Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 5041-5049
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.3133

Using neural networks for the detection of textual patterns in academic texts: A case study at the university of Nariño

Jesús Insuasti[1*], Carlos Mario Zapata-Jaramillo[2], Manuel Bolaños[3]
[1,3]University of Nariño, Pasto, Colombia; insuasti@udenar.edu.co (J.I.) mbolanos@udenar.edu.co (M.B.).
[2]Universidad Nacional de Colombia, Medellín, Colombia; cmzapata@unal.edu.co (C.M.Z.J.).

Abstract: Dealing manually with large volumes of textual information leads to problems in qualitative research, such as the time spent reading and processing the data and the possibility of losing critical details. In education sciences, investigations are commonly based on the qualitative paradigm, requiring substantial amounts of data to be processed textually. In this situation, an alternative for processing text to detect patterns using neural networks in the computational linguistics scenario is proposed. In this regard, qualitative research using a quasi-experimental method is performed. The quasi-experiment is done within the Qualitative Information Analysis course of the master's degree in Higher Education Teaching in 2022 with Class XIV; in such a course, six students evaluated the computational solution using material related to their working master's thesis. The results from the application test are satisfactory due to the positive user acceptance, providing a plausible alternative to commercial computational solutions on the market.
Keywords: *Computational, Linguistics, Neural, Networks, Patterns, Textual.*

1. Introduction

In qualitative research, especially within the Education Sciences, the vastness of textual data can often be overwhelming. With the ever-growing availability of digital text resources, researchers are increasingly confronted with manually shifting through, analyzing, and detecting patterns within these large datasets. The manual nature of this task not only consumes considerable time and effort but also poses risks of potential oversight, leading to missed insights or critical details that could shape the outcomes of a study. As the importance of qualitative research solidifies its place in the academic landscape, so does the necessity for innovative approaches to data handling.

Step into the expanding realm of computational linguistics, which fundamentally aims to comprehend and manipulate human language in a manner that computers find significant. In the last ten years, there has been an impressive increase in incorporating neural networks into computational linguistics, presenting groundbreaking opportunities for automating and enhancing the processing of textual information. This paper delves into the intricacies of leveraging neural networks within computational linguistics to detect patterns in large textual datasets. Beyond merely presenting a theoretical framework, this paper introduces an alternative computational solution and evaluates its practicality and efficiency in real-world scenarios. The research aims to contribute a tool to the academic community and provide a compelling alternative to existing commercial solutions.

This paper has five sections. The first covers the theoretical background, the second deals with the methodological aspects of this research, and the third describes the results. In the fourth section, we present the discussion and conclusions, and finally, the last section depicts some future work related to the main topic of this research.

* Correspondence: insuasti@udenar.edu.co

2. Theoretical Background

In today's digital age, the ability to process and understand vast quantities of text has become more imperative than ever. As the world generates an increasing amount of digital data, including unstructured text data, the importance of qualitative data analysis comes to the fore [1]. Computational linguistics offers techniques and approaches that connect the depth of qualitative research with the vast amounts of Big Data, providing creative solutions to intricate analytical problems [2].

Computational linguistics has evolved from early rule-based systems to sophisticated machine learning models, integrating deeply with the burgeoning field of Natural Language Processing—NLP—[3]. Traditional qualitative analysis, characterized by manual coding and close reading, struggles with scalability [4]. The advent of Big Data necessitates faster, more efficient methods to derive meaningful insights from large text datasets [5].

NLP offers several foundational techniques beneficial for large-scale qualitative analysis. From basic methods like tokenization and stemming to advanced semantic analyses using word embeddings [6], NLP furnishes instruments for exploring the architecture and significance of texts. The latest progress in deep learning, featuring frameworks like transformers and recurrent neural networks (RNNs), has expanded the limits of understanding text [7].

Traditionally, grounded theory has been instrumental in constructing theory from qualitative data [8]. Thematic analysis, another approach, focuses on identifying and analyzing themes within datasets, whereas narrative analysis explores how individuals construct narratives and identities in their speech [9]. The scale of modern datasets has led researchers to seek automated or semi-automated methods to complement these traditional approaches.

