Hate Speech in the Political Domain: The Milei Corpus

Lyna Tahraoui Rarrbo ltahraoui001[1]
Aitor Taboada Muñiz ataboada005[1]
Jon Apaolaza Larraya jonfelix.apaolaza[2]

[1]@ikasle.ehu.eus [2]@ehu.eus

Abstract

We introduce the Milei Corpus, a small but extensible dataset in Argentinian Spanish created to measure Hate Speech in the political domain. It consists of 140 Reddit comments annotated by the authors of this document. Through this workpiece, we attempt to contribute to Hate Speech analysis in social media by following best practices. To achieve this goal, we provide a description of the process and methods carried out: we massively collected Argentina politics-related comments and then manually selected 140 for annotation. We wrote the guidelines for the annotation procedure and tested them by calculating inter-annotator agreement. We also highlight the annotators demographics, the difficulties when creating Hate Speech corpora and the limitations we found during the process.

1 Introduction

In recent years, Hate Speech has become an increasingly important issue in academic research. With the widespread use of social media platforms and online forums, Hate Speech has proliferated across digital spaces, contributing to toxic environments, online harassment, and real-world violence. As a result, researchers have been working to better understand and mitigate the impact of Hate Speech in online contexts. One of the critical steps when addressing this problem is the creation of comprehensive, well-annotated Hate Speech corpora, which can serve as the foundation for developing effective detection and intervention strategies. However, the development of these corpora requires careful consideration of various factors, including the definition of Hate Speech, the acquisition of data, and the need for diverse representation across annotator demographics.

This document examines the creation of such a corpus, describing the whole process and commenting on the decisions taken at each step. Given that the majority of similar resources that can be found are centered exclusively on the English language, we hope the easily imitable process here described, as well as the final resource, will be of interest.

This document is structured as follows: in Section 1 we introduce the work done, defining Hate Speech and offering a comment on what the resource consists of, its aims and applications. In Section 2, we explain how we obtained the raw data to be annotated, and give a brief comment on the characteristics of the final resource. In Section 3, we explore the definition of the guidelines to annotate, from our initial draft to their final form. Finally, we offer some conclusions and ideas for future work in Section 4.

1.1 Hate Speech

Defining Hate Speech is a complex task, primarily due to the subjective nature of what constitutes harmful or discriminatory language. While most people would agree that Hate Speech involves expressions that incite violence, hatred, or discrimination against individuals or groups based on attributes such as race, religion, gender, or sexual orientation, the boundaries of this definition are not always clear. Nonetheless, a clear delimitation of what does and does not constitute Hate Speech is fundamental to the development of our corpus, for all posterior work depends on it.

Upon undertaking this project, we quickly came to the conclusion that the amount of time required to review the relevant literature and define in exact terms what we deem Hate Speech greatly surpasses the scope of this assignment. For this reason, we looked for reputable institutions that regularly work in Hate Speech and took a fitting definition that allowed us to move forward with the project. UNESCO (2024) defines Hate Speech as "*any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on*

the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor". This is the definition and understanding of Hate Speech we will work upon.

We also took great interest in Sachdeva et al. (2022), as they served as a starting point for our guidelines definition. We go more in detail about the relationship between their work and ours in Subsection 3.1.

1.2 Resource Description

The resource we present consists of 140 annotated Reddit comments. Each of these annotation units was manually annotated on 5 different labels: *Target, Topic, Attack/Defend, Violence* and *Hate Speech*. For a more in depth commentary of each of those labels, see Subsection 3.1. The entirety of the corpus consists of posts obtained as explained in Subsection 2.1.

Given the small scope of this assignment, and taking into account that Hate Speech is relatively rare (that is, most of the Internet is not Hate Speech), we chose a very specific source of data: social media posts made in Argentinian Reddit communities discussing Argentina's current president Javier Milei. As a result, our corpus exclusively covers the Argentinian variety of Spanish. We did not find the need to limit the length of the comments, but all of them are demarcated in time from Javier Milei's assumption of office up until late February 2025.

