

LLM - Detect AI Generated Text

Professors:

Prof. Gianlugi Greco

Dott. Carlo Adornetto

Students:

Cristian Ciriaco Campagna 242604

Giovanni Iannuzzi

Pierpaolo Sestito 242707

Contents

1	Introduction	2
1.1	Project Overview	2
1.1.1	Objectives	3
2	Detect AI Generated Text	4
2.1	Dataset Description	4
2.2	Methods	6
2.2.1	Cosine Similarity Filtering	6
2.2.2	Record Generation for decreasing the unbalance	6
2.3	Result : DAIGTv3	9
2.3.1	Preprocessing	10
2.3.2	Architectures	11
3	Modelling	13
3.1	Model description	13
3.2	Model representation	16
4	Results	17
4.1	Model Evaluation	17
4.1.1	Final Considerations	18
4.2	Kaggle Results	19
4.3	Proof of Challenge Participation	19

Chapter 1

Introduction

In the era of rapidly advancing natural language processing (NLP) technologies, the ability to distinguish between human-generated and language model-generated text has become a pivotal challenge. This competition, titled "LLM Detect AI Generated Text," provides a unique opportunity to explore and address this task.

1.1 Project Overview

The competition dataset encompasses two primary components: `train_essays.csv` and `train_prompts.csv`. The former contains essays with identifiers (`id`), associated prompt identifiers (`promptid`), and the actual essay text (`text`). A crucial attribute, `generated`, indicates whether the essay was composed by a human student (marked as 0) or generated by a Language Model (LM, marked as 1). Notably, the objective is to develop a robust classification model capable of discerning the origin of each essay.

The latter dataset, `train_prompts.csv`, provides essential contextual information for the essays. Each prompt is uniquely identified by `promptid` and features metadata such as `promptname`, `instructions`, and the `sourcetext` (`source-text`).

1.1.1 Objectives

The overarching goal of this project is to construct a machine learning model that excels in classifying text as either human-authored or generated by a Language Model. By leveraging the information embedded in the essays and their corresponding prompts, participants are challenged to develop models that generalize well to unseen data and exhibit a nuanced understanding of the distinctive features between human and AI-generated language.

This report delineates the methodology, challenges encountered, and the results achieved during the exploration of this captivating competition. Through a comprehensive analysis, we aim to contribute insights into the evolving landscape of text classification and the intricate interplay between human and artificial intelligence in language generation.

Chapter 2

Detect AI Generated Text

In this phase we describe the dataset and its attributes and we analyze each individual attribute checking its validity and semantics.

2.1 Dataset Description

During the initial exploration of the competition dataset *train_essays.csv*, we observed a significant class imbalance. With only 3 records labeled as **1** (indicating **AI-generated text**) and 1375 labeled as **0** (indicating human-generated text), the dataset exhibited skewed representation.

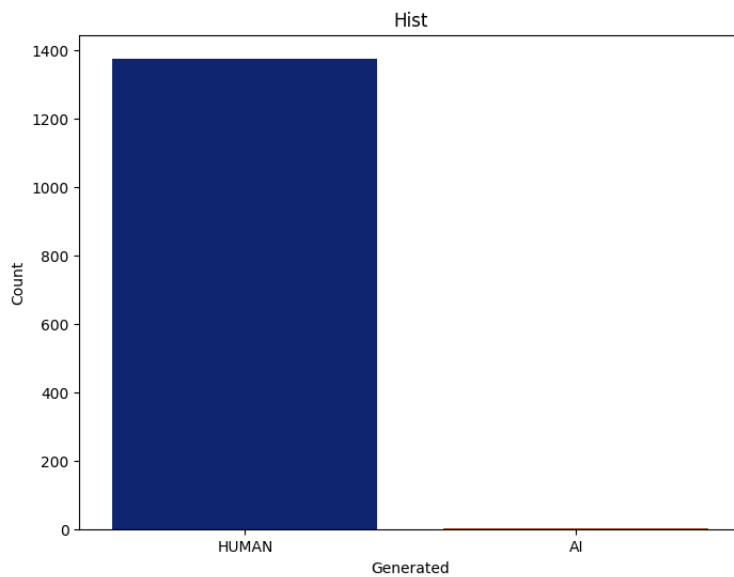


Figure 2.1: train_essays balance

Recognizing the importance of a balanced dataset for model training and evaluation, we embarked on a search for a more comprehensive and balanced alternative.

