Twitbaiter: Model of Clickbait Detection in Spanish

Deyson Gómez Sánchez[1], Jeison D. Jimenez[1], Elizabeth Ruíz Padilla[2], Jairo E. Serrano[1], Juan C. Martinez-Santos[1] and Edwin Puertas[1]

[1]Universidad Tecnológica de Bolívar; School of Engineering, Architecture, and Design; Cartagena de Indias; 130013; Colombia
[2]Universidad de Cartagena; Faculty of Humanities; Linguistics and Literature Program; Cartagena de Indias; 130001; Colombia

Abstract

This paper addresses the binary classification of clickbait in Spanish tweets using three distinct approaches to improve detection. The datasets were provided as part of the TA1C challenge in IberLEF 2025, and the phases of pretraining, data preparation, fine-tuning, and evaluation were structured to train the models. Techniques such as undersampling and oversampling were used to handle class imbalance, aiming to generate a balanced dataset for model training. The three models evaluated were: RoBERTuito with manual features, RoBERTuito with class weighting, and Llama 3.2 – 3B. The results indicated that Model 3, based on Llama 3.2 – 3B, achieved the best performance, with an F1-score of 95.04%, outperforming the other models presented in this work. Furthermore, the model's robustness in a competitive environment was highlighted, as it ranked sixth in the TA1C challenge with a score of 80.115% on the evaluation dataset. While Model 1 showed competitive performance, its reliance on manual features limited its generalization capacity. Model 2 exhibited overfitting, emphasizing the importance of improving balancing and generalization techniques. This study demonstrates the effectiveness of advanced architectures like Llama 3.2 for the clickbait detection task and highlights areas for improvement in future implementations.

Keywords

Binary classification, Clickbait, Spanish tweets, Detection, TA1C challenge, RoBERTuito, Llama 3.2 – 3B, F1-score

1. Introduction

Clickbait detection has gained relevance in the field of Artificial Intelligence, particularly due to the impact this practice has on the information disseminated online. The clickbait phenomenon is characterized by headlines designed to generate curiosity and attract clicks, often at the expense of the quality and truthfulness of the information provided. In the context of the TA1C challenge at the IberLEF 2025 competition, this work focused on developing binary classification models to identify clickbait in Spanish tweets. Using datasets provided by the organizers, three models with different approaches were implemented, including the RoBERTuito model with manual features, RoBERTuito with class weighting, and the Llama 3.2 – 3B model. The evaluation of these models was carried out using performance metrics such as precision, recall, and F1-score, with the goal of finding the most effective model for the task. This study aims to contribute to the advancement of clickbait detection in Spanish, addressing the challenges of class imbalance and the need for models that effectively adapt to the variability of language in social networks.

2. State of the Art

As part of the background related to the linguistic analysis approach for detecting clickbait in Spanish, two studies directly addressing this topic have been identified (Robles, 2020; Loayza, 2024) [1] [2]. These investigations share a concern about unveiling the linguistic-discursive trends used for generating

IberLEF 2025, September 2025, Zaragoza, Spain
deygomez@utb.edu.co (D. G. Sánchez); jalvear@utb.edu.co (J. D. Jimenez); elizabethrp0818@gmail.com (E. R. Padilla); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)
0009-0005-2172-6905 (D. G. Sánchez); 0009-0001-0134-8426 (J. D. Jimenez); 0009-0009-1475-5184 (E. R. Padilla); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

clickbait in Spanish (hereafter CB) and its pragmatic functionality. The following will present the background that constitutes this trend in the order previously outlined.

Robles S. [2] conducts a discourse analysis focused on detecting trends in syntactic categories, which, according to the author, are selected to deploy pragmatic values effectively with the purpose of hooking and manipulating the reader's will. Thus, this article reveals a series of recurring linguistic-discursive configurations in CBs to achieve this goal. To this end, the author examines 540 headlines obtained between 2017 and 2019 from media outlets specializing in clickbait (Buzzfeed, Huffpost, and Upsolc).

