology will be discussed in detail to develop the analysis of sentiments of the social media comments of one of the most innovative Ecuadorian company. Due to nondisclosure agreements the name of the company will not be displayed in this study, nor will its mass consumption products.

2.1 Business Understanding

An Ecuadorian company that manufactures oils, oleochemicals, biofuels, hygiene products, and consumer goods is pioneer in innovation which leads it to keep improving and analyzing the trends and behavior of the Ecuadorian market. This company is well-known to inspire an emotional link between the costumers and the products, where it is intended to “increase the taste of their meals”. That strategic management requires an external and internal continuous analysis to gather information and understand the brand positioning in the minds of their consumers. Therefore, a thorough and detailed analysis should be carried out to identify improvement opportunities (Gürel, 2017), such as:

- Gather customers’ insights from social media to improve the development of new products because digital transformation is setting up a new path of communication.
- Incorporate machine learning techniques to identify the trends of the customers’ behavior because digital transformation requires deeper analysis.
- Complement the traditional market research that is normally carried out.

Therefore, a new opportunity to gather insights of the consumers could be accomplished by the implementation of data mining techniques. The information of social media is a valuable source where customers feel free to give their opinions, feedback, or comments about a product (Lopez and Ruiz, 2008). In this case, the company’s market research could innovate positively by implementing sentiment analysis in their content analysis.

2.2 Data Understanding

In this phase, the extraction and compilation of user comments was carried out. These comments were obtained from various posts on social media, such as Facebook and Instagram of the brands analyzed in this project. It is worth mentioning that, as plan A, a code for the extraction of the data was developed. Unfortunately, at the time of writing this report, some agreements and authorizations were still being signed with the company and their accounts could not be shared for the code extraction. Therefore, it was decided to opt for plan B, which was extracting comments directly from social media as common user of these platforms with its own account, through the following applications:

- Export Comments: It is a Google Chrome extension that allows to manage and execute exports of comments from different social media and, in turn, retrieve data extracted from these comments (ExportComments, 2021). The data can be exported as Excel and CSV, considering that it does not require the use of programming and it has not cost. With this application, 586 Instagram comments of the products were exported, from June 2018 to April 2021.
- FacePager: It is a social media data extraction application made by Jakob Junger. It is built on the use of APIs (Application Programming Interface) and web scrapping tools to get public data from social media (Junger & Till, 2019). It does not require the use of programming and it has not cost. With this application, 4494 Facebook comments of the products were exported, from June 2018 to April 2021.

Once the comments were extracted, they should be labeled as positive, negative, or neutral to train the model, so it would be able to predict them correctly in the future. By the time we finished labeling comments, the dataset was imbalanced because there were far more examples of the majority classes (positive and neutral comments) than of the minority class

(negative comments) (Li et al., 2017). Hence, a technique to balance the dataset must be implemented at the training phase of the model.

2.3 Data Preparation

Data extracted by the previous methods cannot be fed into a machine learning model directly. Not every part of the comment is useful and numerical data is needed instead of textual data. To prepare the data for the models, there are some preprocessing techniques that help to clean the data. Then, numerical features (variables) are extracted from the comments. Finally, the dimension (number of features) is reduced.

2.3.1 Preprocessing Techniques

User-generated content is present as unstructured data in social media platforms such as Facebook and Instagram. Written text in social media is not formal, and the expressions used include characters, or idioms that are not necessarily useful to determine the meaning of the comment. Misspelling, special characters, and mentions are frequent in social media comments. Hence, some preprocessing steps must be followed to ensure that the input to the model developed is the optimal and less features are generated.



There are various studies that have developed experiments to select the best preprocessing techniques to apply in a sentiment analysis project. The preprocessing techniques selected in this project were based on the study performed by Dubiau (2013), in which several combinations of techniques were evaluated. From there, other techniques were added and removed supervising the results of these actions in the models' accuracies. Finally, the preprocessing techniques applied in this project were the following:

- Stop words Elimination: Stop words are defined as those words that have a grammatical function in a sentence but do not contribute to the meaning of it (Willbur, 1992). The NLTK stop words dictionary in Spanish was used to remove them. Some of the stop

words removed include “las”, “por”, “para”, which are mainly articles, prepositions, and conjunctions. These words can be translated to “the”, “by”, and “for” in English.

- Filtration of words with less than 3 characters: Words with less than three characters are removed, for they seldom are relevant in a sentence. Examples of these words include “ser” (“to be” in English) and “una” (“a” or “one” in English).
- Duplicate Characters Removal: Misspelling is frequent in social media comments, and characters may be repeated in a word when they are not supposed to. Duplicate characters were removed to prevent repeated features in the next steps. For example, words like “holaaa” (“hellooo” in English) were changed to “hola” (“hello”).
- Uppercase to Lowercase: For feature selection, it is important to have all words in a text with the same form. Because of that, uppercase letters were transformed into lowercase letters.
- Punctuation Signs Elimination: Punctuation signs are removed because they can bring noise to the models built (Dubiau, 2013).
- Special Characters Elimination: Special characters such as “#”, “%”, or “\$” are removed from texts.
- Emoji Transformation: Emojis are frequently used in social media platforms, and they can be relevant to identify the consumer’s sentiments. For this reason, emojis were transformed into words to preserve their meaning. The Emoji library for Python was used to make this transformation. Examples of this transformation are shown below:

Table 3. *Examples of emoji transformation*

Emoji	Spanish Transformation	English Transformation
	pulgar_hacia_arriba	thumbs_up
	corazon_rojo	red_heart

2.3.2 Feature Extraction

Feature Extraction is the process by which features can be obtained from raw textual data. One way to do it is through vectorization, which consists in getting numerical vectors from texts. Machine Learning algorithms need to work with numerical data as an input, so vectorization is needed to develop the models. In our study, the TF-IDF (frequency–inverse document frequency) method is used to vectorize the comments, for has been proven to be the best vectorization method for NLP (Nguyen, 2008).