With the introduction of algorithms capable of automatic coding, the potential to process large datasets qualitatively has expanded dramatically [10]. Sentiment analysis offers a lens to gauge public opinion and emotions in vast textual data [11]. Similarly, topic modeling allows researchers to identify hidden thematic structures within texts, particularly Latent Dirichlet Allocation [12]. Machine learning, especially text classification, has also paved the way for categorizing qualitative data based on predefined criteria [13].

Despite their promise, computational approaches are not without their challenges. The nuanced intricacies of human language can sometimes be lost, leading to oversimplified analyses [14]. Algorithmic biases are a growing concern, with models potentially reflecting societal biases in training data [15]. Ensuring the reliability and validity of automated analyses is a paramount challenge that researchers must address [16].

The swift advancement of transfer learning models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) carries considerable promise for qualitative research [17]. Active learning systems, integrating human expertise with computational efficiency, offer a promising direction for ensuring quality while managing large datasets [18]. As with all technological advancements, ethical considerations will be essential, especially concerning data privacy and fairness [19].

3. Methodology

This research compares linguistic models to provide a general overview of the main objectives, key features, and primary use cases. Table 1 depicts the comparison.

Table 1.
Comparison of models in natural language processing.

Model/Paper	**Date**	**Purpose**	**Characteristics**	**Appliances**
BERT [17]	2018	Training bidirectional transformers in advance to grasp language	Employs transformer framework. Context is understood in both directions. Utilizes a masked language model.	Text classification, Q&A, NER
GPT [20]	2018	Enhancing comprehension of language through self-taught learning methods	Based on transformer technology. Context is processed from left to right. Pre-training with a generative approach.	Text generation, fine-tuning tasks
Transformers/ Attention [7]	2017	Proposing the Transformer architecture without recurrent layers	Attention mechanism. Scales are better than RNNs or CNNs.	Foundation for many subsequent models
T5 [21]	2019	Investigating transfer learning through a cohesive text-to-text approach	Approaches all NLP challenges as text-to-text tasks. Integrated approach.	Translation, summarization, Q&A
RoBERTa [22]	2019	Optimizing the pretraining of the BERT model	Differences in the size of the model, the amount of training data, and the duration of training. Eliminates Next Sentence Prediction (NSP), utilizes additional data, and extends training period.	Like BERT's use-cases
DistilBERT [23]	2019	Developing a more streamlined version of BERT	Reduced size by 40% while maintaining 95% of BERT's effectiveness. Utilizes the technique of knowledge distillation.	Where BERT is too large or slow
ELECTRA [24]	2020	Introducing a novel approach to pre-training	Substitutes masked tokens and attempts to identify these substitutions. More effective compared to methods based on Masked Language Modeling (MLM).	Text classification, Q&A
Universal Sentence Encoder [25]	2018	Encoding sentences into fixed-size embeddings	- Captures semantic meaning - Trained in various tasks, including SNLI - Provides transfer learning to new tasks	Semantic textual similarity, clustering

ISSN: 2576-8484
Vol. 8, No. 6: 5041-5049, 2024
DOI: 10.55214/25768484.v8i6.3133

According to the above, neural network models based on transformers are the most suitable techniques for computational linguistics. A large amount of textual data is required to work with transformers.

This research creates a linguistic corpus based on the doctoral and master's theses related to education sciences from the University of Nariño's digital repository. The Alberto Quijano Guerrero Library at the University of Nariño provides such information.

In addition, the linguistic corpus is enriched by using some open datasets related to the Word2Vect initiative, which are available in the Hugging Face repository. In this way, model architecture is created to be trained by the dataset due to the linguistic corpus, shown in Figure 1.