This work is relevant as it offers an extensive repertoire of the prototypical word classes in CBs and presents a clear and concise idea of the pragmatic functionality of these syntactic selections.

Loayza E. [1] performs a discourse-pragmatic analysis in which, based on the proposals by Austin, Searle, and their followers, CBs are analyzed as directive illocutionary acts, highlighting the illocutionary force present in them. In this way, after analyzing CBs from videos shared on prominent social media platforms (Facebook, X, YouTube, and Instagram) collected randomly between 2021 and 2023, the author identifies the psychoemotional effects that CBs aim to generate and how these are linguistically designed to attract and retain the reader's attention.

Various studies have addressed the problem of clickbait detection from multiple approaches, incorporating both traditional feature extraction techniques and advanced deep learning models. The following is a chronological review of the most representative works in this field, focusing on those related to the use of Spanish and deep learning for headline classification.

The study by Omidvar et al. [3] was one of the first to apply recurrent neural networks with bidirectional GRU units for detecting clickbait on Twitter, achieving the best performance in the 2017 Clickbait Challenge using only the postText field as input. In the same competition, Wiegmann et al. [4] proposed an alternative approach based on Ridge regression and heuristic manual feature selection, highlighting that these methods, with proper feature engineering, could still be competitive against neural models.

Subsequently, Rajapaksha et al. [5] evaluated Transfer Learning models such as BERT, XLNet, and RoBERTa on Twitter headlines, concluding that RoBERTa consistently outperformed the other models, especially when using hidden outputs and structured fine-tuning strategies. This research laid important foundations for the use of pre-trained models in clickbait classification tasks.

More recently, Broscoteanu et al. [6] presented the RoCliCo corpus for clickbait in Romanian, along with a RoBERTa model trained with contrastive learning. The approach demonstrated a high capacity for modeling the semantic relationship between headlines and the body of the text, achieving notable F1 scores, especially in the non-clickbait class. The central idea of measuring semantic similarity between text components provided an innovative perspective to improve detection accuracy.

Meanwhile, Kydd et al. [7] developed "Deep Breath," a system that integrates machine learning with browser extensions to alert users about potentially misleading headlines. Although its evaluation accuracy was limited, the work stands out for its user-centric approach and real-time interaction.

In the Spanish-speaking context, García-Ferrero et al. [8] introduced NoticIA, a dataset for summarizing articles with clickbait headlines in Spanish, evaluating the performance of LLMs in ultra-summary tasks. Specifically trained models, such as ClickbaitFighter, outperformed LLMs in a zero-shot setup in terms of conciseness and accuracy, highlighting the importance of fine-tuning on domain-specific datasets.

The study by Mordecki et al. [9] significantly advances the field by proposing a revised and operational definition of clickbait based on the concept of curiosity gap. The authors developed TA1C (Te Ahorré Un Click), the first open-source dataset for Spanish clickbait detection, composed of 3,500 annotated tweets from 18 major media outlets. The dataset achieves high annotation consistency (Fleiss' K = 0.825) and supports baseline models reaching 0.84 in F1-score. This work is particularly relevant as it

offers a well-defined framework and benchmark for future models addressing clickbait detection in Spanish-language social media.

Later, Gamage et al. [10] proposed BaitRadar, a multi-model approach for detecting clickbait on YouTube by combining six sources of video information, achieving 98% accuracy. This study emphasizes the usefulness of integrating multiple heterogeneous signals to strengthen detection in audiovisual contexts.

Based on the information presented by Broscoteanu et al. [6], this work will implement a RoBERTa-based model for detecting clickbait in Spanish, applying fine-tuning strategies adjusted to the domain. Additionally, following the approach of Wiegmann et al. [4], manually selected heuristic features will be integrated into the model's input vector to improve its detection capability. Finally, the performance of at least one LLM will also be evaluated to analyze its effectiveness in comparison to the aforementioned strategies.