The vast majority of the comments contained in our corpus are informal in nature, but some more structured ones can also be found. Some examples extracted from the corpus can be found at Appendix B.

Each of the individual comments belongs to their respective owner, and has been used under Reddit's API's terms[1]. The annotations are licensed under CC BY-NC-SA 4.0[2].

The Milei Corpus is publicly available on GitHub[3]. The code for retrieving comments from Reddit and conducting inter-annotator agreement analysis has been made available as well.

[1] https://redditinc.com/policies/data-api-terms

[2] https://creativecommons.org/licenses/by-nc-sa/4.0/

[3] https://github.com/Jonapa/BuildingLanguageResources_AitorT_LynaT_JonA/

1.3 Aim and Application

This kind of corpus is crucial for training machine learning models designed to detect and filter Hate Speech in online content. Such models can be used to automatically identify harmful or discriminatory language, helping to create safer and more inclusive online environments.

Additionally, Hate Speech corpora can be used to study patterns and trends in hateful language, enabling researchers to better understand the psychology behind online toxicity and its societal impacts. This analysis can help governments and organizations take more effective approaches when combating Hate Speech.

2 Data

This section will examine the data collection and filtering process, and present the corresponding Data Statement for the resulting corpus.

2.1 Collection and Filtering

We collected comments from the Reddit social media platform over a period spanning from November 2023, when Milei assumed office as the President of Argentina, to February 2025. The collected data present the characteristics described in Subsection 1.2.

Our data collection methodology has focused on retrieving post comments from two popular Argentinian subreddits: `/r/argentina`[4], which generally features more pro-Milei comments, and `/r/RepublicaArgentina`[5], which leans more towards Peronism. More specifically, we focused on identifying posts (and their comments) containing events likely to trigger polarizing opinions, such as foreign policy or inflation. To achieve this, we manually explored posts within the selected subreddits. This selection aimed to balance comments across the spectrum of Argentinian politics. As a result, a variety of perspectives, including those targeting other political figures and ideologies, are present. It is worth adding that, due to the dynamic nature of social media, comments may be deleted, edited, upvoted, or downvoted, and new ones may be added over time. Consequently, our comment exploration considers only those posts available up until February 22nd. Therefore, when replicating the data retrieval process, this factor should be taken into account, as the collected comments

[4] https://www.reddit.com/r/argentina/

[5] https://www.reddit.com/r/RepublicaArgentina/

may not be fully comparable to a newly performed scrapping.

After scraping approximately $1,000$ comments from about 30 posts in the aforementioned Subreddits, we performed another manual filtering process to select those that best suited our research interests. Moreover, we did not apply any pre-processing, as we consider grammatical features, emoticons, and other linguistic phenomena to be crucial for identifying the intricacies of the matter at hand. Removing these elements could risk losing important context and meaning.

2.2 Data Statement

In order to build the dataset, 1,000 comments written between November 2023 and February 2025 in mainly informal, slang-containing es-AR (Argentine Spanish) were extracted from Reddit, and a further selection of 140 comments was performed to balance potential and non-potential Hate Speech.

In this context, the speech situation is a spontaneous, asynchronous interaction between users, whose intended audience is the Reddit community interested in the Argentinian politics (and the various topics within it). Speakers' age and gender is unknown, but, given the topics discussed, they probably are above the legal age for voting. Speakers are Argentinian and their first language is Argentinian Spanish.

Annotators and annotation guideline developers are a 25-years-old woman with a background in linguistics, and two 24-years-old and 25-year-old men with a computer science training. All of them are native Spanish speakers from Spain.

3 Guidelines, Annotation and Inter-Annotator Agreement

This section presents the guidelines created for annotating the retrieved Reddit comments, along with a qualitative analysis of a small subset of the data. In this subset, we acted as annotators to assess the quality of the guidelines and the resulting annotations, using inter-annotator agreement measures to ensure consistency and accuracy. This annotation process took approximately 40-50 minutes in total for each annotator, with an average of 2 minutes spent on each annotation (20 annotation units, this is, comments). It should be noted that the time spent may vary depending on the length and complexity of the comments and, obviously, the number of samples.

