Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 123-133 recibido 15-12-2022 revisado 03-02-2023 aceptado 12-02-2023

Anticipating the Debate: Predicting Controversy in News with Transformer-based NLP

Anticipando el Debate: Prediciendo la Controversia en Noticias con PLN basado en Transformers

Blanca Calvo Figueras, Asier Gutiérrez-Fandiño, Marta Villegas
Barcelona Supercomputing Center
blanca.calvo@bsc.es

Abstract: Controversy is a social phenomenon that emerges when a topic generates large disagreement among people. In the public sphere, controversy is very often related to news. Whereas previous approaches have addressed controversy detection, in this work, we propose to predict controversy based on the title and content of a news post. First, we collect and prepare a dataset from a Spanish news aggregator that labels the news' controversy in a community-based manner. Next, we experiment with the capabilities of language models to learn these labels by fine-tuning models that take both title and content, and the title alone. To cope with data unbalance, we undergo different experiments by sampling the dataset. The best model obtains an 84.72 micro-F1, trained with an unbalanced dataset and given the title and content as input. The preliminary results show that this task can be learned by relying on linguistic and social features.
Keywords: nlp, controversy prediction, news, spanish.

Resumen: La controversia es un fenómeno social que ocurre cuando un tema genera desacuerdo entre los ciudadanos. En la esfera pública, la controversia se encuentra a menudo relacionada con las noticias de actualidad. Mientras que trabajos anteriores investigaron la detección de la controversia, en este trabajo nos proponemos predecirla basándonos en el título y el contenido de una noticia. En primer lugar, recogemos y curamos un conjunto de datos de un agregador de noticias en castellano que etiqueta las noticias según su controversia mediante las interacciones de la comunidad. Entonces, experimentamos con las capacidades de los modelos de lenguaje para aprender la categoría de controversia mediante el fine-tuneado de modelos que tienen el título y el contenido como contenido de entrada, y también con solo el título. Para lidiar con el desbalanceo de los datos, realizamos experimentos de sampleado de los datos. El mejor modelo obtiene una micro-F1 de 84.72, entrenado con un conjunto de datos desbalanceado y con el título y el contenido como entrada. Los resultados preliminares muestran que esta tarea puede ser aprendida mediante características lingüísticas y sociales.
Palabras clave: pln, predicción de la controversia, noticias, castellano.

1 Introduction

With the rise of digital media, public opinion has increasingly become a political actor (Kshetri and Voas, 2017). In digital spaces, citizens are able to publicly denote their stances and collectively define what topics do not offer consensus. To comprehend these opinion flows, researchers have focused on detecting controversy (Popescu and Pennacchiotti, 2010).

Controversial topics have been defined as topics that generate strong disagreement among large groups of people (Dori-Hacohen, Yom-Tov, and Allan, 2015). Controversial news should not be confused with fake news. While fake news are messages that carry false information (Kshetri and Voas, 2017), controversial news can not be proved false, although its content is being disputed by some. Controversy is a subjective category that lies between what a large group of people might consider true and others false, where there is no objective proof to deny either position. Operationally, whatever people conceive as controversial is controversial (Dori-Hacohen,

ISSN 1135-5948 DOI 10.26342/2023-70-10

Yom-Tov, and Allan, 2015).

From the natural language processing perspective (NLP), controversy detection has been approached as a text classification task (Dori-Hacohen, Jensen, and Allan, 2016). Previous work has focused on the identification of controversy by using edition features (Bykau et al., 2015), interaction features (Coletto et al., 2017), or comment-based features (Hessel and Lee, 2019). While detecting that something has been controversial is a potentially useful task for the study of news, in this work we propose predicting the controversy instead. By predicting we mean forecasting there will be a controversy before the controversy itself has even happened by using the only information we have available as soon as the news are out: its content. We believe this task can be very useful for raising alarms about potentially troublesome news. Our goal in this work is to investigate if controversy in news can be predicted using only the title and the content of the piece of news.

To achieve our goal, we gather a dataset from the news aggregation platform *Menéame*.[1] This website is driven by its community and moderated by senior users. Feedback mechanisms are in place to prevent false or wrong information from being distributed through the platform. While false information is removed, controversial news stay, although they are labeled with the tag *controversial*, with the goal of promoting a critical reading from users. We employ these tags as the annotations of our dataset and we experiment with fine-tuning different language models for the task. We also try different balancing strategies and explore the decisions taken by our best model.