3. Data

For the development of this work, we used the datasets provided by the organizers of Task 1 in the TA1C: Clickbait Detection and Spoiling in Spanish competition [11], which is part of IberLEF 2025 [12]. This dataset is designed to determine whether the content of a tweet linking to a news article is clickbait or not, based on the definition of clickbait established by the organizers. Both the tweet and the linked news article are in Spanish, and the corpus was created with the aim of representing as many Spanish language varieties as possible, including news from media outlets in 12 different countries, as well as international sources. Considering that this is a binary classification problem, the distribution of each class is presented in Table 1.

Dataset	No (class 0)	Yes (class 1)	Total
TA1C_dataset_detection_train	2002	798	2800
TA1C_dataset_detection_dev_gold	497	203	700
TA1C_dataset_detection_test	-	-	700

Table 1
Distribution of datasets by clickbait classification

4. Methodology

This study follows a structured methodology consisting of pre-training, data preparation, fine-tuning, and evaluation phases for the binary classification of clickbait on Twitter. The pipeline used in the implementation of the models for this study is shown in Figure 1.

The pipeline comprises several stages. First, data preprocessing is performed to remove as much information as possible that may introduce noise into the classification model. To this end, URLs, mentions, emojis, hashtags, etc., were handled. Subsequently, class balancing techniques are applied, relevant features are extracted, and the model is trained using different techniques depending on the case.

To evaluate the model's performance, the metrics of accuracy, precision, recall, and F1-score were considered, with the latter being the primary metric. To assess the performance of the fine-tuned model, the dataset was split into 80% for training and 20% for testing.

We initially approached the clickbait detection task using traditional machine learning techniques. In particular, we trained an Extra Trees Classifier with a resampling strategy that combined random

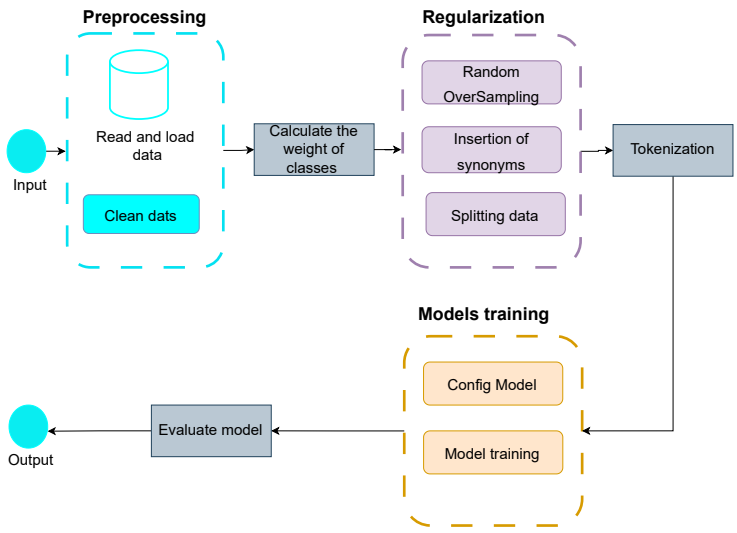

Figure 1: System Pipeline.

oversampling and undersampling to address class imbalance. This model achieved an F1-score of 0.416. We then implemented a CatBoost Regressor under similar conditions, which improved performance to an F1-score of 0.664. However, both results remained below the levels required for a robust solution. Given these limitations, we shifted our focus to transformer-based models, which have demonstrated superior performance in natural language processing tasks.

For this study, we selected the three best-performing transformer architectures. The pipeline in Figure 1 is common to the three evaluated models, although each presents specific variations in data processing, feature extraction, and model architecture. The following section describes the specific details of each implemented model.

4.1. Model 1: RoBERTuito with Hybrid Features

In this work, an initial dataset was used, consisting of a training dataset whose structure is detailed in Table 1. The unequal distribution between the classes was a common challenge in the context of binary classification, as the minority class, "Clickbait," was less represented compared to the majority class. To address this imbalance, two main techniques were used: undersampling and oversampling. The undersampling technique involved reducing the size of the majority class, while oversampling was implemented to increase the number of samples from the minority class. In this case, oversampling generated 702 duplicates of the "Clickbait" class, increasing its representation within the dataset. As a result of these techniques, the final dataset consisted of 3,000 examples, with a balanced 50% distribution for each class.

