Data in Brief 53 (2024) 110220

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of controversial news posts in Spanish from the reader's perspective

Cesar Macias[a], Hiram Calvo[a,*], Omar Juárez Gambino[b]

[a] *Centro de Investigación en Computación, Instituto Politécnico Nacional. Av. Juan de Dios Bátiz, esq. Miguel Othón de Mendizabal, Col. Nueva Industrial Vallejo, Alcaldía Gustavo A. Madero, C.P. 07700, CDMX, México*
[b] *Escuela Superior de Cómputo, Instituto Politécnico Nacional. Av. Luis Enrique Erro S/N, Unidad Profesional Adolfo López Mateos, Zacatenco. Alcaldía Gustavo A. Madero C.P. 07738, CDMX, México*

A R T I C L E I N F O

Article history:
Received 19 December 2023
Revised 14 February 2024
Accepted 14 February 2024
Available online 23 February 2024

Dataset link: Dataset of controversial news posts in Spanish (Original data)

Keywords:
News controversy
News articles
X data
Readers perspective

A B S T R A C T

This paper presents a corpus of Spanish news posts obtained from X with the annotation of controversy made via crowdsourcing. A total of 60 tweets were obtained from 8 different newspapers. For the annotation task, a survey was developed and sent to 31 different participants to answer it with the controversy level they perceived from the news post summary and headline presented on the post. The most frequent selected option was assigned as the initial controversy level of the post. The final annotation of the corpus was made via an analysis of the raw data by computing the Inter Annotator Agreement (IAA). The analysis showed that the binarization of the data was the most convenient way to annotate it. A potential use for this dataset is detailed in further sections.

Specifications Table

Subject	Artificial intelligence
Specific subject area	Controversy detection
Data format	Raw, Analyzed

(*continued on next page*)

* Corresponding author.
E-mail addresses: cmaciass2021@cic.ipn.mx (C. Macias), hcalvo@cic.ipn.mx (H. Calvo), jjuarezg@ipn.mx (O.J. Gambino).
Social media: @MACI_dev_96 (C. Macias)

https://doi.org/10.1016/j.dib.2024.110220

Type of data	Table, CSV file
Data collection	Data crawled on X posts written in Spanish from official Mexican newspaper accounts
Data source location	Institution: Instituto Politécnico Nacional City/Town/Region: Mexico City Country: Mexico
Data accessibility	Repository name: GitHub DOI: https://zenodo.org/doi/10.5281/zenodo.10516303 https://github.com/MACI-dev-96/Corpus/tree/02d9dd0e67552933a2759582b4761307f41fe234/Data_in_Brief
Related research article	Macias, C., Calvo, H., Gambino, O.J. (2022). News Intention Study and Automatic Estimation of Its Impact. In: Pichardo Lagunas, O., Martínez-Miranda, J., Martínez Seis, B. (eds) Advances in Computational Intelligence. MICAI 2022. Lecture Notes in Computer Science(), vol 13613. Springer, Cham. https://doi.org/10.1007/978-3-031-19496-2_7

1. Value of the Data

- Provided data can be used to explore the controversy of X news articles posts.
- This corpus represents a novelty in the controversy classification task, the way to classify controversy until now was by using lexicons like in [1] and [2], or by computing the general polarity of the text [3] and all those perspectives used texts written in English.
- This data can be used to set a benchmark to predict news articles controversy.
- Due to the lack of information about controversy classification of X news articles in Spanish, this is a very valuable resource for further insights, development of experiments or both.

2. Data Description

The GithHub repository contains the file Corpus_of_controversial_news_tweets_in_Spanish_from_the_readers_perspective.csv

Here, the collected and analyzed data is included. This file has two columns for each row. Column names are “TWEET_ID” which contains the ID of the analyzed X post (a.k.a. tweet) and “CLASS” that contains numeric integer indicators for the corresponding class for each analyzed post. Table 1 describes the class label associated to each value.

Fig. 1 displays the distribution of classes in the corpus.

3. Experimental Design, Materials and Methods

To collect the data, we used the official X API. The process followed is described below.

3.1. News Item Selection

Numerous online platforms serve as repositories of information, with predominant instances involving individuals disseminating comprehensive news narratives through social networks or

Table 1
Associated class label to each value.