- TF-IDF Vectorization: TF-ID Vectorization is achieved by multiplying a term's frequency (TF) with its inverse document frequency (IDF). TF and IDF are defined as follows:

$$TF = \frac{\text{Frequency of word in the comment}}{\text{Number of words in the comment}}$$

$$IDF = \log \frac{\text{Total number of comments}}{\text{Number of comments containing the word}}$$

The final output is a vector for each comment containing the TF-IDF scores for each feature.

To prevent irrelevant words to be included as features in the final matrix, the *min_df* parameter of the *CountVectorizer* function is used. This parameter ignores the terms that have a frequency strictly lower than a value called the cut-off point (Scikit-learn, 2020). The final matrix included 6837 features, which correspond to the words used to train the model.

2.3.3 Dimensionality Reduction

The dimensionality of the data refers to the number of attributes or input variables that describe each record in the data. Even though some steps were made to prevent unnecessary

attributes or features to be created, the final matrix still contained 6837 features. Much of these attributes are correlated and are not useful for the model to learn from. Also, a high-dimensional dataset like the one created translates into long computer processing times when the algorithm is learning. In this way, it is very important to apply dimensionality reduction for the processing of a high-dimensional data set (Song et.al, 2013), as it is the case with this project. One of the most common methods to extract features of a dataset is PCA (Principal Component Analysis). However, the matrix obtained from the vectorization is sparse, which means it is large and contains mostly zeros. PCA does not work well with sparse matrices (Hubert et al., 2016), so the following method was used:

- Singular Value Decomposition (SVD): It is a technique used to decompose a matrix into several smaller component matrices or into a smaller number of factors (factorization), preserving the most important information from the original matrix (Ientilucci, 2003).

2.4 Modeling

In this phase, the Support Vector Machine (SVM) and Neural Network (NN) algorithms are selected and applied for sentiment analysis of the brands' social media comments because they achieve better results in terms of classification accuracy levels, and they can better consider the subjectivity of user mentions, according to Medhat et.al (2014).

2.4.1 Support Vector Machine (SVM)

According to Dubiau, it is an algorithm or supervised method of binary classification in which the training of the data consists of finding an optimal hyperplane that separates the vectors, which represent the texts of a data set (vectors of features), into two groups, considering the largest separation as possible. The vectors that establish the margin of the maximum separation are called support vectors (2013):

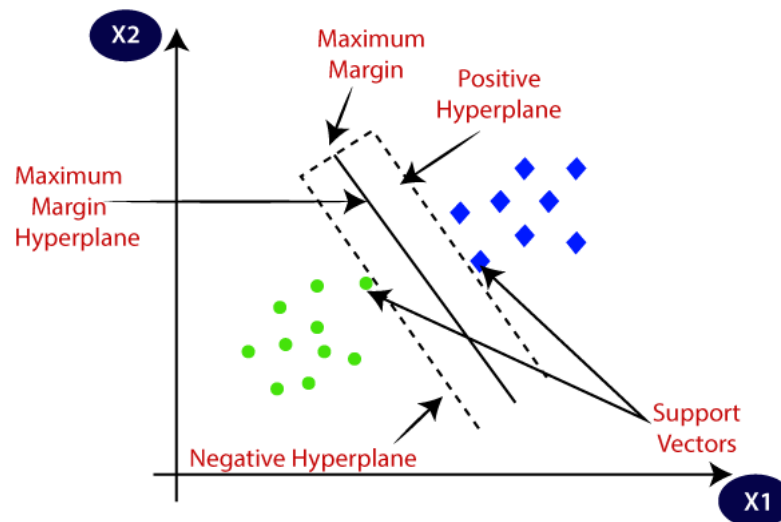


Figure 2: SVM representation (Dubiau, 2013)

Based on Figure 2, it is important to point out that the support vectors are on the positive hyperplane and the negative hyperplane. In Figure 2, it can also be seen two classes (blue and green), considering that a class is the value or values that a variable can take. In this way, based on specific values for each class, it is determined on which side of the hyperplane the vector of features to be classified is located (Dubiau, 2013).

Additionally, it is important to indicate the hyperparameters that compose the `SVC()` module implemented in Python for the development of the SVM model. In this form, based on the information obtained from the official website of Scikit-learn (2020), the hyperparameters used are the following:

- **Kernel:** It is a hyperparameter that transforms the input data of a data set into the required form. There are linear, polynomial or POLY and radial or RBF functions, considering that these last two are useful for a non-linear hyperplane, as in the case of the present project, where there are three classes based on the polarity of the sentiment: positive, negative, and neutral.
- **C:** It is a penalty hyperparameter that represents an incorrect classification, which reflects to the optimization of the model what amount of error is tolerable.

- Max_iter: It is the maximum number of iterations on the training data of the model.
- Degree: It is the degree of the polynomial kernel function.
- Coef0: It is an independent term in the kernel function.
- Class_weight: It is a hyperparameter used to balance data sets, adjusting the weights in the input data inversely proportional to the frequencies of the classes.

It is worth mentioning that to validate and evaluate the effectiveness of the model, it is necessary to use the cross-validation technique, which is useful to verify the lowest test error based on the hyperparameters used (Gupta, 2017).

2.4.2 Neural Networks (NN)

Deep learning algorithms began with the observation and analysis of mammalian brains and how biological neurons send information to each other to detect shapes and control activities (Glorot et al., 2010). Therefore, the purpose of artificial neural networks is to understand information and learn from it as it goes through several layers of neurons. Basically, neurons or nodes are the essential elements of these networks. The transmission of information from one neuron to another is represented by connection weights, which modulate the effect of the input signals and the transfer function that exhibits the characteristics of neurons (Abraham, 2005). According to Bishop (1995), the basic architecture of neural networks consists of three main layers:

- Input layer: the first layer where information flows without performing any computational task.
- Hidden layers: it could be one or multiple layers where the computational tasks are performed. They are used as a connection between the input and output layer.
- Output layer: it is responsible to transfer the results outside of the algorithm.

