Counterhate Arguments - Countering the hate speech on social media

Team Members

Sweta Pati

Swabhi Papneja

1. Introducing the Idea

Objective: In this paper, we identify **authentic** counterhate arguments that address the claims in a hateful tweet towards a **specific** individual.

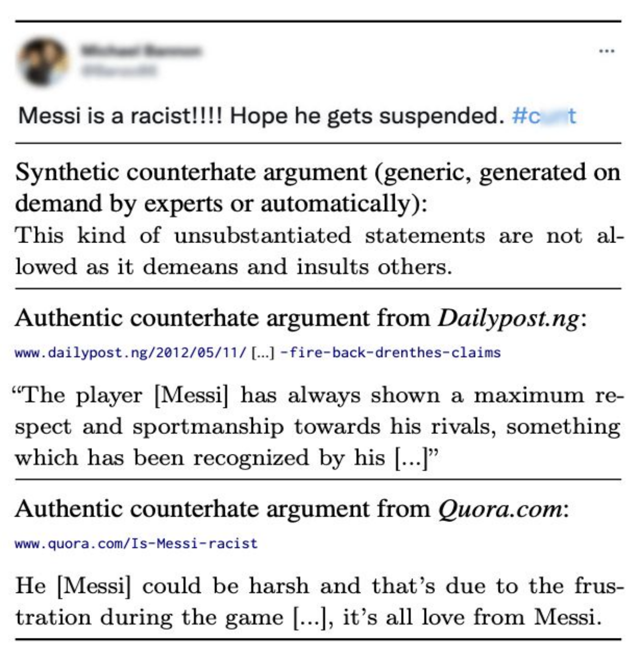
"A hateful speech according to Twitter guidelines is ***any implicit or explicit*** *tweet* ***that attacks an individual's gender, religion, race, ideology, or social class.****"*

"We define ***counterhate*** *as a direct response that counters hate speech and an* ***authentic counterhate argument*** *as a paragraph that appeals to logic by including factual, testimonial, or statistical evidence."*

Let’s look at an example:

Synthetic Vs Authentic Counterhate

- Synthetic counterhate arguments do not address the specific hateful claims towards the targeted individual and as a result generate generic counterhate arguments.
- This study aims at targeting authentic counterhate arguments that appeal to logic by including factual, testimonial, or statistical evidence.
- Such authentic counterhate arguments with sources have the potential to be more effective than generic statements condemning hate.



[redacted]

Messi is a racist!!!! Hope he gets suspended. #c[redacted]t

Synthetic counterhate argument (generic, generated on demand by experts or automatically):
This kind of unsubstantiated statements are not allowed as it demeans and insults others.

Authentic counterhate argument from *Dailypost.ng*:
www.dailypost.ng/2012/05/11/ [...] -fire-back-drenthes-claims

"The player [Messi] has always shown a maximum respect and sportmanship towards his rivals, something which has been recognized by his [...]"

Authentic counterhate argument from *Quora.com*:
www.quora.com/Is-Messi-racist

He [Messi] could be harsh and that's due to the frustration during the game [...], it's all love from Messi.