3.1 Guidelines

The guidelines to annotate the dataset were written as a shorter and simpler version of the items proposal of Sachdeva et al. (2022), fitting it to the scope of this assignment. In the Measuring Hate Speech Corpus, described in said paper, each observation includes 10 ordinal labels: *sentiment, disrespect, insult, attacking/defending, humiliation, inferior/superior status, dehumanization, violence, genocide*, and a 3-valued hate speech benchmark label. Of those, we chose the five we deemed more interesting: *Target*, to know who are the main targets of Hate Speech; *Topic*, to understand what Hate Speech is based on; *Attack-Defend*, to know whether Hate Speech is used to defend or attack ideas; *Violence*, to see the kind of actions Hate Speech consists of; and *Hate Speech*, as detecting it is the main purpose of the annotation. It is worth noting that our *Violence* label includes verbal violence as well.

On the initial set of guidelines, *Hate Speech* and *Violence* could be labeled as "Yes", "No" or "Unclear"; *Attack-Defend* as "Attacking", "Defending" or "Neither attacking nor defending"; *Target* could be labeled as "Group" or "Individual", and *Topic* was the only one allowing multi-labeling and included the labels "Policy", "Origin", "Ideology" and "Sexuality".

As commented above, Hate Speech is a difficult task since it involves subjectivity and interpretation. Thus, in order to achieve a high IAA among the three annotators, the first version of the guidelines included a description of the task, the definition of Hate Speech from UNESCO (2024) and also an explanation of the items and labels to annotate, providing examples for each label. The initial guidelines can be found at Appendix A

After writing the preliminary guidelines, 20 comments were randomly selected to be annotated. Subsequently, the IAA agreement was calculated to test whether the preliminary guidelines were definitive or if changes were needed.

3.2 IAA Measures

To assess the agreement among multiple annotators, we use Fleiss' kappa, as the labels we have defined are categorical. Additionally, since the *Topic* label is formulated as a multi-label category, it is necessary to recalculate Fleiss' kappa for the different values within the class. Thanks to this, we are able to measure the agreement on a numerical

scale, allowing us to better analyze which guidelines may require adjustments or show the most disagreement. In detail, the values range from -1 to 1, with values below 0 indicating performance worse than chance, and values above 0 reflecting increasing values of agreement. We have utilized the R package `irr` (Gamer, 2012) to perform these measurements.

3.3 Evaluation of Annotation Consistency and Guidelines Rewriting

When inspecting the agreements obtained using the first set of developed guidelines (see Table 1, **Initial** column), we observe that all labels exceed the substantial threshold (> 0.61) except for the *Violence* label. We are able to reach agreement on the *Target* and whether it is being *Defended* or *Attacked*, but we encounter discrepancies when determining whether *Hate Speech* and *Violence* are present in the comment. Additionally *Sexuality* shows a noticeable worse agreement than the other topics.

After analyzing the first IAA, we were able to highlight several issues regarding *Hate Speech*, "Sexuality" label and *Violence*.

As represented in Figure 13, when having "Attacking" and "Yes" label in *Violence*, usually there is a "Yes" in *Hate Speech*. However, if there is an attack but no violence, *Hate Speech* is not a "Yes" for the annotators. This is, *Violence* and *Hate Speech* are directly related, and, when reviewing the comments, all the ones containing disagreement in *Violence* also contain disagreement in *Hate Speech*. In this case, we decided to remove the "Unclear" label from *Hate Speech* for it to be guided only by the *Violence* indicator and the Hate Speech definition, hoping to improve agreement when leaving it as "Unclear" is no longer an option.

Topic item remained multi-label, as it didn't create disagreement except for "Sexuality". Subsequently, we decided not to modify this item. In practice, only three comments were labeled on this topic, and the only one with agreement was explicitly using the word *"feminismo"*. Only the woman annotator tagged the other two comments as having "Sexuality" in *Topic*.

