FACULTAD DE
INGENIERÍA Y CIENCIAS

UAI
UNIVERSIDAD ADOLFO IBÁÑEZ

ADULTS ONLY 18+
AO
ESRB

Hate Speech Recognition in Chilean Tweets

Alfonso Tobar-Arancibia
alftobar@alumnos.uai.cl
Sebastián Moreno
Javier Lopatin

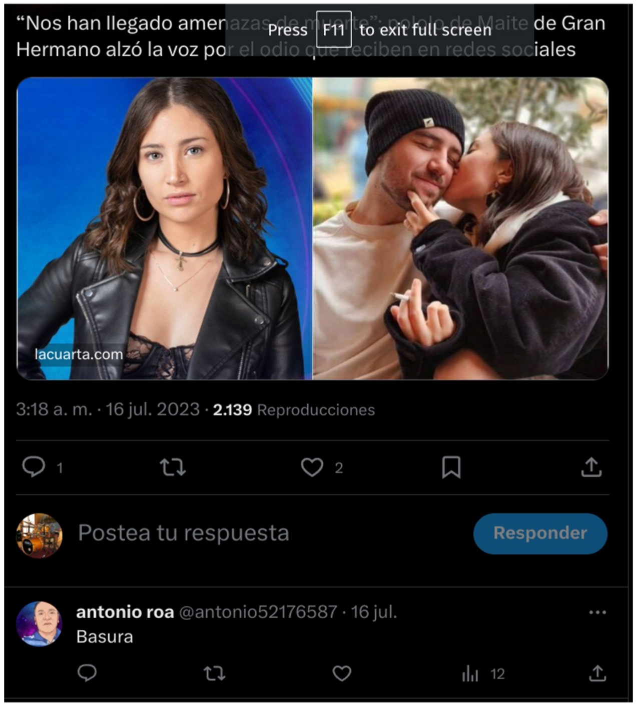
SCCC 2023 42nd IEEE INTERNATIONAL CONFERENCE OF THE CHILEAN COMPUTER SCIENCE SOCIETY

Hate Speech

“...speech or writing that attacks or threatens a particular group of people, based on race, religion or sexual orientation.”

- Hate Speech is a growing concern, especially on social media.
- There are crimes related to hate speech, which have been increased over time.
- It is essential to monitor large-scale hate speech for policy-making.
- Most of the automatic efforts to detect hate speech have been made in English.

Twitter has people dedicated to monitoring tweets, which has been unsuccessful.



“Nos han llegado amenazas de muerte”: pololo de Maite de Gran Hermano alzó la voz por el odio que reciben en redes sociales