As a result of our efforts, we identified and start to analyze the *DAIGTv2* dataset. This new dataset addresses the imbalance concerns present in the initial dataset, **providing a more equitable distribution** between AI-generated and human-generated texts, but **not at all**.

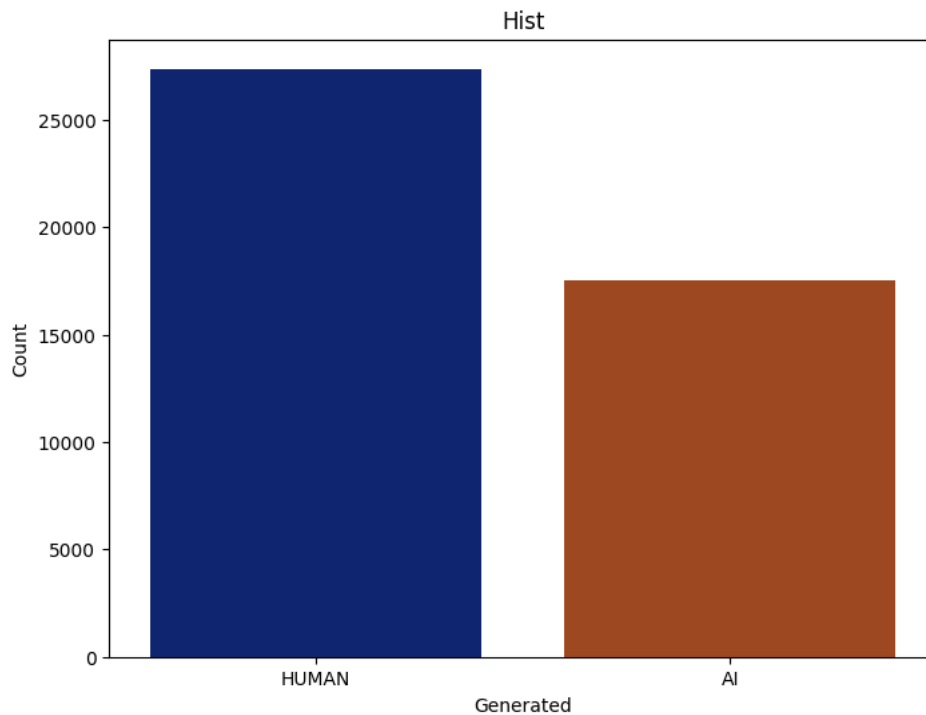


Figure 2.2: DAIGTv2

In the exploration of the DAIGTv2 dataset, our **focus shifted towards the textual contributions of human authors**. To foster a more balanced representation within the dataset, we employed the "*cosine similarity*" metric as a filtering mechanism. This approach aimed to mitigate redundancy and enhance diversity by removing texts with a cosine similarity greater than or equal to 0.9.

2.2 Methods

2.2.1 Cosine Similarity Filtering

The **cosine similarity** was calculated for **pairs of textual entries attributed to human authors**. Entries exhibiting a cosine similarity of **0.9 or higher** were identified as **closely related** and were subsequently **pruned from the dataset**. This method allowed us to retain a diverse set of human-generated texts while **reducing redundancy within the dataset**.

Impact on Dataset balance

The application of the cosine similarity filtering mechanism aligns with our broader objective of achieving a balanced distribution of labels within the dataset. By **selectively removing highly similar texts**, we aimed to **diminish potential biases** introduced by over-represented or redundant human-authored samples.

2.2.2 Record Generation for decreasing the unbalance

In response to the **persistent label imbalance** observed even after cosine similarity filtering, a strategic decision was made to **introduce new records** through the utilization of **language models**.

Prompt Selection Criteria

The generation process commenced by narrowing the **focus to prompts** with a **substantial difference** in the count of **records labeled as 0 and 1**, exceeding a predefined **threshold of 1000**. This criterion aimed to target prompts that inherently displayed an imbalance, thus necessitating the introduction of additional instances.