2. Dataset Overview

Our dataset comprises of

- 250 hateful tweets towards 50 individuals
- 2500 articles
- 54,816 paragraphs

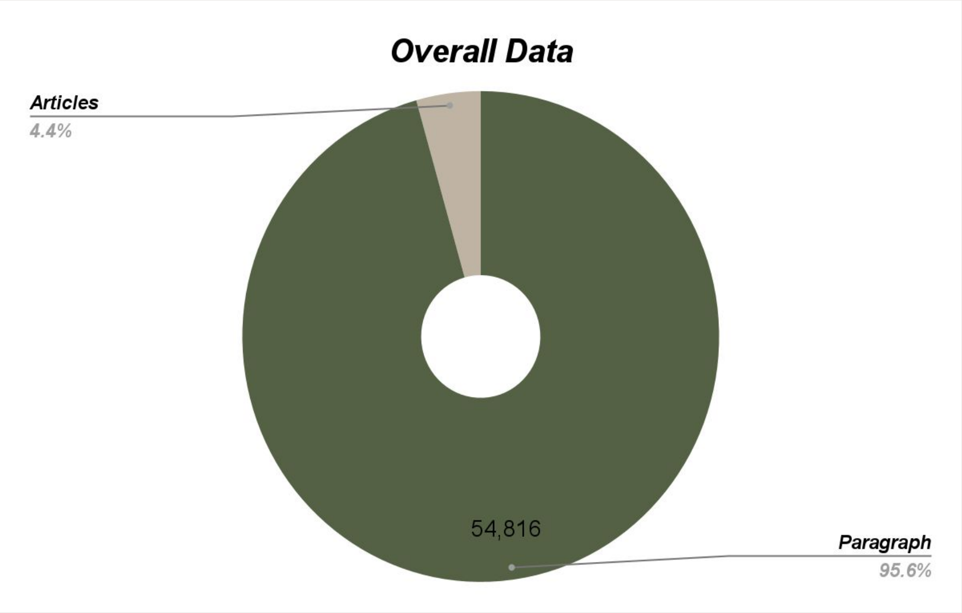
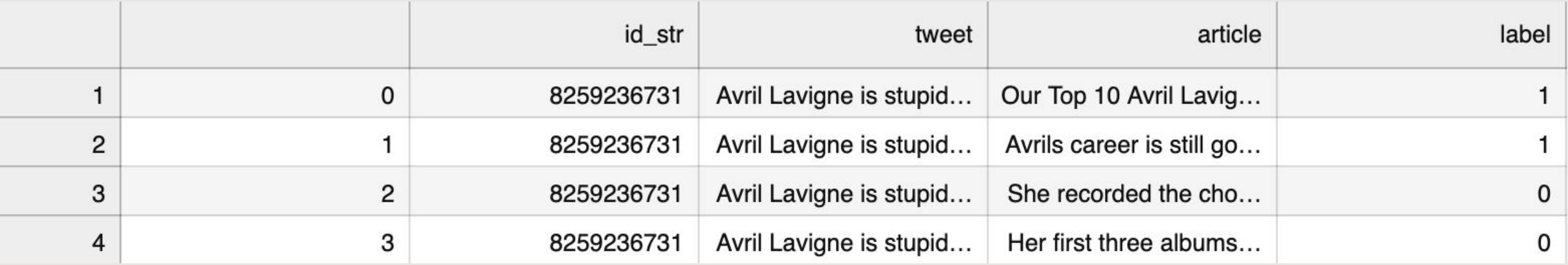


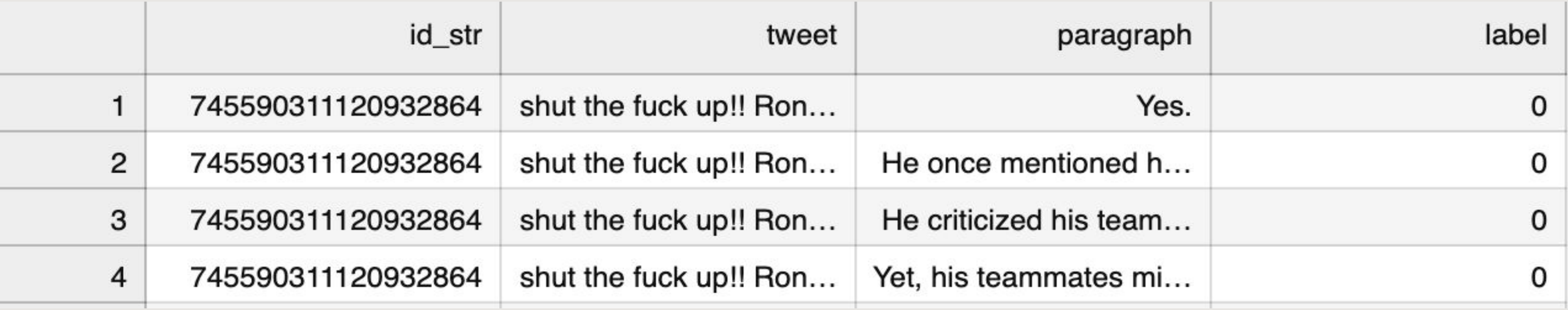
Fig: Distribution of Articles and Paragraphs in the dataset.