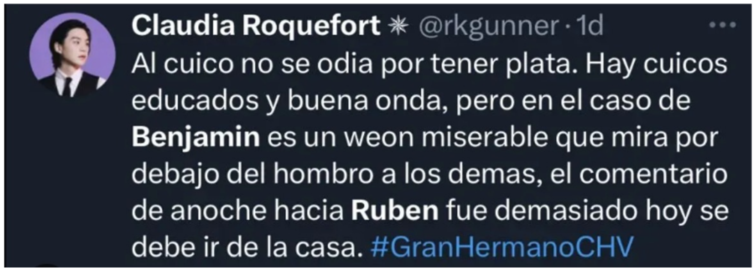
lacuarta.com

3:18 a. m. · 16 jul. 2023 · **2.139** Reproducciones

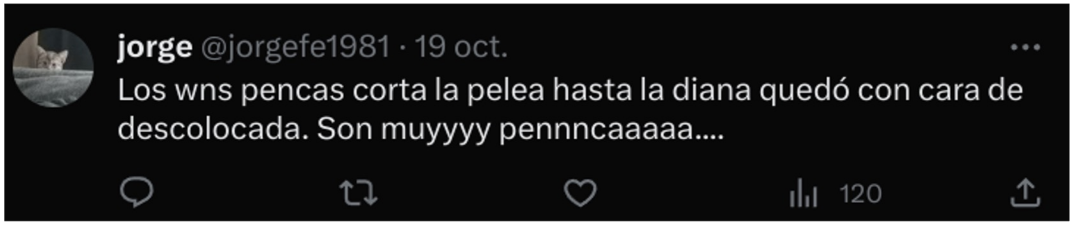
1 2

antonio roa @antonio52176587 · 16 jul.
Basura

12

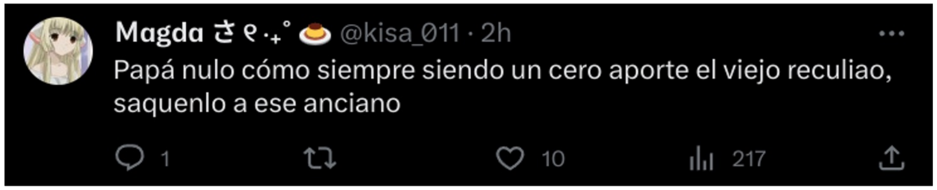


Claudia Roquefort ✷ @rkgunner · 1d
Al cuico no se odia por tener plata. Hay cuicos educados y buena onda, pero en el caso de **Benjamin** es un weon miserable que mira por debajo del hombro a los demas, el comentario de anoche hacia **Ruben** fue demasiado hoy se debe ir de la casa. #GranHermanoCHV



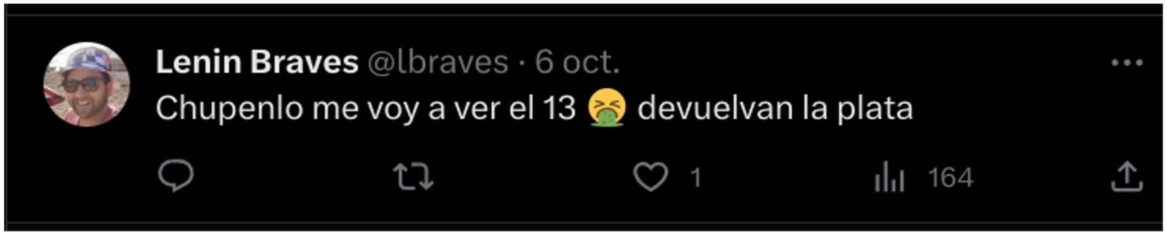
jorge @jorgefe1981 · 19 oct.
Los wns pencas corta la pelea hasta la diana quedó con cara de descolocada. Son muyyyy pennncaaaaa....

120



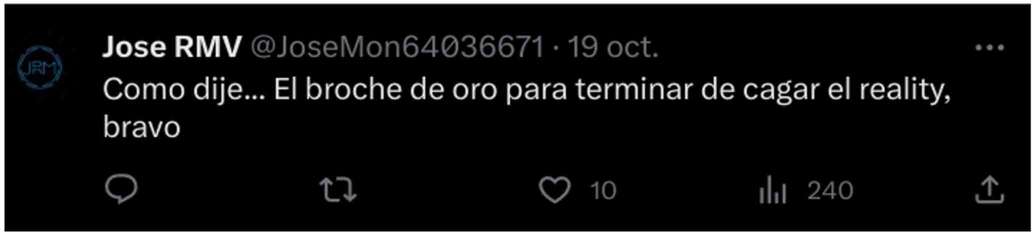
Magda さ ୧ .₊˚ 🍮 @kisa_011 · 2h
Papá nulo cómo siempre siendo un cero aporte el viejo reculiao, saquenlo a ese anciano

1 10 217



Lenin Braves @lbraves · 6 oct.
Chupenlo me voy a ver el 13 🤮 devuelvan la plata

1 164



Jose RMV @JoseMon64036671 · 19 oct.
Como dije... El broche de oro para terminar de cagar el reality, bravo

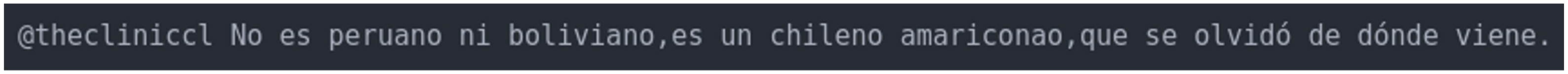
10 240

Problem Definition

Dataset:

- 2256 Training Tweets
- 2291 Test Tweets
- 10 Features (Most of them identifiers)