Model for Text Generation Selection and Analysis

Subsequently, an exhaustive analysis of the **'source'** attribute within the

	prompt_name	human	ai	difference
0	"A Cowboy Who Rode the Waves"	1372	524	848
2	Cell phones at school	1656	463	1193
3	Community service	1542	550	992
6	Driverless cars	1886	364	1522
7	Exploring Venus	1862	314	1548
8	Facial action coding system	2167	917	1250
9	Grades for extracurricular activities	1626	490	1136
11	Phones and driving	1151	415	736
13	Summer projects	1750	951	799
14	The Face on Mars	1583	310	1273

Figure 2.3: sources used in DAIGTv2

DAIGTv2 dataset was undertaken. The 'source' attribute **denotes the models responsible for generating each record.**

source	
persuade_corpus	25996
mistral7binstruct_v1	2421
mistral7binstruct_v2	2421
chat_gpt_moth	2421
llama2_chat	2421
kingki19_palm	1384
train_essays	1378
llama_70b_v1	1172
falcon_180b_v1	1055
darragh_claude_v6	1000
darragh_claude_v7	1000
radek_500	500
NousResearch/Llama-2-7b-chat-hf	400
mistralai/Mistral-7B-Instruct-v0.1	400
cohere-command	350
palm-text-bison1	349
radekgpt4	200

Figure 2.4: Sources used in DAIGTv2

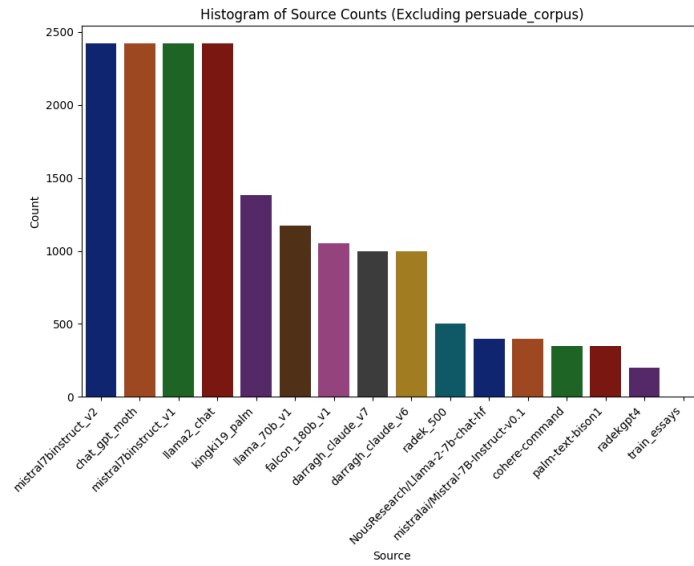


Figure 2.5: Prompt difference

In light of resource constraints and a desire for computational efficiency, a meticulous examination led to the selection of **Cohere** and **Mistral-7b-v0.1** as models for record generation. Both models, chosen for their computational efficiency and ease of use, were deemed suitable alternatives given the limitations in computational resources. Also because they're used less than others.

2.3 Result : DAIGTv3

Our notebook derives from this operations - [GitHub](#)

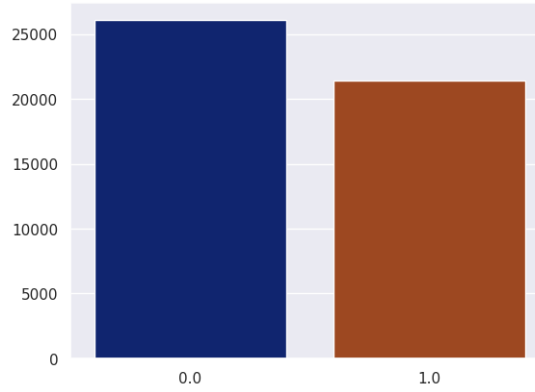


Figure 2.6: Label information

```
Int64Index: 43602 entries, 0 to 43601
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   text             43602 non-null  object
1   label            43602 non-null  float64
2   prompt_name      43602 non-null  object
3   source           43602 non-null  object
4   RDizzl3_seven    43602 non-null  object
5   id               43602 non-null  float64
dtypes: float64(2), object(4)
memory usage: 2.3+ MB

Size : 261612
#Rows : 43602
#Cols : 6
```

Figure 2.7: Other information

During the pre-processing phase, $[prompt_name, source, RDizzl3_seven, id]$ were removed as they were found to be irrelevant or not conducive to the specific objectives of our classification task. The decision to eliminate these columns was made to simplify the DataFrame and focus exclusively on information essential to the ongoing classification task. This data cleaning operation helped optimize the DataFrame structure, making it more suitable for analysis and model training.

2.3.1 Preprocessing

In the initial phase of the **data preparation process**, our focus was on **optimizing the data**.

