A Comparative Analysis of Offensive Discourse in the 2021 Chilean Presidential Campaign on Twitter and WhatsApp

Hernan Sarmiento*, Jorge Ortiz†*‡, Felipe Bravo-Marquez†*‡, Marcelo Santos§¶ and Sebastián Valenzuela‖*ˣ
*Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile
†Department of Computer Science, University of Chile, Santiago, Chile
‡National Center for Artificial Intelligence (CENIA), Santiago, Chile
§School of Communication, Universidad Diego Portales, Santiago, Chile
¶Center for the Study of Media, Public Opinion, and Politics in Chile (MEPOP), Santiago, Chile
‖School of Communications, Pontificia Universidad Católica de Chile, Santiago, Chile
ˣMillennium Nucleus on Digital Inequalities and Opportunities (NUDOS), Santiago, Chile

***Abstract*—Participatory society has often been regarded positively, frequently associated with the ideals of a more democratic and equitable civilization. Nevertheless, the idea of participation may act as a two-sided phenomenon in terms of empowerment, especially in the realm of social media platforms. This dichotomy is evident as increased participation often leads to a rise in offensive and divisive language, reflecting the challenging balance between open dialogue and the maintenance of respectful discourse on these platforms. In this work, we comprehensively examine the use of offensive language during a highly polarizing event on two online platforms, Twitter and Whatsapp. In our study, we focus in the 2021 Chilean Presidential Elections, a political event where candidates from two opposing parties faced each other. Using a state-of-the-art model and all available labeled data in literature, we determine the level of offensive language across platforms and parties. Our results show that Twitter messages contain, on average, up to 15% more of offensive language than Whatsapp.**

***Index Terms*—social media, hate speech, polarization, Twitter, WhatsApp.**

I. Introduction

A society where everyone can participate and share their views is often seen as a positive step towards a more democratic and fair community. This is especially true with social media platforms, where people have the chance to express themselves freely. However, this opportunity for everyone to speak up comes with its own set of challenges, particularly when it comes to the language used on these platforms.

While social media allows for more voices to be publicized, it also leads to an increase in offensive and divisive language [1]. This problem highlights the difficulty in maintaining a space where open discussions can happen while also keeping conversations respectful. Social media can sometimes become a place where hate speech is spread, disagreements become more extreme, and false information is shared [2]. This situation shows how complicated it can be to create a society where everyone can participate effectively.

Sigurbergsson and Derczynski [3] describe offensive language as encompassing various expressions, ranging from simple profanities to more severe forms, including hate speech. The prevalence of toxic and virulent language has long been a central concern in discussions on social media.Studies comparing the use of offensive language in social media are often limited to the study of a phenomenon on a single platform [3]–[9]. Recently, research has also focused on understanding variations in the adoption of offensive language across different online platforms [10], [11].

While many studies are focusing on the differences in toxic behavior across various social media platforms, one area that often gets less attention is the role of *social media affordances* in shaping those differences. This term refers to the communication properties and capabilities provided by social media platforms that influence user interactions and behaviors [12]–[14]. These affordances are shaped by the platforms' functionalities, structures, and operating systems, as well as the motivations and opportunities of users, including non-human agents. They are not merely technological features or are their effects, but rather opportunities for action or possibilities for action that arise from the interplay between users and platforms [15], [16].

In our work, we aim to compare the prevalence of offensive language expressions in political communications across online platforms by adopting an affordances-based approach. We focus on this issue by looking closely at how offensive language was used during a time of strong political disagreement: the 2021 Chilean Presidential Elections. This election, which featured candidates from very different political sides, provides a perfect example to study the patterns of use of offensive language on two different online platforms in a heated political environment. To address an affordances-based approach in online communications, we focus on Twitter (now

We acknowledge the support of ANID - Millennium Science Initiative Program - Code ICN17_002 (IMFD) and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Additionally, MS is partially funded by grant Fondecyt de Iniciación 11230980 and grant ANID NCS2021_063 and SV is funded by grant ANID - NCS2022_046 (NUDOS).

known as X) and WhatsApp, two widely used social platforms for voter mobilization.

Using text classification techniques, specifically Language-Agnostic Sentence Representations (LASER) embeddings [17], we represented the messages as dense vectors. We then employed logistic regression, trained on a large dataset from previous studies on offensive language detection in Spanish, to measure the prevalence of offensive language on these two social media platforms and among different political groups. Our results show a clear difference between the platforms, with Twitter having, on average, up to 15% more offensive language compared to WhatsApp. Additionally, we note that differences in the same platform but supporting distinct candidates, are not conclusive and need to be further inspected. Finally, we use the LIME (Local Interpretable Model-Agnostic Explanations) model interpretability framework [18] in our offensive language detection research, applied to every community and platform, motivated by the need to ensure model transparency, adaptability to diverse linguistic contexts, and robustness in addressing biases. This approach enhanced the reliability of our findings and contributed to a more comprehensive understanding of offensive language usage in various online environments.