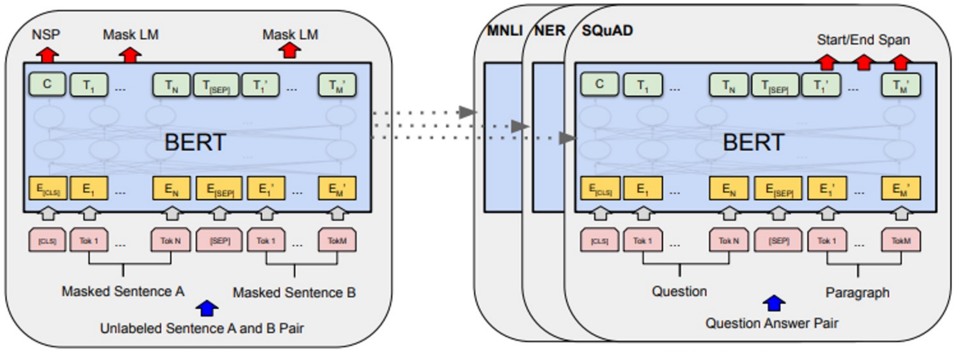
- NLP Problem: Hate Speech Recognition.
- Determine whether a Tweet is Hate or not.
- Detect targeted Community:
    - **Women**
    - **LGBTQ+**
    - **Indigenous**
    - **Immigrants**
- Tweets annotated up to 3 reviewers.
- Not a clear consensus on how hateful a Tweet is.



```
@theclinicc1 No es peruano ni boliviano,es un chileno amariconao,que se olvidó de dónde viene.
```

Hate:2 LGTBQ+:1

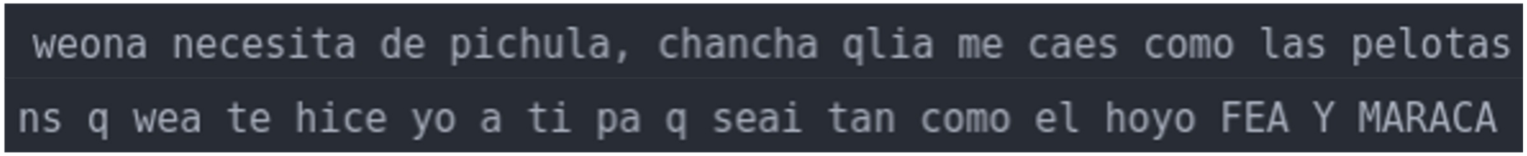
RoBERTuito (Perez et al, 2022)



Pre-training

Fine-Tuning

- Based on Meta’s Roberta Model (Liu et al, 2019).
- Pretrained for several Tasks: Hate Speech Recognition.
- 500 Millions of **Spanish Tweets**.