The main contributions of this work are:

- We present a new approach for developing a controversy prediction dataset that matches our operational definition of controversy based on the algorithm of *Menéame*.
- We show that it is possible to predict the forthcoming controversy using mainly the title and the content of a news post.
- We investigate the relevance of linguistic features for the controversy prediction model.
- Finally, we make the best model available for the natural language community.[2]

In the following sections, we review the previous work on controversy detection (Section 2), we present our dataset and the labeling methodology (Section 3), we explain the models that have been trained on it (Section 4), and we display the primary results and an analysis based on explicability techniques (Section 5). Finally, we reflect on the need for this kind of work and propose future work to be done with these novel resources (Section 6).

2 Previous Work

Popescu and Pennacchiotti (2010) were the first to propose the detection of controversial events on social media. This idea was followed by other researchers, who modeled controversy through social media interactions (Coletto et al., 2017), sentiment analysis, and word matching (Sriteja, Pandey, and Pudi, 2017).

Other approaches investigated controversy in a collaboratively edited database (namely Wikipedia), by relying on the back-and-forth substitutions of content embedded within a similar context (Bykau et al., 2015). This challenge has been addressed as a clustering task (Dori-Hacohen, Jensen, and Allan, 2016) and as a classification task (Jang et al., 2016).

Controversy in the newswire domain was first approached by (Rethmeier, Hübner, and Hennig, 2018), who labeled user comments using up and down votes from other users, collecting 20.5k comments. More recent approaches have gone further and have used the labeled comments to predict the controversy of the post they are commenting on, using manually-labeled data from public forums (Hessel and Lee, 2019; Zhong et al., 2020). Finally, (Kim, 2019) created an explainable model by providing a descriptive sentence of the controversial topic, automatically generated from the comments.

Overall, controversy detection has been overlooked when compared to other NLP tasks in the area of information systems. Our work differentiates from all the previous approaches in both its data collection design and the inputs that we use to train the model. As in

[1] `https://www.meneame.net/`

[2] The model: `https://huggingface.co/PlanTL-GOB-ES/Controversy-Prediction`
The code: `https://github.com/PlanTL-GOB-ES/controversy-detection-model`

previous work, our work relies on community-labeled data, which is essential to identify a social phenomenon such as controversy. Furthermore, our work focuses on predicting controversy (as opposed to detecting it), for which reason we just provide the model with the title and the content of the piece of news.

3 Dataset

Given that there is no dataset for controversy detection in Spanish, we create our own by using available data and sampling it. We also analyze the dataset and give details on its statistics.

3.1 Nature of the Dataset

In this work, we have gathered a dataset of news posts from the platform *Menéame*. The internal design of this website,[3] which is conceived as a social network to promote healthy debate by allowing different views to converge and discuss, provides us with the possibility of labeling our data in an automated community-driven way. The gross dataset that we collected as of February 18th of 2022 contains a total of 236,969 posts.

Menéame is a news aggregator that compiles Spanish news based on users' suggestions. Users can publish the pieces of news they found interesting, and the rest of the users can vote and comment to decide if it is interesting enough to get them into the front page. To prevent the dissemination of fake news, spam, or other issues, the users can also report if there is a problem with the piece of news. The reactions of the users can trigger a warning algorithm that raises the alert sign. If the reports are well-grounded, moderators or the publishers themselves can decide to remove the content. In a middle ground between posts that are reported and posts that are finally removed, we find controversial posts. This is a temporary tag that the website gives to promote further consideration from the readers. After 30 hours, if the post is not removed, it just stays in the historical data of the platform but is marked as *controversial*.

The warning algorithm works in the following way.[4] One hour after the piece of news has been published, the algorithm starts checking the reactions to the post, and marks it as *controversial* if both of the following conditions are met:

- There are more than 4 negative votes or the negative votes represent more than 62.5% of the overall votes. This percentage keeps decreasing over time and is 10% after 6 hours.
- The average karma[5] of the users who voted negatively is higher than the average karma of the users who voted positively multiplied by 0.625. This ratio keeps decreasing over time and the average positive karma is multiplied by 0.1 after 6 hours.

