Sentiment Analysis Model on Twitter About Video Streaming Platforms in Mexico

Rosalia Andrade-Gonzalez[1] and Roman Rodriguez-Aguilar[2](✉)

[1] Facultad de Ingenieria, Universidad Anahuac Mexico, Mexico City, Mexico
[2] Facultad de Ciencias Económicas y Empresariales, Universidad Panamericana, Augusto Rodin 498, 03920 Mexico City, Mexico
rrodrigueza@up.edu.mx

Abstract. This work addresses the analysis of the content of the comments on Twitter in the period from December 2020 to February 2021 on the video streaming platforms in Mexico: Netflix, Disney+ and Prime Video. The analysis involves the extraction of comments on Twitter, cleaning the text and the development of a supervised support model for Text Mining for the sentiment classification of tweets in the categories: Positive, Negative or Neutral (spam); as well as the use of resampling techniques to measure the variability of the model's performance and improve the precision of its parameters. The result allows the measurement of user satisfaction levels and the detection of the most dissatisfied and liked aspects of the platforms. Finally, a business intelligence dashboard was developed in Power BI for the interactive visualization of the results under different information filters.

The results show that there is a large percentage of Neutral tweets (spam) that refer mainly to advertising about new releases. Netflix's satisfaction level is the highest compared to the rest of the platforms due to the liking for its original series, variety, and dynamism of launches; on the contrary, the most unpleasant aspect is removing content from your catalog. For its part, Disney+ has satisfaction lower due to the limited variety of its catalog and the expense involved. In the case of Prime Video, lower levels of satisfaction are observed for removing content from its catalog and for paying more than one platform per month. The application of this methodology could benefit in measurement of satisfaction levels, understanding, decision-making and monitoring of new strategies implemented by the platforms.

Keywords: Text mining · Sentiment analysis · Analytical intelligence · Streaming · Satisfaction levels

1 Introduction

Streaming is a technology that allows you to view an audio or video file directly from an Internet page or a mobile application, without completely downloading it to a device. In other words, it is displayed as it is downloaded to a smart TV, computer, tablet, or phone [1].

J. A. Marmolejo-Saucedo et al. (Eds.): COMPSE 2021, LNICST 393, pp. 73–87, 2021.
https://doi.org/10.1007/978-3-030-87495-7_6

The market in Mexico has been led by Netflix since it arrived in the country in 2011 but the launches of new video streaming services continue, with more and more platforms with different business models being added to the market, such as the launch of Disney+ at the end of 2020. Until June 2020, Netflix covered 50% of the users of the video platforms followed by Claro Video and Amazon Prime [2]. However, for September of the same year, Netflix's share fell to 36%, as did Claro Video's share, moving to third position due to the increase in Prime Video's share with 22% of the market. In addition to this, the launch of the new Disney+ platform, at the end of the same year, increases competition.

According to Patrick O'Neill, co-founder of Sherlock Communications "The concern is not about the quality of the programs, but rather with so many options and new competitors preparing to land in Latin America, the ocean of available content could generate 'fatigue of decision 'on users. However, aware of this challenge, the platforms are now looking to improve their recommendation algorithms, all based on big data". The most used video streaming platforms in the Mexican Republic and which will be analyzed in this paper are Netflix and Prime Video as well as the new platform Disney+ due to its recent launch.

Netflix is the platform with the highest price but in variety and quantity, it is the one with the largest offer. Its offer is aimed at all types of public and even with the arrival of new competitors, it continues to be the main platform. Much of its success lies in its commitment to its own content and dynamism in its launches. One of the closest competitors for Netflix is Prime Video, thanks to the large number of people who have already established a relationship with the online sales giant Amazon, and that unlike other platform, it has decided to compete with its rivals offering subscriptions at low prices to attract new customers. The platform provides access to two types of content, the one included in the membership and the one that can be rented or purchased at affordable prices, it also allows you to watch the first episode of selected series for free. In addition, subscribers have access to an Amazon Music account, priority in order delivery from the online store, among others. However, its interface can be a bit more complex and varies in style from device to device. While this platform has 4,295 titles, these tend to be older and while it has been investing in its own original content and in the long term it is probably the right strategy, it is currently not enough to displace Netflix [3].

On the other hand, Disney+ is a platform focused on family entertainment, aimed mainly at children's audiences with limited adult content, which is one of the most common criticisms that the platform receives in addition to the fact that the pace of its launches is slow and has the fewest titles in its catalog. As of September 2019, Netflix had 6,783,000 active accounts and Prime Video served 465,000 customers [4].

Twitter is a platform that has become a news and trends portal with 240 million users worldwide. It is a social network focused purely on informing and in many countries, it is used mainly to be informed and aware of the latest news or official announcements made by different government entities, political figures, and opinion leaders in the world. In general, the age range that dominates this social network ranges from 18 years to 34 years, which covers 53% [5]. In 2020, Twitter was the fifth most used platform in Mexico with 57% participation, the country ranks number 10 in the top reach of Twitter worldwide and is one of the 5 countries that exceeds the average reach with 10% in January 2020 [6].

