A Comparative Study of NLP and Machine Learning Techniques for Sentiment Analysis and Topic Modeling on Amazon Reviews

Gina V. Acosta Gutiérrez
Esan University, Lima, Perú
14100966@ue.edu.pe

***Abstract*— Nowadays, due to the growth on the offer of online products, a large amount of information is available on internet can be found, especially from the reviews written by users who have purchased products on online platforms. It is possible to analyze these reviews in order to extract useful information about the user's opinion about a product. The present paper consists of the development of a model based on machine learning techniques in order to identify trends patterns of online products based on technical data, which in the present case would be the products that are available in the electronic commerce platform Amazon. Several authors have highlighted the importance of analyzing comments, reviews or any type of feedback obtained from users for sellers or a company that offers its products in online trading platforms. This paper consists of creating a model that uses various machine learning techniques in order to generate useful information for salespeople. In this way, the machine learning model will be able to identify the opinion of a user based on its review and determine if this user qualified the product in a positive or negative way.**

***Keywords*—** Machine Learning, Text Mining, NLP, Sentiment Analysis, Online Reviews

I. INTRODUCTION

The main focus of this paper is to identify trend patterns of online products in the e-commerce sector, specifically in the products of the Electronics section of the Amazon platform. The construction of a trend pattern model that uses machine learning techniques for the summary and analysis of a large volume of reviews will be carried out. Its implementation to summarize and analyze reviews will represent an important contribution to the online sales sector, since the results on the opinion and user experience of the products can be useful for potential buyers and sellers.

Currently, it is important that every company that offers products has a presence in the virtual world, from owning a web page to sell their products, to publishing their products on an e-commerce web page. In this way, it offers its consumers a new way of acquiring their products and expanding their audience since on the Internet this is more easily. According to an ONTSI report [1], the websites that offer their products mainly on the internet are the main purchase channel (69.2%), followed by the websites of brands with physical stores (43%). Regarding the consumer, [2] the main difference between one who goes to a physical store and one who chooses to enter an online store, resides in the quantity and quality of information that the customer accumulates before buying, it should be noted that this is crucial in the case of the online consumer [3].

Both positive and negative reviews are of vital importance and influence the company and the products they offer [4]. In the event that the review is positive, it will help other customers interested in this type of products finally decide to buy from this company, so these comments serve as a kind of showcase to potential consumers. On the other hand, if these comments are negative, they are helpful to improve the products and to identify deficiencies or new needs that are demanded and can be covered by the products of this company.

In this work, all reviews from 4 products from the Electronics department on the Amazon platform were used. Therefore, in order to classify and analyze this large amount of text, contained in the reviews [5], it is necessary to apply different methods of Text Mining [6], as well as Sentiment Analysis [7] and other Machine Learning techniques. These are text mining [8] methods and are used to carry out text analysis. Text mining [9] is the process of deducing high quality patterns, trends and information from a given data set. Hence, this work will use this technique to obtain relevant information in order to identify the desired characteristics of the authors of the product reviews. Text mining [10] can be used for the development of various tasks, such as theme modeling, document categorization, document grouping and text summaries.

Another technique that will be used in this research is Sentiment Analysis, [11] [12] which is the field of study that analyzes people's opinions, feelings, evaluations, assessments, attitudes and emotions towards entities as products, services, organizations, events, themes and their attributes. Regarding on the use of Sentiment Analysis, if a company captures information about the feelings of customers, this analysis can provide insight into the attitudes and opinions they have about the company. In the same way, it is possible to know the general feeling of a customer regarding a product, as well as if their vision is positive or negative. Obviously, these feelings regarding a company or a certain product can provide very valuable information whether a user have bought the product or not, what would be the appropriate actions to take in relation to the website to try to get the customer to buy the product. In addition, [13] another of the important applications of text data is the recognition of patterns, since it allows the capture and ordering of complaints, claims, suggestions and comments made by customers of a company, and this information can be used to identify problems faster and more flexibly. Therefore, it can be affirmed that the application of the Sentiment Analysis technique [14] will allow us to identify how the review author feels about his purchase and the product purchased (if he is satisfied with his purchase or if the product was not what I expected, since complaints can be identified) from the analysis of a review.

Here, the contributions of this paper are describe as follow:

- The main contribution of this paper is to provide a comparison between the different techniques of NLP and machine learning in order to determine which one offers better results for feature extraction of the Amazon online reviews.
- Moreover, Topic Modeling was applied on both positive and negative reviews to evaluate and analyse the results of this technique.
- Since these experiments are constructed on a real-world dataset, which was crawled from electronic devices available on Amazon; so that the experiment result shows actual reference significance.

The remaining part of this paper is organized as follows. Section II contains the related work regarding online reviews analysis. Section III describes the methodology used by the author to collect the data, construct the dataset, the steps to preprocess the data, the feature extraction techniques used, as well as the classification models applied. Section IV shows the techniques used for data visualization and images of the results. Section V provides the results analysis and discussion from the experiments carried out. Finally, Section VI contains conclusions of the present research paper.