Class label	Value
Not controversial	0
Controversial	1

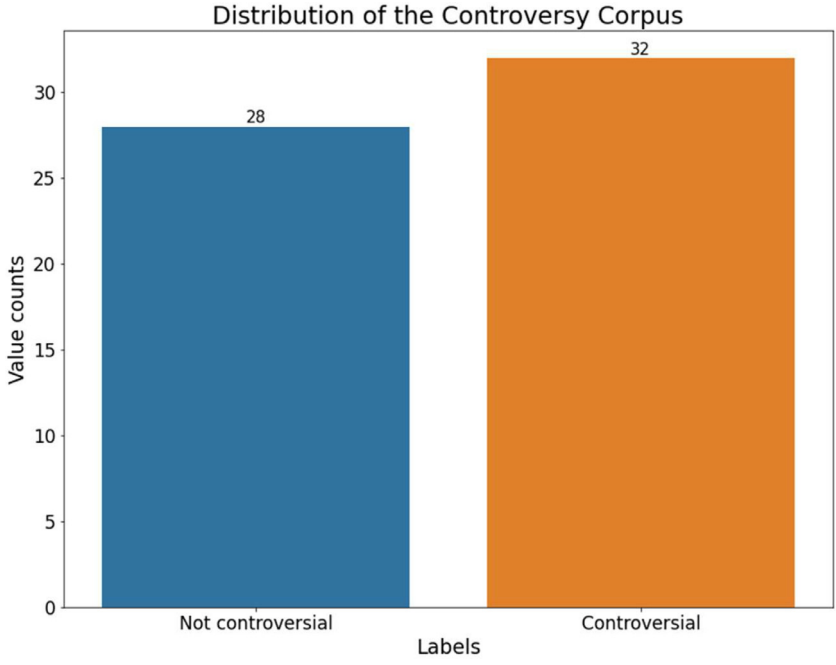

Fig. 1. Distribution of the corpus.

personal blogs and websites. Such entities may consist of independent researchers, as well as individuals who, having perused a news story, recontextualize it before posting their adapted versions, thereby contributing to the proliferation of misinformation or deceptive content. In contrast, established newspapers, typically maintaining an online presence, distribute condensed news segments of pertinent information on social media platforms, complemented by URLs directing readers to the full articles on their respective websites.

The selection of information sources was contingent upon adherence to specific criteria, wherein preference was given to official newspapers (of national or international stature) that communicate news content in the Spanish language. Furthermore, the chosen sources were required to possess dedicated web pages and maintain a presence on the platform X. This selection criterion aimed to concentrate on newspapers characterized by a diminished propensity for disseminating fake news or engaging in disinformation practices. Following a meticulous selection process, the chosen sources included El Universal, La Jornada, CNN en Español, BBC News Mundo, Milenio, Reforma, and Proceso.

After the meticulous curation of information sources, a comprehensive corpus was constructed through manual extraction of 60 news items. The selection process prioritized articles addressing contentious subjects such as politics, economy, and security. Additionally, an inclusive approach encompassed the incorporation of non-controversial news topics, spanning domains such as science, technology, and culture.

3.2. The Annotation Process

Once we had our news collection, the next step was to annotate them. For this task, we developed a Google Forms questionnaire, to simulate the process by which an individual reads a news segment published on X and based on that information, shares their comments. Often the complete news article is not consulted before passing judgement. Therefore, the survey

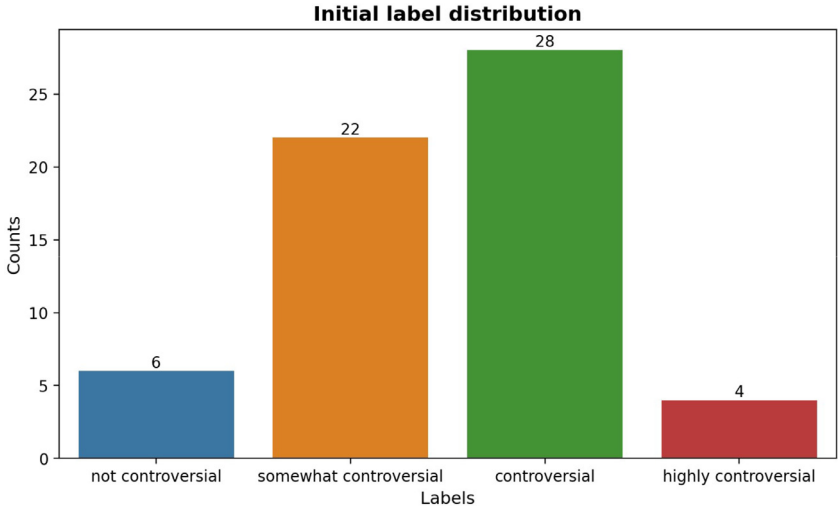

Fig. 2. Initial label distribution of the corpus.