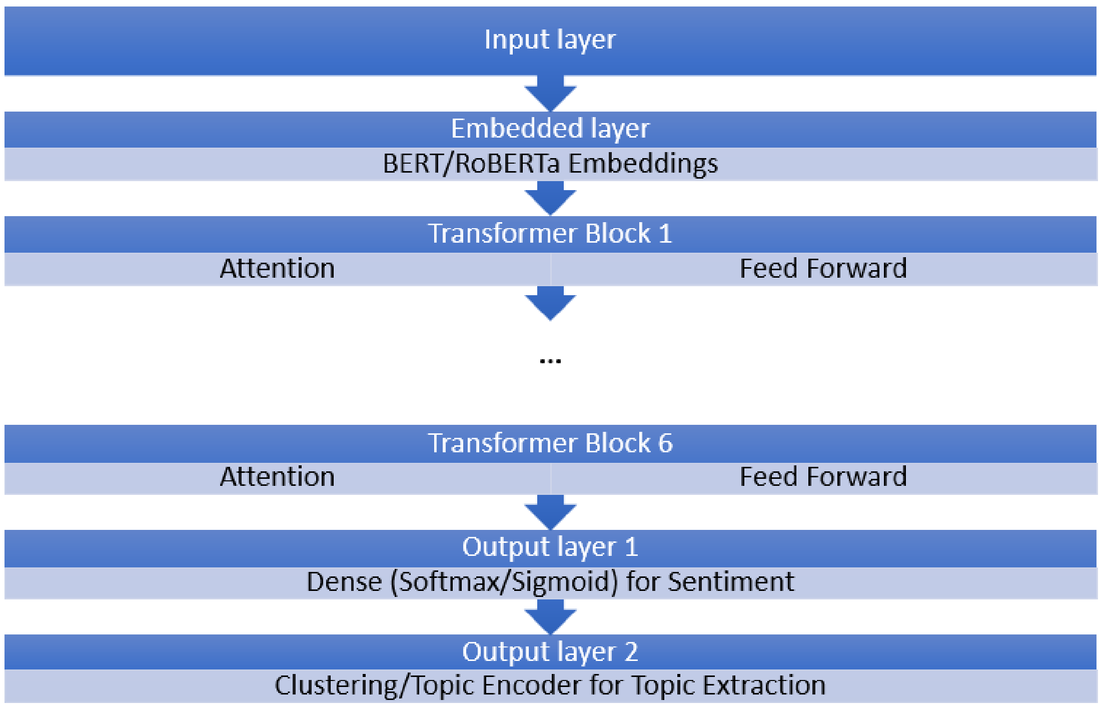

Figure 1.
Model architecture of the neural networks for the computational solution.

Once the neural network architecture has been created and deployed through the Galeras.NET Research Group, the training dataset is inserted into such a neural network model. The necessary links can be established to predict categories through learning iterations—epochs—. There are some general criteria for the training time of a neural network depending on the size of the linguistic corpus. In the case of this research, it has been possible to consolidate the size of the linguistic corpus close to 50 Mb of information. Therefore, the training time of the neural network, 10 minutes or 600 seconds, corresponds to the recommendations specified in Figure 2, which summarizes the average time required to perform well for an example dataset from the linguistic corpus on a local machine.

Dataset size	Average time to train
0 - 10 MB	10 sec
10 - 100 MB	10 min
100 - 500 MB	30 min
500 - 1 GB	60 min
1 GB+	3+ hours

Figure 2.
The average duration of neural network training time.
Source: This research is based on [17].

According to experience, the neural network was trained for approximately 10 minutes, during which up to 43 models based on pre-established algorithms could be evaluated. This information can be observed in Figure 3.