We collected this boolean feature along with relevant metadata and the title and summary of the news documents. Some examples of controversial news are:

- *La pasta podría ser considerada verdura en los comedores escolares de EEUU.*
  Pasta could be considered a vegetable in the school canteens in the US.
  Negative votes: 24
  Positive votes: 129
- *Los ateos, mucho más inteligentes que los creyentes*
  Atheists are way smarter than believers.
  Negative votes: 40
  Positive votes: 331
- *Entramos en el caos de los test Covid a los profesores de Madrid: "Nos llevan como ovejas al matadero"*
  We get into the chaos of teacher's Covid tests in Madrid: "They are carrying us like sheep to the slaughter."
  Negative votes: 32
  Positive votes: 89

[3]The source code of this website is fully open-source in `https://github.com/Meneame/meneame.net`

[4]The code of the algorithm can be found here: `https://github.com/Meneame/meneame.net/blob/60fc5935e46fb72c47945abc63cd062803d030a8/www/libs/link.php#L1441`

[5]The karma of the users is a metric of the overall reliability of each user in the platform. Actions that raise someone's karma are: positive votes to their proposed posts, positive votes to news that are finally published, negative reports to news that are finally deleted, and positive votes to their comments. Actions that decrease someone's karma are: negative votes to their proposed news, negative votes to news that are not deleted after 30 hours, and negative votes to their comments. More detailed information can be found here: `https://github.com/Meneame/meneame.net/wiki/Karma`

Collection	Set	Total	Label
All	-	236,969	5,584 (C)
			231,385 (NC)
Sampled	-	20,386	5,584 (C)
			14,802 (NC)
Unbal.*	Train	18,270	4,950 (C)
			13,320 (NC)
	Valid	1,058	317 (C)
			741 (NC)
Balanced*	Train	9,900	4,950 (C)
			4,950 (NC)
	Valid	634	317 (C)
			317 (NC)
Test*	Test	1,058	317 (C)
			741 (NC)

Table 1: Dataset split counts. Collections marked with '*' are taken from "Sampled". "Sampled" comes from "All". C stands for controversy and NC for Not Controversy.

3.2 Sampling and Splitting

Table 1 shows the instance counts of the whole dataset. A total of 231,385 non-controversial posts and 5,584 controversial posts have been collected. The label imbalance of the whole dataset made the modeling difficult, as it ends up mainly predicting the label of the majority of the instances (non-controversial). To address this problem we undersampled the data and created the *Sampled* dataset, from which we defined a train, valid, and a test set (*Unbalanced* in Table 1). Additionally, we created a balanced set for train and valid, with the same number of instances in the two classes. The test set is shared among datasets.

3.3 Statistics

While most of our corpus comes from news of the general press, we also have instances from social networks, specialized press, blogs, satirical press, sports press, and fact-checking websites. Remarkably, general press has the lowest ratio of controversial posts, while social networks, satirical, and fact-checks are more often controversial. The data can be seen in Figure 1.

We obtained the top ten sources of news in Table 2. The first two sources are social networks, which exhibit that posts without publishing control are more likely to end up in controversy.

Although the collected data has abundant metadata, such as tags, topic, positive votes, negative votes, users' comments, clicks, promoting votes,[6] and karma, we only rely on the title and the summary of the content of the piece of news for our classification.[7] The reason behind this decision is to develop a model that can be used over any set of news, coming straight from the source. In this sense, we do not want to use any platform-specific feature for the prediction, but rather common features that are present on all kinds of news sites. Nevertheless, as an interesting insight, we discuss the relations between some of these metadata and our labels.

Source	Controv.	Total	Ratio
twitter.com	326	2,300	14.17%
youtube.com	306	5,955	5.13%
eldiario.es	293	7,533	3.88%
publico.es	190	5,804	3.27%
20minutos.es	118	6,048	1.95%
elconfidencial	117	5,188	2.25%
elplural.com	103	1,590	6.47%
huffingtonpost.es	85	786	10.81%
elmundo.es	81	6,191	1.30%
elespanol.com	81	1,995	4.06%

Table 2: Top ten sources, sorted by the number of controversial posts.

In Table 3, we show the ten most frequent tags given by the users to controversial news posts along with the total number of occurrences on the whole dataset. Remarkably, the most frequent labels are those related to politics. Some of them (i.e. "ayuso" and "vox") show a notably higher percentage than others.

