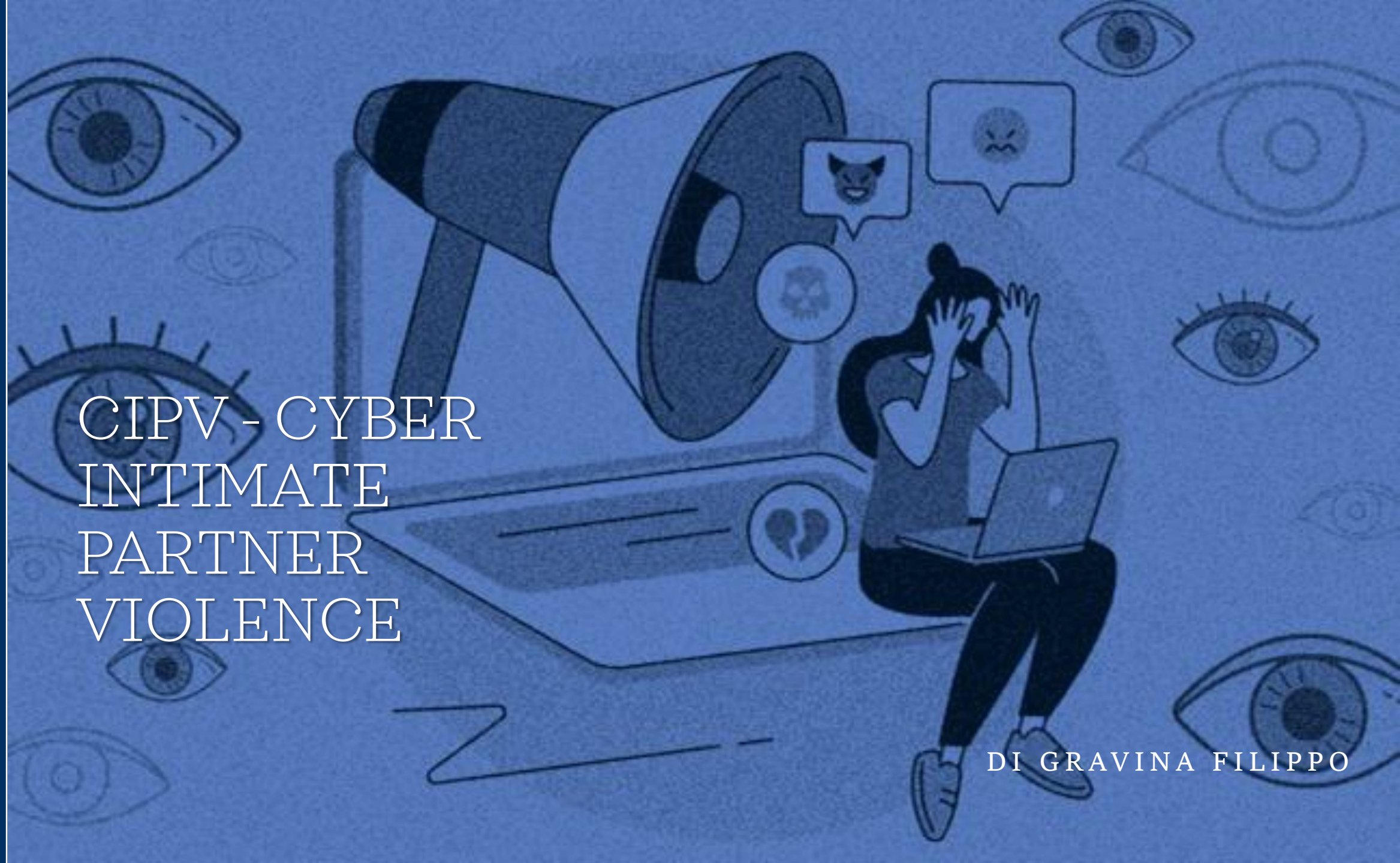
CIPV - CYBER INTIMATE PARTNER VIOLENCE



DI GRAVINA FILIPPO

INTRODUCTION AND MOTIVATIONS

- Cyber Intimate Partner Violence (C-IPV) is a form of abuse directed towards one's partner via social media and digital platforms.
- C-IPV can come in various forms:
  - Emotional abuse
  - Verbal abuse
  - Physical abuse