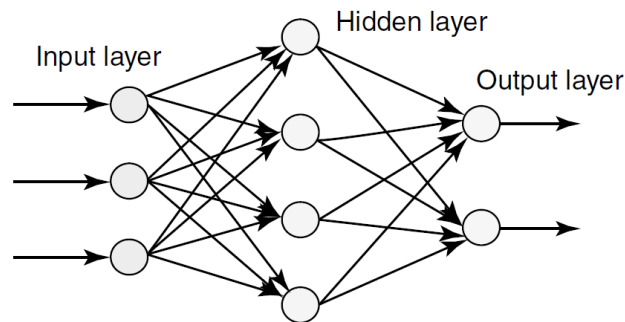


Figure 3: Neural Network representation (Abraham, 2005)

Due to the supervised learning approach, the neural network requires an input vector with its correspondent labeled response or final value. Therefore, as the input vector passes through the layers, the errors between the desired and the actual value for each node are found and the weights could change accordingly to the learning rule (Abraham, 2005).

There are several hyperparameters that should be set along the architecture of the neural network. In Python, each hyperparameter is called in the model and its value is established or declared. Based on the information obtained from the official website of Keras (2020) and Abraham (2005), the hyperparameters used are the following:

- Units: The number of neurons in each layer.
- Initializers: Define how to set the initial random weights of the layers.
- Activation function: It is the function implemented in each neuron to deliver an output based on the input.
- Constraints: They are applied on the model parameters during training.
- Dropout rate: Fraction of input units to drop to avoid overfitting.
- Optimizer: Argument required to compile a Keras model.
- Loss function: It computes the quantity that a model should seek to minimize.
- Learning rate: It controls the speed at which the model is adapted to the problem.
- Epoch: One pass through the training set.

- Batch size: Number of training examples utilized in one epoch.

Once the hyperparameters of the model were considered, it was important to find the best combination of them to achieve the best performance of the model. Therefore, hyperparameters were tuned by using Grid Search, which is one of the most widely strategies implemented to optimize hyperparameters (Bergstra and Bengio, 2012). According to Abraham (2005), Grid Search evaluates every possible combination of a set of values for each hyperparameter. Those values were established based on the range of the hyperparameters and their most common values while training and testing neural networks. For instance, Figure 4 displays the set of values of learning rate and batch size, and how they were considered in Grid Search by using `param_grid`:

```
# define the grid search parameters
learn_rate = [0.001, 0.01, 0.1, 0.2, 0.3]
batch_size = [128, 256, 300, 384]
param_grid = dict(learn_rate=learn_rate, batch_size=batch_size)
```

Figure 4: Set of values for learning rate and batch size for Grid Search

Once all the trials were executed, the best combination of hyperparameters was chosen based on the best scores obtained. Then, the model was set correctly.

Another important consideration to develop a machine learning algorithm, in this case a neural network, is the treatment of imbalanced data in order to predict correctly all the classes of comments, which are: positive, negative, and neutral. For this project, the balance of data was implemented using SMOTE or Synthetic Minority Over-Sampling Technique. This algorithm deals with the imbalanced data by over-sampling the minority class through synthetic examples (IBM, 2020), which was implemented by using the Python library `imbalanced-learn`.

2.5 Evaluation

The performance of the machine learning algorithms implemented in the previous section were evaluated. The evaluation was carried out based on an accuracy metric, because it indicates in a direct way if a classification model was trained correctly and how it works in general. The accuracy represents the number of correct predictions over the total predictions of the test set (Mishra, 2018). This means how exact is the classification of the comments of the products in relation to the polarity of the sentiment that is expressed in each one.

2.5.1 Results of SVM

Different scenarios of the SVM algorithm were run in Python with specific values of the hyperparameters. Each scenario with its respective combination of values for each hyperparameter and the accuracy result are presented below:

Table 4: Hyperparameters and model accuracy results of SVM

Hyperparameters	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Kernel	Rbf	Rbf	Poly	Poly
C	0.1	1	0.1	0.1
Max_iter	12 iterations	10 iterations	7 iterations	7 iterations
Degree	-	-	-	3
Coef0	-	-	-	0.0
Class_weight	-	-	-	Balanced
Accuracy	59%	61%	64%	75%

Based on Table 4, it is evident that the model was optimized from 59% to 75% accuracy, through a specific setting of Kernel, C and Max_iter hyperparameters from scenario 1 to scenario 3, jointly with Degree, Coef0 and Class_weight hyperparameters in scenario 4. Then, to evaluate the model, a confusion matrix was carried out, which basically allows to visualize the proportion of predicted labels in relation to the proportion of true labels of the comments

extracted from the social media of the products after applying the SVM algorithm. The confusion matrix described is presented below:

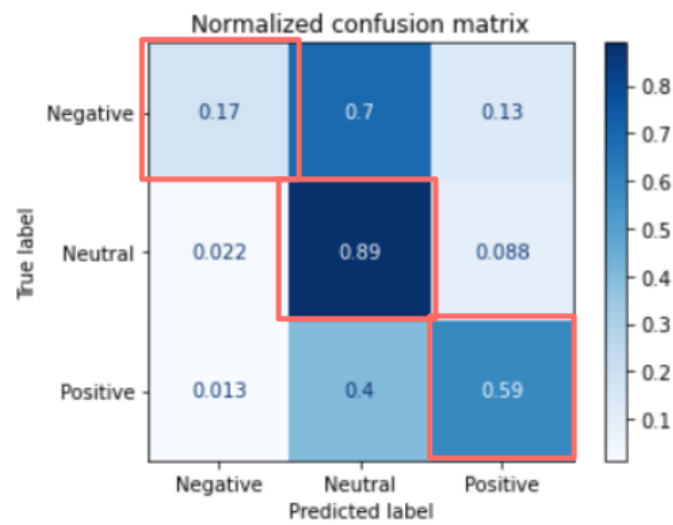


Figure 5: Confusion Matrix of SVM

As it can be observed in Figure 5, the prediction of the neutral labels (0.89) and positive labels (0.59) are more accurate than the prediction of the negative ones (0.17). Due to this lack of reliability regarding the prediction of negative sentiment comments, the SVM model is not appropriate to carry out the sentiment analysis.