The remainder of this paper is structured as follows. In Section II, we review related work on offensive language and hate speech detection in social media. Section II describes the 2021 Chilean Elections and presents our Twitter and WhatsApp datasets comprising messages related to this event. Section IV outlines our methodology for inferring the degree of offensive language in our data to compare the platforms. Section V details our experimental setup. Section VI discusses our results and analysis of the comparison. Section VII provides a discussion of our findings in the context of the affordances of the social media platforms examined in this study. Finally, Section VIII presents our conclusions.

A. Warning: Sensitive Content

Please note that this research may include instances of swear words or strong language as part of its analytical content. Viewer discretion is advised.

B. Ethical Statement

The dataset obtained from WhatsApp group conversations possesses a high level of sensitivity, primarily due to its political relevance and semi-confidential character. Consequently, we took immediate measures to anonymize critical elements of the data, specifically WhatsApp user identifiers. Furthermore, we systematically eliminated all forms of multimedia content, encompassing images, audio files, and video clips. Moreover, following current ethical guidelines [19], our analytical focus was placed on the collective examination of textual content rather than dissecting individual messages. This strategy was employed to maximize the prevention of potential identification of any participant based on their contributed content. Lastly, our intention is to disseminate the dataset exclusively on a selective basis. This policy is to ensure that there is minimal risk of individual identification either through the message content itself or via correlation with other datasets, or any forthcoming technology that could compromise the privacy and safety of the individuals involved.

II. Related Work

The detection of offensive language and hate speech on online social media has been an area of considerable interest, especially as platforms grow in both size and influence. Early works by Davidson et al. [20] and Wasemm and Hovy [21] laid the groundwork for automated hate speech and offensive language detection using machine learning techniques. They employed a variety of models, such as logistic regression and support vector machines, trained on datasets annotated for hate speech and offensive language.

One of the main challenges to detect these types of toxic expressions is that most of the existing resources have been consolidated mainly for the English language [21], [22]. This means that there is much less information and fewer tools available for other languages (e.g., Spanish, Italian, and Arabic), reducing the ability to effectively identify and mitigate the spread of such languages across different linguistic communities. The nuance of regional dialects, idioms, and cultural expressions poses a further challenge, as these subtleties are often lost or misinterpreted by models trained on standard language datasets. Consequently, the development of multilingual and culturally sensitive models is essential, and while there has been progress with initiatives such as the creation of datasets, the scope and depth of these resources remain limited [23]. Therefore, it is crucial to expand research and tool development to include the diverse array of global languages and dialects present in online discourse.

In this context, there is significant work on evaluating multilingual and cross-lingual approaches for detecting hate speech and offensive content on online platforms [24]–[26]. The general strategy involves utilizing English data to classify messages in other languages (e.g., Spanish or Italian) by leveraging multilingual resources such as embeddings and lexicons.

III. The 2021 Chilean Elections

This election featured a contest between José Antonio Kast, a candidate from the far right, and Gabriel Boric, a left-wing candidate. Our choice to focus on this election was due to its distinction as the most polarized in Chile in over three decades. No centrist coalition progressed to the run-off stage [27]. Kast's campaign was grounded in right-wing principles, mirroring the stances of Donald Trump in the USA and Jair Bolsonaro in Brazil. His political agenda included plans to dissolve the Chilean Ministry of Women and Gender Equity, build a barrier along the northern border to curb immigration, introduce substantial tax reductions, and ban all forms of abortion [27], [28].

In contrast, Boric headed the *Frente Amplio*, a coalition of various left-wing parties and groups, including the Socialist Party. His campaign was built on a progressive platform. This

included transitioning from a privately controlled system to a public model for health and pensions, increasing corporate taxes, and implementing a wide-ranging array of feminist policies [29], [30].

Given the significant ideological differences between these campaigns, this election presents a valuable example for the application of the analytical methods described in this article.

A. Dataset description

We collected Twitter and WhatsApp data in the context of the 2021 run-off presidential election in Chile, which confronted the far-right candidate José Antonio Kast with the left-wing candidate Gabriel Boric. The Twitter data was collected using a set of keywords and hashtags related to the campaign via the Twitter Academic API. To identify pro-Kast and pro-Boric communities, we followed Sarmiento et al. [31] and used a stochastic block model [32] on retweet networks. Subsequently, for each retweet network, we extracted original tweets. Then, with all these tweets, we build the Twitter dataset.

On the other hand, the WhatsApp data was obtained from public groups organized by each campaign. Garimella and Eckles [33] define public WhatsApp groups as "any group on WhatsApp which can be joined using a publicly available link" (p. 7). We identified publicly available links to these groups from the candidate's websites and social media (e.g. Twitter messages, public Facebook groups where they were advertised, etc.). We used snowball sampling to detect additional public groups. In this way, the initial groups served as a springboard to enter new public groups related to the ideological communities.

Table I summarizes the collected data across platforms. For Twitter, we retrieved $429,371$ messages posted by $7,837$ unique Twitter users. This collection comprises 56% $(242,007)$ posted from users in the Pro-Kast community, while 44% $(187,364)$ were collected from Pro-Boric. On the other hand, Whatsapp dataset contains $465,200$ messages written by $35,200$ unique users. In detail, 23% $(107,299)$ of the messages were collected from Pro-Kast groups and 77% $(357,901)$ were retrieved from Pro-Boric. For the rest of the article, we refer to this collection as the *election dataset*.

