Sentiment Analysis of Amazon Reviews using Deep Learning Techniques

Experiment Findings · June 2023

DOI: 10.13140/RG.2.2.35547.75046

4 authors, including:

Isabella Noriega
Tecnológico de Monterrey
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Francisco J. Cantu-Ortiz
Tecnológico de Monterrey
117 PUBLICATIONS **863** CITATIONS
SEE PROFILE

Sentiment Analysis of Amazon Reviews using Deep Learning Techniques

Isabella Noriega Quirós
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, Nuevo León
A01250571@tec.mx

Héctor Valdez Alvarez
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, Nuevo León
A01252068@tec.mx

Carlos Ernesto Mendoza Ramírez
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, Nuevo León
A01251967@tec.mx

Francisco J. Cantú-Ortiz
School of Engineering and Sciences
Tecnológico de Monterrey
Monterrey, Nuevo León
fcantu@tec.mx

***Abstract*—In 2020, with the rise of the COVID-19 pandemic, e-commerce, commercial transactions conducted electronically on the internet, increased 43%, and has kept increasing ever since. Amazon is a multinational technology company with a focus on e-commerce. Reviews are of importance when buying a product, hence they can offer valuable and honest insight regarding the selected article; nevertheless, it can take a big amount of time to read each of the reviews. On the other hand, buyers are not the only ones interested in reading reviews, but also, sellers that wish to distinguish and make their own products successful. This paper aims to apply sentiment analysis to a database gathered from Amazon reviews from different departments, with the purposes of classifying a review according to its sentiment, and also, identifying the characteristics that distinguish the best products from the rest. The language models used are the Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, and XLNet for sentiment analysis.**

***Index Terms*—amazon, electronics, sentiment analysis, reviews.**

I. Introduction

In the year 2020 humanity faced a terrible problem, the pandemic. With the arrival of COVID-19, our society as we knew it was forced to undergo drastic changes. This global event forced us to shut ourselves in to minimize contact with the outside world. It was an event that affected us as a society physically, emotionally and economically without precedent on this scale. Fortunately, thanks to the vaccines and the health measures that were taken, this terrible disease that took thousands of lives now no longer affects us as it did in the beginning. Hospitalizations dropped drastically thanks to vaccines, infections were still reduced by them, because of this our doors reopened and we can coexist again as a society almost like before the pandemic. The pandemic brought many changes with it, most of them forced by the situation, but not all of them necessarily bad. One of the important changes was with electronic commerce. [8].

E-commerce was just beginning to gain strength before the pandemic, multiple online stores emerged and large companies like Amazon gradually gained strength. However, many were still skeptical about buying online, for fear of fraud, phishing and in general there was a great distrust of putting our personal data and credit cards in the hands of an electronic company. The transition from conventional commerce to electronic commerce was something that was happening with small steps, but the pandemic forced us to take a gigantic step as electronic commerce became one of the best ways to avoid contact with covid-19 in those difficult times. With the pandemic, online consumption increased worldwide, in the year of the pandemic in Brazil the sale of computers increased by 112%, in just March in Chile online sales increased by 119% and in a matter of weeks in Bogota electronic commerce increased 23%. This sudden increase in sales changed the industry, at Amazon they had to hire 100,000 employees to be able to deliver the many orders they were receiving. [8]. For their part, physical stores in Mexico also underwent drastic changes. The two and a half years of the health emergency and the accelerated penetration of e-commerce in Mexico, caused La Comer, Soriana and Chedraui to pause the opening of new physical stores, to instead focus on improving the experience on its different sales platforms. [11].

When the transition from web 1.0 to web 2.0 is made, one of the main changes was that now people had an online presence, one of the first companies to take this step was Amazon, by giving users a voice with the implementation of reviews and ratings for purchased products. Now that we have Big Data 2.0, we have the tools and resources to go beyond than just having to read millions of reviews to know if the products will be useful to us, this boosted the development of our research. The numbers in the sales of the products give us statistical information about those products, but the reviews give us something else, how customers feel with what they bought and how satisfied they are with them, for this reason we decided to use that information to make a sentiment analysis of amazon products reviews using deep learning techniques. [9]. A sentiment analysis, also referred as opinion mining, can be defined as a contextual mining of text which identifies and extracts subjective information in source material. It is a natural language processing (NLP) technique used to determine if the data has a positive sentiment or a negative sentiment. It is usually performed on textual data to monitor the reaction to a certain topic in social media or to help businesses have customer feedback for their products. [9].