The model was a combination of representations generated by the RoBERTuito uncased model described by Pérez et al. (2022) [13] and manually extracted features from the text. Specifically, eight numerical features were extracted based on 128 keywords or phrases associated with clickbait. These words were determined through the analysis of the works of Loayza Maturrano (2024) and Sararobles Ávila (2020) [1] [2], who identified the most representative terms in the context of clickbait.

The manually extracted features were as follows:

1. **Clickbait word count**: The number of times a word from the list appears in the text.
2. **Clickbait ratio**: The percentage of words in the text that are part of the clickbait list.

3. **Boolean presence**: A binary indicator that determines whether the text contains at least one clickbait word.
4. **Numbers**: A binary indicator indicating whether the text contains numeric digits.
5. **Exclamations**: A binary indicator indicating whether the text contains exclamation marks (!).
6. **Question marks**: A binary indicator indicating whether the text contains question marks (?).
7. **Uppercase letters**: The proportion of uppercase letters in the text.
8. **Length**: The total number of words in the text.

Once these eight manual features were extracted, they were combined with the 768 dimensions generated automatically by the RoBERTuito model. The concatenation of both representations resulted in a final 776-dimensional vector, which was used for the binary classification of the texts.

The RoBERTuito uncased model corresponds to a variant of the RoBERTa architecture. This model was pre-trained using a corpus of 500 million Spanish-language tweets, providing it with an excellent ability to capture specific patterns from the Twitter platform, such as mentions, emojis, hashtags, and diverse content. Additionally, RoBERTuito has demonstrated its effectiveness in previous classification tasks, such as hate speech detection in the SemEval 2019 Task 5, HatEval dataset, sentiment and emotion analysis in the TASS 2020 dataset, and irony detection in the IrosVa 2019 dataset.

4.2. Model 2: RoBERTuito + Class Weighting

The dataset used, like in the previous model, consists of the training dataset shown in Table 1. In the preprocessing stage, as in the previous model, URLs and representations of mentions, emojis, hashtags, and diverse content were left to the RoBERTuito uncased tokenizer. To address this imbalance, the class weighting technique was used instead of increasing the data through data augmentation techniques. The class weights were calculated using the compute_class_weight function from sklearn [14], with the "balanced" strategy to adjust to the class distribution in the training set.

4.3. Model 3: Llama 3.2 - 3B

In this study, the meta-llama/Llama-3.2-3B model [15] was fine-tuned. The dataset used consisted of merging the training and dev gold datasets, shown in Table 1. Data cleaning was carried out through a function that encompasses all the aspects of the text mentioned earlier. To address the class imbalance, the class weights were first calculated for the original dataset to apply class weights, disproportionately penalizing errors made in the minority class.

Subsequently, random oversampling was used, randomly duplicating 499 examples from the minority class to raise its representation to 1,500 examples. As a third strategy, semantic data augmentation was applied based on synonym replacement using WordNet, with a modification rate of 30%, focusing only on words longer than four letters. This technique generated 600 additional synthetic examples, increasing the "Clickbait" class to a total of 2,100 entries. The model was quantized to 4-bit (QLoRA) with alpha=32 over 7 transformer modules for efficiency, with 2.94% of trainable parameters (97M of 3.31B).

5. Evaluation

We evaluated the effectiveness of the models based on the metrics previously mentioned, along with the training and validation losses. Once the data was read and processed, the class weighting obtained is as stipulated in Table 2.