2.5.2 Results of Neural Network (NN)

Different scenarios of the Neural Network algorithm were implemented in Python, where each hyperparameter was set up according to the combination of the Grid Search. The hyperparameters tuned were batch size, epochs, optimizer, learning rate, initialization, activation function, dropout regularization, and the number of neurons in each layer. The values for each hyperparameter were established based on the values proposed by Brownlee (2016) and some trial-and-error tests. At the end, approximately 130 combinations were tested. The path of improvement (the first and last model tested) is shown in Table 4 where the values of the hyperparameters are displayed with its respective accuracy:

Table 5: Hyperparameters and model accuracy results of NN

Initial Layout		Hidden Layer	
Hyperparameters	Scenario	Hyperparameters	Scenario
Units	1200	Units	1000
Kernel_initializer	-	Kernel_initializer	-
Function activation	Relu	Function activation	Relu
Kernel_constraint	-	Kernel_constraint	-
Dropout rate	0.9	Dropout rate	0.9
Output Layer		Compile Model	
Hyperparameters	Scenario	Hyperparameters	Scenario
Units	3	Loss	Categorical_crossentropy
Kernel_initializer	-	Optimizer	Adam
Function activation	Softmax	Learning rate	0.01
Accuracy		52%	

Initial Layout		Hidden Layer	
Hyperparameters	Scenario	Hyperparameters	Scenario
Units	1500	Units	1200
Kernel_initializer	Uniform	Kernel_initializer	Uniform
Function activation	Relu	Function activation	Relu
Kernel_constraint	Maxnorm(3)	Kernel_constraint	Maxnorm(3)
Dropout rate	0.9	Dropout rate	0.9
Output Layer		Compile Model	
Hyperparameters	Scenario	Hyperparameters	Scenario
Units	3	Loss	Categorical_crossentropy
Kernel_initializer	Zero	Optimizer	Adamax
Function activation	Softmax	Learning rate	0.001
Accuracy		79%	

Based on the Table 5, the model was optimized from 52% to 79% accuracy. Then, the confusion matrix was carried out as below:

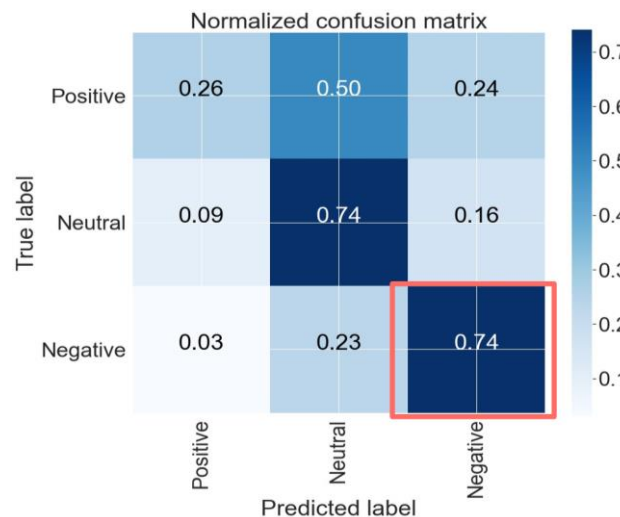


Figure 6: Confusion Matrix of NN

As it can be observed in Figure 6, the prediction of negative labels (0.74) was improved from the one displayed in the SVM model. Therefore, Neural Network algorithm was chosen as the best model to develop sentiment analysis.

A strong reliability regarding the prediction of negative sentiment comments is important because improvement opportunities can be addressed by tackling the comments that show complaints, doubts, and claims from the customers.

3. Data Presentation

In this last section, based on the results of the Neural Network (NN) model, a text analysis was carried out to obtain meaningful insights from the comments extracted of the products. Then, the model was deployed to predict the sentiments of these comments. It is worth mentioning that, specifically, 2571 comments were used from product A and 8671 comments from product B.

As seen in Figure 7, word clouds of the most frequent words have been created for each product. In addition, histograms showing the 20 most frequent unigrams (one word) and bigrams (two words) are present in Appendixes A and B.



Figure 7: Word Clouds for Product A (left) and Product B (right)

These visualizations give some insights about the consumers perspective of the products. For instance, Product A most frequent words refer to the product’s flavor and presentation. Attributes of the product are addressed directly. The most frequent words include ‘tasty’, ‘delicious’, ‘love’, and ‘want’. In contrast, the word cloud for Product B includes words that refer to moments and sharing. Words like ‘family’, ‘dreams’, ‘brothers’ and ‘moments’ are present.

The most frequent words in social media platforms for both products match the purpose of the posts present in them. Posts present in Product A’s social media emphasize on the actual product. Pictures of the product are shown alongside with some caption describing how tasty it is. However, the posts published for Product B portray different scenarios. Here, pictures of families sharing and cooking together are frequent.