The aim of this paper is to apply a sentiment analysis to a database gathered from amazon reviews from different departments and products with the objective of classifying a review according to its sentiment and identifying the characteristics that distinguish the best product from the rest using deep learning techniques.

II. Related Work

Sentiment analysis has advanced significantly in recent years, thanks to the development of deep learning models, particularly transformer-based architectures like BERT, GPT, and their variants [1]. These models have achieved top performance in various NLP tasks, including sentiment analysis, by leveraging large-scale pre-training on diverse textual data and fine-tuning on task-specific labeled data [2].

An important development in this field is the introduction of RoBERTa (A Robustly Optimized BERT Pretraining Approach), which builds on BERT and improves its performance by modifying key hyperparameters, training strategies, and using a larger dataset for pre-training [4]. Another notable development is the development of the ELECTRA model (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), which demonstrates improved efficiency and performance compared to BERT by employing a more effective pre-training task called replaced token detection [5]. These deep learning models have demonstrated their effectiveness in sentiment analysis tasks, such as classifying opinions expressed in reviews, social media, and other text sources.

There are several works related to sentiment analysis through deep learning techniques, one of them is the mentioned architecture BERT (Bidirectional Encoder Representations from Transformers) which is a groundbreaking deep learning model introduced by Devlin et al. BERT represents a significant advancement in various NLP tasks, including sentiment analysis, by capturing contextual information in both directions (left-to-right and right-to-left) through its masked language model (MLM) pre-training objective. BERT is pre-trained on large text corpora (such as BooksCorpus and English Wikipedia) and fine-tuned for specific tasks using task-specific labeled data. This pre-training and fine-tuning approach allows BERT to effectively understand the relationships between words and phrases [3].

ULMFiT (Universal Language Model Fine-tuning) is an approach for text classification tasks introduced by Howard and Ruder (2018). ULMFiT leverages transfer learning to fine-tune pre-trained language models on specific tasks, such as sentiment analysis. The approach consists of three main steps: 1° pre-training a language model on a large corpus of text, 2° fine-tuning the language model on the target task's data, and 3° training a classifier using the fine-tuned language model. ULMFiT incorporates several innovative techniques, such as discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing, which together enable the model to achieve good performance on a variety of text classification tasks with relatively small amounts of labeled data [6].

Another deep learning model is XLNet, a generalized autoregressive pretraining model proposed by Yang et al. (2019) that improves upon BERT by addressing some of its limitations. One major limitation of BERT is the inability to effectively model bidirectional contexts due to the use of the masked language model (MLM) pre-training objective. XLNet addresses this issue by using a permutation-based training objective, which models the joint probability of a sequence of tokens by conditioning each token on all other tokens in the sequence. Additionally, XLNet integrates segment recurrence mechanisms and relative positional encodings to better capture long-range dependencies in text. As a result, XLNet achieves one of the most advanced performance on several NLP benchmarks, including sentiment analysis tasks [7].

III. Method and Data

A. Data Validation

The data used for this paper will be a dataset gathered from a github meant for sentiment analysis for Amazon reviews. The github [10] is already divided into training and testing labeled 0 or 1 corresponding to a bad or good review relatively, and comes with the exact text written by the clienet. The size of the training data is 150,000 entries, and the size of the testing data is of 30,000 entries; having a 83-17 percentage distribution.

B. Statistical Analysis

When using any dataset for experimentation it is always important to keep in mind that not only the train-test distribution should be valid, but the distribution among labels as well. Here we present images III-B and III-B that correspond to the validation that the data can be correctly used for this analysis.

The labels are 50 and 50 percent distributed, which means the dataset is ready for experimentation.

C. Exploratory Data Analysis

TF-IDF and K-means clustering will be firstly used to eliminate the words that are not exactly useful for the purpose aimed. In Figure III-C and III-C we can see a first visualization of the most common words used in each type of the reviews. However, some words mentioned a specific product, whereas we are interested on the words that make those products either good or bad, thus the filtering must be applied.