As for *Violence*, we analyzed the comments with "Yes" full agreement and those ones contained words as *"basura", "ladrones"* or *"pelotudo"*. Comments with "No" and full agreement contained no explicit insults, but could contain words as *"caradura"* and usually represented an opinion or a positive sentiment. In addition, as the "Unclear" label didn't have a full agreement in any comment but was the most used label when there was an annotator annotating differently, we decided to remove it as well.

Finally, in order to fine-tune our guidelines table, we redefined the question in *Violence* as it had the lowest agreement, we put the *Hate Speech* item as the last one to annotate (to be able to label first and take into account *Attack-Defend* and *Violence*), and we modified the explanations in *Target* to clarify whether referencing multiple individual people fell under "Individual" or "Group".

Label	**Initial**	**Final**
Hate Speech	0.652	0.907
Target	0.886	0.931
Topic: Policy	1	0.794
Topic: Origin	1	1
Topic: Ideology	0.866	0.688
Topic: Sexuality	0.63	0.732
Attack / Defend	0.856	0.815
Violence	0.406	0.794

Table 1: Fleiss' Kappa values for each label using the initial and the final set of guidelines.

Considering these factors, we revised our initial draft into our final version (see Table 1, **Final** column). We can observe an overall increase in agreement values; however, some labels experienced a partial decline, though the agreement remains substantial in all cases. The topics of *Policy* and *Ideology* are the most affected, indicating that the initial consensus was not successfully transferred to the final draft of the guidelines. We theorize that these two labels might have some overlap that makes it difficult for annotators to choose one, the other, or both.

4 Conclusions and Future Work

Annotating Hate Speech is a challenging task because of its complex and subjective nature. This subjectivity makes it difficult to create agreed-upon criteria for annotation, as one must not only define what constitutes Hate Speech, but its components as well. Additionally, annotators' personal biases can influence their perception of each individual item, and poorly selected guidelines will result in low annotator agreement.

Given that the whole process has been detailed and is easily reproducible, and that the code for ob-

taining the Reddit comments is available, it could be interesting to follow a similar process and expand the corpus to cover other varieties of Spanish or even other areas of interest rather than just the politics one. In particular, although "Sexuality" and "Origin" are considered in the corpus, their presence has been quite limited, and further exploration of these concepts could be an interesting line for future work.

Acknowledgments

We would like to thank our classmate Aimar Sagasti for providing the necessary resources to support the writing of this work. Additionally, we extend our gratitude to our instructors Itziar and Izaskun for their valuable insights and constructive feedback.

Limitations

While we are satisfied with the final state of the Milei corpus, we wish to acknowledge some of its limitations.

As of now, the corpus is fairly small, consisting of 140 items. It is also very specific, exclusively containing comments about Javier Milei written in Argentinian Spanish. For these reasons, we do not advise utilizing the Milei on its own to train machine learning models.

We would also like to acknowledge the possibility of annotator bias in our data. The annotators are of similar age (24/25 years old) and were all raised in Spain. In addition, none of us speaks the Argentinian variety of Spanish, nor has academic formation on Hate Speech.

References

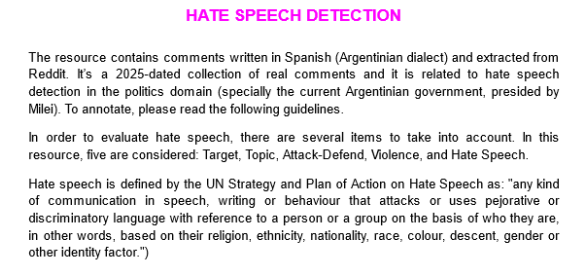
Python Reddit API Wrapper Development. 2025. Praw: The python reddit api wrapper.

Matthias Gamer. 2012. Various coefficients of interrater reliability and agreement.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

UNESCO. 2024. What you need to know about hate speech.

A Initial Guidelines



HATE SPEECH DETECTION

The resource contains comments written in Spanish (Argentinian dialect) and extracted from Reddit. It's a 2025-dated collection of real comments and it is related to hate speech detection in the politics domain (specially the current Argentinian government, presided by Milei). To annotate, please read the following guidelines.