The number of monthly active users of Twitter exceeded 9.5 million, which represents an increase of almost 24% compared to that registered in the same month of the previous year. It has also been identified that most of the user's access through devices with Android systems with 6.7 million users (70%), although the proportion of users of iOS systems has increased since April 2020 [7]. Twitter began using the hashtag as a method of indexing keywords that would help facilitate good search results to see the top posts surrounding a specific hashtag and participate in the latest trends. The concept was first used in 2007 and since then most social networks have taken advantage of hashtags for the same purpose.

Hashtags that spread quickly and are used by a wide variety of users become trending. This means that a keyword is popular and is being used by many people online. They are an effective way to increase interactions and build a brand when they are using promotional material, when announcing new launches or to generate interest towards the business, they are also helpful in finding a target audience [8]. One way to connect with other users on Twitter is to use mentions, these are dealt with by putting the at symbol (@) in front of the name of a particular user within the content of the tweet. With mentions, users can quote people related to their publication and attract their attention [9].

Text Mining is an interdisciplinary field that involves the modeling of unstructured data to extract useful and high-quality information or knowledge from text by creating patterns and trends taking advantage of numerous statistical techniques of machine learning and computational linguistics [10].

Sentiment analysis refers to the different methods of computational linguistics that help identify and extract subjective information from existing content in the digital world (social networks, forums, websites, etc.). It extracts a tangible and direct value, such as determining whether a text taken from the Internet contains positive or negative connotations [11].

In studies about sentiment analysis, Machine Learning algorithms such as Naive Bayes, Linear Regression, Support Vector Machines and Neural Networks are commonly used. Most of the research and dictionaries are developed in the English language, so there is a need to develop ad hoc tools for the case of each language, also considering the idioms of each region and the expressions used in social networks.

There are studies on the analysis of sentiment of comments on Twitter in the Spanish language applied to other topics, such as Customer Voice Analysis [12]. However, there is no research on sentiment analysis in Twitter comments about streaming platforms in Mexico.

In recent years, the number of platforms that offer new video streaming services in Mexico with different business models has increased due to the profitability of this technology and the great acceptance by users as it is a practical and economical option. Increased competition implies a new challenge for platforms with a significant positioning in the current market such as Netflix or Prime Video, and on the other hand it implies a significant challenge for new platforms seeking to position themselves in the market such as Disney+. For this, it is important to consider the perception of current and potential clients of the platforms through social networks and measure satisfaction levels by identifying the aspects of greatest liking, dissatisfaction or even failures of each platform in the face of the demanding demand for content by users and the strategies of their competitors. The objective of this work is to measure satisfaction levels through customer sentiment towards the main video streaming platforms, using a sentiment classification model for posts on Twitter in Mexico. The work is structured as follows, in section two the materials and methods are described, section three presents the main results and finally the conclusions and references.

2 Materials and Methods

2.1 Description of the Data

For the compilation of the information and construction of the database, Twitter Archiver was used, which is a tool that searches for tweets by means of a keyword or hashtag, saving the matching tweets in a Google spreadsheet automatically. So that every hour it searches and extracts all the new tweets accumulating them in the spreadsheet [13]. The tool was used to compile the databases of the tweets from each of the platforms separately in the period from December 2020 to February 2021. Later they were downloaded and by programming in R, they were consolidated into a single database with 436,110 records in total.

In general, texts can be grammatically complex and in the case of comments on social networks, such as those on Twitter, it can find many empty words, punctuation marks, spelling mistakes and words with a lot or little frequency that do not contribute to decipher the actual content of the comment. For this reason, any data analysis process begins with a preliminary step that includes pre-processing, cleaning, and exploratory analysis. To get the meaning of a text, a measure is needed so quantitative data is first extracted by processing the text with various transformation methods. Although there is no single methodology to perform text preprocessing and cleaning, each technique seeks to discard unnecessary information and there are various methods, packages, APIs, and software that can transform text into quantitative data. In the case of R, there are packages that clean the text mainly in English, so in this work we will develop a process of cleaning the text itself, focused on the data with which it is working and with the flexibility of be able to adapt it to other texts.

For the purposes of this work, a V-Corpus will be used, with the purpose of converting each of the tweets into a different document and that the entire collection is temporarily stored in memory. Once the separation has been carried out in different documents, the Tokenization method will be used, which is the process of separating sequence of characters or a defined document unit, into phrases, words, symbols, or other useful elements called tokens. For the quantitative analysis of the text, they are considered as a collection of words or bag of words and the key words, frequencies of occurrences and the importance of each word in the text are extracted.