articles.csv



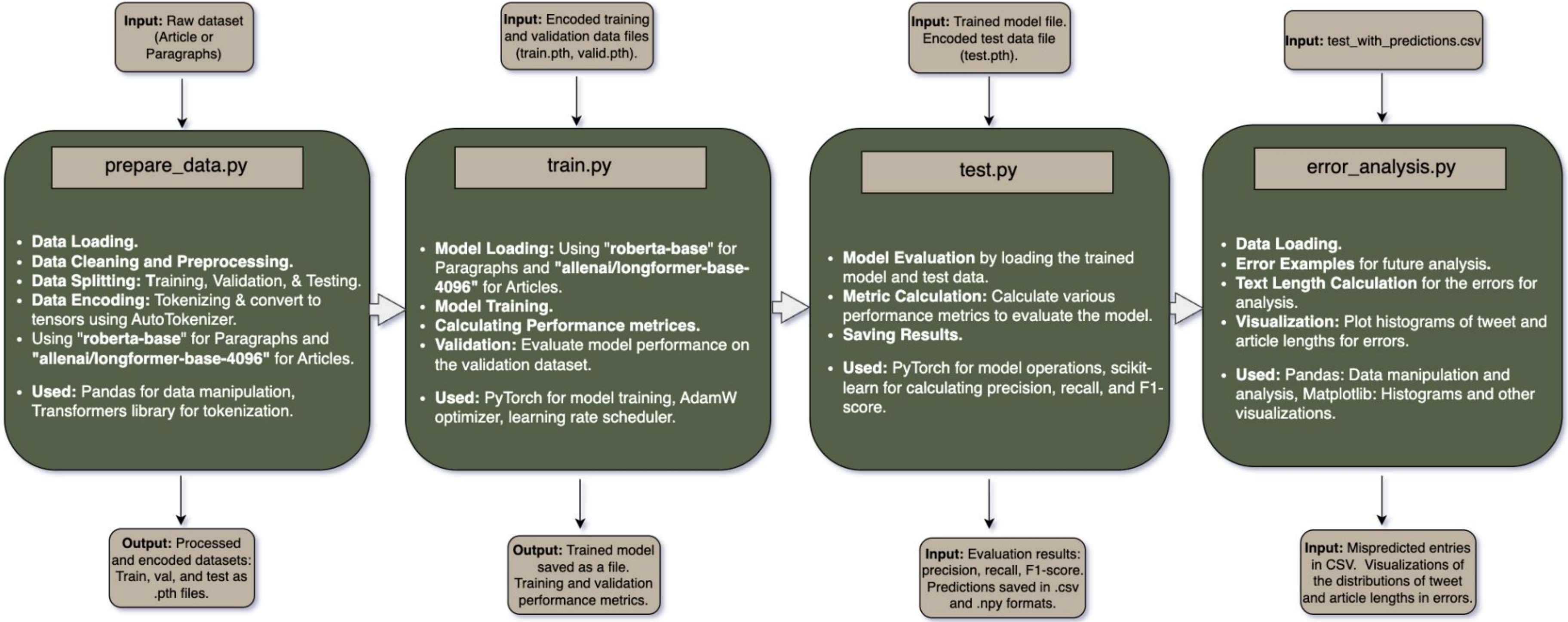
		id_str	tweet	article	label
1	0	8259236731	Avril Lavigne is stupid…	Our Top 10 Avril Lavig…	1
2	1	8259236731	Avril Lavigne is stupid…	Avrils career is still go…	1
3	2	8259236731	Avril Lavigne is stupid…	She recorded the cho…	0
4	3	8259236731	Avril Lavigne is stupid…	Her first three albums…	0

paragraphs.csv



	id_str	tweet	paragraph	label
1	745590311120932864	shut the fuck up!! Ron…	Yes.	0
2	745590311120932864	shut the fuck up!! Ron…	He once mentioned h…	0
3	745590311120932864	shut the fuck up!! Ron…	He criticized his team…	0
4	745590311120932864	shut the fuck up!! Ron…	Yet, his teammates mi…	0

3. Approach



Input: Raw dataset (Article or Paragraphs)

prepare_data.py

- **Data Loading.**
- **Data Cleaning and Preprocessing.**
- **Data Splitting:** Training, Validation, & Testing.
- **Data Encoding:** Tokenizing & convert to tensors using AutoTokenizer.
- Using "**roberta-base**" for Paragraphs and **"allenai/longformer-base-4096"** for Articles.