In order to evaluate hate speech, there are several items to take into account. In this resource, five are considered: Target, Topic, Attack-Defend, Violence, and Hate Speech.

Hate speech is defined by the UN Strategy and Plan of Action on Hate Speech as: "any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.")

Figure 1: Text appearing on top of the Guidelines, including the definition of Hate Speech.

Hate Speech **[Does this comment contain hate speech?]**	Yes	The intention of harmful attack is clear, the features of the comment are described in the definition above	JAJAJJA CHUPENSE UNA JAPI KUKAS PUTOS
	Unclear	The annotator is not sure about whether the comment is hate speech or not	Da vergüenza. Pensé que ésta clase de presidentes solo aparecían en ficción. Lo peor es que no importa las boludeces que haga, los tenés a los liberbobos diciendo que lo votarían otra vez 🤦‍♀️
	No	The comment may be an attack or not, but cannot be identified as hate speech according to the definition above	que vergüenza me da que sea mi presidente.

Figure 2: Initial Guideline for annotating the Hate Speech label.

Target **[Who is/are the target(s) of this comment?]**	Group	The target of a comment are non-specific people, but a identifiable group	Está lleno de zurditos mugrosos este sub
	Individual	The target is named	ponete a laburar milei lcdtm

Figure 3: Initial Guideline for annotating the Target label.

Topic (multi-label) **[Why is/are the target(s) commented for?]**	Policy	The target is adressed because of their economic measures, regarding foreign issues...	El problema del liberalismo es que todos los liberalismos del mundo trabajan para EEUU & UK.
	Origin	The target is addressed because of their ethnicity, colour, descendt, nationality...	Entonces, cuando se van los judios?
	Ideology	The target is addressed by religion or political ideology	Kukitas lloronas llorando, que lindo jajaja me mueeee 🤣🤣🤣
	Sexuality	The target is addressed by aims of gender or sexual orientation	no lo enfocan nunca pero notaron que milei camina como si tuviera el culo roto! tiene algun problema en las piernas, la cadera o esta paspado?

Figure 4: Initial Guideline for annotating the Topic label.

Violence **[This comment calls for using verbal or physical violence against the target(s)?]**	Yes	There are clear and explicit forms of violence (insults, suggestions about harmful actions against the target...)	Me parece re mal que Lali le haga buling al minusválido mental, que se meta con alguien de su tamaño.
	Unclear	There annotator cannot say if the expressions used in the comment are violence in an explicit form	Milei es la representación de qué pasaría si un tuitero tuviera mayor poder más allá del escribir pelotudeces en una red social
	No	There's no form of violence expressed in the comment	A la izquierda, porque aunque puedan decir alguna pavada de vez en cuando, tienen mejores valores, y se necesita alguien que piense en el trabajo de la gente, no en las ganancias de las multinacionales. (Hay muchas razones para no votar a milei, cualquier razón es buena).

Figure 6: Initial Guideline for annotating the Violence label.

B Final Guidelines

Attack-Defend **[Is this comment attacking or defending the target(s)?]**	Attacking	The comment is attacking the target, even if it's defending other people that cannot be identified as the main target of the comment	Massa no puede ser tan HIJO DE PUTA
	Neither attacking nor defending	The comment is rather a neutral opinion or statement and it's not attacking nor defending the target	Te puede gustar o no Milei, pero dejando la política aparte, creo que todos podemos empatizar con él en el momento en que no sabe dónde meterse después de entrar y no entender una goma de inglés.
	Defending	The aim of the comment is defending the target	Excelente el javo emperador

Figure 5: Initial Guideline for annotating the Attack-Defend label.

HATE SPEECH DETECTION

The resource contains comments written in Spanish (Argentinian dialect) and extracted from Reddit. It's a 2025-dated collection of real comments and it is related to hate speech detection in the politics domain (specially the current Argentinian government, presided by Milei). To annotate, please read the following guidelines.

In order to evaluate hate speech, there are several items to take into account. In this resource, five are considered: Target, Topic, Attack-Defend, Violence, and Hate Speech.