Regarding the topics, we perform a similar analysis and show the distribution in Table 4. We decided to obviate the tenth topic as it is "rude language" and yet negligible in terms of controversy. Remarkably, "current issues" and "issues of social interest" are on the top of the list.

A point-biserial correlation was run to determine the relationship between the number of comments and the controversy label. There was a positive correlation between the variables ($r_{pb}29.90$, $p < 0.05$).

Finally, we also analyzed the number of likes and dislikes, as well as the difference between the likes and dislikes with respect

[6] What the website calls *meneos*.

[7] We use the summary of the piece, which is given by the users. In shorter posts, this is usually the whole content. In traditional news articles, this is almost always the first paragraph of the article. This is useful for training, as we have a length limitation given by the language model.

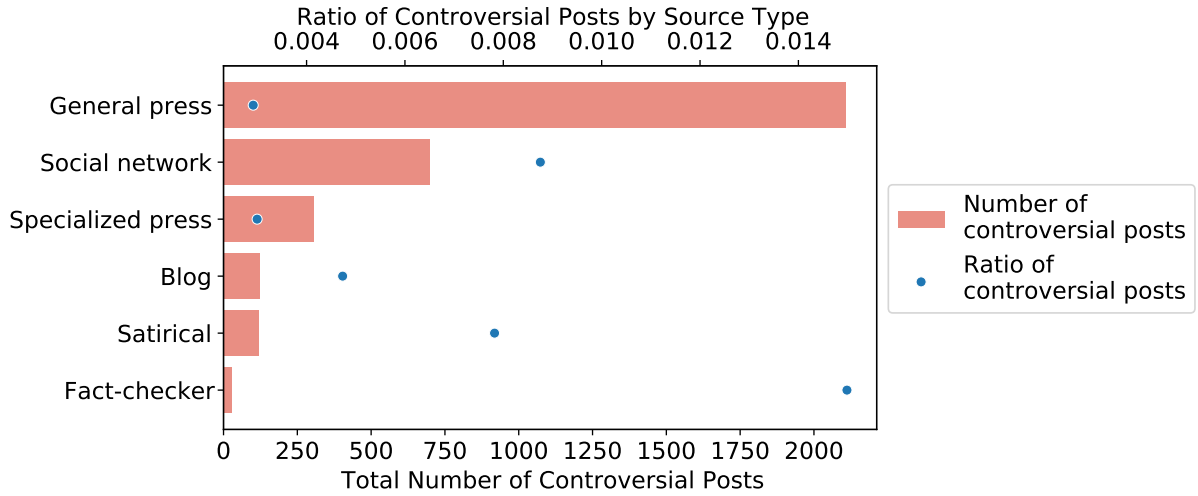

Figure 1: Controversial posts for each source type. The blue dots indicate the ratio between the number of controversial posts and the overall number of posts of each source type.

Tag	Explanation	Controversial	Total	Ratio
madrid	Spanish city/county	231	4,644	4.97%
pp	Political party	226	8,036	2.81%
españa	Spain	190	7,185	2.64%
vox	Political party	181	919	19.69%
podemos	Political party	173	1,515	11.41%
humor	Humour	170	3,726	4.56%
coronavirus	Coronavirus	153	2,596	5.89%
ayuso	Political leader	109	523	20.84%
psoe	Political party	84	2,505	3.35%
covid	Covid	74	1,174	6.30%

Table 3: The ten most common tags, sorted by the number of controversial posts.

to the controversy label. We run the point-biserial experiments and the results are the following:

- Positive votes: There is no correlation between the positive votes and the controversy label ($r_{pb}3.86$, $p < 0.05$).
- Negative votes: There is a strong positive correlation between the negative votes and the controversy label ($r_{pb}80.97$, $p < 0.05$).
- Positive-Negative difference: There is a negative correlation between the positive-negative difference and the controversy label ($r_{pb}10.51$,$p < 0.05$).

4 *Experimental setup*

To train a baseline for this task we selected the Spanish RoBERTa-base (Gutiérrez-Fandiño et al., 2022), as it has been trained on a large and clean corpus and it is the best performing model in Spanish to date.

Given that the training model only supports up to 512 input tokens, we used a truncation strategy for our data. When fine-tuning the model with title only, the truncation strategy does not apply, as titles are never long enough. In contrast, using the title and the summary concatenated,[8] the input data ends very often truncated.