Model	Class	Weight
Model 2	Class 0	0.699
	Class 1	1.754
Model 3	Class 0	0.700
	Class 1	1.748

Table 2
Class weights by model

Taking into account this data, it is noted that the model will indeed give more weight to errors in class 1 because it is the class with fewer data. Model 1 was trained for 5 epochs with a decay rate of 0.01, while models 2 and 3 were trained for 10 epochs with the same decay rate and early stopping with a patience of 3 evaluations. The training results of the three models are as follows:

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.5271	0.4508	0.8050	0.8047	0.8037	0.8041
2	0.2623	0.2493	0.8983	0.8985	0.8998	0.8983
3	0.1668	0.2783	0.9167	0.9233	0.9134	0.9156
4	0.0416	0.2623	0.9450	0.9462	0.9438	0.9447
5	0.0080	0.3224	0.9467	0.9477	0.9455	0.9464

Table 3
Training metrics - Model 1

The results presented in Table 3 show that the model consistently improved in performance. The training loss decreased from 0.5271 in the first epoch to 0.0080 in the fifth, while the validation loss dropped from 0.4502 to 0.3224. Accuracy, recall, and F1 also progressively improved, reaching 94.67%, 94.57%, and 0.9464, respectively, by the end of training. These metrics reflect a good model fit, with a reduction in both training and validation loss, indicating an improvement in the model's ability to generalize.

Feature	Clickbait	Normal	Diff
Clickbait words	1.809	1.658	0.159
Has questions	0.204	0.021	0.183
Has numbers	0.253	0.362	-0.109
Number of words	23.484	24.522	-1.037
Has clickbait	0.852	0.832	0.020
Clickbait ratio	0.085	0.072	0.013
Uppercase ratio	0.047	0.054	-0.007
Has exclamations	0.018	0.013	0.005

Table 4
Feature importance analysis - Class averages

Additionally, Table 4 shows that the most decisive features for distinguishing clickbait content include both the frequency of specialized vocabulary words, with an average of 1.8 per text compared to 1.65 in normal texts. Regarding emotional and formatting patterns, there is a higher usage of exclamation marks (81.8% vs 61.3%) and uppercase letters (4.7% vs 5.4%). In contrast, non-clickbait texts tend to include more figures (36.2% vs 25.3%), suggesting a more informative and factual approach.

According to the data in Table 5, the training of Model 2 showed consistent improvement in accuracy, F1-score, and recall, with a notable decrease in training loss (from 0.2207 to 0.0001), indicating memorization, and validation loss (from 0.3293 to 0.6235). The model stopped at step 600 (epoch 8/10). The key metrics remained stable and high, indicating that the model generalizes well to validation data and has achieved a good balance between the classes.

Finally, according to the data in Table 6, Model 3 showed rapid convergence, reducing the training loss

Step	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
100	0.220700	0.332985	0.910714	0.910378	0.910714	0.910532
200	0.065300	0.398182	0.908929	0.914590	0.908929	0.910483
300	0.017500	0.527177	0.914286	0.913964	0.914286	0.914111
400	0.000400	0.613735	0.908929	0.908751	0.908929	0.908836
500	0.000700	0.669383	0.903571	0.902391	0.903571	0.902756
600	0.000100	0.623510	0.910714	0.911097	0.910714	0.910892

Table 5
Training metrics - Model 2

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.341900	0.392726	0.888043	0.848352	0.919048	0.882286
2	0.346200	0.308347	0.928261	0.933824	0.907143	0.920290
3	0.094400	0.309100	0.933696	0.899777	0.961905	0.929804
4	0.010500	0.323472	0.944565	0.924138	0.957143	0.940351
5	0.021400	0.376201	0.954348	0.943662	0.957143	0.950355
6	0.000400	0.362376	0.951087	0.931034	0.964286	0.947368
7	0.000000	0.372603	0.946739	0.922551	0.964286	0.942957

Table 6
Training metrics - Model 3

to nearly zero. The validation loss reached its minimum during the second and third epochs before showing a slight increase, which may indicate potential overfitting. Nevertheless, the final validation metrics F1-score: 94.73%, recall: 96.43%, along with corresponding accuracy and precision demonstrate solid performance and good generalization ability on the validation set.

5.1. Model testing

Once the training was completed, the trained model was evaluated using the test set, corresponding to 20% of the data initially reserved, resulting in the metrics shown in Table 7.