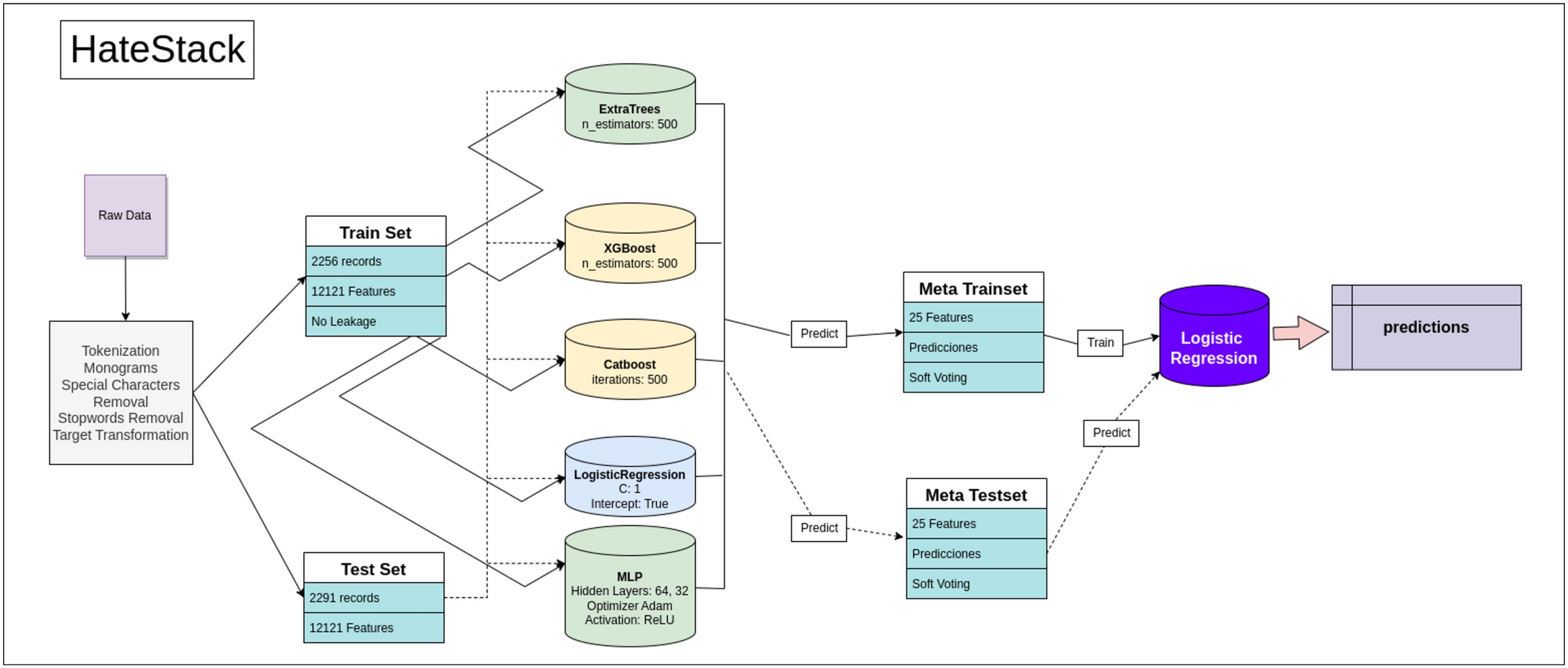


```
weona necesita de pichula, chancha qlia me caes como las pelotas
ns q wea te hice yo a ti pa q seai tan como el hoyo FEA Y MARACA
```

- Very specific problem due to Vocabulary and Chilean Slang.
- Spelling Errors
- How to tokenize?

HateStack

We propose HateStack a two-level ensemble model comprising a feature extraction process, five Level-1 models, and a logistic regression as a second-level model.



HateStack

Raw Data

Tokenization
Monograms
Special Characters Removal
Stopwords Removal
Target Transformation