The next step is to put together and assume that all tweets are a collection of unique words (bag of words) where frequency and order are irrelevant. It will seek to standardize all words by converting them to lowercase and removing accents, punctuation, special characters, and stop-words. On the other hand, the grammar in any language allows the use of derivationally related words with similar meanings, which are nothing more than a different form of the same word. Therefore, to reduce the inflectional forms and the derived words to the common or infinitive base form a stemming was applied. To find the similarity between words, it is necessary to evaluate how similar they are through distance measure. One way to find the similarity between two words is through "edit distance", which refers to the number of operations required to transform a string of characters, in this case one word, into another, such as the Levenshtein distance. Once the words have been reduced to their inflectional form, another way to unify the words is through the n-grams. An n-gram is a group of n words that are written together frequently.

The initial database consists of 436,110 records of tweets in Spanish language on video streaming platforms: Netflix, Prime Video and Disney+, in the period from December 20, 2020, to February 27, 2021. It has 21 fields provided by the Twitter Archiver tool plus 4 variables calculated from these (Table 1).

Table 1. Description of variables

Name	Description
Date	Tweet publication date
Screen Name	Name of the user who posted the tweet
Full Name	Full name of the user who posted the tweet
Tweet Text	Tweet text
Tweet ID	Tweet id
Link(s)	Link related to the tweet
Media	Link to the image contained in the tweet
Location	Location of the user who posted the tweet
Retweets	Number of retweets from other users to the published tweet
Favorites	Number of users who added the published tweet to their favorites
App	Application used to post the tweet
Followers	Number of user followers
Follows	Number of users the user follows
Listed	Number of user lists
Verified	Indicates if the account is verified or not
User Since	User creation date on Twitter
Location	User's location
Bio	Biography of the user in his profile
Website	User website
Time zone	Time zone
Profile Image	User profile photo link

(*continued*)

Table 1. (*continued*)

Name	Description
Years of antiquity	Years of seniority of the user on Twitter
Platform	Streaming platform mentioned in the tweet
Frequency	Number of times the same tweet was published in the analysis period
Unique users' frequency	Number of different users who published the same tweet in the analysis period

2.2 Structure of the Proposed Model

The structure of the process for measuring customer satisfaction levels on video streaming platforms in Mexico is based on the results of the sentiment prediction model of posts on Twitter, consists of collecting and downloading tweets about the platforms in a certain period with Google's Tweet Archiver tool. The databases will be read and consolidated with the R program, which will also load all the dictionaries required for cleaning text and creating predictor variables. The R program will carry out the text cleaning process, the creation of the variables required for the model and will apply the model to predict the sentiment of the tweets. The construction of the sentiment model started from the manual classification of each twit according to the content in three levels: positive, negative, and neutral sentiment. To select the sentiment classification model to be used, a comparison was made of a set of generally accepted methodologies in machine learning:

- Support vector machines
- K-nearest neighbors
- Decision trees
- Neural networks
- Boosting
- Random forest

The base with the variables provided by the Tweet Archiver tool, those created for the model and the sentiment prediction, will be read by the Power BI tool that generates an interactive visualization board in which the results can be visualized by performing different information filters that will allow to monitor the results over time (Fig. 1).

It is important to highlight that for the definition of the dictionaries, custom dictionaries based on the Spanish language were used. The base with the variables provided by the Tweet Archiver tool, those created for the model and the sentiment prediction, will be read by the Power BI tool that generates an interactive visualization board in which the results can be visualized by performing different information filters (Fig. 1).

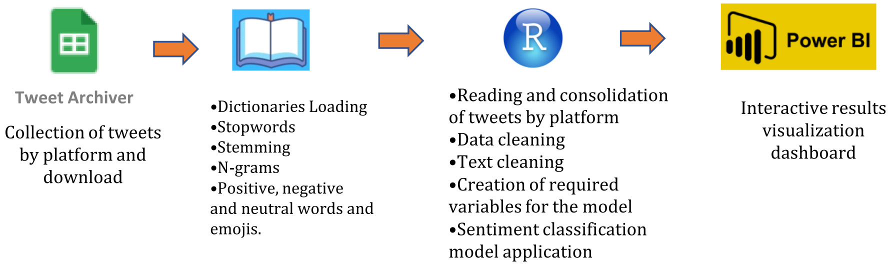

Tweet Archiver
Collection of tweets by platform and download
•Dictionaries Loading
•Stopwords
•Stemming
•N-grams
•Positive, negative and neutral words and emojis.
•Reading and consolidation of tweets by platform
•Data cleaning
•Text cleaning
•Creation of required variables for the model
•Sentiment classification model application
Power BI
Interactive results visualization dashboard

Fig. 1. Proposed model.