- **Used:** Pandas for data manipulation, Transformers library for tokenization.

Output: Processed and encoded datasets: Train, val, and test as .pth files.

Input: Encoded training and validation data files (train.pth, valid.pth).

train.py

- **Model Loading:** Using "**roberta-base**" for Paragraphs and **"allenai/longformer-base-4096"** for Articles.
- **Model Training.**
- **Calculating Performance metrices.**
- **Validation:** Evaluate model performance on the validation dataset.

- **Used:** PyTorch for model training, AdamW optimizer, learning rate scheduler.

Output: Trained model saved as a file. Training and validation performance metrics.

Input: Trained model file. Encoded test data file (test.pth).

test.py

- **Model Evaluation** by loading the trained model and test data.
- **Metric Calculation:** Calculate various performance metrics to evaluate the model.
- **Saving Results.**

- **Used:** PyTorch for model operations, scikit-learn for calculating precision, recall, and F1-score.

Input: Evaluation results: precision, recall, F1-score. Predictions saved in .csv and .npy formats.

Input: test_with_predictions.csv

error_analysis.py

- **Data Loading.**
- **Error Examples** for future analysis.
- **Text Length Calculation** for the errors for analysis.
- **Visualization:** Plot histograms of tweet and article lengths for errors.

- **Used:** Pandas: Data manipulation and analysis, Matplotlib: Histograms and other visualizations.

Input: Mispredicted entries in CSV. Visualizations of the distributions of tweet and article lengths in errors.

4. Experiments and Results

CHECKPOINT 1: EXPERIMENTS

- Dataset consists of labeled examples split into Training, Validation and Test sets.
- We compare our model's performance to the projected results in the paper to validate reproducibility.
- Metrics used include Precision, Recall, and F1-score.

Experiments on Article Level

”tweet”

”tweet”,”article”

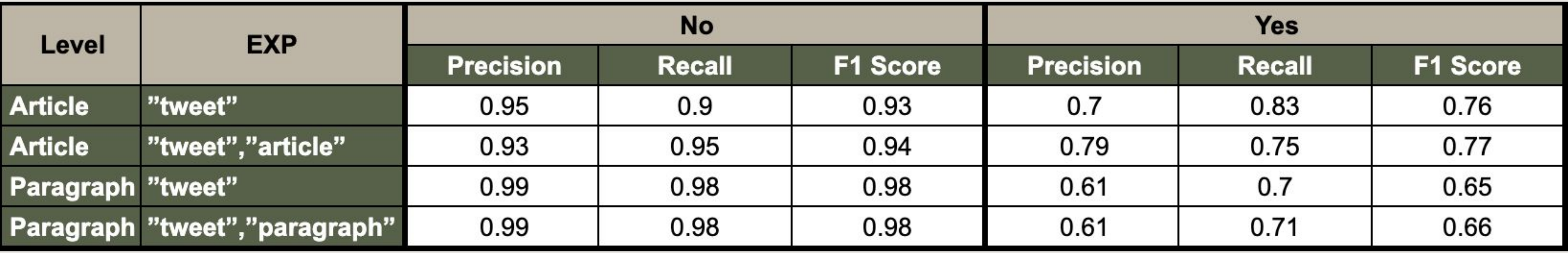
Experiments on Paragraph Level

”tweet”

”tweet”,”paragraph”