II. RELATED WORK

Ever since there are several online platforms which contain a lot of information about user opinion about products, and all this information is available online, therefore, there is an interest on analyze this large amount of data in order to obtain valuable information.

Meanwhile, on websites like Amazon or Ebay, users can submit their comments along with a specific polarity rating (from 1 to 5 stars). There is the possibility of mismatches between the written review and the polarity of the rating, for instance, a user can send a very good review, but still give a low rating. Therefore, in a previous work [15] in which the objective of the thesis was the development of a web service application that can be used to deal with this situation. Thus, if there is a discrepancy between the predicted rating score and the rating score sent, a warning or notification is generated. The data was preprocessed to remove the noise, then this data was converted to a vector using the unsupervised learning algorithm Paragraph Vector: 3.5 million product reviews are converted to 300 dimensional vectors of fixed length. So each review text is replaced with its corresponding feature vector in the complete data set. For each unique product, the review vectors were ordered according to the publication date of the review. These vectors of reviews ordered together with their corresponding qualifications will be the training sequence for that particular product. Then, these sequences are sent to a GRU to learn the representation of 128-dimensional feature vectors of product information.

The GRU is trained using a dropout or dropout rate of 0.25, Adam as the stochastic optimization method, categorical cross entropy as a function of loss, dense layer distributed over time and 128 hidden units. During the training phase, the product incorporation sequence for each unique product is retrieved and stored in a file for later efficient recovery.

Then, SVM is implemented using “Linear” kernel and the default value of other parameters provided by the scikit tool library. All the inlays of reviews were concatenated with the inlays of their products to create a feature vector of 428 dimensions and the SVM was trained using this final inlay vector. As a last step, a web service was developed that will avoid inconsistent reviews and ratings, by using the trained SVM model to predict the kind or type of feeling for a given review. In the event that the predicted class and the rating class sent by the user do not match, a feedback will be sent to the user so that they can correct their rating if desired. Finally, the classifier prepared in this thesis gives a prediction accuracy of 81.82%.

III. METHODOLOGY

Since there are several online platforms which contain a lot of available information about user opinion about products, web scraping was used to obtain the database. Then, in order to preprocess the data, NLP and machine learning techniques were used to extract features from the text. Next, the classification models were built with the extracted features as input, and the next step is to evaluate the performance of the models and analyze the results obtained.

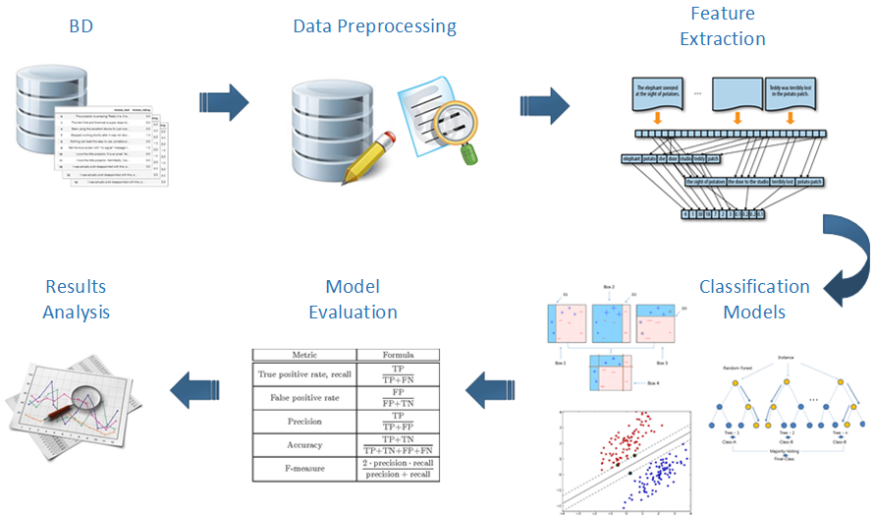

Figure 1. Methodology.

A. *Dataset*

For the acquisition of the database necessary to carry out this work, the Web Scraping technique was implemented to obtain all the reviews of 4 products from the Electronics department of the Amazon platform. These products are the following:

- "1080P wireless security camera with night vision" available on Amazon through the following link https://www.amazon.com/Wansview-Wireless-Security-Surveillance-Audio-K3/dp/B075KGNB58.
- "ASUS Chromebook C202SA-YS02 11.6" Ruggedized and Water Resistant Design with 180 Degree "available on Amazon through the following link https://www.amazon.com/Chromebook-C202SA-YS02-Ruggedized-Resistant-Celeron/dp/ B01DBGVB7K.
- "RIF6 Cube Full LED Mini Projector - 1080p Supported Portable Projector with Built-in Speakers, HDMI Input for Smartphone, Laptop, and Home Theater - Includes Tripod and Remote" available on Amazon through the following link https://www.amazon.com/RIF6-CUBE-Mobile-Pico-Projector/dp/B00QXS8L6I.
- "Samsung Gear S2 Smartwatch - Dark Gray" available on Amazon through the following link https://www.amazon.com/Samsung-Gear-S2-Smartwatch-Dark/dp/B015JQ62RY.