3.1 Model Deployment

To analyze more deeply the customers’ perspective for both products, the neural network developed was deployed to obtain the sentiments of the unlabeled comments extracted for the text analysis. The following results were obtained:

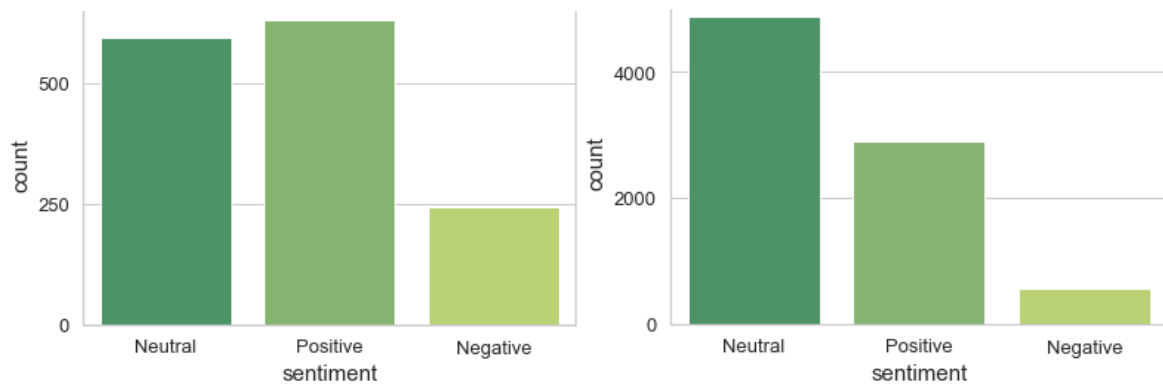


Figure 8: Count of Comments by Sentiment for Product A (left) and Product B (right)

As seen in Figure 8, the comments extracted for Product A reflect more positive sentiments than neutral or negative. Nevertheless, negative comments still represent a significant proportion of the total. On the other hand, negative comments for Product B are relatively few comparing to the quantity of positive and neutral comments. Overall, the sentiments associated with the comments present in the social media platforms for Product A and Product B are mostly neutral and positive.

4. Applications

The methodology developed for sentiment analysis across this study could be implemented in the market research of Ecuadorian companies, specifically while developing content analysis. Even though the collection of information based on Traditional Market Research (TMR) methods can help to address some costumers' needs through the questions presented in focus groups, interviews, or surveys, the development of content analysis aims to get information about their needs from the content generated in social media. Instead of implementing these methods separately, the following mixed methods enhance the performance of market research. The potential value of mixing these approaches lies in uncovering insights of how and why people engage with the product in a controlled

environment, such as an interview, or in a free of judgement environment, such as social media (Snelson, 2016).

4.1 Mixed Methods Research

The first mixed method is called Convert Parallel Design, which is a triangulation approach whereby quantitative and qualitative results are brought together to analyze and understand the market. In both strands, content analysis is an important input. Therefore, sentiment analysis is a powerful tool to understand the perception of the brand and the feelings that they are evoking in the customers.

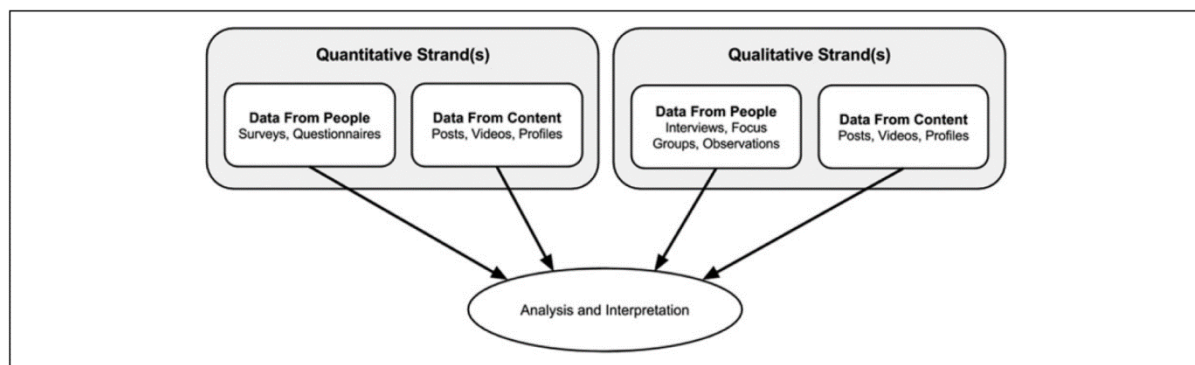


Figure 9: Convert Parallel Design (Snelson, 2016)

The second mixed method is called Exploratory Sequential Design, which is an approach that tackles the qualitative analysis first, and then the quantitative one as it is shown of Figure 10. Therefore, it starts with the content analysis to gather frequencies of appearance of information, intention, sentiment of the post, and so on. Then, these results are analyzed in a quantitative way, so they could be brought together to understand the market.

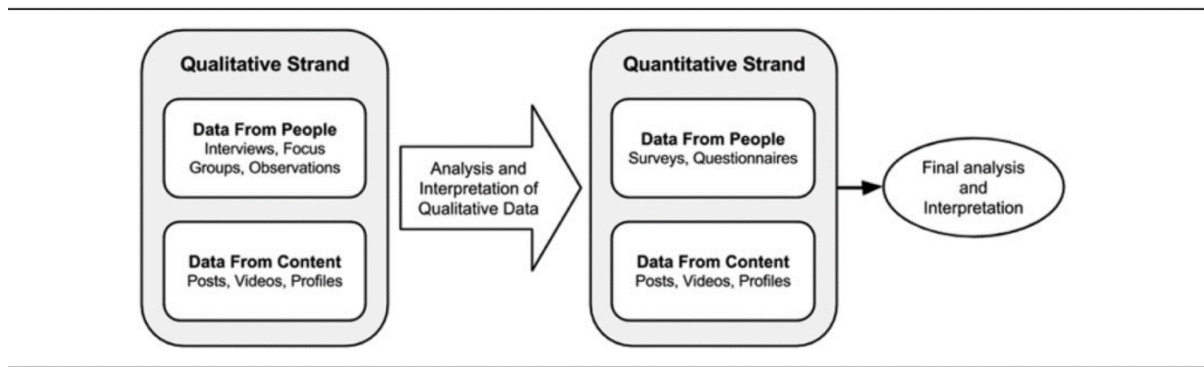


Figure 10: Exploratory Sequential Design (Snelson, 2016)

As it is display in Figure 9 and 10, content analysis is part of both quantitative and qualitative research due to the valuable insights that it offers across the market research. Therefore, sentiment analysis is fundamental tool to develop content analysis that helps to understand the market.