Stop Words Removal and Lowercase Conversion

The first operation performed was the **removal of "stop words"**, commonly occurring words that contribute little semantic value to the text (*eg. the, and, is, of, in, it, for*). To implement this operation, three different functions were considered: the first based on the **Natural Language Toolkit (NLTK)**, the second utilizing the **SpaCy** library, and the third leveraging **Gensim**. Through a series of experiments, it was found that the **SpaCy-based approach yielded the most satisfactory results** in terms of preserving the text's meaning while simultaneously reducing noise caused by low semantic value words. Finally, **to standardize the text and reduce the dataset's dimensionality**, all sentences were **converted to lowercase**. This operation **prevents** the duplication of features due to **different representations of words with uppercase and lowercase letters**, simplifying the model's learning phase.

Other attempts

In addition to the previously mentioned methods, other preprocessing techniques were employed. However, during the training phase, they yielded inferior results, leading to their exclusion. Below is a list:

- **Strip punctuation**
- **Lemmatization** (using NLTK)

2.3.2 Architectures

In our pursuit of achieving the goal of distinguishing between human and AI-generated text, we embarked on a series of attempts, each aimed at finding the most suitable architecture for our specific dataset characteristics.

- **BERT-model based**

Our initial approach involved leveraging a **pre-trained BERT** (Bidirectional Encoder Representations from Transformers) model. BERT is renowned for its **contextual understanding of language**, making it a promising candidate. However, our **dataset's** substantial **size surpassed the 512-token limit** imposed by BERT. Recognizing the inefficiency of discarding such a significant amount of data, we concluded that utilizing a pre-trained BERT model would not be a viable solution for our specific case.

- **LongFormerModel variant**

To overcome the token limit constraint, we explored a **variant of BERT** known as the **LongFormerModel**, which **provides the capability** to handle sequences **of up to 4096 tokens**. Regrettably, due to **computational resource saturation**, we had to halt our efforts to use this pre-trained model. The demanding resource requirements rendered it impractical for our current infrastructure.

- **RNN final version**

Faced with the challenges posed by pre-trained models, we turned our attention to building a custom solution. Our third and final attempt involved the implementation of a **Recurrent Neural Network (RNN)**. RNNs are **well-suited for sequential data** and exhibit the **ability to capture dependencies** over time. This approach offers more **flexibility in handling** datasets with **varying lengths** of text sequences.

Hyperparameter Tuning

Upon transitioning to the development of our final solution using a Recurrent Neural Network (RNN), we recognized the **critical importance of fine-tuning the model's hyperparameters** to optimize its performance. To **address the various trial and error** phases in parameter selection during model construction, such as the *number of neurons* in layers, *dropout rates*, *activation function* in output layer. We relied on the use of **Keras Tuner**. The goal of this technology is to fine-tune hyperparameters. It **was not used for the final model selection** but facilitated **monitoring the accuracy and loss trends**, ultimately **providing a baseline model** for more **meaningful experiments**. For our specific use case, we chose to use the **Random Search** algorithm, one of the search strategies available within Keras Tuner.

Chapter 3

Modelling

3.1 Model description

The architecture incorporates various layers to effectively capture and represent intricate patterns within sequential textual data.

- **Tokenization Layer**

The initial layer of the model utilizes a **text tokenization layer**, represented as *encoder*. This layer **pre-processes** input text data, **converting words into numerical tokens**. The tokenization step is crucial for translating the textual information into a format suitable for subsequent neural network layers.

```
tf.keras.layers.TextVectorization(  
    output_sequence_length=500, standardize=None, max_tokens=8000)
```

– **Embedding Layer**

Following the tokenization layer, an **embedding layer** is employed to **transform the discrete tokenized representation into continuous vector space**. The Embedding layer has an *input_dim* equal to the vocabulary size obtained from the tokenization layer. Each token is embedded into a vector of size *output_dim*, set to 3000 in this architecture. The *mask_zero=True* parameter is utilized to handle sequences of varying lengths, effectively masking padded values.

```
tf.keras.layers.Embedding(  
    input_dim=len(encoder.get_vocabulary()),  
    output_dim=3000,  
    mask_zero=True),
```

– **Bidirectional LSTM Layers**

Two bidirectional **Long Short-Term Memory (LSTM)** layers are **employed for sequence modeling**.