The Web Scraping technique was used to obtain and download in a JSON document all product reviews with its attributes such as the text of the review, the date of publication of the review, the title of the review, the rating from 1 to 5 of the review and the author of the review, as can be seen in Figure 2.

```
    {
        "review_text": "The RIF6 is a perfect size for what I
needed and wanted. For its size, it has amazing sound and
range. It is small but mighty. I would highly recommend this
speaker.",
        "review_posted_date": "22 Sep 2016",
        "review_header": "Small but mighty",
        "review_rating": "5.0 ",
        "review_author": "Amazon Customer"
    },
    {
        "review_text": "Great speaker, Unbelievable bass.",
        "review_posted_date": "14 Feb 2017",
        "review_header": "Five Stars",
        "review_rating": "5.0 ",
        "review_author": "Amazon Customer"
    },
    {
        "review_text": "Stopped working shortly after it was
not returnable. I would love to have it replaced because it
was great while it worked for a month or two. Now it's just a
paper weight",
        "review_posted_date": "02 Apr 2016",
        "review_header": "expensive paper weight",
        "review_rating": "1.0 ",
        "review_author": "Matthew Mahany"
    },
```


Figure 2. Sample JSON file of collected reviews.

B. Data Preprocessing

The JSON file obtained in the previous stage will be converted into a CSV file to be able to extract the reviews and create a dataframe and thus be able to carry out the preprocessing of its content.

Now, the preprocessing of its content can be carried out, which will consist of the elimination of all types of characters that are not letters, punctuation marks and numbers. Also, all the contractions present in the reviews were expanded ("I'll" for "I will", "wasn't" for "was not", etc.), as well as the elimination of 'stopwords' such as: This, me, they, that, and, the, etc. In addition, in this phase the entire dataframe content was converted to lowercase for better data processing. As a result of this preprocessing, the result that can be observed in Figure 3.

	review_text	review_rating
0	projector amaze really tiny pack huge punch am...	5.0
1	item first foremost super dope lol bought item...	5.0
2	would not recommend buying own one two month use t...	1.0
3	easy link great sound size	5.0
4	use excellent device year die return first one...	5.0

Figure 3. Dataframe result of the preprocessing phase.

Finally, the lemmatization of each word in the data was carried out, which replaces the word with its base form after performing the POS Tagging to identify if the word is an adjective, a noun, a verb or an adverb according to its context.

C. Feature Extraction

Two techniques were used to extract features: DTM + TF-IDF and Word2Vec, which will be described below.

1) DTM + TF-IDF: A Document-Term Matrix is used, which uses the TF-IDF scores, so the DTM can create a numerical matrix based on the weights of the TF-IDF instead of the frequency of terms that appear in the data.

2) Word2Vec: The Word2Vec [16] model is built to be trained with the content of the reviews. Then, the function to generate the feature vectors is defined, which combines the values of the word2vec vectors of each word in each review and is divided by the total number of words. Likewise, the function is defined to obtain the average characteristics vector, which iterates on the reviews, adds the sum of the vectors of each review using the

aforementioned function, to the predefined vector whose size is the number of total reviews and the number of features in Word2Vec. From these functions the training data is obtained for the validation in the form that can be seen in Figure 4.

3) N-Grams: The types of N-Grams to be used are defined: bigrams and trigrams. Here, the range for the generation of the N-Grams is set. First, it is defined as "ngram_range = (1,2)", which will result in bigrams. Then, the same procedure is performed to obtain trigrams indicating "ngram_range = (1,3)".

```
X_train_Vec
```

```
array([[-0.04555708, -0.00486659,  0.00556902, ..., -0.00517262,
        -0.08671413, -0.00286379],
       [-0.04541216, -0.00395017,  0.00434157, ..., -0.00415998,
        -0.08722883, -0.00390796],
       [-0.04634353, -0.00479589,  0.0058531 , ..., -0.00551935,
        -0.08762593, -0.00282803],
       ...,
       [-0.0451892 , -0.00505304,  0.00575377, ..., -0.00509765,
        -0.08653592, -0.00339814],
       [-0.04500055, -0.00730082,  0.00730946, ..., -0.00667912,
        -0.08641034, -0.00193931],
       [-0.04402219, -0.00648898,  0.00590209, ..., -0.00539651,
        -0.08631831, -0.00323595]], dtype=float32)
```

Figure 4. Training data result of the Word2Vec techniques.