Train Set
2256 records
12121 Features
No Leakage

Test Set
2291 records
12121 Features

ExtraTrees
n_estimators: 500

XGBoost
n_estimators: 500

Catboost
iterations: 500

LogisticRegression
C: 1
Intercept: True

MLP
Hidden Layers: 64, 32
Optimizer Adam
Activation: ReLU

Predict

Meta Trainset
25 Features
Predicciones
Soft Voting

Predict

Meta Testset
25 Features
Predicciones
Soft Voting

Train

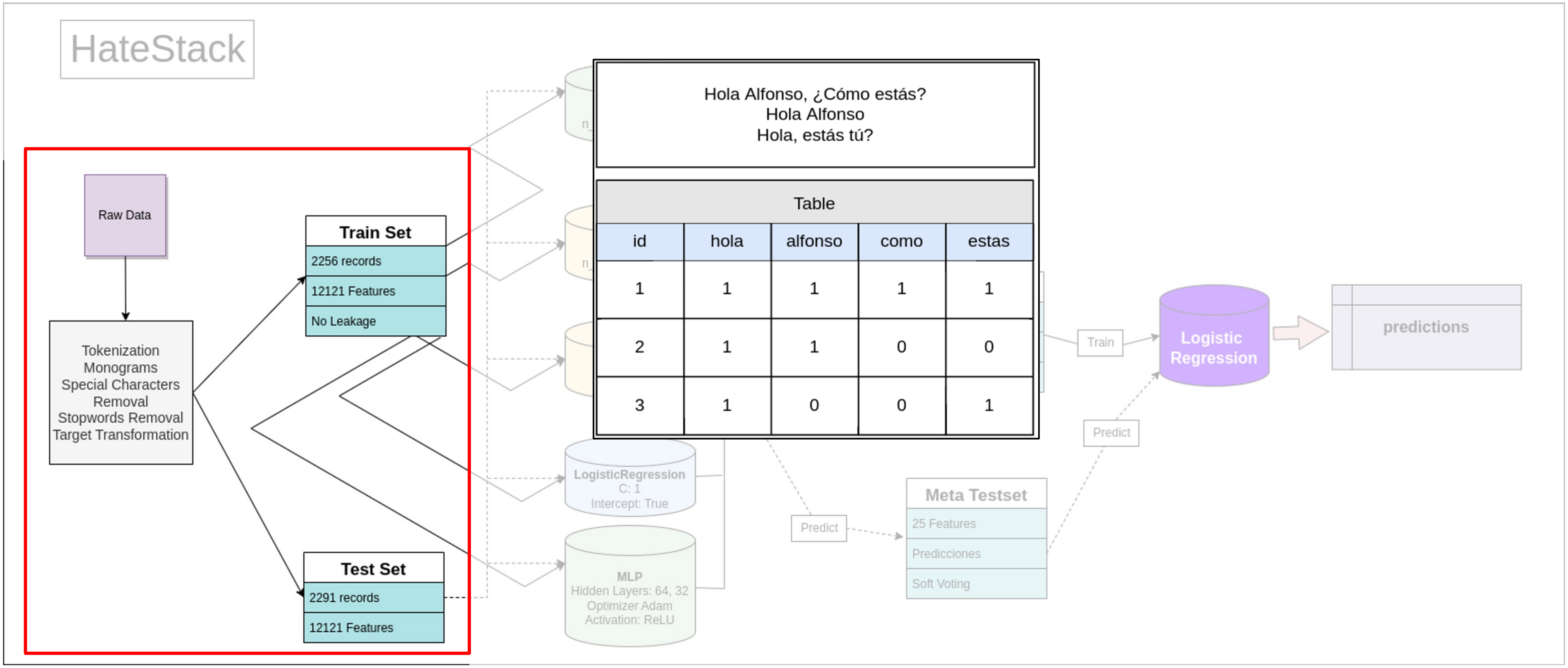
Predict

Logistic Regression

predictions

Data Preprocessing

The data was preprocessed using common Text Cleaning Techniques such as: Tokenization on Monograms, Special Characters and Stopwords removal. Target was converted to Binary Labels.