Community	**Twitter**		**WhatsApp**		**TOTAL**	
	Msgs	%	Msgs	%	Msgs	%
Pro-Kast	242,007	56	107,299	23	349,306	39
Pro-Boric	187,364	44	357,901	77	545,265	61
TOTAL	**429,371**		**465,200**		**894,571**	
	Users	%	Users	%	Users	%
Pro-Kast	4,038	52	5,112	19	9,150	26
Pro-Boric	3,799	48	22,251	81	26.050	74
TOTAL	**7,837**		**27,363**		**35,200**	

TABLE I
NUMBER OF MESSAGES AND UNIQUE USERS PER COMMUNITY.

IV. METHODOLOGY

In our work, we aim to understand the level of offensive language that emerges in highly polarized online discussions. Given the lack of resources in non-English languages related to offensive language (e.g., language-specific models) and the cost of obtaining high-quality labeled data for a specific event to train supervised models, we aim to explore methodologies used in other studies for effectively identifying and analyzing offensive language. Considering the challenges posed by linguistic diversity and the scarcity of specialized resources in non-English languages, we have focused on understanding the use of cross and multilingual text representations that assist in identifying offensive language in social media messages.

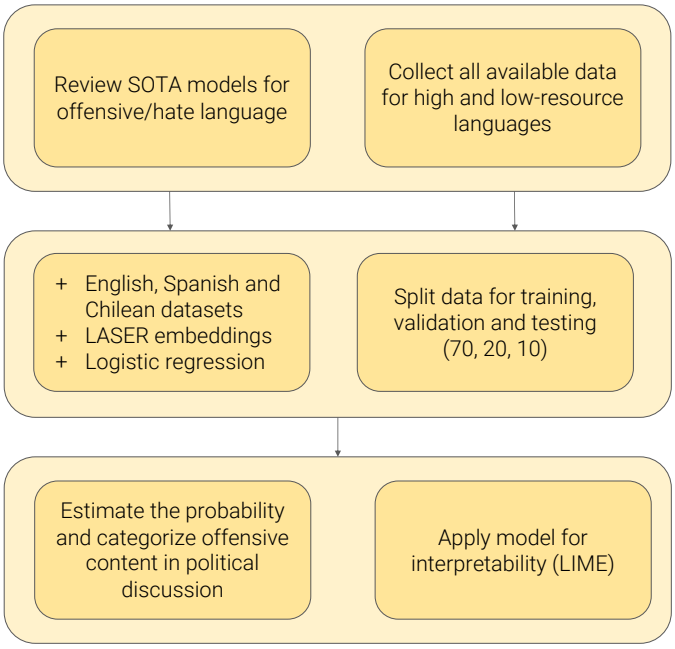

Fig. 1. Proposed Methodology.

Figure 1 provides a general overview of our proposed workflow. Firstly, we reviewed the literature related to the identification of offensive messages in social media, with an emphasis on the Spanish language. This review enabled us to describe which models and data are available for our research. Secondly, we focused on developing a classification model that automatically categorizes messages into offensive and non-offensive classes. Based on our previous review, we trained a classifier using a multilingual approach that includes both English and Spanish data. Additionally, we represented text data using LASER embeddings [17] and employed the logistic regression algorithm as a classifier. Ultimately, we classified the entire dataset into respective categories. We estimated the probability of each message being offensive and analyzed the aggregated results for each platform and candidate. For a better understanding of how the offensive class is determined in our classifier, we incorporated the use of LIME [34]. This model is widely utilized for the interpretability of supervised approaches, including in the field of hate speech detection [35].

V. EXPERIMENTAL SETUP

In this section, we detail the setup of our experiments, including the collection of data to train an offensive classifier,

the model used for representing data and the evaluation of the offensive classifier.

A. Offensive language datasets

Most published approaches and resources for detecting offensive language and hate speech are tailored for English. Furthermore, a significant limitation of existing methods is the highly contextual nature of hate speech and offensive language detection. This implies that data is often labeled according to the specific characteristics of an event and the requirements related to the dimension being analyzed (e.g., minority groups).

In light of the aforementioned challenges in automatically detecting offensive language and the well-documented difficulties associated with the cost of labeling data for training supervised models, we leveraged existing data from previous studies to create a large-scale labeled dataset. We reviewed available repositories[1] and collected data meeting the following criteria: 1) the dataset contains microblogging messages labeled as offensive language or hate speech (and their negative classes); 2) the dataset includes messages written in Spanish or English; and 3) the dataset is publicly accessible.

Language	**Source**	Offensive	Non-offensive
Spanish	Pereira-Kohatsu et al. [36]	4,433	1,567
	Arango et al. [37]	6,516	3,318
	Basile et al. [38]	2,055	2,895
	Total	13,004	7,780
English	Basile et al. [38]	4,210	5,790
	Zampieri et al. [39]	4,400	8,840
	Total	8,610	14,630
Merged	***Total***	21,614	22,410

TABLE II
DATASETS CONSIDERED IN THIS WORK.