The main task is to identify the type of toxicity in a toxic conversation

given a dataset comprised of conversations between a perpetrator and a victim

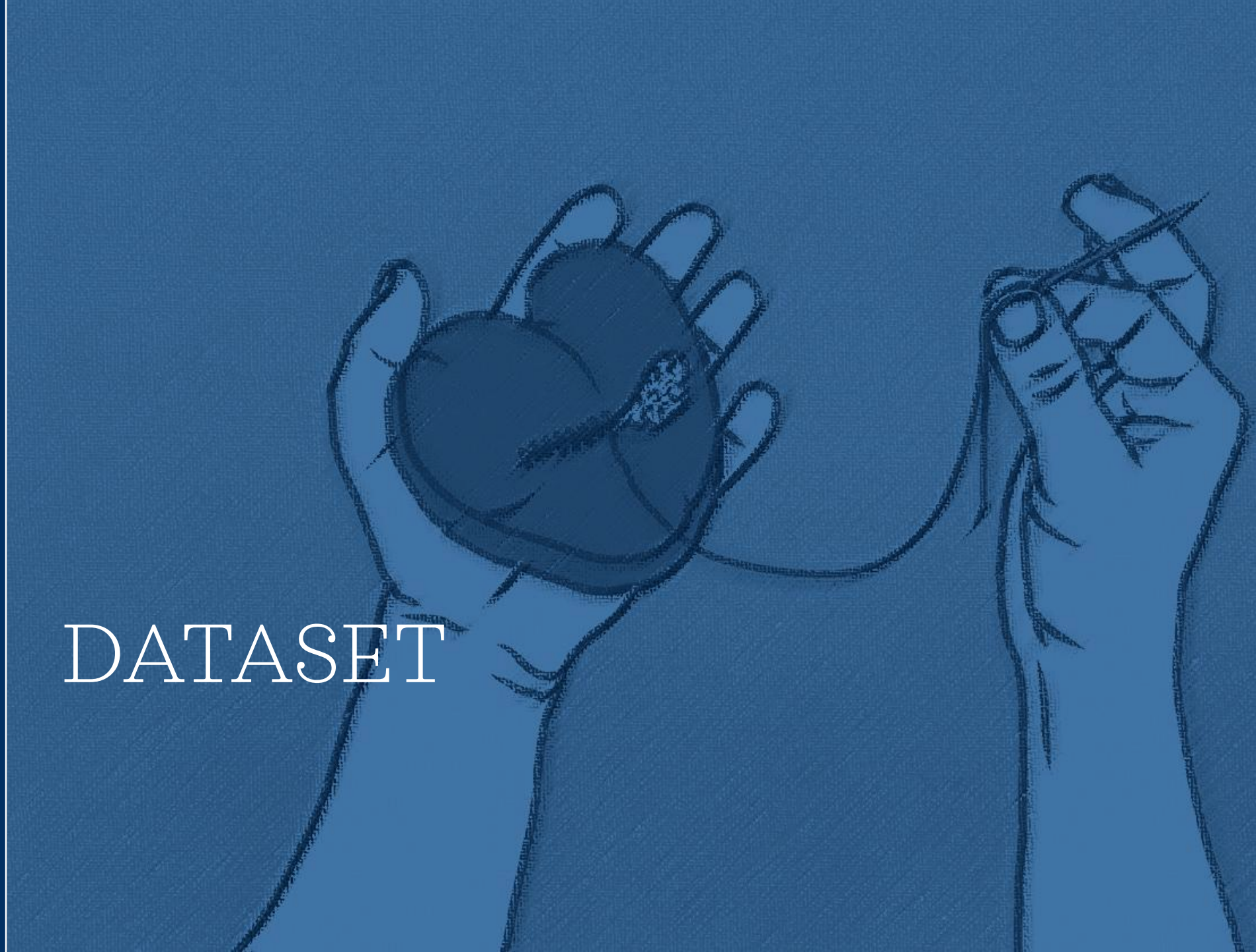
RELATED WORK

- Classification of Cyber Intimate Partner Violence has recently been studied by the paper "LLaMAntino against Cyber Intimate Partner Violence"

- The authors answered the two research questions:
  - RQ1: is the model able to recognize toxic sentences?
  - RQ2: are the explanations provided with 2-shot prompting similar to the "gold standard" provided by experts?