presented only the news excerpt published by the information source, and participants were instructed to respond to the survey considering the following instructions.

1. In accordance with the following news segments, please select an option that best describes, in your opinion, the level of controversy (contentiousness) of the news given the following definition. Controversiality: discussion of opposing opinions among two or more individuals.
2. Controversiality arises from the divergence of opinions in the comments or responses to the news. In other words, the more divided you believe the comments will be, the higher the level of controversiality.

3.3. Annotators Could Choose Between Four Options

- Not controversial
- Somewhat controversial
- Controversial
- Highly controversial

A total of 31 individuals participated in the labelling process by responding to the survey. While a preference was given, it was not restrictive, for the educational attainment to be at least high school level. The average age of the respondents was 25 years. No personal or sensitive data of the survey participants were collected.

3.4. Data Analysis

Once the annotation process finished, we proceeded to collect the answers to analyse them. To allocate the initial label to each text, the methodology adopted involved assigning the label with the highest frequency of votes. After the establishment of the initial label, their distribution was as shown in Fig. 2.

In this stage we compute the Inter Annotator Agreement (IAA) to assign the final class to each post. To evaluate the IAA, we used the Cohen's kappa coefficient. This coefficient evaluates

Table 2
IAA metrics.

Metrics	κ (quaternary)	κ (binary-1)	κ (binary-2)
Mean	0.074	0.143	0.145
Standard deviation	0.096	0.155	0.180
Minimum agreement	-0.179	-0.358	-0.223
Maximum agreement	0.386	0.596	1.0

the agreements between the annotators of a corpus and is defined by the following equation.

$$\kappa = (p_o - p_e)/(1 - p_e)$$

Where p_o is the empirical probability of agreement in the label assigned to any sample and p_e is the expected agreement from the labelling process. This coefficient varies in the range $[-1,\ 1]$ and the value of the coefficient indicates whether there was a good level of agreement $\kappa \geq 0.8$ or that there was no agreement at all between annotators $\kappa \leq 0$. This statistic can only be calculated for the binary case (two annotators), so we had computed the κ value for every possible combination of annotators.

We contemplated three potential scenarios for the final label assignment. These include the distribution of labels in the initial corpus (quaternary-label case), and two binary redistributions: binary-1 and binary-2. The redistribution process for the binary cases is delineated as follows.

Binary-1:

- {not-controversial, somewhat controversial, controversial} → {not-controversial}
- {highly controversial} → {controversial}

Binary-2:

- {not-controversial, somewhat controversial} → {not-controversial}
- {controversial, highly controversial} → {controversial}

One drawback of binary-1 distribution was the considerable data imbalance, with 56 instances in the not-controversial class compared to only 4 instances in the controversial class. In contrast, the binary-2 distribution is more balanced, with 28 instances in the not-controversial class and 32 instances in the controversial class. The results of the IAA analysis are detailed in Table 2.

After a thorough analysis of the Inter-Annotator Agreement (IAA) results, it was observed that the level of agreement exhibited an increment upon binarizing the corpus. Notably, between binary-1 and binary-2, both the mean and maximum agreement levels demonstrated an augmentation, while the minimum agreement level experienced a reduction (positive for this analysis). Consequently, the decision was made to adopt the binary-2 distribution as the definitive label assignment for the corpus, as illustrated in Fig. 1.

3.5. Comparative Analysis

Here, a comparison is presented with existing datasets and the methods employed in their creation. The datasets presented herein are comprised of corpora of written text in the English language.