3 Results

3.1 Text Cleaning and Debugging

After the exploratory analysis for a better understanding of the information and cleaning of both variables and records, there is a final base with 35,173 records to continue cleaning the text of the tweets. The text cleaning process consisted of the extraction of hashtags and mentions, text standardization, emoji translation, definition of Stop words, Stemming, definition of Bigrams and Trigrams and selection by criteria of Frequency of words. Later all the text cleaning actions, only 17% of the initial number of unique words (15, 068) is preserved, which are only the most relevant words that will help to build the sentiment model. To identify if users mention more than one platform in the same tweet to make a comparison or promotion between them, 3 indicator variables were created, one for each platform, which indicate whether the pattern sought corresponding to each platform. It is found within the text of the tweet using regular expressions. With this, it was possible to identify that 97% only refer to 1 platform while 3% mention 2 to 3 platforms in the same tweet. A new variable was created with the percentage of characters removed per tweet with the text cleanup calculated as Eq. 1.

$$\left(\frac{\text{Length of original tweet} - \text{Length of clean tweet}}{\text{Length of original tweet}}\right) * 100 \tag{1}$$

That indicates the percentage of irrelevant and discarded words from the original tweet, which can be an important predictor of sentiment. When analyzing the distribution of the variable, it was found that on average 51% of the characters of each tweet were eliminated with the cleaning actions (Fig. 2).

To analyze the content of the tweets, word clouds were created for each platform and variable: Tweet text, Hashtags, Mentions (@) and Translation of emojis to text. The most common terms in the tweets about Netflix are about its series, documentaries, and new releases; while in publications about Disney+ and Prime video, the terms related to its catalog of films and premieres are more frequent (Fig. 3).

In the case of emojis, it is a variable that can provide relevant information for the sentiment classification, so they were transformed into text and will be used as an additional explanatory variable for the sentiment classification model.

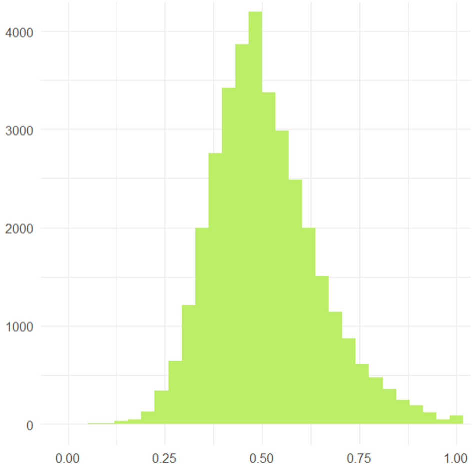


Fig. 2. Distribution of the percentage of character removal by cleaning.

3.2 Sentiment Classification Model

A correlation analysis was performed between the base variables with the 35,173 records, to identify the correlated variables, that is, those that contain the same information to select independent variables for model training. A manual classification of sentiment was performed: Positive, Negative or Neutral, of a random sample of 30% (10,487 tweets) of the total base tweets with the aim of training the model. With the results of the classification, a preliminary analysis was carried out where it was found that 57% of the tweets have Neutral sentiment (spam), 22% Negative and 20% Positive.

By platform, Netflix has the same proportion of Positive and Negative tweets (21%) while Disney+ has a higher proportion of Negative comments (24%) compared to Positive ones (19%) and Prime Video has a higher proportion of Neutral tweets (spam) (62%) and an almost equal proportion between Positive and Negative tweets (18–19%) (Fig. 4).

Word clouds were made by platform and sentiment to identify the main words of each one. As a result of the word clouds, it is observed that the main reasons for Negative tweets about Netflix are due to removing movies/series from its catalog, not having them, being bad or due to the payment of the platform without seeing its content or the expense that caused by being the most expensive. On the other hand, Positive tweets refer to the fact that they like the content and recommend it for being good options, they also express appreciation for documentaries, seasons, and award nominations. For its part, Disney+ has a greater variety of Negative words, since users express various reasons, such as canceling the subscription for not wanting to pay for a new platform in addition to the existing offer, limited/bad catalog, or dissatisfaction with the fall of the platform due to saturation due to the weekly premiere (such as new chapter of "WandaVision"). However, the Positive comments refer to the fact that the platform has the best/favorite films of the users that provoke positive emotions and childhood memories or show liking for new film success stories (such as "The Mandalorian"). In the case of the Prime Video platform, users publish Negative tweets mainly because of movies they remove from the