We trained all the dataset combinations for 5 epochs, using a batch size of 4 per Graphical Processing Unit (GPU), a warmup of 0.06, a weight decay of 0.01, and a learning rate of 1e−5. For the optimizer, we chose Adam (Kingma and Ba, 2015), as it has been proved by the community to offer strong results.

The models were trained on our HPC premises on a machine with 2 IBM Power9 8335-GTH @ 2.4Ghz processors, 512GB of

[8]We concatenate with two light horizontal line symbols ("- -").

Topic	Explanation	Controversial	Total	Ratio
actualidad	current issue	2,671	59,673	4.47%
mnm	social interest	1,207	133,961	0.90%
ocio	leisure	616	7,378	8.34%
cultura	culture	577	21,492	2.59%
politica	politics	252	2,040	12.35%
tecnología	technology	170	8,779	1.93%
ciencia	science	18	1,156	1.55%
Podemos	political party	18	44	40.90%
Hemeroteca	news archive	12	96	12.50%

Table 4: The nine most common topics sorted by controversy.

Random Access Memory, and 4 NVIDIA V100 GPUs with 16GB of HBM2 memory.

5 Results and Analysis

The results of our fine-tuning experiments are shown in Table 5. They are displayed by dataset and by training setting.

Overall, our best model achieves a micro-F1 of 84.72 and an accuracy of 76.65, proving that the task can be effectively learned by a model with only the title and the summary. The addition of the summary does not provide much improvement, since using only the title in the unbalanced dataset already gives reasonably positive results. By contrast, the best model trained on the unbalanced dataset is around 4 points better on F1 than the model trained on the balanced one, showing that the model profits from more negative examples. We also experiment with other language models, such as mBERT (Devlin et al., 2019) and BETO (Cañete et al., 2020), using the same experimental setup and getting similar results.

Explainability analysis. To further analyze our results, we compute SHAP values (Lundberg and Lee, 2017), which assign contribution values to the tokens of each input.[9] We use the best model obtained, set up a SHAP explainer with it, and feed it a balanced set of 11k posts. Then, we aggregate the SHAP values for each token by part of speech (POS) and look at the open categories, namely: verbs, nouns, proper nouns, adjectives, and adverbs.[10] The used model for POS tagging shares the same vocabulary as our controversy model to ensure the tokenization is the same. We aggregate the positively and negatively contributing tokens for each POS tag and observe that proper nouns are, on average, the most contributing tokens to both the controversy and the non-controversy classes (Table 6).

Table 8 shows the 10 most influential words as per part of speech. When looking at proper nouns, we find that the most divergent parties of Spain (*Vox* and *Podemos*) correspond to the most controversial proper nouns of the dataset. Followed by *Ayuso* and *Pablo Iglesias*, two Spanish politicians well-known for having a large number of followers and haters and for being popular targets of memes (Paz, Mayagoitia-Soria, and González-Aguilar, 2021). On the other side of the table, contributing to the non-controversy category, we find technological companies, such as *Google*, *Linux*, and *Microsoft*, former Spanish presidents, like *Rajoy* and *Zapatero*, and geographical entities, such as *Reino Unido*, *Europa* and *Estados Unidos*. In sum, the use of proper nouns related to current politicians and parties is highly related to controversy, while the mention of companies, countries, or former politicians contributes negatively to the controversy class. Although this kind of information would also be highly valuable for humans trying to predict a controversy, it is very dependent on knowledge about the current cultural and political reality of Spain. To observe more enduring patterns, we train another model dropping the proper nouns.

Removing proper nouns. We use the same POS tagger to remove all proper nouns from the dataset and substitute them by the token *PROPN*. Then, we train a model with the exact same experimental setup as our best model: a Roberta-base model fine-tuned with the Unbalanced dataset and the

[9] Positive values mean that the token is contributing to the label "Controversy", while negative values mean a contribution towards Not Controversy. The furthest the value is from 0, the stronger the contribution.

[10] The part-of-speech model we are using can be found in `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne-capitel-pos`.