In [4], the dataset was generated with the aim of identifying controversy on the web. This methodology involved the utilization of Wikipedia articles as seeds to retrieve web pages associated with the general theme of the article through the acquisition of their nearest neighbours. Seed articles were selected based on their degree of controversy, ranging from non-controversial to clearly controversial. In total, 377 pages related to the 41 seed topics were obtained. The corpus was designated as follows: 1 – clearly controversial, 2 – possibly controversial, 3 – possibly non-controversial, or 4 – clearly non-controversial. For this approach, a minimum of two

annotators were enlisted for each of the pages and the topics addressed. Based on this work, the authors automated the process of labelling the Wikipedia articles and the extraction of web controversial pages in [5], this dataset contains 377 webpages and 8755 Wikipedia articles.

In [2], they developed a lexicon of controversial words. They compiled a list of words considering the 2000 most frequent words after filtering out stop words. Subsequently, they performed labelling using crowdsourcing platforms. The words were extracted from news articles published on NewsCred, and annotators, instructed to be familiar with U.S.A. news, were tasked with determining whether a word was Strongly Controversial, Somewhat Controversial, Less Controversial, or Non-Controversial within that context, following a set of instructions. In this manner, they obtained their lexicon.

It is evident that the available datasets for the classification of controversial texts are limited. Despite efforts to detect the controversial nature of news articles, there is a notable absence of initiatives specifically targeting the identification of controversy in Spanish-language news. Furthermore, research considering the readers' perspective during the labelling of controversy is lacking. Therefore, we assert that our contribution holds significance for the scientific community. Notably, labelling techniques have predominantly employed crowdsourcing, whether for the annotation of entire texts or the development of lexicons based on words.

3.6. Future Work

The dataset is inherently limited by its size, raising considerations about the generalizability of controversy through its application. Nevertheless, it provides a groundwork for future research endeavours. By utilizing this dataset and adhering to the procedural methodology outlined for its construction, the scientific community has the opportunity to introduce innovative approaches to tackle the task of identifying controversy in news. Certain limitations stem from the terms and conditions governing the use of the X API, as well as constraints related to the quantity and nature of extractable data. It is advised to harness news thread discussions for additional refinement of classification models, as exemplified in the related research article.

Limitations

Not applicable.

Ethics Statement

The dataset from X is fully anonymized, ensuring that individual users cannot be identified. We followed the ethical guidelines provided by X. The privacy and confidentiality of users are thereby preserved, preventing any impact on the individuals from the data used in this research. Regarding the annotation, no personal or sensitive data of the survey participants was collected. Consequently, the nature of this anonymized data collection did not necessitate ethical approval. This sample lacks representativeness for generalizing conclusions or applying solutions, which limits the validity and ethics of their implementation in wider contexts. Caution is therefore required when considering any practical applications based solely on this dataset.

Data Availability

Dataset of controversial news posts in Spanish (Original data) (GitHub).

CRediT Author Statement

Cesar Macias: Conceptualization, Data curation, Writing – original draft, Visualization, Formal analysis, Methodology; **Hiram Calvo:** Conceptualization, Methodology, Supervision, Writing – review & editing, Resources; **Omar Juárez Gambino:** Conceptualization, Methodology, Supervision, Writing – review & editing, Resources.

Acknowledgements

This research was funded by CONAHCYT and Instituto Politécnico Nacional (IPN).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2024.110220.

References

[1] M. Pennacchiotti and A.M. Popescu, "Detecting controversies in Twitter: a first study," in NAACL HLT 2010 workshop on computational linguistics in a world of social media, 2010.
[2] Y. Mejova, A. X. Zhang, N. Diakopoulos and C. Castillo, Controversy and sentiment in online news, arXiv, 2014.
[3] J. Hessel and L. Lee, "Something's brewing! Early prediction of controversy-causing posts from discussion features", arXiv, 2019.
[4] S. Dori-Hacohen, J. Allan, Detecting controversy on the web, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13), Association for Computing Machinery, New York, NY, USA, 2013, pp. 1845–1848, doi:10.1145/2505515.2507877.
[5] S. Dori-Hacohen, J. Allan, Automated controversy detection on the web, in: A. Hanbury, G. Kazai, A. Rauber, N. Fuhr (Eds.), Advances in Information Retrieval. ECIR 2015. Lecture Notes in Computer Science, Springer, Cham, 2015 vol 9022, doi:10.1007/978-3-319-16354-3_46.