D. Classification Models

The collected reviews are divided and classified in such a way that the reviews that obtained a rating of more than 3 stars are assigned a label of 1 (positive review), and for those that obtained a score less than 3 stars they are cataloged with a label of 0 (negative review).

Then, classification models are built for training and validation. The classification models that were used are the following: Logistic Regression model, Random Forest classifier with 100 trees, AdaBoost classifier, an XGBoost classifier, a Gradient Boosting classifier with 100 as the number of boosting stages to perform, and a Support Vector Machine classifier with a linear kernel.

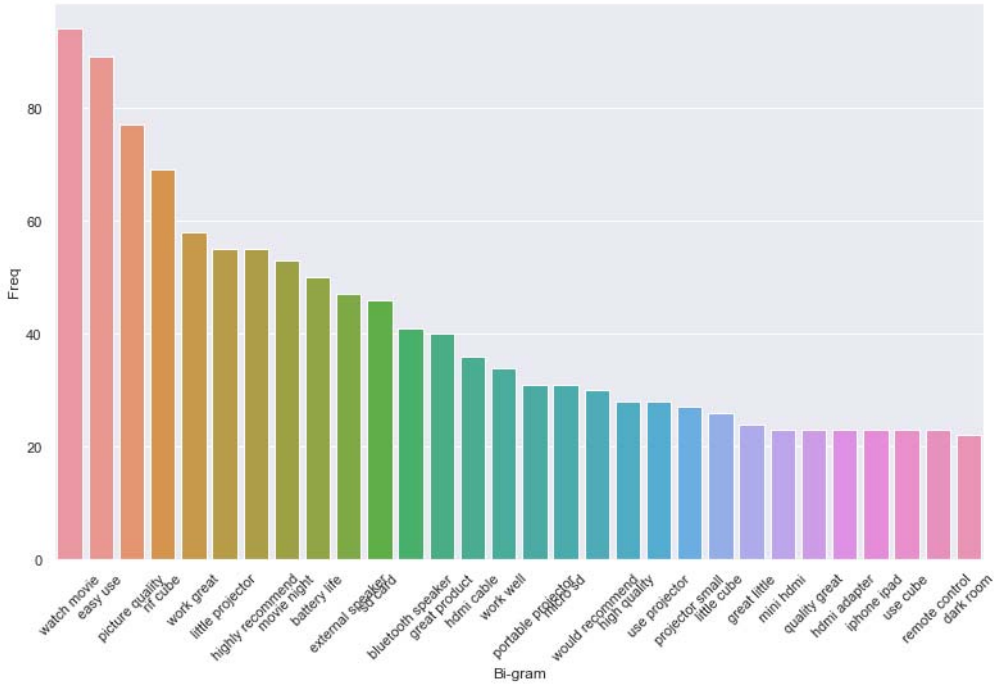

Figure 5. Frequency chart of bigrams of the mini projector.

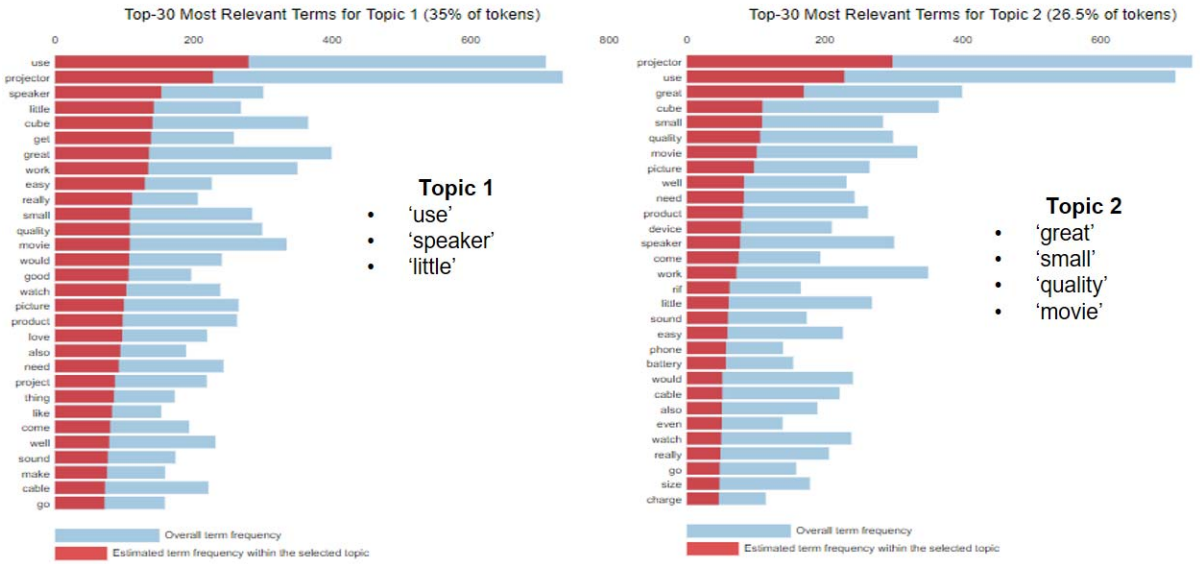

Figure 6. Topic Modeling of positive reviews of mini projector.