Model	Accuracy	Precision	Recall	F1
Model 1	0.9467	0.9477	0.9455	0.9464
Model 2	0.9143	0.9140	0.9143	0.9141
Model 3	0.9543	0.9437	0.9571	0.9504

Table 7
Model evaluation results comparison

The performance analysis on the test set reveals that Model 3 (Llama 3.2 – 3B) achieved the best overall results, with an F1 score of 0.9504, outperforming Models 1 and 2, which achieved values of 0.9464 and 0.9141, respectively. Model 1 (RoBERTuito with Hybrid Features), although not reaching the performance of Model 3, demonstrated competitive performance (F1 = 0.9464), suggesting that the combination of RoBERTuito model embeddings provides significant value in detecting clickbait patterns. This hybrid approach enhances the model's ability to capture subtle signals that are not easily representable in the embedding space without the strategy.

On the other hand, Model 2 (RoBERTuito with Class Weighting), which used class weighting to mitigate the dataset imbalance, showed the lowest performance (F1 = 0.9141). Although the weighting technique aims to favor the minority class during training, in this case, it was not as effective as the previous strategies, possibly due to an underrepresentation of relevant patterns that were not sufficiently captured by the adjusted weights.

5.2. Competition Evaluation

The results revealed that the strategies implemented by our team, VerbNex, identified as "gsdeyson," ranked 6th among the participating teams in the TA1C challenge at IberLEF 2025 Subtask 1. The model's performance metrics on the evaluation dataset provided by the competition organizers are shown in Table 8. The Table 9 presents the official ranking of participants in Subtask 1 of the challenge.

Model	Result
Model 1	0.7666
Model 2	0.7878
Model 3	0.8011

Table 8
Model results on competition evaluation dataset

Rank	Participant ID	User	Task 1 Score
1	973378	tomasbernal01	0.81564
2	972330	escom	0.81525
3	972709	viahes	0.81482
4	973125	dcere	0.80480
5	973423	julian_zsa	0.80347
6	973069	gsdeyson	0.80115

Table 9
Competition ranking - Team performance

The results confirm the trend observed previously in the validation sets. Model 3 (Llama 3.2 – 3B) achieved the best performance, with a score of 0.8011, establishing it as the most effective architecture for the task in our implementations. It was followed by Model 2 (RoBERTuito + Class Weighting) with a score of 0.7878, showing an improvement compared to its previous performance, suggesting that the class weighting adjustment may have had a more positive effect in this setting. Finally, Model 1 (RoBERTuito with Hybrid Features) reached a result of 0.7666, the lowest among the three, indicating that while its hybrid approach was competitive in the test environment, it failed to maintain the same level of effectiveness against a more diverse competitive set. These findings reinforce the robustness of Model 3 and its adaptability for clickbait classification.

According to the data in Table 9, the difference with the fourth and fifth positions was marginal—less than 0.003 points—which highlights the competitiveness of the proposed system. This result supports the effectiveness of Model 3 (Llama 3.2 – 3B) and its consistency compared to systems developed by other research groups, consolidating a solid foundation on which incremental improvements can be applied to climb rankings in future editions of the competition.

6. Conclusions

The work developed for the binary classification of clickbait on Twitter has shown interesting results in terms of the performance of the evaluated models. Although the technique employed to handle class imbalance, specifically through undersampling and oversampling, achieved an acceptable balance in the dataset, it was observed that the implementation of these methods could be improved to maximize their effectiveness.

Model 1, which combines the RoBERTuito model with manually defined features based on a predefined vocabulary, achieved acceptable performance during internal testing. However, its performance in the competitive evaluation was lower compared to other models, reflecting limitations in its generalization

capability. The reliance on the predefined keyword vocabulary turned out to be a restrictive factor, as not all relevant expressions in the competition dataset were covered. Despite these drawbacks, the model shows high improvement potential through the dynamic incorporation of new words and the integration of strategies such as embeddings or contextual synonyms, which would allow better adaptation to new domains and greater robustness against linguistic variations in clickbait.