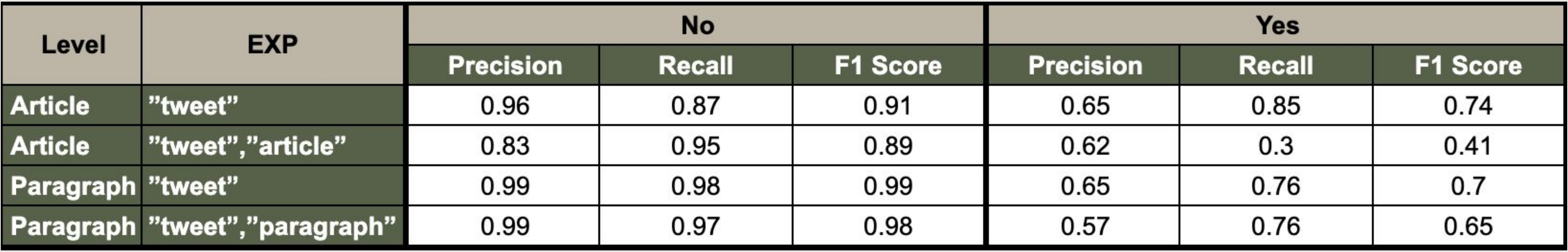
CHECKPOINT 1: RESULTS

Table 1: Our results



Level	EXP	No			Yes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Article	”tweet”	0.95	0.9	0.93	0.7	0.83	0.76
Article	”tweet”,”article”	0.93	0.95	0.94	0.79	0.75	0.77
Paragraph	”tweet”	0.99	0.98	0.98	0.61	0.7	0.65
Paragraph	”tweet”,”paragraph”	0.99	0.98	0.98	0.61	0.71	0.66

Table 2: Published results



Level	EXP	No			Yes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Article	”tweet”	0.96	0.87	0.91	0.65	0.85	0.74
Article	”tweet”,”article”	0.83	0.95	0.89	0.62	0.3	0.41
Paragraph	”tweet”	0.99	0.98	0.99	0.65	0.76	0.7
Paragraph	”tweet”,”paragraph”	0.99	0.97	0.98	0.57	0.76	0.65

CHECKPOINT 1: Major Takeaways

We were able to reproduce significantly close results to that presented in the original paper.

Identification of authentic counterhate arguments specific to target individuals countering the **hate claim is very essential for promoting healthier online interactions.**

5. Current Work: Checkpoint 2

CHECKPOINT 2: Work Done!

- We are focusing on **Multilinguality** dimension to perform a deeper analysis and improvement upon our existing work.

Data Translations

Found top 10 Languages used in twitter.

Languages: Japanese, Spanish, Portuguese, Arabic, French, Indonesian, Russian, Turkish, Hindi

Translated article level data to these 10 languages.

Code Modification

Performed language specific preprocessing

Changed tokenizer & model to “xlm-roberta-base” from “roberta-base” & “longformer-base-4096”

For article level, did appropriate parameter tuning.

We chose XLM-RoBERTa for our multilinguality model because it has been trained on 100 languages making it capable for handling languages with different scripts.

Our dataset comprises of

- 250 hateful tweets towards 50 individuals
- 25000 articles and 164,448 paragraphs

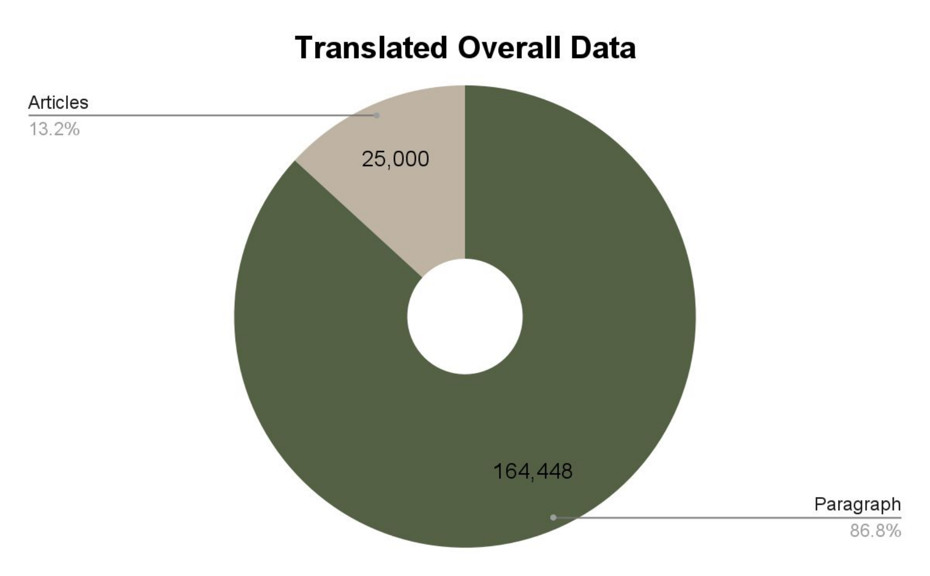


Fig: Overall Distribution of Translated Articles and Paragraphs.