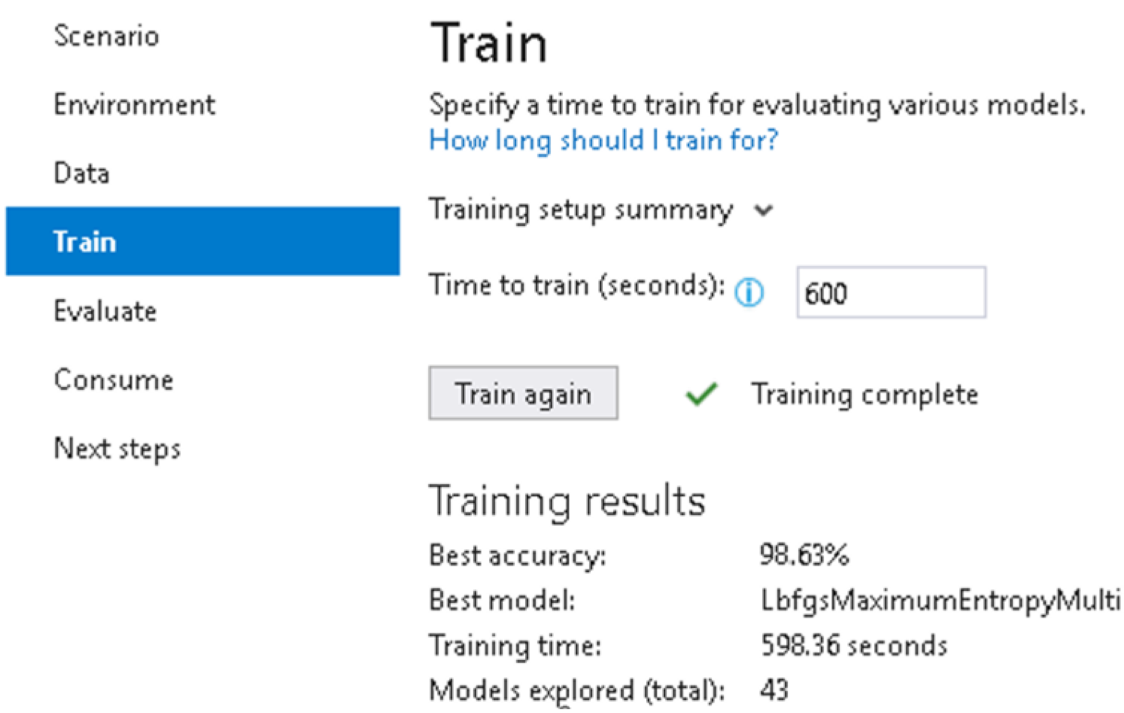

Figure 3.
Training of the neural network through the list of identified corpus structures.

The *FbfgsMaximumEntropyMultiClass* algorithm has been the best for the proposed computational solution applied to the proposed model in the training process. This algorithm embodies a maximum entropy model, an extension of linear logistic regression. The key distinction between the maximum entropy model and logistic regression lies in the classification problem's class capacity. Logistic regression is limited to binary classification, whereas the maximum entropy model accommodates several classes [26]. According to the datasets used in this research, such an algorithm is suitable for sentiment analysis and topic detection.

4. Results

The effectiveness of the computing solution is assessed via a usability study, which involves a range of tests varying in complexity. Six participants participated in this study. Nielsen, Turner, and Lewis stated that most issues will be discovered by a small group of users [27]. These people are developing their qualitative research and have some results from applying data collection instruments. Every participant underwent a structured test following a predefined procedure.

These guided tests consisted of 2 exercises using the computing solution, one for predicting feelings—sentiment analysis—and the other for predicting categories—topic detection—. After completing each guided test, each participant answered a survey regarding their experience with the computing solution. After completing the usability test, the participants completed a survey.

For the usability test, six students interacted with the computing solution for a specific test time between 15 and 20 minutes. To do this, students were asked to take transcripts of their interviews, and these text fragments were exposed to a computing solution. When processing the information, the computing solution showed results in detecting sentiments and identifying categories. A screen capture is shown in Figure 4.