Table II summarizes the datasets retrieved for our work. The merged dataset comprises $44,024$ messages, where 52% and 48% correspond to English and Spanish messages, respectively. With respect to the Spanish messages, 47 $\%$ correspond to Spanish from a Chilean Twitter dataset. The inclusion of this collection could be important to include cultural-specific characteristic of the messages, which is important for cross-lingual and cross-cultural classification as highlighted by [37]. Regarding the class distribution, 49% of messages were labeled as *offensive* and 51% as *non-offensive*. For the rest of the article, we refer to this collection as the *merged dataset*.

B. Feature Extraction

Given that our *merged dataset* contains messages written in both English and Spanish, we focused on multilingual and language-agnostic representations of text data. Previous work indicates that LASER embeddings [17] have demonstrated good performance for detecting offensive language and hate speech in low-resource scenarios (e.g., Spanish messages) [35], [37]. LASER operates by converting any input sentence into a vector in a 1024-dimensional space. The vector representations of sentences by LASER are generic with respect to both the input language and the task. This means that once a sentence is converted to a vector, we can construct a downstream model that works on any input language for any downstream task.

[1]https://github.com/aymeam/Datasets-for-Hate-Speech-Detection/

C. Model Evaluation

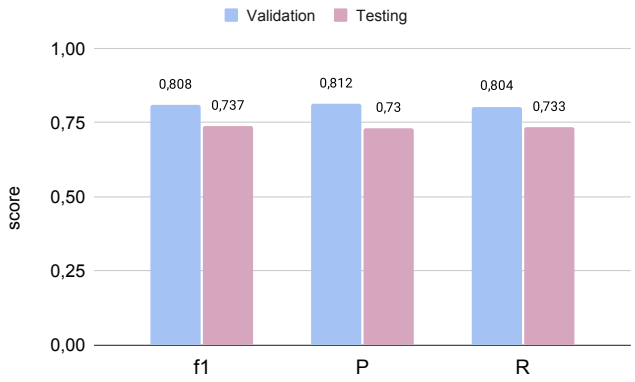

Fig. 2. Classification Performance on Chilean Data.

For training our model, we used the logistic regression algorithm. In the context of offensive language detection, this algorithm has been used in conjunction with LASER embeddings obtaining good classification performance [35], [37]. For our experiments, we considered the `scikit-learn` implementation and same hyper-parameters of previous work.

Given that our goal is to infer the level of offensive language across platforms during the Chilean elections, messages published on these platforms may contain language characteristics specific to the country. To evaluate the performance in identifying offensive messages that include these culturally specific characteristics, we decided that our testing sample should only contain messages written in Spanish and related to Chilean posts. This was feasible considering that the dataset from [37] was obtained exclusively in the Chilean context. Therefore, we reserved 30% of this dataset for inclusion in the testing set. The remainder of the merged data was used for training.

We assessed the classification performance using traditional metrics such as f1-score, recall, and precision. Additionally, we applied 10-fold cross-validation to gauge the model's stability and generalizability across different subsets of the data. The 10-fold cross-validation involved dividing the training set into training and validation samples, and averaging the corresponding metrics to determine the classification performance.

Figure 2 displays the classification performance obtained from our experiments. In the validation scenario, we achieved f1-score, precision, and recall values of 0.808, 0.812, and 0.804, respectively. After training a classifier with all the data from the validation scenario, we evaluated it on the testing set. In this case, our metrics decreased by approximately 5%, indicating good performance and stability when identifying offensive language in messages related to the Chilean context.

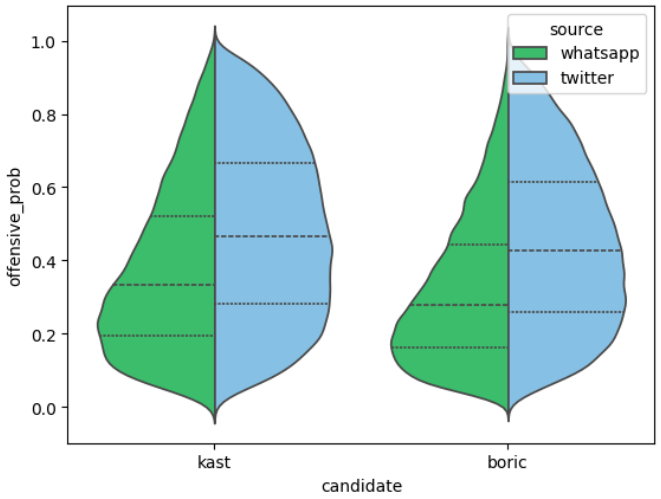

Fig. 3. Violin plots displaying the distribution of offensive content probability in messages related to candidates pro-Kast and pro-Boric from WhatsApp and Twitter sources. Horizontal dashed lines represent quantiles.