- They got the following results:
  - few-shot prompting always outperforms zero-shot prompting
  - even with few-shot prompting, the LLM learns to provide good explanations and learns to detect the presence or the absence of the toxicity

DATASET



DATASET

- The dataset used in the task is provided by the study by Vito Nicola Losavio called “IDaToC: An Italian Dataset of Toxic Conversation”
- This dataset provides:
  - a thousand of examples of synthetic conversations generated by large language models
  - the explanation of the toxicity
  - a binary feature representing if the conversation is toxic or not
  - a column representing the roles of the two people involved in the conversation

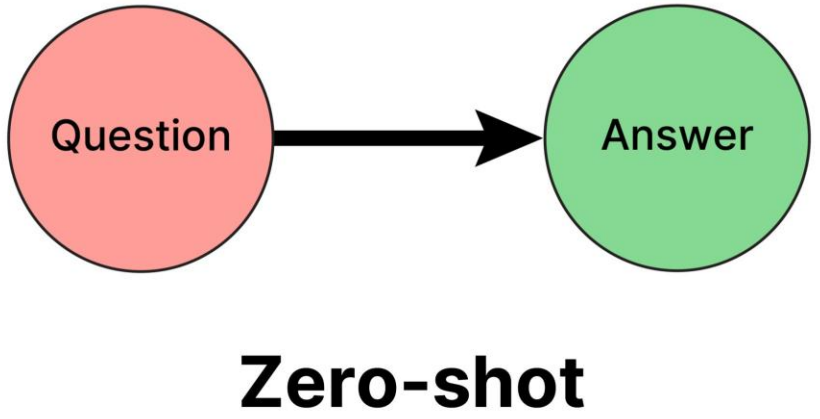
TARGET VARIABLE

- This column contains 10 different values:
  - Controllore e Isolata
  - Dominante e Schiavo emotivo
  - Geloso-Ossessivo e Sottomessa
  - Manipolatore e Dipendente emotiva
  - Narcisista e Succube
  - Perfezionista Critico e Insicura Cronica
  - Persona violenta e Succube
  - Psicopatico e Adulatrice
  - Sadico-Crudele e Masochista
  - Vittimista e Crocerossina

FIRST APPROACHES

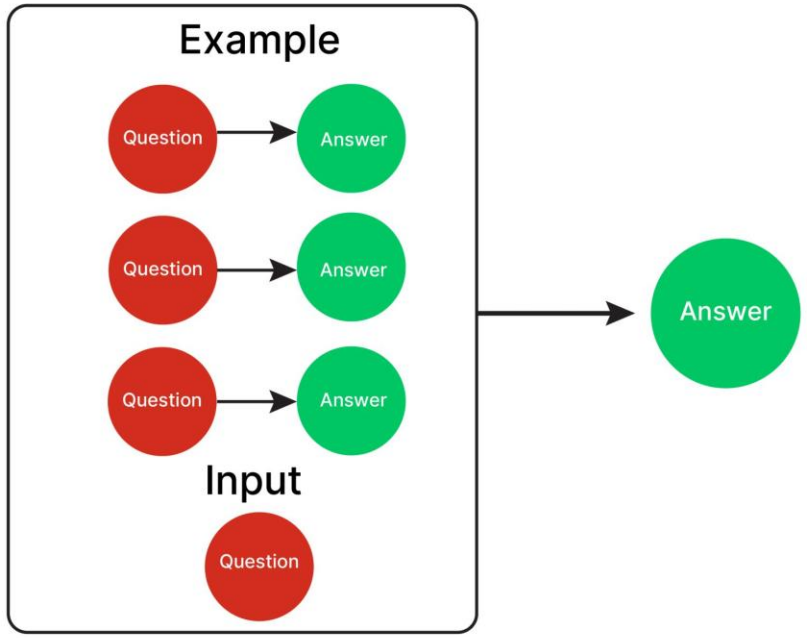
ZERO-SHOT PROMPTING

The zero-shot prompting classification led to very poor results, giving an accuracy score slightly higher than a random guess. Various LLMs have been tested, with the best result in this phase equal to a poor 14% of accuracy. The LLM that gave the best result was meta-llama/llama-4-scout-17b-16e-instruct.