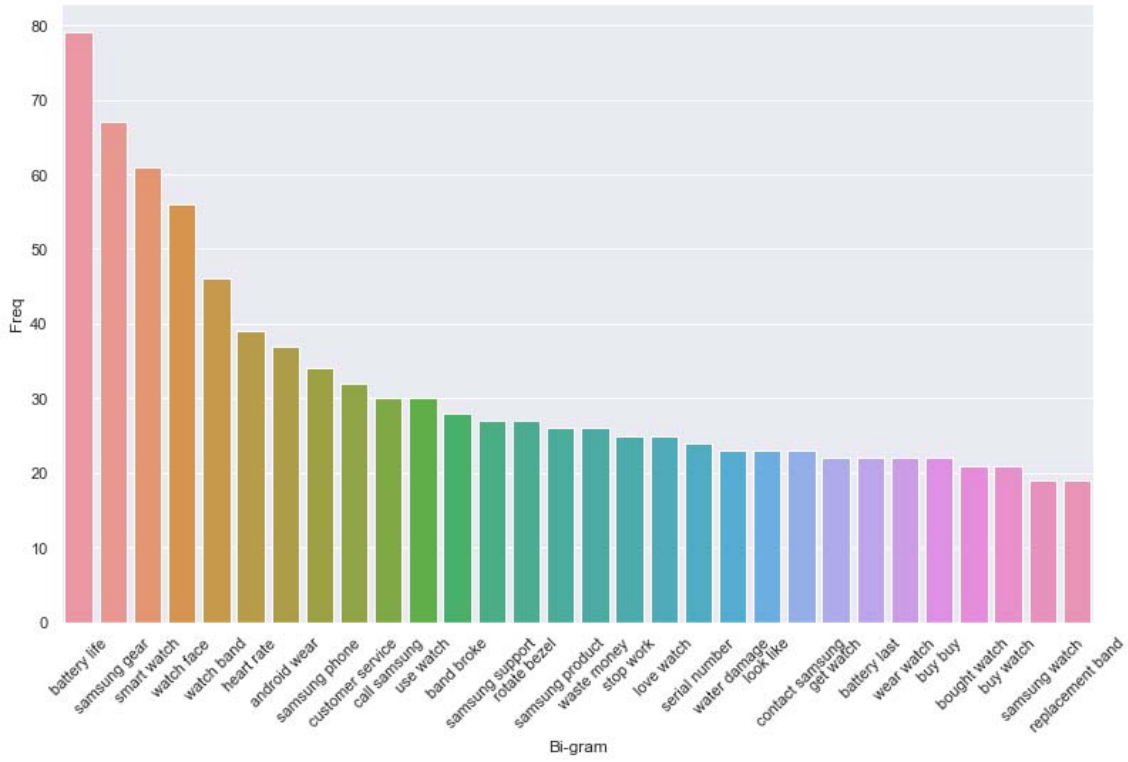

Figure 7. Frequency chart of bigrams of the smartwatch.

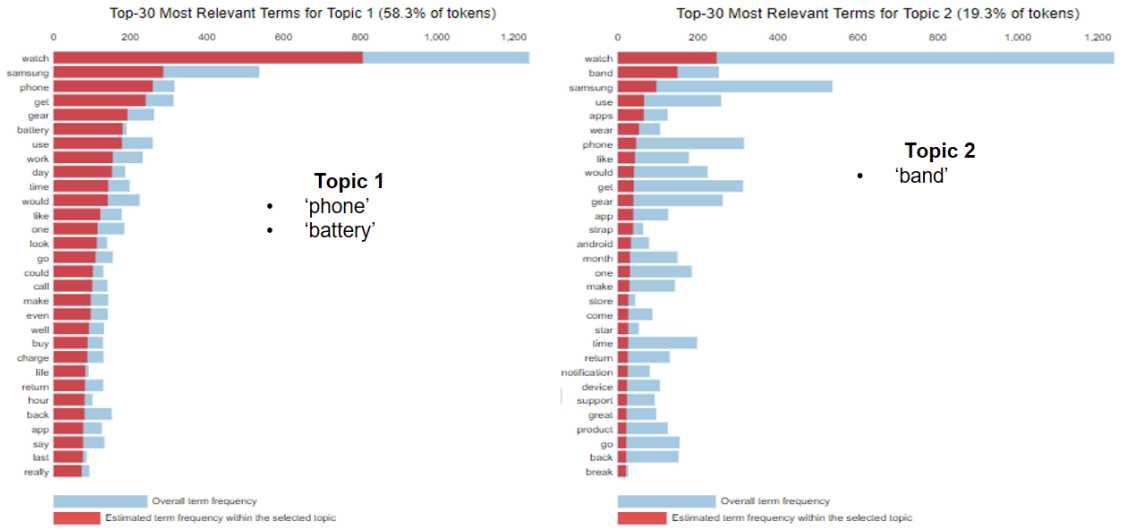

Figure 8. Topic Modeling of negative reviews of smartwatch.