CHALLENGES

During the development of this project, several challenges were faced across each stage of the process. First, the treatment of unbalanced datasets is considered one of the main concerns of any machine learning project due to the nature of the gathered information. It is known that the algorithms could be less effective at predicting examples of a minority class. In consequence, there relies on the importance of balance the dataset, which means that all the labeled results should have the similar presence across the data set. Consequently, a SMOTE (Synthetic Minority Oversampling Technique) algorithm was applied to the training set, which oversampled the minority class (Negative) in the comments' sentiment.

Another challenge was the incorporation and understanding of the Ecuadorian lexicon. Since this algorithm for sentiment analysis will be deployed in Ecuadorian companies, the algorithm should be able to recognize Ecuadorian words and phrases in order to predict the sentiment of the comment correctly. Initially, the construction of the dataset included texts and sentiments from the TASS and MAS corpora. However, these corpora included texts written by users in Peru, Mexico, and Spain. To address the challenge, it was decided to only include the MAS corpus and labeled comments from the Ecuadorian companies of this project. The MAS corpus was included because it was developed for marketing purposes specifically (Navas, 2019). On the other hand, the comments from the Ecuadorian companies were labeled manually by the team. The labelling was made by assigning "Negative" and "Positive" sentiments only to those comments in which those sentiments were obvious.

Finally, the optimization of the hyperparameters of each model was also a challenge. Finding the correct combination of hyperparameters that maximizes the performance of the analyzed models was a time-consuming activity because the algorithm had to analyze all the possible combinations of the hyperparameters' values given for each model. Grid Search and

Random Search are two of the most common methods for hyperparameters tuning. Grid Search evaluates the performance of different combinations of the hyperparameters specified on a grid with a cross validation (Scikit Learn, 2020). In contrast, the hyperparameters evaluated in Random Search are not selected, but they follow a defined probability distribution (Jordan, 2017). As several trials of the model were made, an idea of the best hyperparameters values was already present, so it was necessary to evaluate them between that range of values. Hence, the Grid Search method was selected to tune the hyperparameters of the models.

CONCLUSIONS

Even though the sentiment of each comment can be labelled manually, it is recommended to implement machine learning techniques to exchange man-hours to computational time at executing this activity. Comments were labelled through this project in order to train the models and avoid further manual labelling by the final users of this tool. This is a step forward to achieve process automation, which leads to improve the performance of the processes involved by reducing human errors and the time required to finish an activity. Also, using these techniques, specifically supervised learning approaches, the model can learn from the training data, which means that it could be implemented to analyze the performance of several products from different countries if the team is able to gather comments that show how the costumers express their feelings and perceptions of products in their homeland.

If a supervised learning approach is implemented to perform sentiment analysis, a Neural Networks (NN) model is recognized as one of the best deep learning algorithms to tackle this problem (Medhat et al., 2014). Along this project, it was proved that it has the better performance compared to the SVM algorithm because it was able to classify the data at a high-rate speed with better accuracy. However, this powerful technique of Artificial Intelligence

should be trained correctly in order to display good results. Therefore, the pre-processing techniques are extremely important at preparing the data, so the algorithm could recognize the patterns and label each one of the comments accordingly to the learning rule across the model.

The results of the Neural Network algorithm should be displayed in a practical way, which means that the data visualization techniques implemented should be easy to understand by the decision-makers. These dashboards include a word cloud that shows the frequently used words across all the comments analyzed and they are displayed in different shapes according to its frequency, which means that the bigger words are the ones that appear the most across all the comments. In addition, n-grams bring some insight to understand the topics that were most talked about because they can display n words that are commonly together. Finally, a graph that shows the count of positive, neutral, and negative comments is displayed. It helps to understand how the brand is perceived across its customers and the sentiments that their recall.

RECOMMENDATIONS

Based on what has been done in this project, the following recommendations can be considered for future research on this topic:

The comments generated by users in Facebook and Instagram of Products A and B do not address product-specific attributes directly unless the post generates a path to gather this information. Instead, most of these comments address promotions or topics present in the posts published. However, the remaining comments give an overall perspective of the product and customers' experiences with them. Therefore, it is recommended to aim critical product characteristics or attributes in future posts to obtain more specific comments that lead the company to obtain meaningful insights for product development and improvement. By doing

this, the information can be analyzed around perspectives and sentiments about specific attributes.

Even though the Neural Network model had the best accuracy compared to the SVM model, it can still be improved. As it is visible in Figure 6, there is still confusion between positive and neutral labels predictions. Other preprocessing techniques combinations from the one established by Dubiau (2013) may work better and improve the model's performance, such as negations' treatment and recognition of sarcasm. However, this activity requires further analysis and much more time for research and trial-and-error tests.

Tools such as Facepager not only extract social media comments, but also their corresponding posts, shares, reactions, and so on. The results of the sentiment analysis jointly with these extra data can be used as input for a more complete content analysis. The union of all these data can boost marketing research and help researchers find more useful insights to improve future marketing strategies and the products involved.

During the preprocessing steps applied in the construction of the models, hashtags and the words that followed them were eliminated. The impact of this elimination is unknown. Nevertheless, these hashtags and texts can be separated and stored for further analysis, or they can be processed to keep the words in the text and build the model including them.