FEW-SHOT PROMPTING

Few-shot prompting classification gave slightly better results, leading to a 30% accuracy in the best case. These results have been improved by adding some additional information, such as the result of the keyword extraction algorithm, the topic labeling output and giving in input the explanation of the violence available in the dataset.



KEYWORD EXTRACTION AND TOPIC LABELING

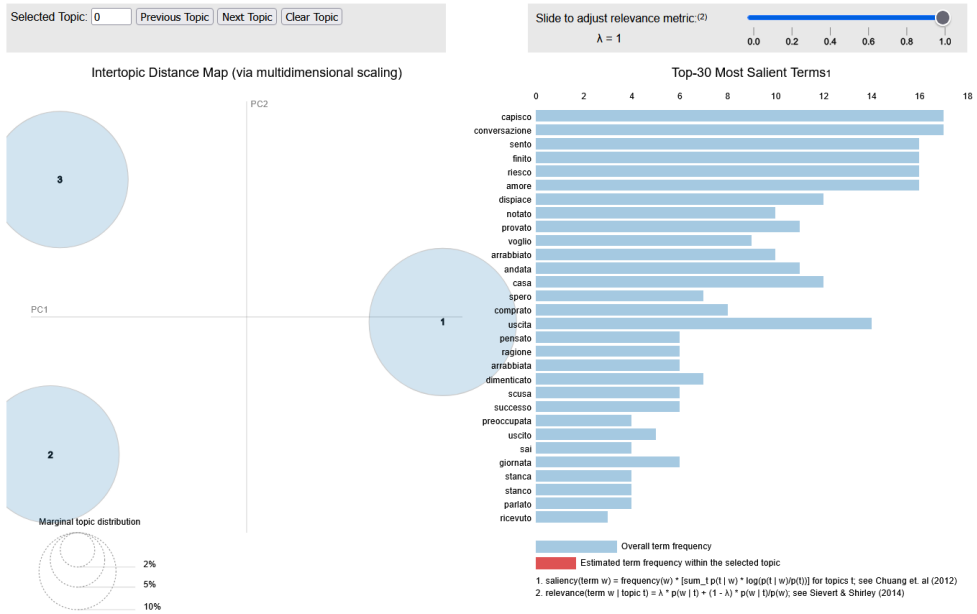
Process

The keyword extraction algorithm is used to find the most relevant words in the conversation, in order to focus more attention on the relevant parts of the sentences.

For the same reason, topic labeling has been implemented. The tested approaches are the LSA and the LDA method. The LDA method gave the best results.

The results of the LDA are printed on interactive web pages for improving the readability of the output.