Hate speech is defined by the UN Strategy and Plan of Action on Hate Speech as: "any kind of communication in speech, writing or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.")

Figure 7: Text appearing on top of the Guidelines, including the definition of Hate Speech.

Target **[Who is/are the target(s) of this comment?]**	Group	The comment may include specific people (named), but the target is a whole identifiable group	Está lleno de zurditos mugrosos este sub
	Individual	The target(s) of the comment is/are named, non-named people are not included or are secondary to the comment	ponete a laburar milei lcdtm

Figure 8: Final Guideline for annotating the Target label.

Topic (multi-label) **[Why is/are the target(s) commented for?]**	Policy	The target is adressed because of their economic measures, regarding foreign issues...	El problema del liberalismo es que todos los liberalismos del mundo trabajan para EEUU & UK.
	Origin	The target is addressed because of their ethnicity, colour, descendt, nationality...	Entonces, cuando se van los judios?
	Ideology	The target is addressed by religion or political ideology	Kukitas lloronas llorando, que lindo jajaja me mueeee 🤣🤣🤣
	Sexuality	The target is addressed by aims of gender or sexual orientation	no le enfocan nunca pero notaron que milei camina como si tuviera el culo roto! tiene algun problema en las piernas, la cadera o esta paspado?

Figure 9: Final Guideline for annotating the Topic label.

Attack-Defend **[Is this comment attacking or defending the target(s)?]**	Attacking	The comment is attacking the target, even if it's defending other people that cannot be identified as the main target of the comment	Massa no puede ser tan HIJO DE PUTA
	Neither attacking nor defending	The comment is rather a neutral opinion or statement and it's not attacking nor defending the target	Te puede gustar o no Milei, pero dejando la política aparte, creo que todos podemos empatizar con él en el momento en que no sabe dónde meterse después de entrar y no entender una goma de inglés.
	Defending	The aim of the comment is defend the target	Excelente el javo emperador

Figure 10: Final Guideline for annotating the Attack-Defend label.

Violence **[This comment calls for using verbal or physical violence, or uses pejorative terms against the target(s)]**	Yes	There are clear and explicit forms of violence (insults, suggestions about harmful actions against the target...)	Me parece re mal que Lali le haga buling al minusválido mental, que se meta con alguien de su tamaño.
	No	There's no form of violence expressed in the comment against the target.	A la izquierda, porque aunque puedan decir alguna pavada de vez en cuando, tienen mejores valores, y se necesita alguien que piense en el trabajo de la gente, no en las ganancias de las multinacionales. (Hay muchas razones para no votar a milei, cualquier razón es buena).

Figure 11: Final Guideline for annotating the Violence label.

Hate Speech **[Does this comment contain hate speech?]**	Yes	The intention of harmful attack is clear, the features of the comment are described in the definition above	JAJAJJA CHUPENSE UNA JAPI KUKAS PUTOS
	No	The comment may be an attack or not, but cannot be identified as hate speech according to the definition above	que vergüenza me da que sea mi presidente.

Figure 12: Final Guideline for annotating the Hate Speech label.

C Agreement on the First Annotation

	Attack-Defend			Violence			Hate Speech		
Item	Attack	Neither	Defend	Yes	Unclear	No	Yes	Unclear	No
2		X				X			X
3	X			X			X		
4			X			X			X
5	X			X		X		X	X(j)
6	X			X			X		
7	X				X	X			X
8	X					X			X
9	X					X			X
10	X			X		X		X(l)	X
11	X			X	X			X	X(j)
12			X			X			X
13		X				X			X
14	X			X		X		X(l)	X
15	X			X			X		
16	X			X		X		X(l)	X
17			X			X			X
18	X		X			X	X		
19	X	X			X	X			X
20	X				X	X			X
21	X				X	X			X

Figure 13: Agreement on the First Annotation. In Red: There is agreement that the comment does not contain Hate Speech. In Green: There is agreement that the comment does contain Hate Speech. In Blue: There is disagreement on whether the comment contains Hate Speech.