Model	Dataset	Training setting	micro-F1	Accuracy	Time (s)
Roberta-base	Balanced	Title	0.7026	0.6295	1653
		Title + Summary	0.8093	0.7353	1267
	Unbalanced	Title	0.8197	0.7268	2631
		Title + Summary	**0.8472**	**0.7665**	2615
BETO	Unbalanced	Title	0.8398	0.7533	2568
		Title + Summary	0.8361	0.7429	2562
mBERT	Unbalanced	Title	0.8309	0.7287	3221
		Title + Summary	0.8347	0.7448	3277

Table 5: Model results by dataset and training setting.

POS	Positive SHAP	Negative SHAP
PROPN	0.0155	-0.0233
VERB	0.0053	-0.0122
ADV	0.0038	-0.0089
NOUN	0.0054	-0.0134
ADJ	0.0058	-0.0125

Table 6: SHAP values of the tokens, aggregated per part-of-speech.

POS	Positive SHAP	Negative SHAP
PROPN-token	0.0054	-0.0119
VERB	0.0060	-0.0120
ADV	0.0044	-0.0086
NOUN	0.0063	-0.0147
ADJ	0.0067	-0.0134

Table 7: SHAP values of the tokens aggregated per part-of-speech for the model without proper nouns. The PROPN-token value corresponds to the substitution token we used.

Title+Summary input. The obtained results are remarkably good, with an F1 of 83.53 and an accuracy of 74.76. The aggregated SHAP values in Table 7 show that the impact of the PROPN-token is much lower than the aggregated proper nouns were, and the rest of the POS categories have increased in relevance. These results suggest that controversy can be predicted by relying mainly on linguistic features.

We identify some linguistic patterns in the table with the top influencing tokens by POS for this new model (Table 9). When looking at verbs, we observe that actions in the third singular person of the perfect tense (e.g. *ha hecho*, *ha sido*, *ha convertido*, *ha respondido*, etc.) are often associated with controversial posts. Instead, verbs in the simple present tense (e.g. *es*, *hay*, *pide*, *tienen*, etc.) are associated with non-controversial posts. Looking at the posts of our dataset, we identify that while the third singular person of the perfect tense is often used to speak about people in a rather informal tone, like in the sentence "*La portavoz adjunta de Compromís Mónica Oltra ha vuelto a lucir esta mañana en el primer pleno ordinario del nuevo periodo de sesiones de Les Corts una de sus famosas camisetas.*" (This morning Monica Oltra has worn again another of her famous t-shirts in the Parliament); the present simple is used for more factual information, like in the sentence "*Los trabajadores de RTVE rodean la mesa de edición con carteles que dicen: #Vergüenza #Vergonya*" (The workers from RTVE surround the edition office with signs that say: #Shame).

Finally, we observe some other linguistic patterns that seem to indicate controversy, such as the use of adverbs at the beginning of the sentences, and the use of adjectives indicating political positioning (e.g. *feminist* or *independentist*).

6 Conclusion and Future Work

In this work, we have built a dataset for controversy prediction in Spanish and we have characterized it in many dimensions. Controversy has been labeled in a community-driven manner, which matches the operational definition of controversy itself, given in the introduction.

With this dataset, we have experimented creating two different collections: a balanced one and an unbalanced one. In this particular experiment, we have shown that the amount of samples is more important than balancing the labels.

The models we trained have provided positive results and have shown that they can effectively capture the insights of the title (and the summary). These models can be used

as predictors of the controversy generated in news. In the explainability analysis, we have shown that this model is capturing differences in the linguistic register of controversial posts, as well as the social and political reality of Spain. We have been able to highlight some characteristics of the controversial register, such as using the third singular person of the perfect tense.

In the future, this model can be used for media monitoring, by trying to understand how controversy evolves in media as a function of time. It can also be used for media analysis, by running it against online media to observe editorial inclinations toward controversy. Additionally, one could analyze how the perception of controversy evolves in Spanish society, as what generates controversy today might not be controversial in the future, or the other way around. This dataset will continue to grow, as the community behind it is still highly active. We set as a future goal to keep expanding it to capture possible shifts in the perception of a controversy.

Finally, the model could also be used to highlight disputed posts as soon as they are published, as this has been suggested as a mitigation strategy for the impact of disinformation (Dori-Hacohen, 2015). In this line, previous work has indicated certain relation between highly disputed news and fake news (Shu et al., 2019). We suggest executing this model against a fake news database to study this phenomenon.