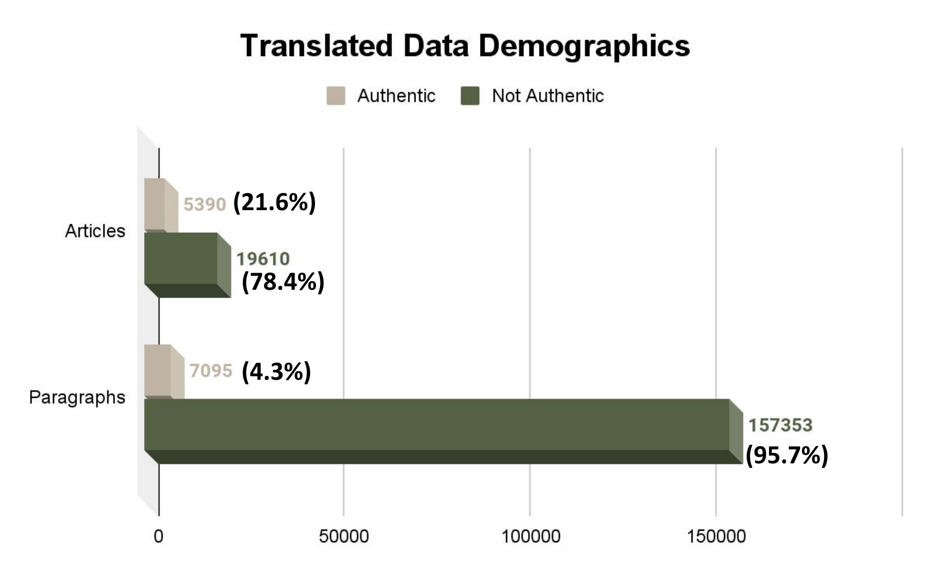
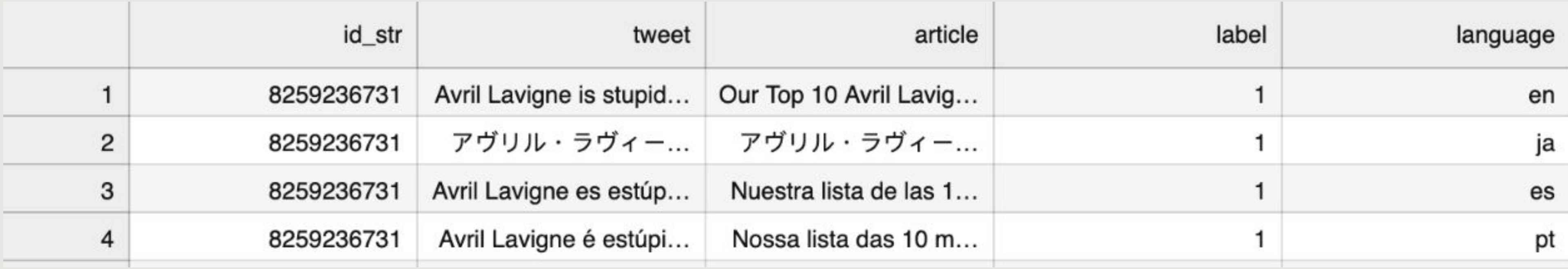