IV. DATA VISUALIZATION AND EXPLORATION WITH N-GRAMS AND TOPIC MODELING

In this phase, unigrams, bigrams and trigrams are generated for both positive and negative reviews, and Topic Modeling [17] for positive and negative reviews is implemented. In this way, it becomes easier to identify which words (unigrams) or phrases (bigrams and trigrams) are most frequent in both positive and negative reviews.

From the bar chart on the frequency of the bigrams of the positive reviews it can be obtained both the features of the product that are most mentioned and the opinion of the users on these features. In the case of the RIF6 Cube mini projector, it can be seen from Figure 5, that the 'easy use' bigram has one of the highest frequencies, so it can be said that it is a feature that users value.

Furthermore, the presence of the bigrams 'watch movie' and 'movie night' in this graphic give an idea of the use that users give to this projector.

Also, when doing the Topic Modeling to the positive reviews of the mini projector, the terms that include the first two topics found with this technique appear. In Figure 6 it can be seen that in the topic 1 the terms' use, 'speaker', 'little', have a higher frequency, while in topic 2 the words 'great', 'small', 'quality', 'movie'. Because Topic Modeling is an unsupervised technique, its results are interpreted by the user, so it can be deduced from these results that the first topic talks more about the use of the mini projector and its speaker or speaker, while in topic 2 the image quality predominates.

On the other hand, from the bar chart on the frequency of the bigrams of the negative reviews, the features of the product that are most mentioned in this group of reviews can be seen. In the case of the Samsung Gear S2 Smartwatch, it can be seen from Figure 7, that the 'battery life' bigram has one of the highest frequencies, so it can be stated that the smartwatch's battery life did not meet the expectations of the users who rated the product negatively.

Similarly, when doing the Topic Modeling to the negative reviews of the smartwatch it can be observe the terms that include the first two topics found with this technique. In Figure 8 it can be seen that in topic 1 the terms 'phone' and 'battery' have a higher frequency, while in topic 2 the word 'band' stands out among the others. So it can be deduced that the first topic talks more about the battery that the smartwatch has, while in topic 2 the band or strap that has this device predominates.

Another way to visualize the data is through a wordcloud, which shows the words most frequently in the text, giving a larger size to those words or phrases that are more frequent.

Figure 9. Wordcloud of the security camera.

As can be seen in Figure 9, through a wordcloud some of the most commented features in the reviews can be identified, such as ‘customer service’, ‘night vision’, ‘easy set’, etc. Which indicates that these features are important for users. Similarly, users' appreciation of the product can be identified through the phrases ‘work great’ and work well’.

V. Results Analysis and Discussion

In this stage the classifiers will be evaluated through statistical metrics such as accuracy, precision, recall and the F1 score. This will allow a better analysis of the results.

Table 1. Model metrics for the DTM+TF-IDF Technique

Miniproyector				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.885	0.890	0.980	0.933
Random Forest	0.885	0.890	0.980	0.933
Adaboost	0.873	0.916	0.930	0.923
XGBoost	0.881	0.883	0.985	0.931
Gradient Boosting	0.873	0.912	0.935	0.923
SVM	0.869	0.912	0.930	0.921
Laptop				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0. 908	0.9	0.990	0.942
Random Forest	0.924	0.920	0.986	0.952
Adaboost	0.895	0.899	0.973	0.934
XGBoost	0.892	0.881	0.993	0.934
Gradient Boosting	0.892	0.893	0.976	0.933
SVM	0.908	0.919	0.966	0.942
Smartwatch				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.858	0.858	0.969	0.910
Random Forest	0.877	0.893	0.950	0.920
Adaboost	0.866	0.877	0.954	0.914
XGBoost	0.856	0.858	0.967	0.909
Gradient Boosting	0.866	0.879	0.952	0.914
SVM	0.879	0.905	0.934	0.919
Security Camera				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.890	0.905	0.970	0.936
Random Forest	0.912	0.913	0.987	0.949
Adaboost	0.887	0.892	0.983	0.935
XGBoost	0.883	0.888	0.983	0.933
Gradient Boosting	0.876	0.900	0.957	0.928
SVM	0.908	0.941	0.949	0.945

As can be seen from Table 1, the DTM + TF-IDF technique is the feature extraction technique with better results, since using this technique a greater accuracy was obtained than the results from the Word2Vec technique shown in Table 2. Likewise, it can also be observed that, using the first technique, the best results are obtained with the Random Forest Classifier model. Similarly, using the second technique it can be seen that the best results are obtained with the Logistic Regression model.