VI. Inferring Offensive Language

Using our previously trained and evaluated classifier, we obtained the probability of being offensive for every message in our *election dataset*. Figure 3 shows the probability distribution of messages divided into candidates and platforms. On the one hand, WhatsApp messages display a positively skewed distribution, having an interquartile range of 0.372 ($q1 = 0.271$ and $q3 = 0.643$). On the other hand, Twitter messages show a different distribution where we obtained an interquartile range of 0.253 ($q1 = 0.210$ and $q3 = 0.463$). This indicates that Twitter messages exhibit a narrower range of variation in terms of offensive language usage compared to WhatsApp. The lower interquartile range on Twitter, with $q1 = 0.210$ and $q3 = 0.463$, suggests that the majority of Twitter messages fall within a relatively constrained offensive language usage spectrum. In contrast, WhatsApp messages, with an interquartile range of 0.372 and $q1 = 0.271$ and $q3 = 0.643$, display a wider spread of offensive language usage, with a larger proportion of messages exhibiting varying degrees of offensiveness. This distinction underscores the platform-specific differences in how offensive language is employed, possibly reflecting variations in user behavior, content moderation, or the nature of discussions on these platforms during polarized events Additionally, the median probability value for Twitter ($q2 = 0.447$) was higher than that of WhatsApp ($q2 = 0.292$), implying that, on average, Twitter messages tend to have a greater likelihood of containing offensive language during polarized events. This finding underscores the need for platforms and communities to consider tailored strategies for mitigating offensive language and promoting constructive dialogue, particularly during events marked by heightened polarization, to foster a more inclusive and respectful online environment.

Regarding the comparison between candidates, there are no clear differences to suggest a prevalence of offensive language related of supporting a specific community when we compared the same source. In these scenarios, we identified that offensive probabilities were no greater than 5% when comparing candidates in the same platform. For instance, the median of the offensive probability distribution in pro-Kast community for WhatsApp displays a value of 0.466, while pro-Boric in the same platform shows a value of 0.435.

To comprehend the type of messages identified for our classifier as offensive, we selected the most and least likely messages in each platform and candidate. Tables III and IV present these examples of messages identified by our classifier. The messages with high offensive probabilities (Table III) are characterized by explicit insults, aggressive tones, and derogatory labels, reflecting polarized and contentious interactions. These messages often contain swear words, personal attacks, or derogatory terms targeting specific groups or individuals. On the other hand, the messages in Table IV, with low probabilities of being offensive, showcase a more civil and respectful communication style. These messages are marked by expressions of gratitude, well-wishes, and positive affirmations, indicating a more constructive and harmonious discourse.

A. Offensive language interpretability

Interpretability in the context of offensive language is crucial for ensuring model transparency, understanding contextual nuances, reducing biases, improving model accuracy, ensuring ethical AI practices, facilitating human oversight, and its adaptability to various models. These factors contribute to a more robust and reliable approach to handling offensive language in digital communication.

In previous work, LIME has been used to calculate the average importance given to words by a particular model in the task of hate speech detection [35], [40]–[42]. Given the computational resource and time constraint of applying LIME model, we selected, in each community and platform, the top 100 most offensive messages as well as the top 100 least offensive. Thus, for every message we applied the LIME model, computed the top 5 most predictive words and their attention for each sentence in the these samples. The total score for each word in each community and platform is calculated by summing up all the attentions for each of the sentences where the word occurs in the top 5 LIME features. The average predictive score for each word is calculated by dividing this total score by the occurrence count of each word [35].

Figure 4 shows the the top 5 most and least probable words given their average predictive score (APS) identified by LIME. One of the main observations across the candidates and platforms is that derogatory and swear terms are weighted more heavily in determining how likely a message is to be offensive. For instance, the terms *asqueroso (nasty or disgusting)*, *cerdos (pigs)* and *criminales (criminals)* appear with an APS greater than 0.2%. This means that, if one of these terms was removed from the messages, the probability of being offensive will

Source	Community	Message	Prob
Whatsapp	pro-Kast	Hopefully they lock up those shitty left-wing scumbags, resentful liars :swearing_emoji :swearing_emoji :swearing_emoji	0.991
Whatsapp	pro-Boric	Yellow jackets are useless, damn right-wing scumbags	0.974
Twitter	pro-Kast	You are a damn communist pig, bourgeois bribe-taker, harasser, ignorant liar, thief, drug addict	0.996
Twitter	pro-Boric	This right-wing idiot is also blind, Piñera is an abuser, thief	0.994

TABLE III
EXAMPLES OF MESSAGES WITH THE HIGHEST PROBABILITIES OF BEING OFFENSIVE IDENTIFIED IN EACH COMMUNITY AND PLATFORM. MESSAGES WERE TRANSLATED TO ENGLISH FOR A BETTER COMPREHENSION.

Source	Community	Message	Prob
Whatsapp	pro-Kast	Glory to the Father, glory to the Son, and glory to the Holy Spirit	0.005
Whatsapp	pro-Boric	Thank you very much to everyone for the work and effort in this beautiful campaign	0.007
Twitter	pro-Kast	Hear us, Lord, we pray for the well-being of our country, amen	0.006
Twitter	pro-Boric	May it go very well for you, Gabriel, you have done a good campaign	0.012

TABLE IV
EXAMPLES OF MESSAGES WITH THE LOWEST PROBABILITIES OF BEING OFFENSIVE IDENTIFIED IN EACH COMMUNITY AND PLATFORM. MESSAGES WERE TRANSLATED TO ENGLISH FOR A BETTER COMPREHENSION.

decrease. Another interesting observation in these results is that a few hashtags, for instance *borisesviolenciayterrorismo (boricisviolenceandterrorism)* also exhibit positive average predictive score, specifically in the pro-Kast community in Twitter. Furthermore, there are several terms, specifically in the pro-Boric community in WhatsApp, displaying a APS higher than 0.2%, showing the importance that these could have in this community.