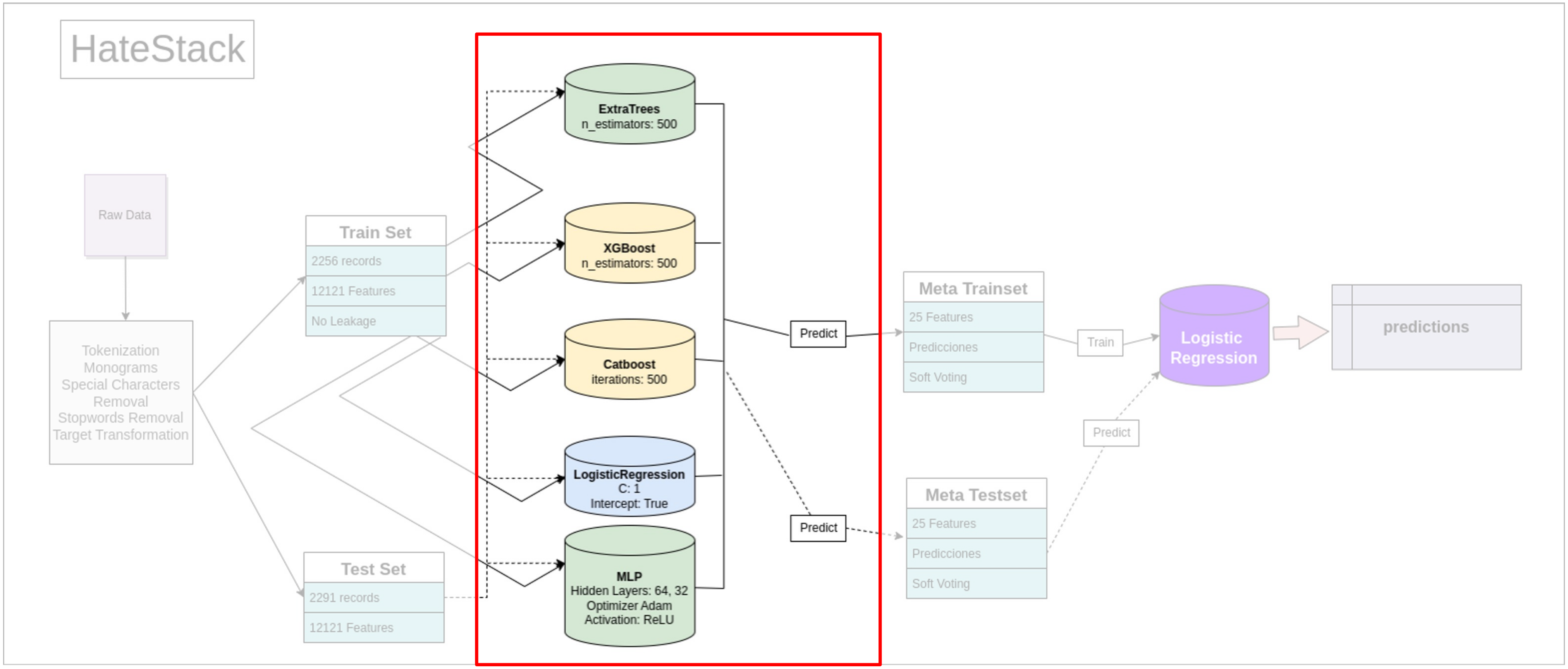
Hola Alfonso, ¿Cómo estás?
Hola Alfonso
Hola, estás tú?