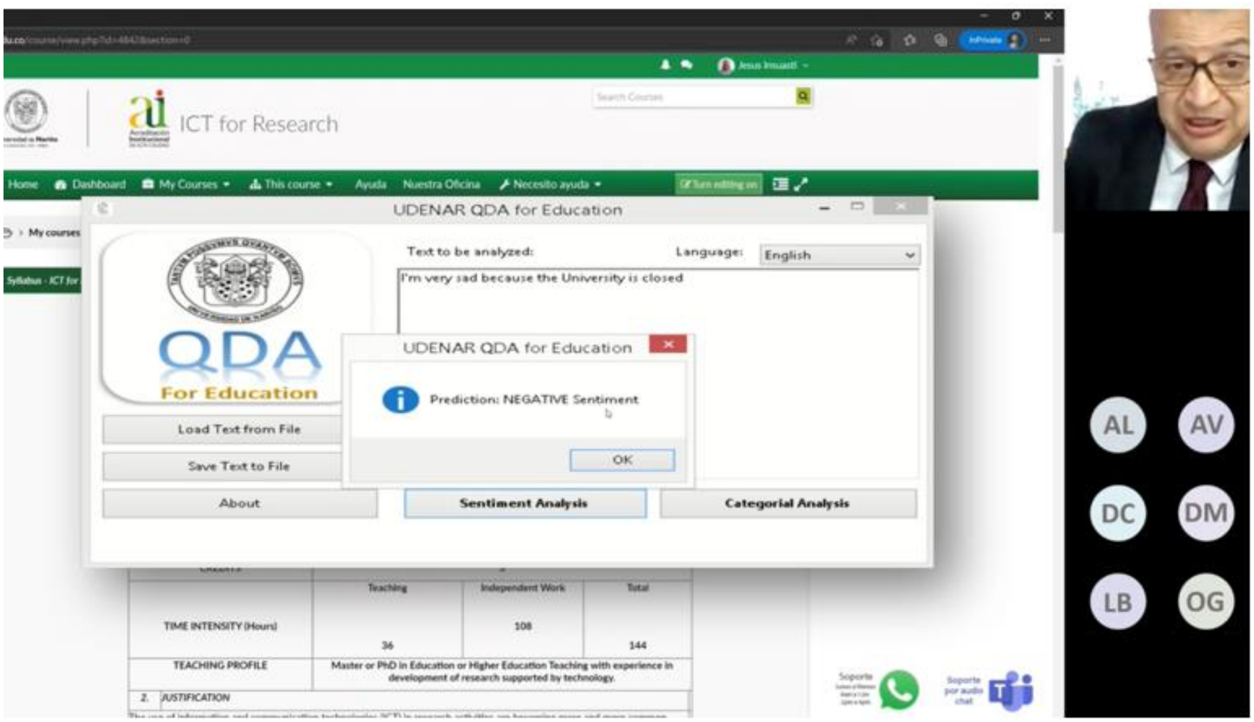

Figure 4.
Screen capture about the use of the computing solution in the classroom.

The students repeated this procedure until the agreed-upon time was fulfilled. At the end of the experiences, the students responded to an information collection instrument described below with four probes—questions—from P1 to P4.

P1. Evaluate the degree of functionality of the tool ({1} Poor, {2} Regular, {3} Good {4} Excellent)

P2. Did you like how the tool handled the information? ({1} It was chaotic, {2} It was

confusing, {3} It was neat, {4} I definitely liked it)
P3. How accurate is the tool? ({1} Very imprecise, {2} Imprecise, {3} Accurate, {4} Very accurate)
P4. Overall, how easy was it to use the tool? ({1} Very difficult, {2} Difficult, {3} Easy, {4} Very easy)

Subsequently, Figure 5 presents a summary of the usability test. The findings indicate that six master's students took part in this assessment. The coding uses sequential numbers for each participant, ranging from 01 to 06, based on their testing sequence, and their initials are denoted by capital letters.