Acknowledgements

This work has been funded by the Spanish State Secretariat for Digitalization and Artificial Intelligence (SEDIA) within the framework of the Plan-TL, and the IBERIFIER project funded by the European Union (action number 2020-EU-IA-0252).

References

Bykau, S., F. Korn, D. Srivastava, and Y. Velegrakis. 2015. Fine-grained controversy detection in Wikipedia. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1573–1584, April. ISSN: 2375-026X.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Coletto, M., K. Garimella, A. Gionis, and C. Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3-4:22–31, October.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dori-Hacohen, S. 2015. Controversy Detection and Stance Analysis. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 1057, New York, NY, USA, August. Association for Computing Machinery.

Dori-Hacohen, S., D. Jensen, and J. Allan. 2016. Controversy Detection in Wikipedia Using Collective Classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 797–800, New York, NY, USA, July. Association for Computing Machinery.

Dori-Hacohen, S., E. Yom-Tov, and J. Allan. 2015. Navigating Controversy as a Complex Search Task. page 5.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Hessel, J. and L. Lee. 2019. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. 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)*, pages 1648–1659, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jang, M., J. Foley, S. Dori-Hacohen, and J. Allan. 2016. Probabilistic Approaches to Controversy Detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2069–2072, New York, NY, USA, October. Association for Computing Machinery.

Kim, Y. a. 2019. Unsupervised Explainable Controversy Detection from Online News. *Proceedings of the European Conference on Information Retrieval*.

Kingma, D. P. and J. Ba. 2015. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kshetri, N. and J. Voas. 2017. The Economics of "Fake News". *IT Professional*, 19:8–12, November.

Lundberg, S. M. and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 4765–4774.

Paz, M. A., A. Mayagoitia-Soria, and J.-M. González-Aguilar. 2021. From Polarization to Hate: Portrait of the Spanish Political Meme. *Social Media + Society*, 7(4):205630512110629, October.

Popescu, A.-M. and M. Pennacchiotti. 2010. Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1873–1876, New York, NY, USA, October. Association for Computing Machinery.

Rethmeier, N., M. Hübner, and L. Hennig. 2018. Learning Comment Controversy Prediction in Web Discussions Using Incidentally Supervised Multi-Task CNNs. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 316–321, Brussels, Belgium, October. Association for Computational Linguistics.

Shu, K., L. Cui, S. Wang, D. Lee, and H. Liu. 2019. dEFEND: Explainable Fake News Detection. *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 395–405, July.

Sriteja, A., P. Pandey, and V. Pudi. 2017. Controversy Detection Using Reactions on Social Media. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 884–889, November. ISSN: 2375-9259.

Zhong, L., J. Cao, Q. Sheng, J. Guo, and Z. Wang. 2020. Integrating semantic and structural information with graph convolutional network for controversy detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 515–526, Online, July. Association for Computational Linguistics.