The **first LSTM** layer, configured with 16 units and *return_sequences=True*, **processes input sequences bidirectionally while maintaining the temporal sequence information**.

```
tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(  
    units=16,  
    return_sequences=True))
```

Subsequently, a **dropout layer** with a *rate* of 0.5 is introduced to **mitigate overfitting**.

```
tf.keras.layers.Dropout(rate=0.5)
```

The **second** bidirectional **LSTM** layer, **mirroring the configuration of the first**, further captures contextual dependencies within the sequence. The absence of *return_sequences* in this layer indicates that it provides a consolidated representation rather than a sequence.

```
tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(
    units=16))
```

– Dense Layers

A densely connected layer with 8 *units* and Rectified Linear Unit (*ReLU*) activation is incorporated for **feature extraction** and **dimensionality reduction**. Additionally, a kernel regularization term with L1 penalty (0.1) is applied to the weights of this layer.

```
tf.keras.layers.Dense(
    units=8,
    activation = 'relu',
    kernel_regularizer = regularizers.l1(0.1))
```

The final layer is a dense layer with a single unit and Sigmoid activation function, producing the binary classification output. The sigmoid activation is chosen for its suitability in binary classification tasks.

```
tf.keras.layers.Dense(1, activation='sigmoid')
```

This comprehensive architecture is tailored to address the complexities handling long-sequences in natural language text, encompassing tokenization, embedding, bidirectional sequence modeling, and dense layers for feature extraction and prediction.

3.2 Model representation

At finally we have:

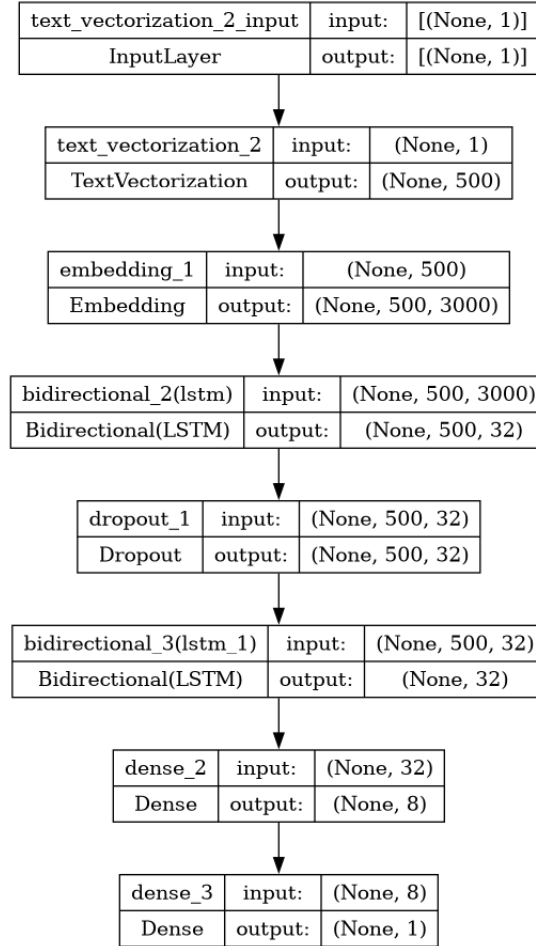


Figure 3.1: Recurrent Neural Network

Chapter 4

Results

4.1 Model Evaluation

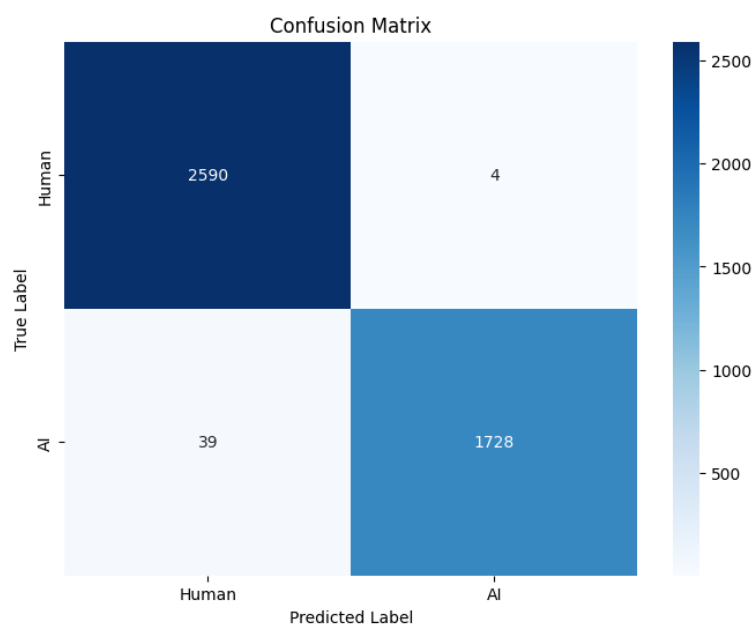


Figure 4.1: Confusion Matrix

Test Loss: 0.18071378767490387
Test Accuracy: 0.990139901638031

Figure 4.2: Evaluation on Test Set

	precision	recall	f1-score	support
0	0.99	1.00	0.99	2589
1	0.99	0.98	0.99	2162
accuracy			0.99	4751
macro avg	0.99	0.99	0.99	4751
weighted avg	0.99	0.99	0.99	4751

0.9905283098295096

Figure 4.3: Predict on Test Set

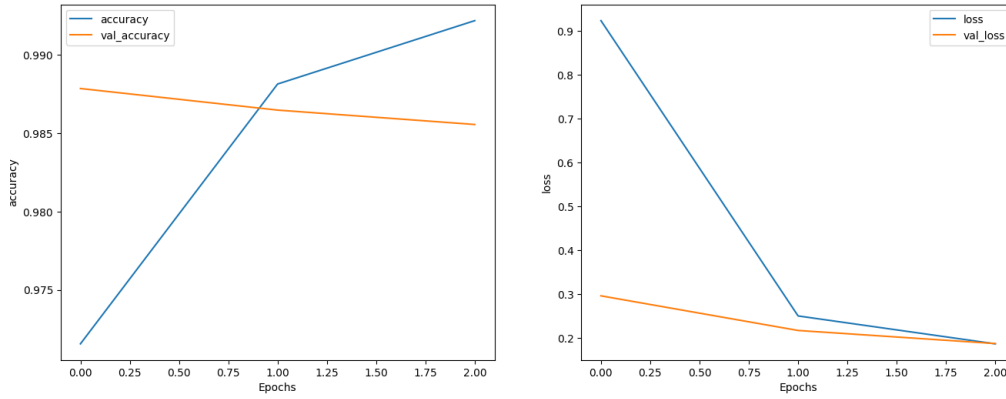


Figure 4.4: Loss and Accuracy

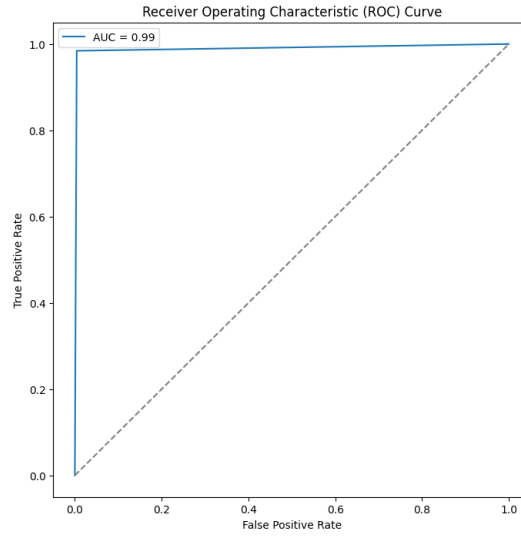


Figure 4.5: ROC Curve

4.1.1 Final Considerations

In the light of the results obtained, some considerations emerged on how the various technologies and resources obtained during the work could be used, in order to obtain a more consistent, balanced and general dataset.

- Use the found models for the text generation to generate a number of different prompts (1000) and then use the same models to generate a number of records (100) labeled with 1 (AI). In the course of the studies and the development of the project we found some generative models (*persuade-corpus*, used also in DAIGT-v2), pre-trained on datasets containing only texts written by humans. After several considerations, we realized that we could use them to generate the remaining records (100), for each prompt, labeled with 0 (Human). This would allow us to have a balanced and more general dataset.
- Generate more AI records with other generative models to balance at all (maybe with others models that respect our computational resources limitation).

4.2 Kaggle Results


#	Team	Members	Score	Entries	Last	Join
2927	G15		0.855	10	11h	

Figure 4.6: Accuracy on LLM Detect AI Generated Text Kaggle

4.3 Proof of Challenge Partecipation

- [GitHub Repository](#) 
- [Link to competition](#) 