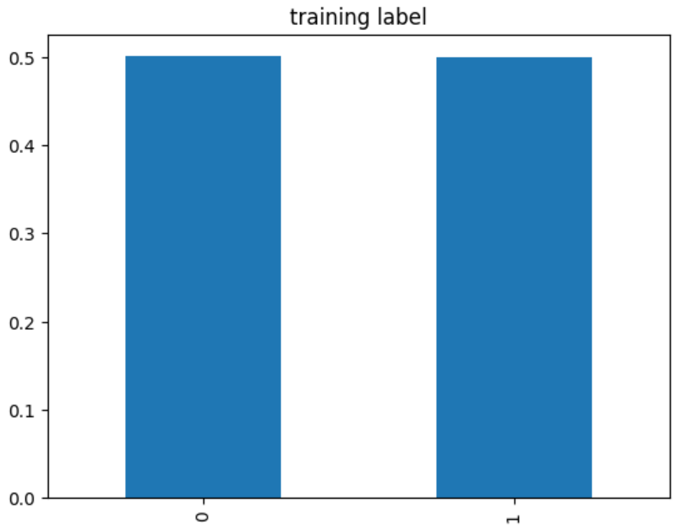

Fig. 1. Distribution of labels among training data

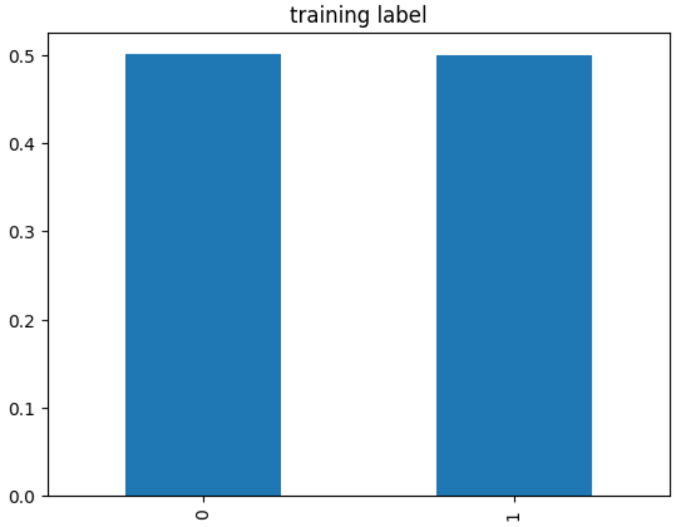

Fig. 2. Distribution of labels among testing data

D. Approach

The chosen techniques to be applied for sentiment analysis after applying the TD-IDF and K-means clustering are: BERT, RoBERTa, XLNet, and ULMFIT.

1) BERT: BERT represents a significant advancement in various NLP tasks, including sentiment analysis. BERT is pre-trained on large text corpora and fine-tuned for specific tasks using task-specific labeled data.
2) ULMFiT: ULMFiT leverages transfer learning to fine-tune pre-trained language models on specific tasks, such as sentiment analysis. ULMFiT incorporates several innovative techniques, such as discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing.
3) XLNet: That improves upon BERT by addressing some of its limitations.XLNet addresses this issue by using a permutation-based training objective.
4) RoBERTa: It is based on the BERT model as its name suggest, It builds on BERT and modifies key hyperparameters.

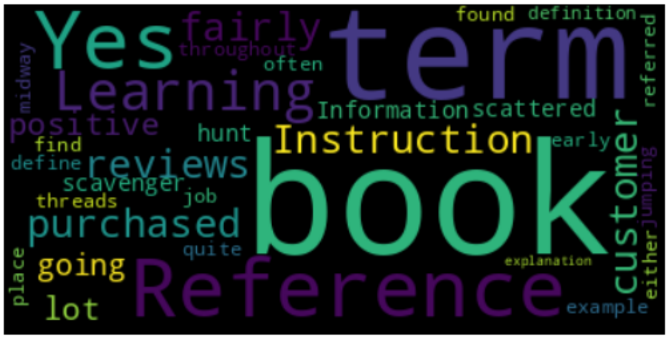

Fig. 3. Most common words found in negative reviews.