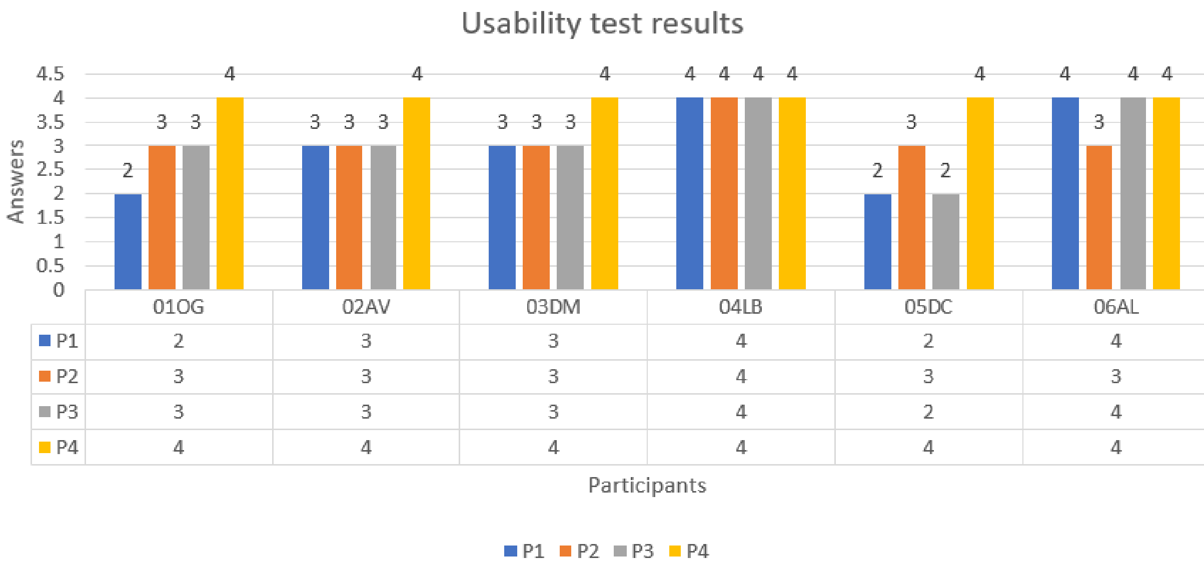

Figure 5.
Usability test results with tabular information.

5. Discussion & Conclusions

Considering the computational solution as an initial state, it is possible to extend its functionality and capacity according to the developments that can be made from this built solution. The fact that the computational solution has been deployed as a desktop application is a feature that can make its use massive through setup wizards for download. The nature of this research supports the processes of interpreting the results of data collection instruments that can be presented in qualitative research scenarios. In this order of ideas, the computational solution can be consumed by distinct types of users related to qualitative research. Only in the context of the University of Nariño is there a great possibility of using the computational solution beyond the master's program in Education since the University has an academic offer of both undergraduate and postgraduate programs that are related to Social Sciences, Human Sciences, Arts and Letters, among other types of programs where qualitative research is constituted as a fundamental axis in consequence with the epistemology of such sciences. Thus, the perspectives of the product regarding the possibility of use and growth of its functionality are feasible, not only at the University of Nariño but also at other universities.

Overall, the usability assessment yielded positive results. The findings indicate that the participants found the computing solution acceptable. Based on their responses, there was a consensus of positive feedback, indicating that the proposed computing solution is feasible and highly intuitive.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6: 5041-5049, 2024
DOI: 10.55214/25768484.v8i6.3133

6. Future Work

Considering the results of the experience in real scenarios, it is necessary to enhance the computational solution by expanding its capacity to work with plain text; that is, the tool must be able to process large volumes of textual information, such as the compiled results of all the individuals who were submitted to fill out data collection instruments.

On the other hand, increasing the functionality of the computational solution is considered necessary in the sense of allowing sentiment analysis and topic detection in non-textual sources of information. In this scenario, significant advantages could be exploited from using the computational solution when conducting linguistic analysis, multimedia digital sources of information such as audio and video recordings of interviews, and audio and video recordings from field observations, among others.

The computational solution is available for mass consumption through a website, and its setup wizard for desktop environments is available. In this sense, considerable expectations regarding future developments may arise.

Acknowledgment:

The authors of this paper wish to extend their appreciation to the computational language research group at the Universidad Nacional de Colombia in Medellin and the Galeras.NET research group at the University of Nariño. This research was self-funded, and the computational resources provided by the Galeras.NET research group at the University of Nariño were utilized. Furthermore, the authors would like to thank the master's students in Higher Education Teaching at the University of Nariño for their involvement in conducting the quasi-experiment.

References

[1] R. Kitchin, "Big Data, new epistemologies, and paradigm shifts," *Big Data & Soc.*, vol. 1, n.º 1, p. 205395171452848, April 2014. Disponible: https://doi.org/10.1177/2053951714528481

[2] E. Klein, S. Bird y E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

[3] C. D. Manning, *Foundations of Statistical Natural Language Processing*. Cambridge, Mass: MIT Press, 1999.