LDA interactive web page

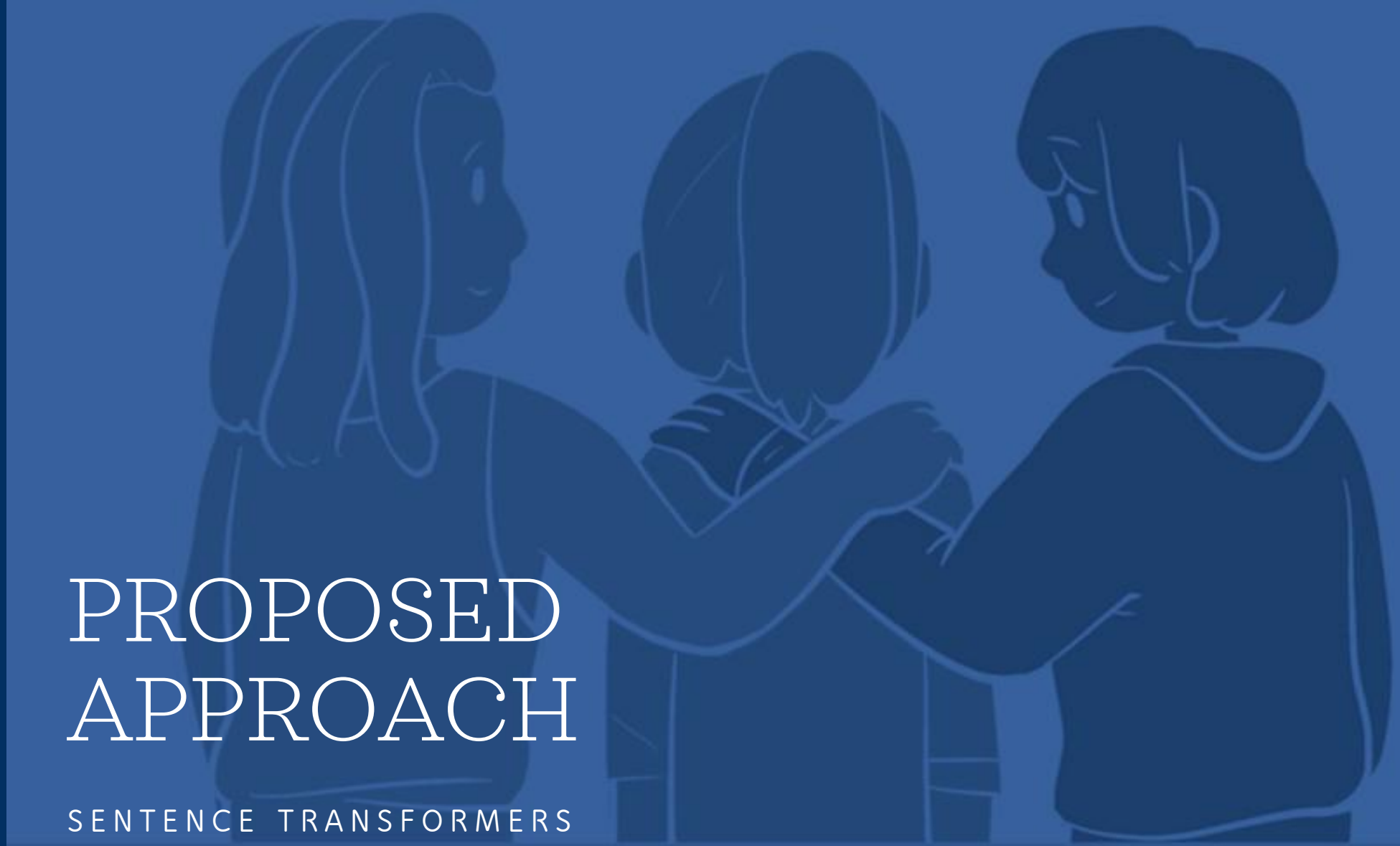


PROMPTING RESULTS

Even with few-shot prompting and additional information, such as:

- Keyword extraction
- Topic labelling
- Explanation

The best result only achieved 45% of accuracy. Even in this case, the LLM that gave the best result was meta-llama/llama-4-scout-17b-16e-instruct.