BIBLIOGRAPHIC REFERENCES

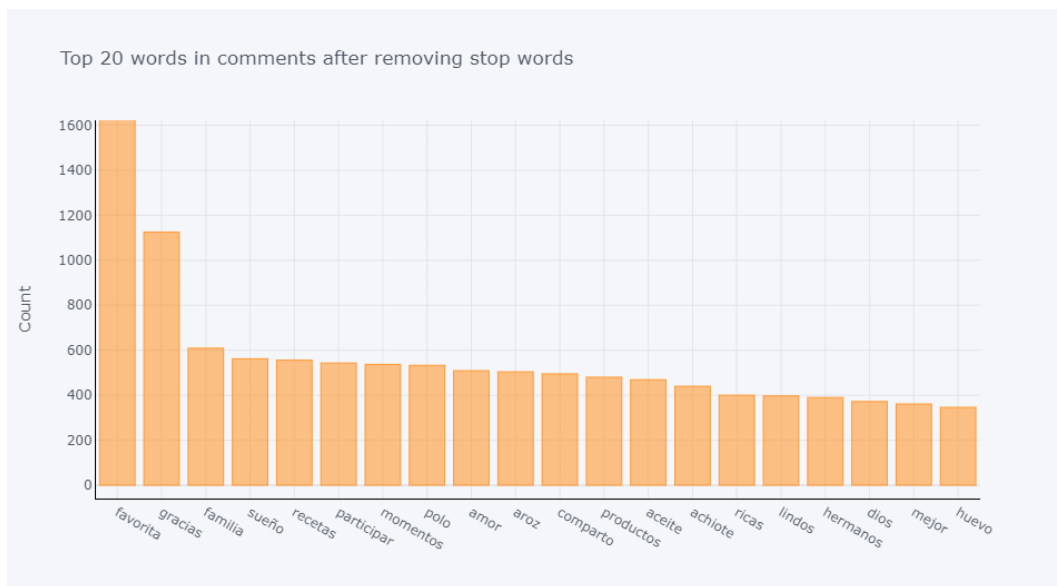
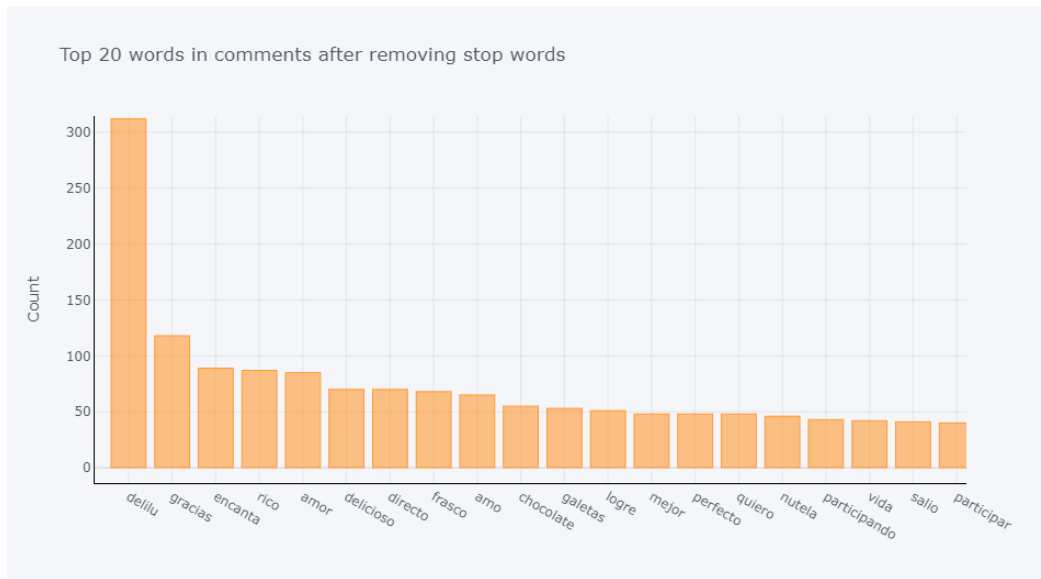
- Abraham, A. (2005). *Artificial Neural Networks*. Handbook of Measuring System Design.
doi:10.1002/0471497398.mm421
- Aggarwal, C., and Zhai, C. (2012). *Mining Text Data*. Springer Science & Business Media.
- Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis.
NursingPlus Open, 2, 8–14. <https://doi.org/10.1016/j.npls.2016.01.001>
- Bergstra, J., and Bengio, Y. (2012). *Random Search for Hyper-Parameter Optimization*.
Journal of Machine Learning Research 13.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press,
Oxford, UK.
- Botero Villada, M. (2019) METODOLOGÍAS PARA EL DESARROLLO DE PROYECTOS
DE MINERÍA DE DATOS, Mario Botero Villada's personal collection.
https://www.scipedia.com/public/Botero_Villada_2019
- Cunningham P., Cord M., Delany S.J. (2008). Supervised Learning. In: Cord M.,
Cunningham P. (eds) Machine Learning Techniques for Multimedia. Cognitive
Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2
- Dubiau, L. (2013). *Procesamiento de Lenguaje Natural en Sistemas de Análisis de
Sentimientos*. Universidad de Buenos Aires, 31-32.
- Espinosa-Zúñiga, Javier Jesús. (2020). Aplicación de metodología CRISP-DM para
segmentación geográfica de una base de datos pública. *Ingeniería, investigación y
tecnología*, 21(1), e00008. Epub 03 de agosto de 2020.
<https://doi.org/10.22201/fi.25940732e.2020.21n1.008>
- ExportComments. (2021). API Documentation. ExportComments.
<https://exportcomments.com/api>