POS	Word	Explanation	SHAP	Word	Explanation	SHAP
	explica	explains	0.40	es	is	-7.10
	desmonta	dismantles	0.34	son	are	-3.15
	responde	answers	0.32	hay	there are	-2.94
	muestra	shows	0.31	tiene	has	-2.77
VERB	explicando	explaining	0.24	era	was	-2.25
	habla	talks	0.24	pide	asks	-2.22
	Desmontando	Dismantling	0.19	tienen	have	-2.00
	recuerda	remembers	0.18	dice	says	-1.68
	voy	coming	0.15	está	is	-1.60
	ha respondido	has answered	0.14	hacer	do/make	-1.57
	Vox	political party	6.05	Google		-2.45
	Podemos	political party	5.96	PP	political party	-2.43
	Pablo Iglesias	politician	5.76	Zapatero	ex-politician	-1.82
	Ayuso	politician	5.30	Reino Unido	UK	-1.69
PROPN	Madrid		3.88	Rajoy	ex-politician	-1.54
	VOX	political party	1.65	Linux		-1.51
	Pedro Sánchez	politician	1.28	SGAE	institution	-1.45
	Isabel Díaz Ayuso	politician	1.07	Europa		-1.32
	Pablo Casado	politician	0.79	Estados Unidos	United States	-1.30
	Ada Colau	politician	0.65	Microsoft		-1.28
	Así	Like this	1.26	no	no	-21.78
	Cómo	How	0.76	más	more	-15.17
	Cuando	When	0.31	hoy	today	-3.68
	Además	Morover	0.30	también	also	-2.34
ADV	Aquí	Here	0.16	muy	very	-2.18
	literalmente	literally	0.09	después	after	-1.88
	Anoche	Last night	0.07	ahora	now	-1.82
	consecuentemente	consequently	0.06	ya	already	-1.72
	brutalmente	brutally	0.06	donde	where	-1.54
	sexualmente	sexually	0.06	casi	almost	-1.44
	vídeo	video	3.80	años	years	-7.93
	derecha	right	0.80	millones	milions	-5.74
	mentiras	lies	0.60	Gobierno	Government	-5.07
	bulo	fakes	0.54	presidente	president	-3.56
NOUN	discurso	discourse	0.50	personas	people	-3.16
	ultraderecha	far-right	0.44	euros	euros	-2.87
	izquierda	left	0.44	ministro	minister	-2.73
	respuesta	answer	0.43	mundo	world	-2.57
	tuit	tweet	0.41	juez	judge	-2.48
	monarquía	monarchy	0.38	Policía	Police	-2.38
	feminista	feminist	0.67	nuevo	new	-2.31
	extrema	extreme	0.44	gran	big	-1.93
	independentista	independentist	0.37	mayor	older/higher	-1.85
	independentistas	independentists	0.28	nueva	new	-1.71
ADJ	ultraderechista	far-rightist	0.27	pasado	past	-1.48
	española	Spanish	0.26	Nacional	National	-1.44
	morada	purple	0.25	grandes	big	-1.42
	ultraderechistas	far-rightists	0.24	últimos	last	-1.31
	mediática	media	0.23	mejor	better	-1.21
	política española	Spanish politics	0.22	general	general	-1.18

Table 8: Top 10 influencing words per Part-of-Speech with no named entities.

POS	Word	Explanation	SHAP	Word	Explanation	SHAP
	muestra	shows	1.04	es	is	-5.25
	desmonta	dismantles	0.49	hay	there is	-2.61
	responde	answers	0.47	pide	asks	-1.86
	ha hecho	has done	0.45	tienen	have	-1.62
VERB	explica	explains	0.42	son	are	-1.61
	ha sido	has been	0.35	hace	does	-1.51
	ha convertido	has converted	0.34	era	was	-1.41
	ha respondido	has answered	0.28	pagar	pay	-1.32
	ha declarado	has declared	0.27	Fallece	Dies	-1.24
	ha publicado	has published	0.26	tiene	has	-1.24
	Así	Like this	1.84	no	no	-16.52
	Además	Morover	1.11	más	more	-12.99
	Cómo	How	0.76	hoy	today	-5.26
	Sin	Without	0.54	ahora	now	-1.41
ADV	Cuando	When	0.41	ayer	yesterday	-1.23
	Ahora	Now	0.33	muy	very	-1.21
	No	No	0.25	antes	before	-1.17
	No obstante	Nevertheless	0.25	menos	less	-1.09
	Sin embargo	Nevertheless	0.24	sólo	only/just	-1.09
	Después	After	0.16	casi	almost	-1.00
	vídeo	video	4.31	años	years	-6.83
	partido	party	2.12	ministro	minister	-5.51
	respuesta	answer	1.70	millones	milions	-4.66
	líder	leader	1.51	Gobierno	Government	-4.36
NOUN	formación	formation	1.22	ciudad	city	-3.40
	bulo	fake	0.98	mundo	world	-3.35
	discurso	discourse	0.96	personas	people	-2.73
	mensaje	message	0.86	países	countries	-2.71
	pandemia	pandemic	0.83	país	country	-2.64
	redes	networks	0.82	presidente	president	-2.42
	feminista	feminist	0.85	mayor	older	-2.01
	independentista	independentist	0.67	nuevo	new	-1.83
	morada	purple	0.50	nueva	new	-1.68
	oficial	official	0.45	Europea		-1.55
ADJ	ultraderechista	far-rightist	0.44	gran	big	-1.50
	falso	false	0.43	grandes	big	-1.37
	independentistas	independentists	0.42	general	general	-1.35
	extrema	extreme	0.37	Civil	Civil	-1.17
	madrileña	from Madrid	0.36	Nacional	National	-1.15
	madrileño	from Madrid	0.33	nuevas	new	-1.07

Table 9: Top 10 influencing words per Part-of-Speech in the model with no proper names.