Table 2. Model metrics for the Word2Vec Technique

Miniproyector				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.831	0.834	0.995	0.907
Random Forest	0.826	0.850	0.960	0.902
Adaboost	0.806	0.846	0.935	0.888
XGBoost	0.835	0.854	0.965	0.906
Gradient Boosting	0.798	0.863	0.895	0.879
SVM	0.835	0.835	1.000	0.910
Laptop				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.879	0.881	0.990	0.933
Random Forest	0.826	0.851	0.960	0.902
Adaboost	0.806	0.851	0.950	0.898
XGBoost	0.831	0.852	0.960	0.903
Gradient Boosting	0.800	0.859	0.890	0.874
SVM	0.835	0.833	1.000	0.909
Smartwatch				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.855	0.857	0.970	0.910
Random Forest	0.802	0.826	0.945	0.881
Adaboost	0.781	0.827	0.940	0.880
XGBoost	0.808	0.831	0.945	0.884
Gradient Boosting	0.775	0.833	0.875	0.854
SVM	0.810	0.811	1.000	0.896
Security Camera				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.874	0.877	0.985	0.928
Random Forest	0.821	0.846	0.950	0.895
Adaboost	0.801	0.847	0.945	0.893
XGBoost	0.826	0.848	0.950	0.896
Gradient Boosting	0.795	0.852	0.880	0.866
SVM	0.830	0.830	1.000	0.907

Similarly, in the case of N-Grams, it can be observed that the technique with the best results is 1-2 ngrams (Table 3), since when implementing this technique, a higher accuracy value was obtained. Also, it can be seen from Table 3 and Table 4 that the logistic regression model predominates along with the SVM models, obtaining higher accuracy in these two models for N-Grams techniques.

Table 3. Model metrics for the 1-2 NGram Technique

Miniproyector				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.902	0.904	0.985	0.943
Random Forest	0.848	0.844	1.000	0.915
Adaboost	0.857	0.891	0.940	0.915
XGBoost	0.877	0.879	0.985	0.929
Gradient Boosting	0.848	0.865	0.965	0.913
SVM	0.873	0.916	0.930	0.923
Laptop				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.908	0.898	0.993	0.943
Random Forest	0.879	0.866	0.997	0.927
Adaboost	0.905	0.910	0.973	0.940
XGBoost	0.895	0.887	0.990	0.935
Gradient Boosting	0.866	0.866	0.976	0.918
SVM	0.905	0.905	0.979	0.941
Smartwatch				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.892	0.902	0.959	0.930
Random Forest	0.849	0.845	0.977	0.906
Adaboost	0.872	0.882	0.956	0.918
XGBoost	0.859	0.858	0.971	0.911
Gradient Boosting	0.846	0.857	0.952	0.902
SVM	0.874	0.897	0.938	0.917
Security Camera				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.922	0.918	0.996	0.955
Random Forest	0.873	0.875	0.987	0.928
Adaboost	0.894	0.899	0.983	0.939
XGBoost	0.887	0.895	0.979	0.935
Gradient Boosting	0.855	0.882	0.953	0.916
SVM	0.922	0.931	0.979	0.954

Therefore, according to the results obtained from the 4 techniques implemented, the DTM + TF-IDF technique gives better results than the N-Grams and Word2Vec techniques and the predominant models are Logistic Regression and Random Forest Classifier for its higher accuracy in the 4 techniques overall.

Overall, it was possible to verify that through the use of unigrams, bigrams and trigrams applied to a significant set of reviews, it is possible to identify not only the features most commented on by users but also their opinion about them or the product in general.

However, it is necessary to emphasize that, when using user reviews, classified only as positive and negative, it should be taken into account that both positive and negative reviews can have partial opinions in favor and against the characteristics of the product. That is, a user who qualifies the product with a positive review has a positive opinion regarding the product in general, however, it may also include in the review some characteristics that he did not like about the product.

Table 4. Model metrics for the 1-3 NGram Technique

Miniproyector				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.902	0.900	0.990	0.943
Random Forest	0.828	0.829	0.995	0.905
Adaboost	0.857	0.891	0.940	0.915
XGBoost	0.877	0.879	0.985	0.929
Gradient Boosting	0.857	0.873	0.965	0.917
SVM	0.869	0.912	0.930	0.921
Laptop				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.900	0.890	0.993	0.939
Random Forest	0.834	0.826	0.993	0.902
Adaboost	0.905	0.910	0.973	0.940
XGBoost	0.895	0.887	0.990	0.935
Gradient Boosting	0.853	0.855	0.973	0.910
SVM	0.889	0.896	0.969	0.931
Smartwatch				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.888	0.896	0.961	0.927
Random Forest	0.833	0.823	0.988	0.898
Adaboost	0.868	0.879	0.954	0.915
XGBoost	0.862	0.861	0.971	0.913
Gradient Boosting	0.836	0.856	0.938	0.895
SVM	0.865	0.894	0.929	0.911
Security Camera				
	Acc	**Prec**	**Recall**	**F1**
Logistic Regression	0.908	0.903	0.996	0.947
Random Forest	0.862	0.866	0.987	0.922
Adaboost	0.894	0.899	0.983	0.939
XGBoost	0.887	0.895	0.979	0.935
Gradient Boosting	0.869	0.890	0.962	0.924
SVM	0.915	0.924	0.979	0.950