Table

id	hola	alfonso	como	estas
1	1	1	1	1
2	1	1	0	0
3	1	0	0	1

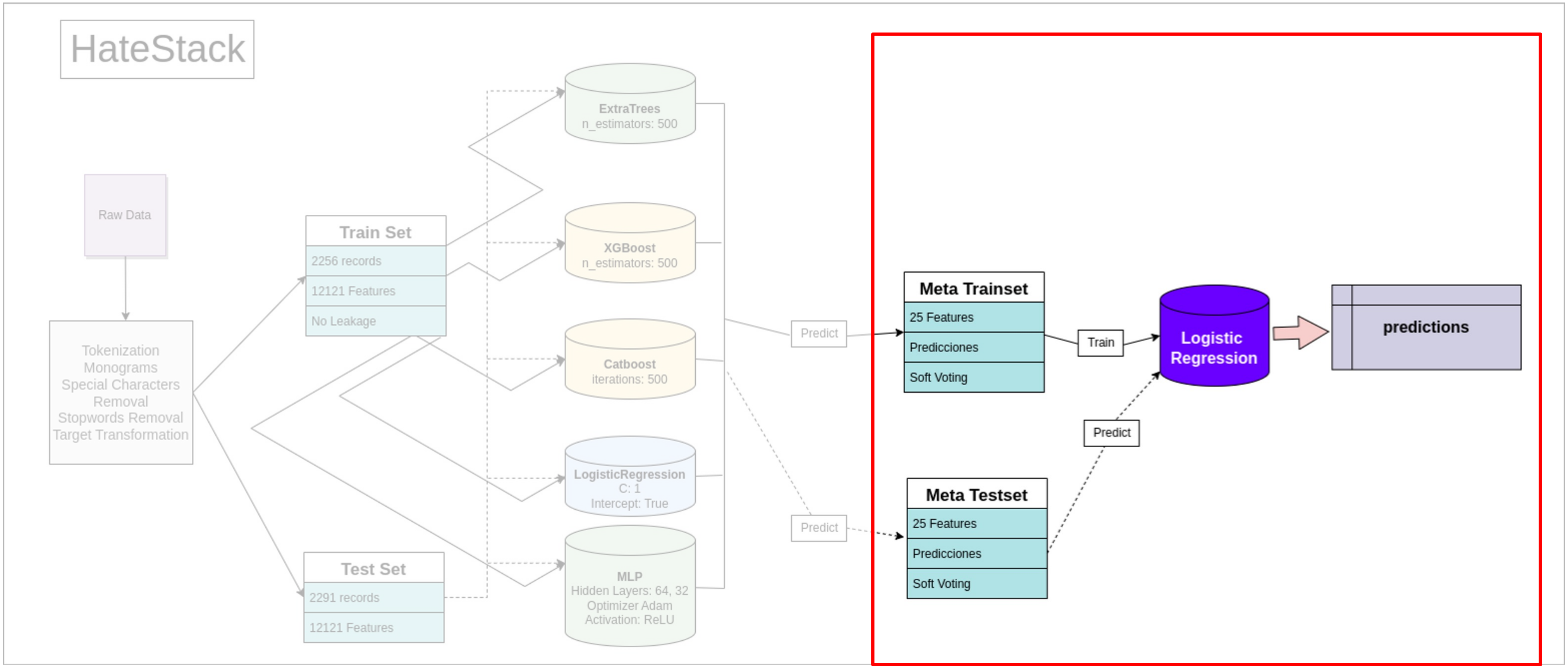
Single Models

Every Level-1 model was trained on the whole data using optimal Hyperparameters. Level-1 models provide their own predictions for every record in the dataset.



Stacking (Volpert, 1992)

Level-2 model uses Level-1 predictions as features. Then, using a Cross Validation Prediction it ensembles final predictions.

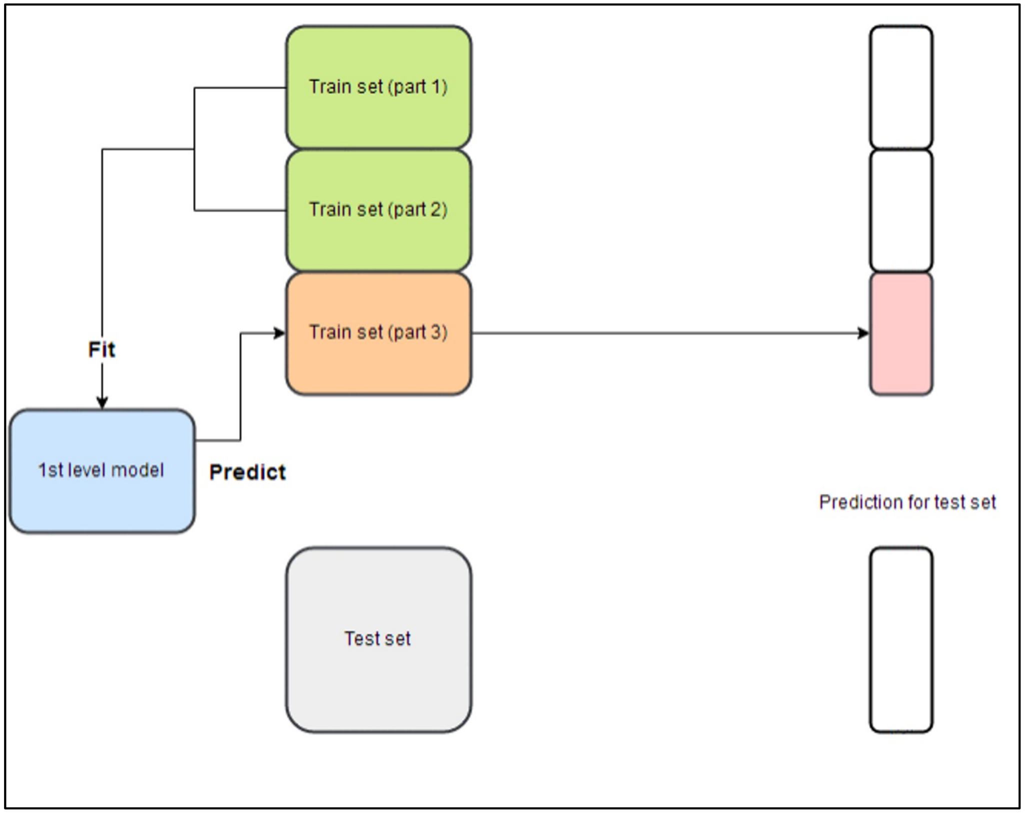


Stacking Inner Working

- Model trained as a Multilabel Classification Task.
- Best Hyperparameters found with GridSearch.
- Model validated using 5-Fold CV.
- Stacking used 3-Fold inner CV.