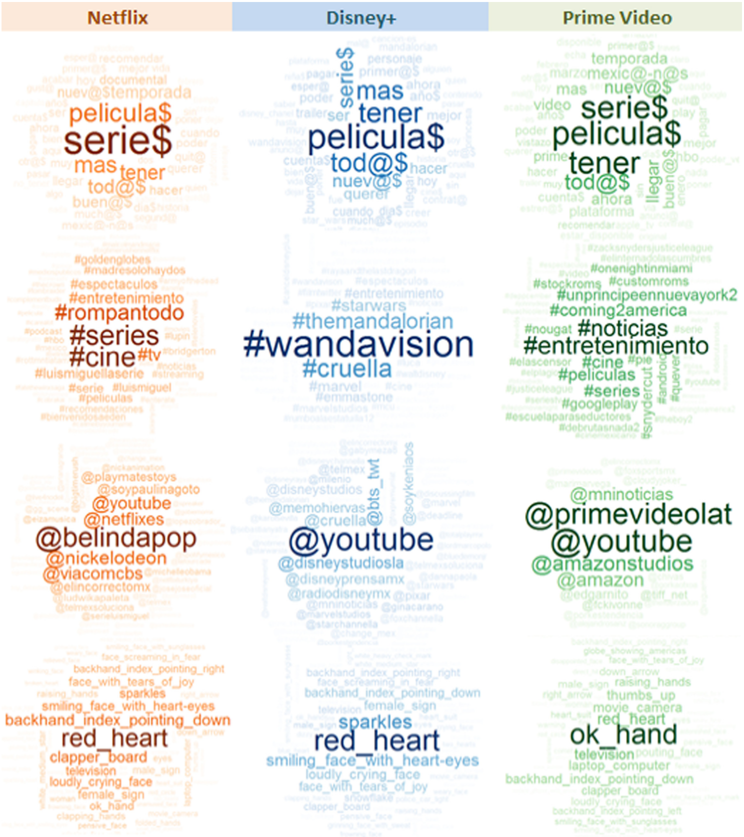


Fig. 3. Word clouds by platform and text of the tweet, hashtags, mentions and emojis.

catalog or because they are paying for the platform without occupying it, having bad content or because of the ads that are shown. On the other hand, Positive tweets refer to the fact that the platform contains good options for movies that users consider to be the best (Fig. 5).

From the sample of the base with the manual classification of the sentiment of the tweets, the main words that influence the classification of the sentiment were identified as a "catalog" of words by sentiment. As a criterion for the selection of the main words of each feeling, those with the highest frequency were considered from a certain percentile. Similarly, a "catalog" of Positive, Negative and Neutral (spam) emojis was established based on the frequency of the emojis to assign them in a sentiment category.

Finally, there is a base with 23 variables for calculating the optimal percentile to establish the "catalogs" of feeling words with which to obtain greater effectiveness in training the sentiment classification model. To determine the optimal percentile to establish the "catalog" of each sentiment, a cross-validation process was carried out with all possible combinations of percentiles in multiples of 25%, in which the classification models were tested: Support Vector Machines, KNN, Naive Bayes, Decision Tree and

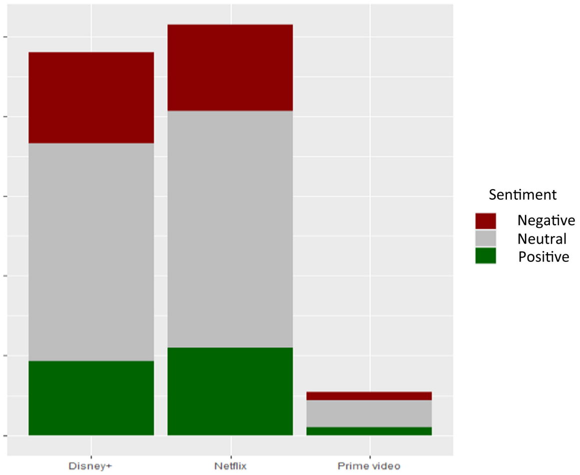


Fig. 4. Sentiment ratio per platform

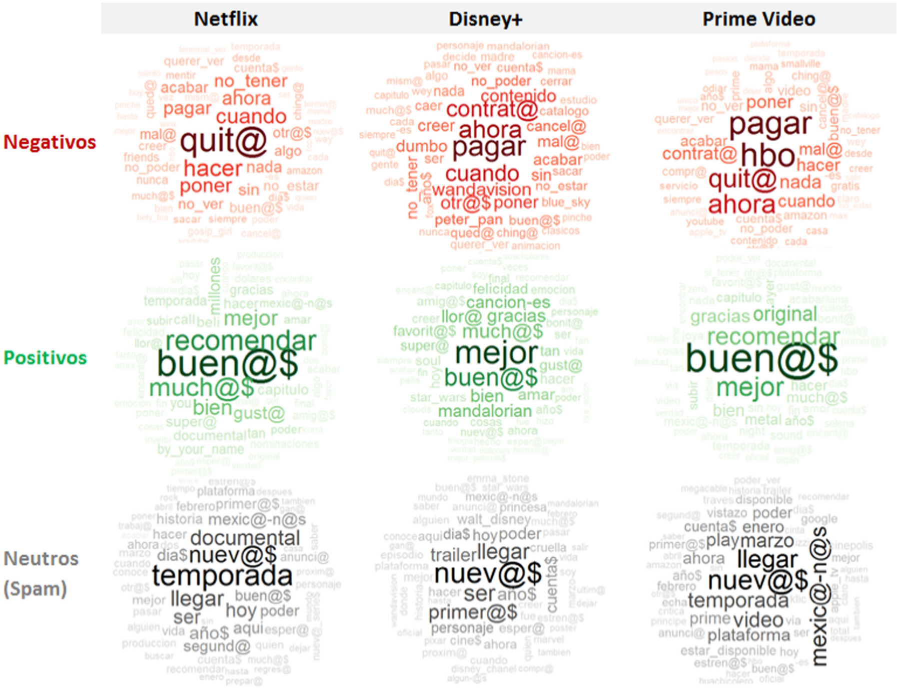


Fig. 5. Word clouds by platform and sentiment of the base sample with manual classification for model training.