VI. CONCLUSION

In the present work, a machine learning model was implemented to detect trend patterns of technical data in online products. It is important to emphasize that while this work is true, it was carried out using Amazon platform reviews, its methodology and implementation are not limited to these. Likewise, it is worth highlighting the importance of the role of the data preprocessing phase, especially in the elimination of characters or words, in lemmatization and in the elimination of stopwords. This is due to the fact that if proper lemmatization is not carried out with the use of Pos tagging, the meaning or meaning of the words or phrases could be altered. The same happens with the elimination of stopwords, since some denials are considered as stopwords, such as (would not, will not, is not, etc.). However, for the development of this work, the elimination of these words would have a negative effect on the identification of their polarity, since the words that characterize the negative reviews would be eliminated. Another important factor is the technique of extracting features to use, since although it is true there are multiple techniques to be able to extract features from a text, it is important to identify those whose results are the best, especially for the construction of the models of the classification and in its subsequent evaluation.

REFERENCES

[1] M. SI, 'Estudio sobre comercio electrónico B2C', ONTSI, 2017. [Online]. Available: http://www.ontsi.red.es/ontsi/es/content/estudio-sobre-comercio-electr%C3%B3nico-b2c-edici%C3%B3n-2017

[2] A. López-Quesada, 'La importancia del mundo digital en las ventas', Conexion Esan, 2013. [Online]. Available: https://www.esan.edu.pe/conexion/actualidad/2013/03/04/importancia-mundo-digital-ventas/

[3] G. Lackermair, D. Kailer and K. Kanmaz; K, "Importance of Online Product Reviews from a Consumer's Perspective", Advances in Economics and Business, Vol.1, Issue.1, pp.1-5, 2013.

[4] A. Macario, 'La importancia de las reseñas online', Puro Marketing, 2018. [Online]. Available: https://www.puromarketing.com/30/29952/importancia-resenas-online.html

[5] J. McAuley, and J. Leskovec, "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text", RecSys 2013 - Proceedings of the 7th ACM Conference on Recommender Systems, 2013.

[6] J. Han, J. Pei, and M. Kamber, "Data Mining: Concepts and Techniques", Elsevier, 2011.

[7] I. Hemalatha, P. S. Varma, and A. Govardhan, "Sentiment Analysis Tool using Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science, Vol.2, pp.105-109, 2013.

[8] C. Aggarwal, and C. Zhai, "Mining Text Data", Springer, 2012.

[9] P. Rochina, '¿Qué es y cuáles son las aplicaciones del Text Mining?', Revista Digital INESEM, 2017. [Online]. Available: https://revistadigital.inesem.es/informatica-y-tics/text-mining/

[10] S. Das, 'Text Mining and Topic Modeling Using R', DZone, 2017. [Online]. Available: https://dzone.com/articles/text-mining-using-r-and-h2o-let-machine-learn-lang

[11] E. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications, Vol.112, pp.44-48, 2015.

[12] Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. doi:10.2200/s00416ed1v01y201204hlt016

[13] L. Joyanes, "Big Data, Análisis de grandes volúmenes de datos en organizaciones", Alfaomega Grupo Editor, 2016

[14] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and Combining Sentiment Analysis Methods", COSN, 2013

[15] N. Shrestha, "Deep Learning Implementation for Comparison of User Reviews and Ratings", M.S. Thesis, Department of Computer Science, University of Nevada, Las Vegas, 2013.

[16] A. Thakker, 'Introduction to Word2Vec & How it Works', Aditya Thakker, 2017. [Online]. Available: https://www.adityathakker.com/introduction-to-word2vec-how-it-works/

[17] U. Malik, 'Python for NLP: Topic Modeling', Stack Abuse, 2019. [Online]. Available: https://stackabuse.com/python-for-nlp-topic-modeling/

Author Profile

Gina Violeta Acosta Gutiérrez is an Information Technology and Systems Engineering student at Esan University in Lima-Perú and will be graduating in 2020. She has strong interests in Business Intelligence, Data and Web mining, and Computer Vision. Currently, she participates proactively in research projects related to the field of Big Data Analytics, Computer Vision and Game-Based Learning.