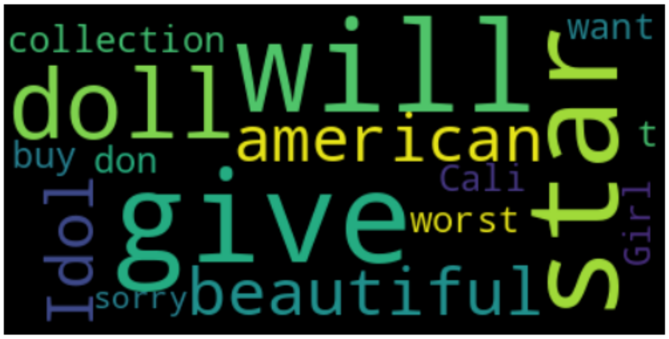

Fig. 4. Most common words found in positive reviews.

IV. Results

The project was carried out as established in the methodology section. The four models were generated and trained with our database. Once we had the results of the performance metrics for each model, they were compared in a table.

We can appreciate the results obtained from the four sentiment analysis models applied in Table I.

Model	**Accuracy**	**F1-Score**	**Sensitivity**	**Specificity**
BERT	79%	0.879	0.959	**0.05**
ULMFIT	81%	0.875	0.945	**0.05**
XLNet	79%	0.883	**1**	0
RoBERTa	**82%**	**0.901**	**1**	0

TABLE I
Results obtained from the sentiment analysis models

With the results presented once the metrics were analyzed it can be concluded that the model that obtained the best performance is: RoBERTa.

V. Discussion

One of the things worth discussing is the results. As a team, and first time performing sentiment analysis, we expected the

ULMFIT model to do better; nevertheless, as results point out, it seems that for this particular task, the RoBERTa model was able to achieve better scores as of the performance metrics.

Some of the limitations this work had, was of course, the dataset. It cannot be taken as a model that would always results the same, since it has a particular purpose.

Another limitation was that, for this case, reviews can be starred between 0 and 5, which was later classified as negative (0-2) and positive (3-5) from a subjective point of view.

VI. Conclusion

To conclude we want to express that the first two objectives are similar, in summary, the most common words found for each sentiment can aid either a client or a seller to make better decisions faster regarding which item they should buy, or which aspects they should take special care of when making a sell.

- The most common words that describe a positive review are: easy, sound, screen, quality, time, clear, and light.
- The most common words that describe a negative review are: problem, issue, power, button, update, return, and company.

The third objective was to find the best model to perform sentiment analysis for amazon reviews of the electronics department.

- The best model for sentiment analysis based on the results is: RoBERTa.

We hope this work motivates other computer science students to keep improving scores and exploring sentiment analysis tools derived from written sources. We would also like to say that we hope, at least for people involved in the selling and buying industry of e-commerce, that this work could offer helpful insights as to what they should focus on when buying or selling an item.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 5998-6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[2] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[5] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3707-3720). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.290

[6] Howard, J., & Ruder, S. (2018). Fine-tuned language models for text classification. In Proceedings of the 2018 Conference of the Association for Computational Linguistics (ACL) (pp. 2545-2554). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1248

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems 32 (NeurIPS 2019) (pp. 5754-5764). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/1dfb7fdf55ceb9dbeadf833729670b07-Paper.pdf

[8] Enrico, C. (2020). El efecto de COVID-19 en el ecommerce. Forbes. https://www.forbes.com.mx/el-efecto-de-covid-19-en-el-ecommerce/

[9] Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

[10] Muhammed Buyukkinaci. (2019). Github. Retrieved from https://github.com/MuhammedBuyukkinaci/TensorFlow-Sentiment-Analysis-on-Amazon-Reviews-Data

[11] Rodríguez, A. (2022). ¿Adiós tiendas físicas? Ventas en línea y COVID 'pausan' apertura de comercios. El Financiero. https://www.elfinanciero.com.mx/empresas/2022/10/20/e-commerce-y-el-covid-pausan-la-apertura-de-nuevas-tiendas/

[12] Gupta, S. (2018). Sentiment Analysis: Concept, Analysis and Applications. Towards Data Science. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17