Neural Network, with the default parameters and considering a base segmentation of 50% for training and 50% for validation. To measure the effectiveness of the models in each combination, the misclassification rate was calculated to choose the combination of percentiles per sentiment that minimizes the error in most of the models and in turn maximizes the percentiles so that the "catalogs" contain the fewest possible words.

With the result of the cross-validation, the optimal 50th percentile was determined for each sentiment and the three "catalogs" were formed, consisting of 1,525 Positive words, 1,035 Negative words and 3,548 Neutral words. The definition of the optimal percentile to generate the "catalogs" of words by sentiment and form the table with the final variables of best effectiveness in all the models with the default parameters, the optimization of the parameters will be sought, minimizing the misclassification rate by validation with 50% of the base as training and 50% for validation.

The misclassification rate of each model was calculated under different seeds and percentages of the size of the training and validation base to measure the variability of each model under different scenarios and to determine the average effectiveness. As a result, the models with the lowest average rate were Boosting (18.0%) and Polynomial Support Vector Machine (18.4%). However, when applying the model to the validation data with the different percentages of sample and seeds, it was obtained that the best model is the Linear Support Vector Machine with an average misclassification rate of 22.7% (Table 2).

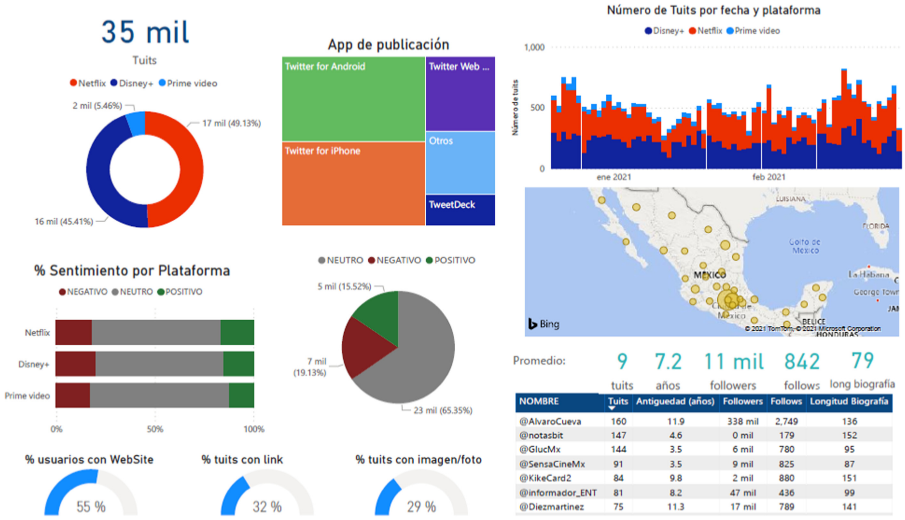
Table 2. Misclassified rate (%) of validation by percentage of sample and seed

		Validation Sample							
%		SVM Lineal	SVM Radial	SVM Polynomial	KNN	NN	Pruned tree	Random Forest	Boosting
50%	Seed 1	23.0	24.3	26.3	27.0	23.6	28.6	22.8	24.0
	Seed 2	23.0	24.3	26.3	27.1	24.3	28.6	23.2	23.4
	Seed 3	23.0	24.3	26.3	26.9	24.3	28.6	23.1	23.6
40%	Seed 1	23.6	24.4	26.6	27.7	23.3	28.1	23.5	23.9
	Seed 2	23.6	24.4	26.6	27.4	23.9	28.1	23.5	23.6
	Seed 3	23.6	24.4	26.6	27.5	23.8	28.1	23.7	23.9
30%	Seed 1	22.8	23.0	25.2	26.2	22.8	27.3	23.4	24.6
	Seed 2	22.8	23.0	25.2	26.2	22.7	27.3	23.2	24.0
	Seed 3	22.8	23.0	25.2	26.1	22.8	27.3	23.5	24.4
20%	Seed 1	22.3	23.4	24.4	25.8	22.8	27.3	23.1	21.7
	Seed 2	22.3	23.4	24.4	25.8	22.0	27.3	23.4	22.4
	Seed 3	22.3	23.4	24.4	25.7	22.6	27.3	23.5	22.5
10%	Seed 1	22.0	23.0	25.1	28.7	23.5	26.5	23.4	22.5
	Seed 2	22.0	23.0	25.1	28.6	23.1	26.5	23.8	22.6
	Seed 3	22.0	23.0	25.1	28.6	23.0	26.5	23.5	23.0
Average		22.7	23.6	25.5	27.0	23.2	27.6	23.4	23.2

The most effective model to model the sentiment of tweets is the Linear Support Vector Machine with a cost of 0.01, which has a 77% average effectiveness, which implies a misclassification error of 23%. When applying the model to the total base of tweets, it was determined that in general, 65% of tweets are Neutral (spam), 19% have Negative sentiment and 16% Positive. Analyzing the proportion by platform it was found that there is a greater proportion of Negative comments compared to Positive in the case of Disney+ with 20% and 15%, respectively. On the other hand, Netflix shows a similar proportion of Positive and Negative (17% and 18%). Compared to the other platforms, Prime video has a higher proportion of Neutral tweets (spam) with 71% and presents a higher percentage of Negative tweets (17%) compared to Positive ones (12%).