In contrast, we also displayed those terms that have negative average predictive score. In other words, the terms that make a message less likely to be offensive according to our model. Our results suggest that some terms like *hijo (son)*, *gloria (glory)* and *bendiciones (blessing)* became a message 20% lesser when they appear.

VII. DISCUSSION

We studied communication of political activists in the context of the 2021 run-off presidential election in Chile, which confronted the far-right candidate José Antonio Kast with the left-wing candidate Gabriel Boric. This case provides an excellent opportunity to study online incivility as it was the most polarized election in over 30 years, as none of the centrist coalitions made it to the run-off.

It is essential to acknowledge that our collected data may contain inherent biases due to limitations such as the specific period, keywords chosen, and the nature of public WhatsApp groups. One of the candidates (Boric) had a notoriously more active campaign activity mediated by WhatsApp. However, our approach has been tailored to ensure the broadest possible coverage of relevant online conversations connected to the candidates and their political communities. This methodological choice, which aligns with best practices in the field [33], [43], yields a dataset that, though not perfect, provides us with the best available data for analyzing and comparing the prevalence of hate speech on Twitter and WhatsApp during the 2021 Chilean presidential election.

In Section VI, our results suggest that offensive language was more prevalent on Twitter than on WhatsApp. To explain the reason for this finding, we describe the following differences in social media affordances that support it:

a) Public vs. Private Nature: The public nature of Twitter allows for greater visibility and potential for amplification, making it more conducive to the spread of toxic content as strategies of viralization or political confrontation. Conversely, WhatsApp is essentially a private messaging app designed for communication within closed or niche groups or among individuals -either way, a more controlled audience.

b) Audience Size and Diversity: The extensive user base and the potential for interaction with a diverse range of people on Twitter can increase exposure to differing opinions and the likelihood of encountering hate speech. In contrast, in chat platforms, such performativity may be less relevant as relationships tend to happen with more like-minded people in more symmetric networks [44]. This may result in less exposure to diverse perspectives and reduced opportunities for offensive content.

c) Anonymity and Accountability: Twitter provides users with a higher degree of anonymity compared to WhatsApp, making it easier for individuals to engage in toxic behavior without fearing immediate consequences or personal accountability. In contrast, WhatsApp generally requires users to have identifiable profiles tied to their phone numbers, which can increase accountability and potentially discourage hate speech.

d) Communication Dynamics: Twitter is designed to facilitate rapid, real-time communication and public conversations among users. The brevity of Twitter content (called tweets), the ease of sharing and retweeting content, and the directionality, that is, the ability to mention and reply to others, can help create an environment where messages can quickly escalate into offensive exchanges. On the other hand, WhatsApp, with its emphasis on private one-to-one or group conversations, often fosters more personal and intimate interactions, which may discourage the propagation of hate speech.