[4] V. Braun y V. Clarke, "Using thematic analysis in psychology", *Qualitative Res. Psychol.*, vol. 3, n.º 2, pp. 77–101, January 2006. Disponible: https://doi.org/10.1191/1478088706qp063oa

[5] D. Boyd y K. Crawford, "Critical questions for big data", *Inf., Communication & Soc.*, vol. 15, n.º 5, pp. 662–679, June 2012. Disponible: https://doi.org/10.1080/1369118x.2012.678878

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, y J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2013.

[7] A. Vaswani et al., "Attention is All you Need", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[8] B. G. Glaser y A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. in Observations (Chicago, Ill.). Aldine, 1967

[9] C. K. Riessman, "Narrative methods for the human sciences." *Narrative methods for the human sciences*. Sage Publications, Inc, Thousand Oaks, CA, US, 2008.

[10] S. M. Mohammad y P. D. Turney, "Crowdsourcing a word-emotion association lexicon", *Comput. Intell.*, vol. 29, n.º 3, pp. 436–465, September de 2012. Disponible: https://doi.org/10.1111/j.1467-8640.2012.00460.x

[11] B. Pang y L. Lee, "Opinion mining and sentiment analysis", *Found. Trends® Inf. Retrieval*, vol. 2, n.º 1–2, pp. 1–135, 2008. [online]. Disponible: https://doi.org/10.1561/1500000011

[12] D. Blei, A. Ng, y M. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, May 2003, Doi: 10.1162/jmlr.2003.3.4-5.993.

[13] F. Sebastiani, "Machine learning in automated text categorization", *ACM Comput. Surv.*, vol. 34, n.º 1, pp. 1–47, March 2002. [Online]. Disponible: https://doi.org/10.1145/505282.505283

[14] E. Hovy y J. Lavid, "Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics", *International Journal of Translation Studies*, vol. 22, pp. 13–36, January 2010.

[15] K. Crawford y T. Paglen, "Excavating AI: The politics of images in machine learning training sets", *Ai & soc.*, June 2021. Disponible: https://doi.org/10.1007/s00146-021-01162-8

[16] A. Joulin, E. Grave, P. Bojanowski, y T. Mikolov, "Bag of Tricks for Efficient Text Classification", *ArXiv*, vol. abs/1607.01759, 2016

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[18] B. Settles, "Active Learning Literature Survey". University of Wisconsin, Madison. 52, 2009

[19] J. Metcalf and K. Crawford, "Where are human subjects in Big Data research? The emerging ethics divide," *Big Data & Soc.*, vol. 3, n. º 1, p. 205395171665021, January 2016. Disponible: https://doi.org/10.1177/2053951716650211

[20] A. Radford, K. Narasimhan, T. Salimans y I. Sutskever. "Improving language understanding with unsupervised learning." OpenAI. [Online]. Disponible: https://openai.com/research/language-unsupervised

[21] B. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *J. Mach. Learn. Res.*, vol. 21, pp. 140:1-140:67, 2019, [Online]. Disponible in: https://api.semanticscholar.org/CorpusID:204838007

[22] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *ArXiv*, vol. abs/1907.11692, 2019

[23] V. Sanh, L. Debut, J. Chaumond, y T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", *ArXiv*, vol. abs/1910.01108, 2019.

[24] K. Clark, M.-T. Luong, Q. V Le, y C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators", *arXiv preprint arXiv:2003.10555*, 2020.

[25] D. Cer et al., "Universal Sentence Encoder", *ArXiv*, vol. abs/1803.11175, 2018.

[26] H.-F. Yu, F.-L. Huang, & C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," Machine Learning Journal, vol. 85, pp. 41–75, 2011. Disponible: https://doi.org/10.1007/s10994-010-5221-8

[27] C. Turner, J. Lewis, y J. Nielsen, "Determining Usability Test Sample Size", in *International Encyclopedia of Ergonomics and Human Factors*, vol. 12, 2006, pp. 3084–3088.