3.3 Interactive Display Board

An interactive visualization dashboard was developed with the Microsoft business analysis tool Power BI, to facilitate the visualization of the results, perform ad hoc filters and monitor the content of tweets on video streaming platforms through weather. The board consists in two sections. General section contains the total number of tweets and graphs with the proportion of tweets by platform, sentiment, and APP as well as the trend of the number of daily tweets by platform and a map of the Republic of Mexico that shows the number of tweets by state. The percentage of users with the website and the percentage of tweets with links and images or photos in the publication is displayed. It also shows the average number of tweets, antiquity, followers, follows and length of biography per user and the table with the details of each user (Fig. 6).

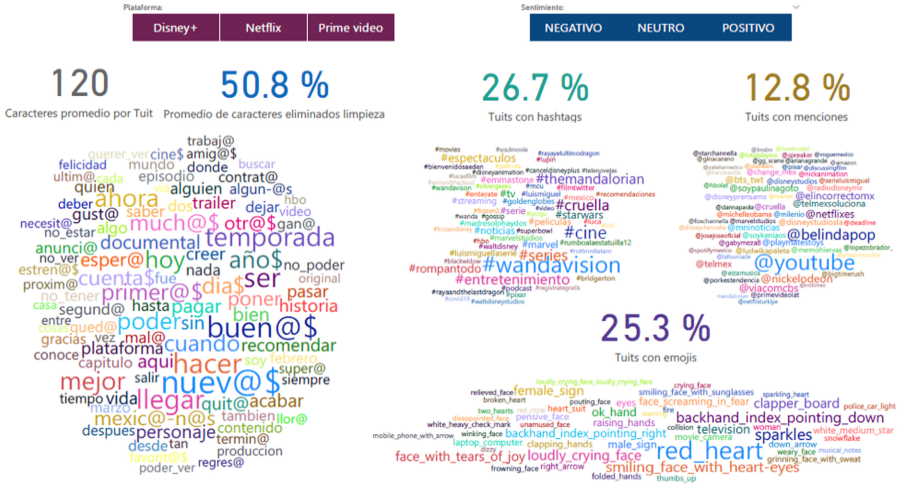


Promedio: 9 tuits 7.2 años 11 mil followers 842 follows 79 long biografía

NOMBRE	Tuits	Antiguedad (años)	Followers	Follows	Longitud Biografía
@AlvaroCueva	160	11.9	338 mil	2,749	136
@notasbit	147	4.6	0 mil	179	152
@GlucMx	144	3.5	6 mil	780	95
@SensaCineMx	91	3.5	9 mil	825	87
@KikeCard2	84	9.8	2 mil	880	151
@informador_ENT	81	8.2	47 mil	436	99
@Diezmartinez	75	11.3	17 mil	789	141

Fig. 6. General dashboard power BI.

The word cloud section contains the word clouds of tweets, hashtags, at and emojis with the option to filter them by platform and sentiment. It will also display the average number of characters of the tweets originally and the percentage that is removed after the cleaning process; as well as the percentages of tweets that contain at least one hashtag, mentions and emojis (Fig. 7).



Plataforma:

Disney+ | Netflix | Prime video

Sentimiento:

NEGATIVO | NEUTRO | POSITIVO

120 Caracteres promedio por Tuit

50.8 % Promedio de caracteres eliminados limpieza

26.7 % Tuits con hashtags

12.8 % Tuits con menciones

25.3 % Tuits con emojis

Fig. 7. Word cloud dashboard power BI.

4 Conclusions

The Text Mining application focused on the analysis of sentiment in social networks on streaming video platforms is a new proposal that helps to extract useful and easy to understand information about what users are expressing in relation to the service and content offered with the objective to help decision-making by analyzing the impact of increasing competition in the sector. In the development process, text cleaning is the most complex part due to the absence of catalogs in different languages or idioms used by different regions and in social networks, which implies developing an ad hoc solution with the analyzed context. In the present work, Spanish-language dictionaries applicable to texts related to video streaming platforms were developed because of a process designed for the pre-processing and cleaning of texts for periodic updating or applicable to other types of texts in the Spanish language.

Netflix has led the video streaming platform market in Mexico for being the first since its arrival in 2011 with the largest number of users, variety, and original content; However, with the increase in competition, its percentage of market share has decreased, and it has diversified among the rest of the platforms in recent years; one of the main reasons is because the platform has the highest price in the market. Its main competition is Prime Video, a platform that has penetrated the market due to its relationship with Amazon and its attractive price; However, the platform does not register a significant number of tweets for sentiment analysis and most of the posts are advertising. On the other hand, the launch of the new Disney+ platform has registered a high volume of publications like Netflix and its main attraction is its children's content and the launch of new film success stories.