- Garcia-Osorio, C. (2019). Metodologías de Desarrollo de proyectos de minería de datos – Una visión centrada en CRISP-DM. Seminario para el Digital Innovation Hub de Burgos – Industria 4.0. DOI: 10.13140/RG.2.2.34208.02566
- Glorot, X., Bordes, A., and Bengio, Y. (2010). *Deep Sparse Rectifier Neural Networks*. *Journal of Machine Learning Research* 15.
- Gupta, P. (2017). *Cross-Validation in Machine Learning*. Towards Data Science. <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>
- Gürel, E. (2017). *SWOT ANALYSIS: A THEORETICAL REVIEW*. *Journal of International Social Research* 10(51):994-1006. DOI: 10.17719/jisr.2017.1832
- Hu, Y., and Li, W. (2011). *Document sentiment classification by exploring description model of topical terms*. *Computer Speech & Language*, 25(2), 386–403. doi:10.1016/j.csl.2010.07.004
- Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics*, 58(4), 424–434. doi:10.1080/00401706.2015.1093962
- IBM. (2020b). Nodo SMOTE. IBM Documentation. Retrieved from <https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=nodes-smote-node>
- Ientilucci, E. J. (2003). *Using the singular value decomposition*. Rochester Institute of Technology, New York, United States, 1-2.
- Jordan, J. (2018). Hyperparameter tuning for machine learning models. Jeremy Jordan. <https://www.jeremyjordan.me/hyperparameter-tuning/>
- INEC. (2020). Tecnologías de la Información y Comunicación-TIC. Instituto Nacional de Estadística y Censos. <https://www.ecuadorencifras.gob.ec/tecnologias-de-la-informacion-y-comunicacion-tic/>
- Kang, H., Yoo, S. J., and Han, D. (2012). *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews*. *Expert Systems with Applications*, 39(5), 6000–6010. doi:10.1016/j.eswa.2011.11.107
- Kaufmann, M. (2012). *JMaxAlign: A Maximum Entropy Parallel Sentence Alignment Tool*. ACL Anthology, Proceedings of COLING 2012: Demonstration Papers. <https://www.aclweb.org/anthology/C12-3035/>
- La Fabril. (2021). *Quienes somos – La Fabril*. La Fabril. Recovered from <https://www.lafabril.com.ec/quienes-somos/>
- Li, Y.-M., and Li, T.-Y. (2013). *Deriving market intelligence from microblogs*. *Decision Support Systems*, 55(1), 206–217. doi:10.1016/j.dss.2013.01.023
- Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies.

- Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge, England: Cambridge University Press.
- Lopez, I., and Ruiz, S. (2008). *Las respuestas cognitivas y emocionales del consumidor como determinantes de la eficacia del sitio web*. *Revista Española de Investigación de Marketing ESIC* vol. 12, n.º 1 (43-63)
- Medhat, W., Hassan, A., and Korashy, H. (2014). *Sentiment analysis algorithms and applications: A survey*. *Ain Shams Engineering Journal*, 5(4), 1093–1101. doi:10.1016/j.asej.2014.04.011
- Mishra, A. (2018). *Metrics to Evaluate your Machine Learning Algorithm*. Towards Data Science. <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
- Moraes, R., Valiati, J. F., and Gavião Neto, W. P. (2013). *Document-level sentiment classification: An empirical comparison between SVM and ANN*. *Expert Systems with Applications*, 40(2), 621–633. doi:10.1016/j.eswa.2012.07.059
- Navas, M., & Rodriguez, V. (2019). Corpus for Marketing Analysis in Spanish. MAS Corpus. <http://mascorpus.linkeddata.es/>
- Nguyen, Heidi; Veluchamy, Aravind; Diop, Mamadou; and Iqbal, Rashed (2018) "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review: Vol. 1 : No. 4* , Article 7. Available at: <https://scholar.smu.edu/datasciencereview/vol1/iss4/>
- Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. KDnuggets. <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
- Price, R. A., Wrigley, C., & Straker, K. (2015). *Not just what they want, but why they want it*. *Qualitative Market Research: An International Journal*, 18(2), 230–248. doi:10.1108/qmr-03-2014-0024
- Scikit-learn. (2020). *sklearn.feature_extraction.text.CountVectorizer* — *scikit-learn 0.24.1 documentation*. Scikit-Learn Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.feature_extraction.text.CountVectorizer
- Scikit-learn. (2020). *sklearn.svm.SVC* — *scikit-learn 0.24.1 documentation*. Scikit-Learn Documentation. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

- Snelson, C. L. (2016). *Qualitative and Mixed Methods Social Media Research: A Review of the Literature*. International Journal of Qualitative Methods.
<https://doi.org/10.1177/1609406915624574>
- Song, M., Yang, H., Siadat, S. H., and Pechenizkiy, M. (2013). *A comparative study of dimensionality reduction techniques to enhance trace clustering performances*. Expert Systems with Applications, 40(9), 3722–3737. doi:10.1016/j.eswa.2012.12.078
- Tsytsarau, M., and Palpanas, T. (2011). *Survey on mining subjective data on the web*. Data Mining and Knowledge Discovery, 24(3), 478–514. doi:10.1007/s10618-011-0238-6
- Voramontri, Duangruthai & Klieb, Leslie. (2018). Impact of Social Media on Consumer Behaviour. International Journal of Information and Decision Sciences. 11.
10.1504/IJIDS.2019.10014191.
- Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1), 45–55. <https://doi.org/10.1177/016555159201800106>
- Zhi-Hua Zhou. (2018). A brief introduction to weakly supervised learning, *National Science Review*, Volume 5, Issue 1, Pages 44–53, <https://doi.org/10.1093/nsr/nwx106>

ANNEX A: MOST FREQUENT UNIGRAMS FOR PRODUCT A AND PRODUCT B



ANNEX B: MOST FREQUENT BIGRAMS FOR PRODUCT A AND PRODUCT B