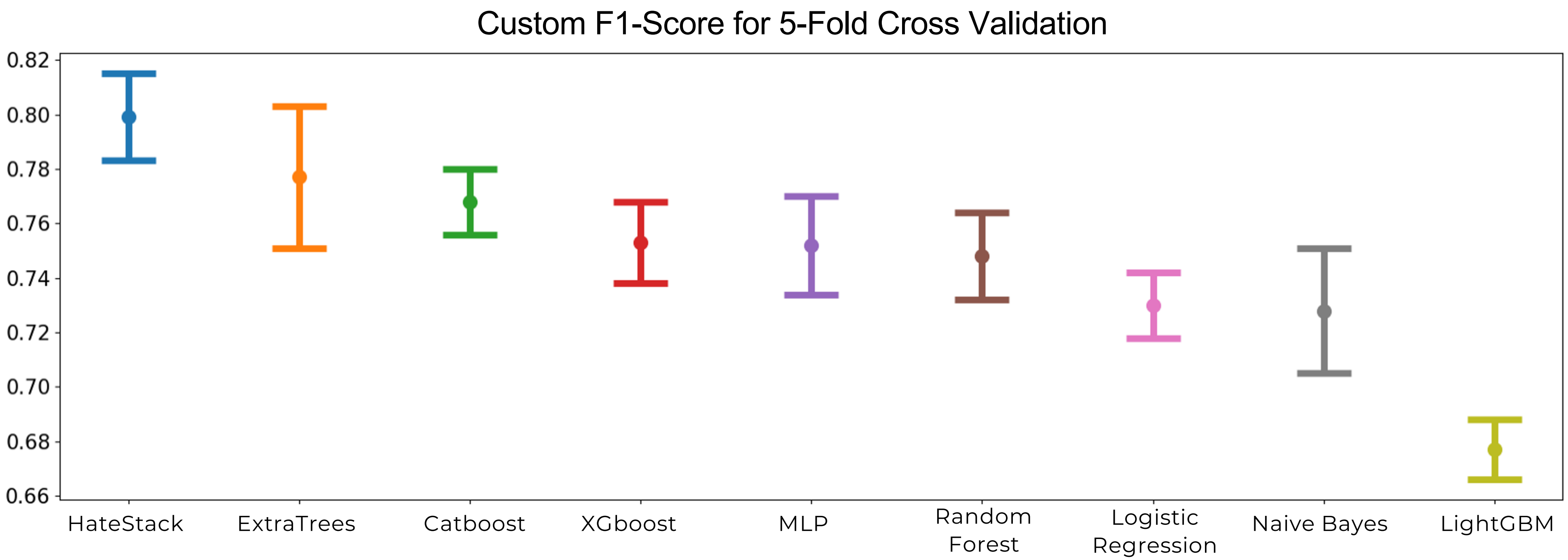
- Model Evaluated with Custom F1-Score:

$$F_{1Custom} = 0.50 \cdot (F_{1Hate} + \text{Macro } F_{1communities})$$



Our Results

HateStack is significantly better than other 8 baselines in Validation and Test Time.



HateStack achieved a **0.799** on Validation and **0.818** on Test set.

Surprisingly, RoBERTuito had the worst performance among all the baselines with **0.293**

Conclusion

- We propose a two-level ensemble model comprising a feature extraction process, five Level-1 models, and a logistic regression as a Level-2 model.
- HateStack was able to outperform classical ensembles (XGBoost, Catboost, and LightGBM), and state-of-the-art Deep Learning models, such as RoBERTuito.
- RoBERTuito, a transformer-based architecture pre-trained in Spanish Tweets for Hate Speech, didn’t work as expected.
- We believe, that this poor performance could be explained by the Chilean slang, which is very uncommon to be part of the vocabulary of pre-trained models.