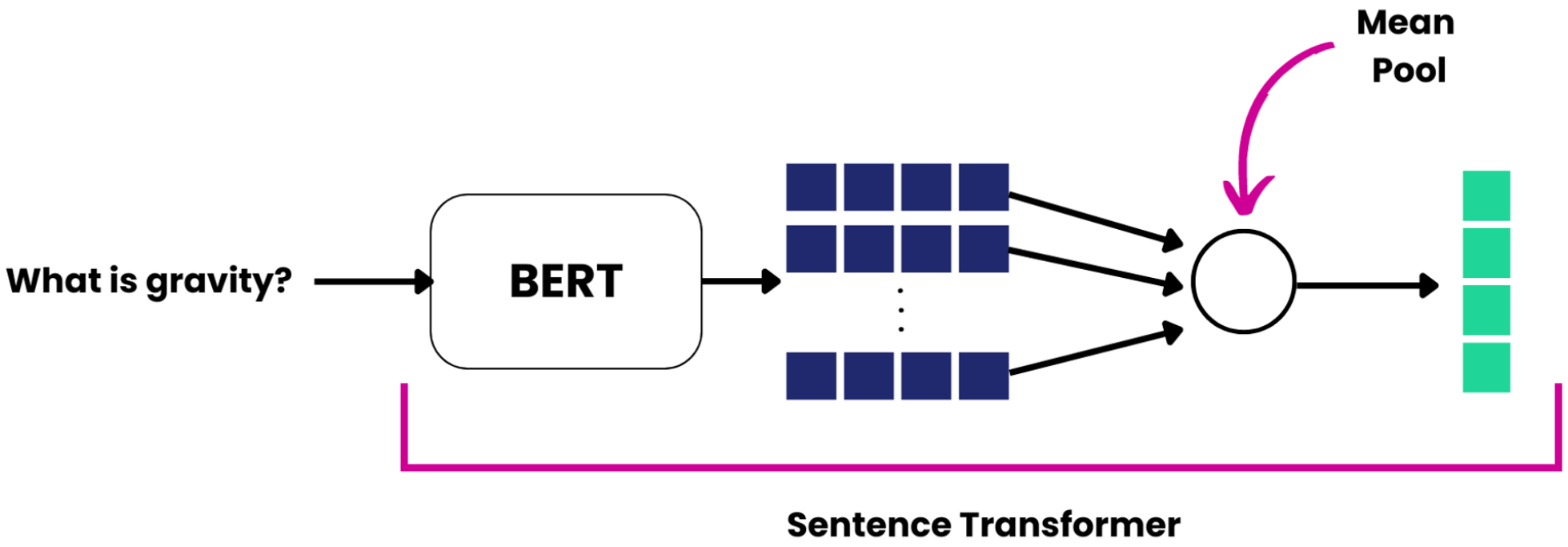


PROPOSED APPROACH

SENTENCE TRANSFORMERS

SENTENCE TRANSFORMERS

Sequence transformers are used to encode the input sentences and give these encodings in input to machine learning models that classify these new instances. About ten sequence transformers were tested, and the best sequence transformer turned out to be the model intfloat/multilingual-e5-large.



TESTED MACHINE LEARNING MODELS

KNN

- K-nearest neighbours is a non-parametric, supervised learning classifier. It is a fast algorithm, but in this case it performs the worst in terms of accuracy.

- Best result: 53% accuracy.

Random forest

- A random forest is a machine learning algorithm that uses an ensemble of decision trees to make predictions.

- Best result: 58% accuracy

TESTED MACHINE LEARNING MODELS

Multi-layer perceptron

- MLP is a feed-forward neural network consisting of fully connected neurons with nonlinear activation functions.

- Best result: 60% accuracy

Multinomial logistic regression

- It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable.

- Best result: 62% accuracy

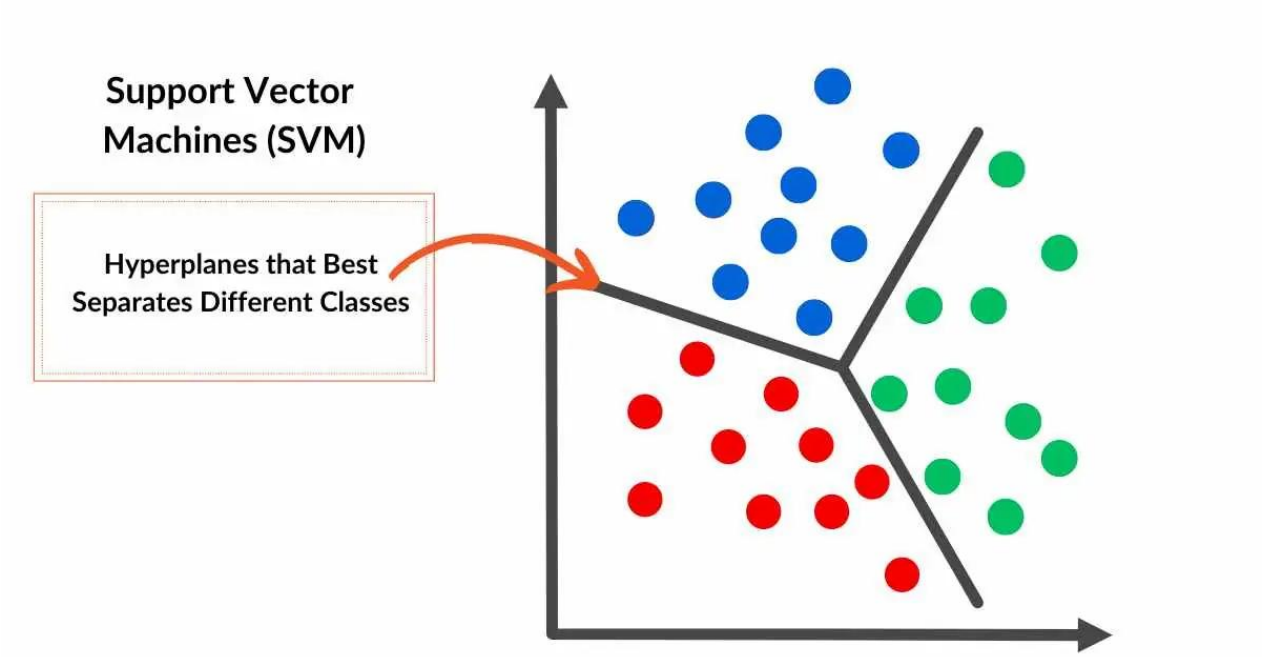
SUPPORT VECTOR CLASSIFIER

The chosen model is the Support Vector Machine.