Model 2 showed good overall performance, with stable accuracy, particularly in the weighted F1 score, which remained close to 91%. However, the model exhibited clear overfitting, reflected in the increase of validation loss while the training loss decreased. This phenomenon suggests that the model may have memorized specific patterns from the training set, which affects its ability to generalize. Furthermore, it showed a bias towards the majority class, with suboptimal performance in detecting the minority class (clickbait), where recall only reached 83.77%. To improve, it would be crucial to implement additional techniques such as threshold optimization and the use of SMOTE to increase the diversity of examples. Additionally, performing a detailed error analysis will help identify patterns in incorrect predictions, which could fine-tune the detection ability for complex cases.

Model 3, based on the Llama 3.2 model and trained with the QLoRA strategy, demonstrated excellent performance in clickbait classification, achieving a near-perfect balance between precision and recall for both classes. The use of an effective data balancing strategy significantly improved the model's ability to handle imbalanced classes. Although the model only completed 7 of the 8 planned epochs due to early stopping activation, no performance degradation was observed.

7. Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 for grammar, spelling, and translation assistance. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. Loayza Maturrano, Los títulos de los clickbait en las redes sociales desde la teoría de los actos de hablathe clickbait titles in social networks from speech act theory, Tierra Nuestra 18 (2024) 35–46. doi:10.21704/rtn.v18i1.1867.

[2] S. R. Ávila, El clickbait: clases de palabras para la construcción de un titular engañoso, in: [2] 107–124.

[3] A. Omidvar, H. Jiang, A. An, Using neural network for identifying clickbaits in online news media, 2018. URL: https://arxiv.org/abs/1806.07713. arXiv:1806.07713.

[4] M. Wiegmann, M. Völske, B. Stein, M. Hagen, M. Potthast, Heuristic feature selection for clickbait detection, 2018. URL: https://arxiv.org/abs/1802.01191. arXiv:1802.01191.

[5] P. Rajapaksha, R. Farahbakhsh, N. Crespi, Bert, xlnet or roberta: The best transfer learning model to detect clickbaits, IEEE Access 9 (2021) 154704 – 154716. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120472076&doi=10.

1109%2fACCESS.2021.3128742&partnerID=40&md5=b335f674b85faab3690a249980e74b11. doi:10.1109/ACCESS.2021.3128742, cited by: 46; All Open Access, Gold Open Access.
[6] D.-M. Broscoteanu, R. T. Ionescu, A novel contrastive learning method for clickbait detection on roclico: A romanian clickbait corpus of news articles, 2023. URL: https://arxiv.org/abs/2310.06540. arXiv:2310.06540.
[7] M. Kydd, L. A. Shepherd, Deep breath: A machine learning browser extension to tackle online misinformation, 2023. URL: https://arxiv.org/abs/2301.03301. arXiv:2301.03301.
[8] I. García-Ferrero, B. Altuna, Noticia: A clickbait article summarization dataset in spanish, 2024. URL: https://arxiv.org/abs/2404.07611. arXiv:2404.07611.
[9] G. Mordecki, G. Moncecchi, J. Couto, Te ahorré un click: A revised definition ofnbsp;clickbait andnbsp;detection innbsp;spanish news, in: Advances in Artificial Intelligence – IBERAMIA 2024: 18th Ibero-American Conference on AI, Montevideo, Uruguay, November 13–15, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2025, p. 387–399. URL: https://doi.org/10.1007/978-3-031-80366-6_32. doi:10.1007/978-3-031-80366-6_32.
[10] B. Gamage, A. Labib, A. Joomun, C. H. Lim, K. Wong, Baitradar: A multi-model clickbait detection algorithm using deep learning, 2025. URL: https://arxiv.org/abs/2505.17448. arXiv:2505.17448.
[11] G. Mordecki, L. Chiruzzo, R. Laguna, J. Prada, A. Rosá, I. Sastre, G. Moncecchi, Overview of TA1C at IberLEF 2025: Detecting and Spoiling Clickbait in Spanish-Language News, Procesamiento del Lenguaje Natural 75 (2025).
[12] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.
[13] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.
[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[15] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

A. Online Resources

The GitHub repository containing the implementation and resources of this work is available via:

- GitHub