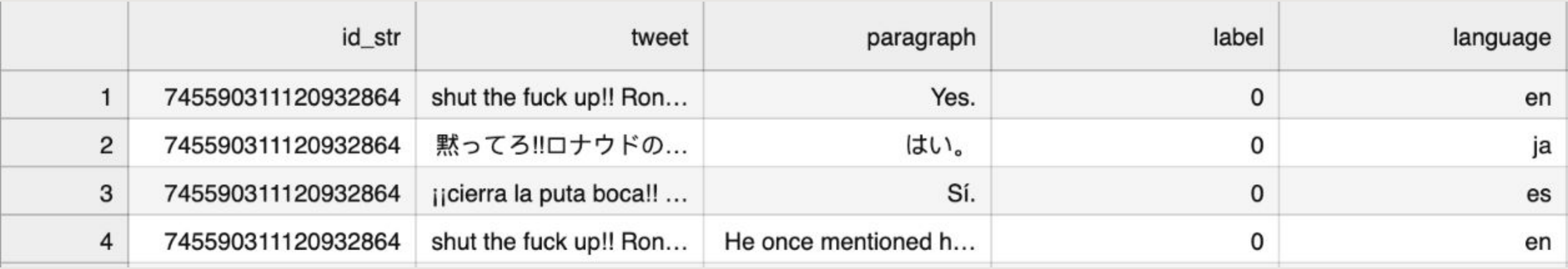
Fig: Distribution of Authentic and Non Authentic Counterhate responses within Translated Articles and Paragraphs.

articles.csv



	id_str	tweet	article	label	language
1	8259236731	Avril Lavigne is stupid…	Our Top 10 Avril Lavig…	1	en
2	8259236731	アヴリル・ラヴィー…	アヴリル・ラヴィー…	1	ja
3	8259236731	Avril Lavigne es estúp…	Nuestra lista de las 1…	1	es
4	8259236731	Avril Lavigne é estúpi…	Nossa lista das 10 m…	1	pt

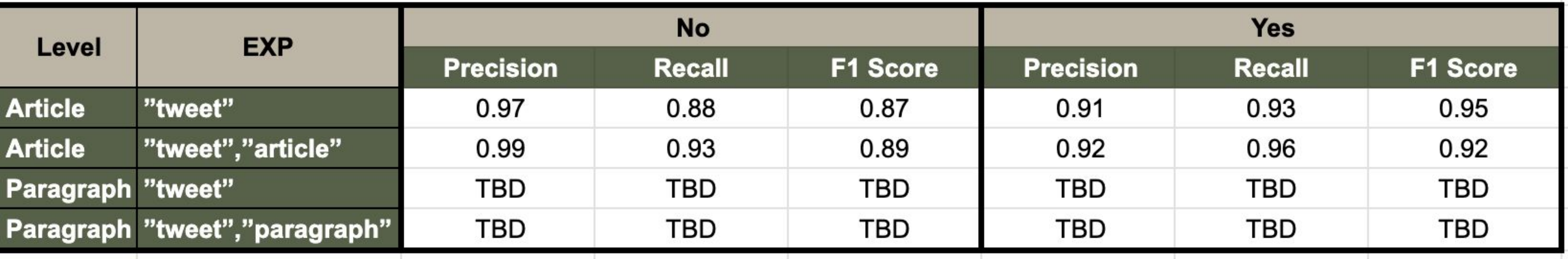
paragraphs.csv



	id_str	tweet	paragraph	label	language
1	745590311120932864	shut the fuck up!! Ron…	Yes.	0	en
2	745590311120932864	黙ってろ!!ロナウドの…	はい。	0	ja
3	745590311120932864	¡¡cierra la puta boca!! …	Sí.	0	es
4	745590311120932864	shut the fuck up!! Ron…	He once mentioned h…	0	en

CHECKPOINT 2

Table 3: Our results



Level	EXP	No			Yes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Article	”tweet”	0.97	0.88	0.87	0.91	0.93	0.95
Article	”tweet”,”article”	0.99	0.93	0.89	0.92	0.96	0.92
Paragraph	”tweet”	TBD	TBD	TBD	TBD	TBD	TBD
Paragraph	”tweet”,”paragraph”	TBD	TBD	TBD	TBD	TBD	TBD

6. Future Work: Checkpoint 2 & Beyond

Future Work Plan

- We will look into collecting real world hateful tweet and counterhate data in other languages apart from translation of our current dataset.

Social media has become an integral part of our daily lives, with an increasing number of users worldwide.

The impact of social media on society and culture is undeniable, **hence identifying counterhate arguments is essential for reducing online hate speech, fostering more respectful interactions, and creating a safer, more inclusive environment on internet platforms**.

Thank You