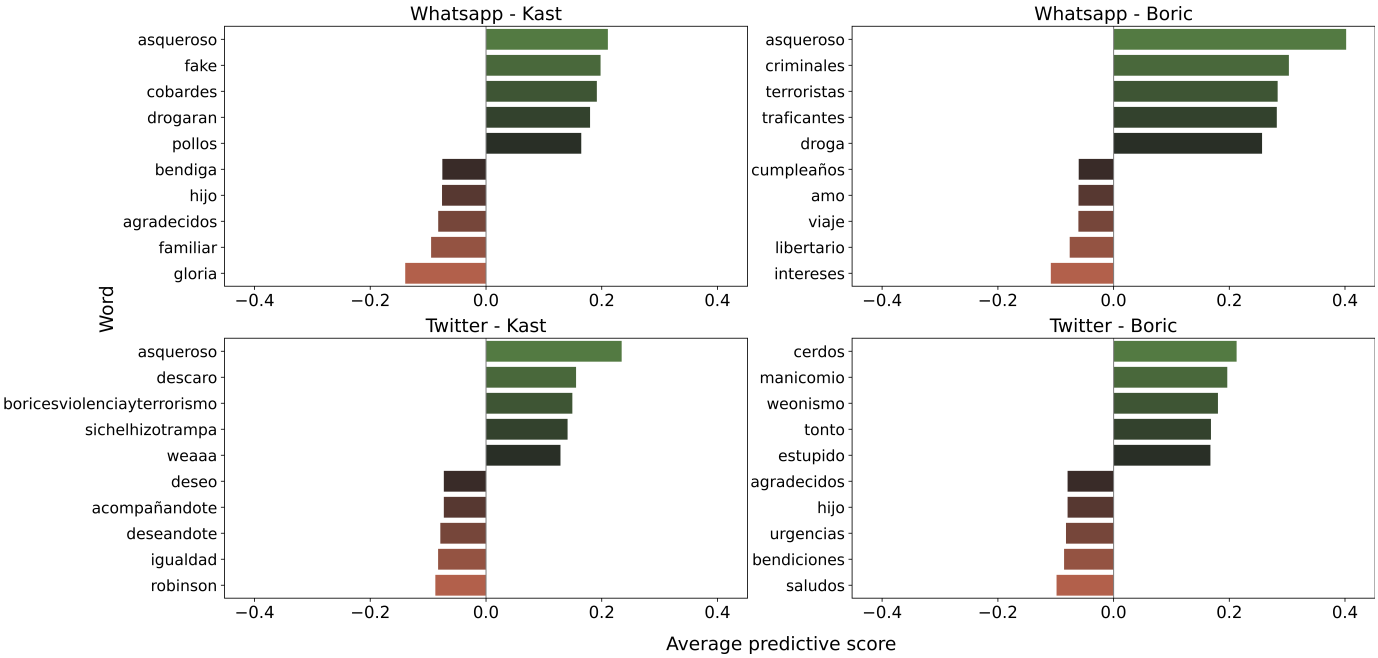

Fig. 4. Top-5 most and least likely words to determine if a message is offensive using LIME. The model interpretability was applied on every platform and community.

This could explain why platforms that prioritize content by engagement metrics, such as Twitter, tend to privilege incendiary content, producing a vicious circle of online toxicity [45], [46].

e) Algorithmic Influence: The algorithms employed by Twitter and WhatsApp to curate and display content may play a role in the difference in toxic behavior. Twitter's algorithm prioritizes content based on engagement metrics, which can inadvertently amplify provocative or controversial messages [47], [48]. In contrast, WhatsApp generally displays messages chronologically within groups or private conversations, without employing extensive algorithmic content recommendations.

Overall, considering these variations in affordances, Twitter is expected to have higher toxicity levels compared to WhatsApp in the realm of political communication.

VIII. Conclusion

Our study provides valuable insights into offensive language dynamics during a polarized political event across social media platforms. To our knowledge, this is the first study to systematically compare offensive discourse on Twitter and WhatsApp for the same event. Using state-of-the-art text classification methods and a comprehensive dataset, we uncovered significant disparities, with Twitter exhibiting up to 15% more offensive content on average. This finding underscores the importance of platform-specific affordances in shaping online discourse. While differences within platforms among supporters of distinct political candidates were less pronounced, they warrant further investigation. These results have important implications for platform governance, political communication strategies, and efforts to foster healthier online discourse during polarized events.

References

[1] I. Gagliardone, M. Pohjonen, Z. Beyene, A. Zerai, G. Aynekulu, M. Bekalu, J. Bright, M. A. Moges, M. Seifu, N. Stremlau *et al.*, "Mechachal: Online debates and elections in ethiopia-from hate speech to engagement in social media," *Available at SSRN 2831369*, 2016.

[2] C. Machado, B. Kira, V. Narayanan, B. Kollanyi, and P. Howard, "A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections." in *Companion proceedings of the 2019 World Wide Web conference*, 2019, pp. 1013–1019.

[3] G. Sigurbergsson and L. Derczynski, "Offensive language and hate speech detection for danish," in *Proceedings of the International Conference on Language Resources and Evaluation: LREC 2020*. European Language Resources Association, 2020, pp. 3498–3508.

[4] J. Guberman, C. Schmitz, and L. Hemphill, "Quantifying toxicity and verbal violence on twitter," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 2016, pp. 277–280.

[5] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, 2017, pp. 86–95.

[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE access*, vol. 6, pp. 13 825–13 835, 2018.

[7] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," *arXiv preprint arXiv:1809.08651*, 2018.

[8] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in *Proceedings of the third workshop on abusive language online*, 2019, pp. 46–57.

[9] T. Wijesiriwardene, H. Inan, U. Kursuncu, M. Gaur, V. L. Shalin, K. Thirunarayan, A. Sheth, and I. B. Arpinar, "Alone: A dataset for toxic behavior among adolescents on twitter," in *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*. Springer, 2020, pp. 427–439.

[10] P. Rossini and R. Maia, "Characterizing disagreement in online political talk: Examining incivility and opinion expression on news websites and facebook in brazil," *Journal of Deliberative Democracy*, vol. 17, no. 1, pp. 90–104, 2021.

[11] E. Sydnor, "Platforms for incivility: Examining perceptions across different media formats," in *Studying Politics Across Media*. Routledge, 2020, pp. 97–116.

[12] J. L. Davis and J. B. Chouinard, "Theorizing affordances: From request to refuse," *Bulletin of science, technology & society*, vol. 36, no. 4, pp. 241–248, 2016.

[13] S. Faraj and B. Azad, "The materiality of technology: An affordance perspective," *Materiality and organizing: Social interaction in a technological world*, vol. 237, no. 1, pp. 237–258, 2012.

[14] D. H. Kim and N. B. Ellison, "From observation on social media to offline political participation: The social media affordances approach," *New media & society*, vol. 24, no. 12, pp. 2614–2634, 2022.

[15] S. K. Evans, K. E. Pearce, J. Vitak, and J. W. Treem, "Explicating affordances: A conceptual framework for understanding affordances in communication research," *Journal of computer-mediated communication*, vol. 22, no. 1, pp. 35–52, 2017.