To clean the text of the tweets, dictionaries of stop words, stemming, n-grams were elaborated, and various techniques were applied to reduce the number of irrelevant words in the tweets for the sentiment analysis, keeping on average only half of the words in every tweet. A manual classification of a sample of tweets was also carried out in each sentiment to develop catalogs of words and emojis per sentiment for the training of the model based on a cross-validation process that minimized the classification error in all the models.

It was determined that the best sentiment classification model is a Linear Support Vector Machine with an average classification error of 22.7%. Applying the model, 65% of the tweets have Neutral sentiment (spam) and most are published by official accounts of users with a greater number of publications, followers, character length in their biography, hashtags within the tweets and a greater proportion of words removed with the text cleaning process.

Netflix's satisfaction levels are the best as it has an equitable ratio between Positive and Negative tweets, this due to its liking for its content, mainly from original series; on the contrary, the aspect of greatest discontent is removing content from its catalog. For its part, the entry of Disney+ to the Mexican market was well received; However, the level of satisfaction is lower due to the limited variety of its catalog and the expense involved, considering that users already have some other platform. In the case of Prime Video, lower levels of satisfaction are observed for removing content from its catalog and for paying more than one platform per month. All the results can be observed interactively through a visualization board (dashboard) developed in Power BI in which ad hoc filters, analysis can be carried out, and the content of tweets on video streaming platforms can be tracked through time.

References

1. Selectra. Streaming: qué es, cómo funciona, precios y periodo de prueba 2020 (2020). Recuperado de https://selectra.mx/streaming#que-es-el-streaming
2. EL CEO. Netflix perderá concentración de mercado en México por más plataformas de streaming (2020). Recuperado de https://elceo.com/tecnologia/netflix-perdera-concentracion-de-mercado-en-mexico-por-mas-plataformas-de-streaming/
3. Sin embargo, Dávila, A.L.: El streaming convierte 2020 en su mejor año en la historia, y prepara grandes sorpresas para 2021 (2020). Recuperado de https://www.sinembargo.mx/24-12-2020/3913008
4. El Universal, Lucas, N.: México cerró el 2019 con más de 10 millones de cuentas OTT pagadas (2019). Recuperado de https://www.eleconomista.com.mx/empresas/Mexico-cerro-el-2019-con-mas-de-10-millones-de-cuentas-OTT-pagadas-20200122-0052.html
5. Yi MiN Shum Xie. Resumen de Twitter 2020 (2020). Recuperado de https://yiminshum.com/twitter-digital-2020/
6. Statista. Redes sociales con el mayor porcentaje de usuarios en México en 2020 (2020). Recuperado de https://es.statista.com/estadisticas/1035031/mexico-porcentaje-de-usuarios-por-red-social/

7. Statista. Número de usuarios activos mensuales (MAU) de Twitter en México de enero de 2019 a agosto de 2020, por sistema operative (2020). Recuperado de https://es.statista.com/estadisticas/1172236/numero-de-usuarios-activos-mensuales-twitter-mexico-sistema-operativo/
8. Hootsuite, Adame, A.: Cómo utilizar hashtags: una guía rápida y sencilla para cada red social (2019). Recuperado de https://blog.hootsuite.com/es/hashtags-la-guia-completa/
9. Postcron, Skaf, E.: Cómo usar Twitter: 15 tips indispensables (2019). Recuperado de https://postcron.com/es/blog/como-usar-twitter/#:~:text=Para%20esto%2C%20basta%20con%20anteponer,asegurarte%20de%20que%20la%20vean
10. Kumar, A., Paul, A.: Mastering Text Mining with R. Packt Publishing, Birmingham (2016)
11. ITELLIGENT. Análisis de sentimiento, ¿qué es, cómo funciona y para qué sirve? (2017). Recuperado de https://itelligent.es/es/analisis-de-sentimiento/
12. Gonzalez, R.A., Rodriguez-Aguilar, R., Marmolejo-Saucedo, J.A.: Text mining and statistical learning for the analysis of the voice of the customer. In: Hemanth, D.J., Kose, U. (eds.) ICAIAME 2019. LNDECT, vol. 43, pp. 191–199. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-36178-5_16
13. Digital inspiration. Save Tuits in Google Sheets (2015). Recuperado de https://digitalinspiration.com/product/twitter-archiver
14. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Stanford (2008)
15. del Carmen, V.P.M., Covarrubias, C.C.: Identificación del color en video en tiempo real solución estadística a un problema computacional. Tesis Universidad Anáhuac México, México (2017)
16. Decision Tree Modeling. Course Notes. Ed. SAS Education ISBN 978-1-59994-280-3
17. Neural Network Modeling. Course Notes, Ed. SAS Education ISBN 978-1-59047-771-7