[16] I. Hutchby, "Technologies, texts and affordances," *Sociology*, vol. 35, no. 2, pp. 441–456, 2001.

[17] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the association for computational linguistics*, vol. 7, pp. 597–610, 2019.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] N. Herrada Hidalgo, M. Santos, and S. Barbosa, "Affordances-driven ethics for research on mobile instant messaging: Notes from the global south," *Mobile Media & Communication*, p. 20501579241247994, 2024.

[20] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[21] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[22] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the international AAAI conference on web and social media*, vol. 12, no. 1, 2018.

[23] A. Arango Monnar, J. Perez, B. Poblete, M. Saldaña, and V. Proust, "Resources for multilingual hate speech detection," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Seattle, Washington (Hybrid): Association for Computational Linguistics, Jul. 2022, pp. 122–130. [Online]. Available: https://aclanthology.org/2022.woah-1.12

[24] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, 2019, pp. 363–370.

[25] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5838–5844.

[26] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, pp. 477–523, 2021.

[27] BBCNews, "Kast vs. Boric: las principales propuestas de los rivales más antagónicos que ha tenido Chile en las últimas décadas," https://www.bbc.com/mundo/noticias-america-latina-59383712, 2021, [Online; accessed 19-May-2023].

[28] FastCheck, "Programa presidencial de José Antonio Kast tiene un apartado que se llama "Coordinación Internacional Anti-Radicales de Izquierda"," https://tinyurl.com/fastcheck2021, 2021, [Online; accessed 19-May-2023].

[29] BBCNews, "Gabriel Boric: en qué consiste la agenda transformadora con la que llega a la presidencia de Chile," https://www.bbc.com/mundo/noticias-america-latina-59723286, 2021, [Online; accessed 19-May-2023].

[30] ElPaís, "Chile consolida el primer Gobierno feminista latinoamericano," https://elpais.com/internacional/2022-03-11/la-consolidacion-del-primer-gobierno-feminista-de-chile.html, 2021, [Online; accessed 19-May-2023].

[31] H. Sarmiento, F. Bravo-Marquez, E. Graells-Garrido, and B. Poblete, "Identifying and characterizing new expressions of community framing during polarization," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 841–851.

[32] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, p. 011047, 2014.

[33] K. Garimella and D. Eckles, "Images and misinformation in political groups: Evidence from whatsapp in india," *arXiv preprint arXiv:2005.09784*, 2020.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[35] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "A deep dive into multilingual hate speech classification," in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*. Springer, 2021, pp. 423–439.

[36] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors*, vol. 19, no. 21, p. 4654, 2019.

[37] A. Arango, J. Pérez, B. Poblete, V. Proust, and M. Saldaña, "Multilingual resources for offensive language detection," in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 2022, pp. 122–130.

[38] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 54–63.

[39] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1415–1420. [Online]. Available: https://aclanthology.org/N19-1144

[40] D. N, R. Avireddy, A. Ambalavanan, and B. R. Selvamani, "Hate speech detection using LIME guided ensemble method and distilbert," in *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Gandhinagar, India, December 13-17, 2021*, ser. CEUR Workshop Proceedings, P. Mehta, T. Mandl, P. Majumder, and M. Mitra, Eds., vol. 3159. CEUR-WS.org, 2021, pp. 396–411. [Online]. Available: https://ceur-ws.org/Vol-3159/T1-39.pdf

[41] M. A. Ibrahim, S. Arifin, I. G. A. A. Yudistira, R. Nariswari, A. A. Abdillah, N. P. Murnaka, and P. W. Prasetyo, "An explainable ai model for hate speech detection on indonesian twitter," *CommIT (Communication and Information Technology) Journal*, vol. 16, no. 2, pp. 175–182, 2022.

[42] A. Maronikolakis, P. Baader, and H. Schütze, "Analyzing hate speech data along racial, gender and intersectional axes," in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2022, pp. 1–7.

[43] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto, "(mis) information dissemination in whatsapp: Gathering, analyzing and countermeasures," in *The World Wide Web Conference*, 2019, pp. 818–828.

[44] C. D. Soares, L. A. Joia, D. Altieri, and J. G. L. Regasso, "What's up? mobile instant messaging apps and the truckers uprising in brazil," *Technology in Society*, vol. 64, p. 101477, 2021.

[45] J. A. Frimer, H. Aujla, M. Feinberg, L. J. Skitka, K. Aquino, J. C. Eichstaedt, and R. Willer, "Incivility is rising among american politicians on twitter," *Social Psychological and Personality Science*, vol. 14, no. 2, pp. 259–269, 2023.

[46] L. Munn, "Angry by design: toxic communication and technical architectures," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–11, 2020.

[47] A. Hasell, "Shared emotion: The social amplification of partisan news on twitter," *Digital Journalism*, vol. 9, no. 8, pp. 1085–1102, 2021.

[48] G. Corsi, "Evaluating twitter's algorithmic amplification of low-trust content: An observational study," *arXiv preprint arXiv:2305.06125*, 2023.