SVC is a specific implementation of the Support Vector Machine algorithm that is designed specifically for classification tasks.

All the models were tested with grid search, which helped to find the best hyperparameters for the models.

Best result: 64% accuracy

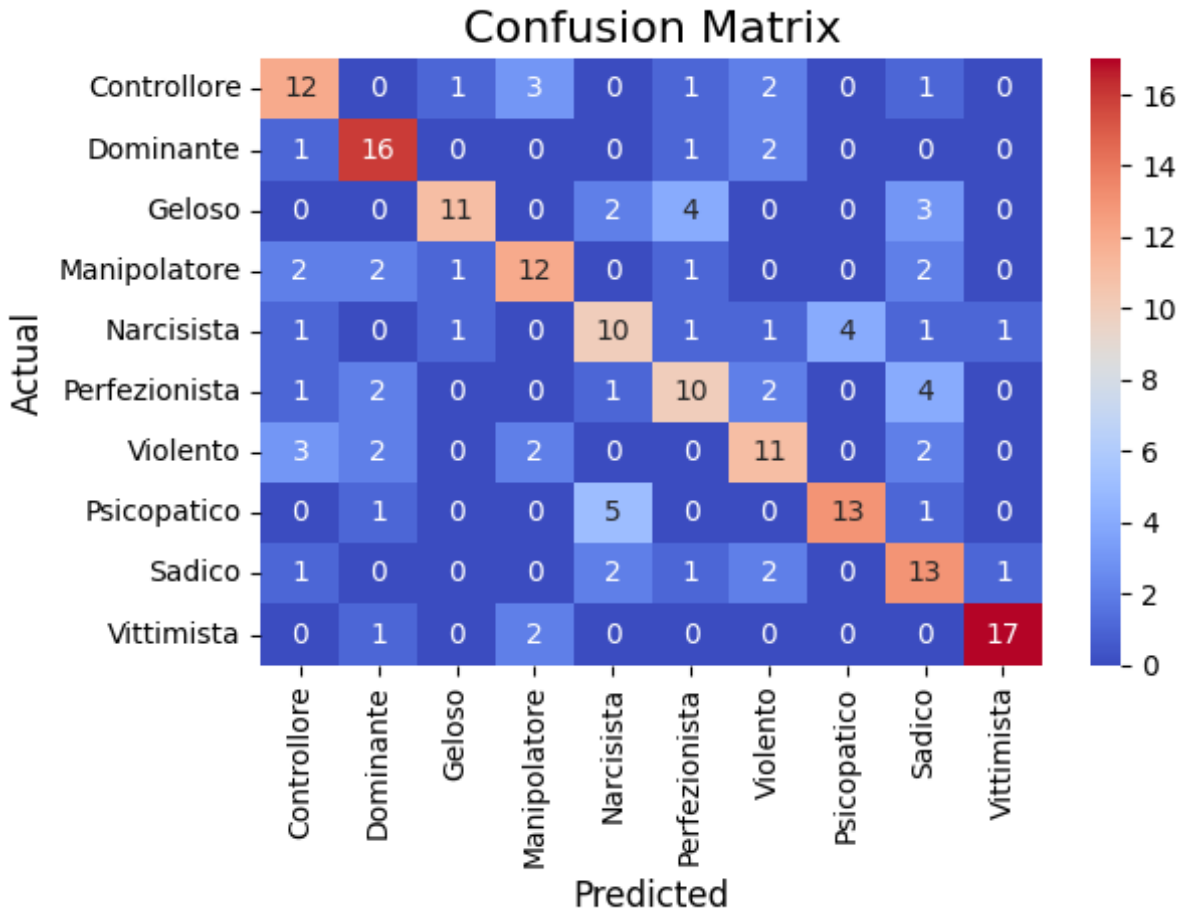


EVALUATION

RESULTS

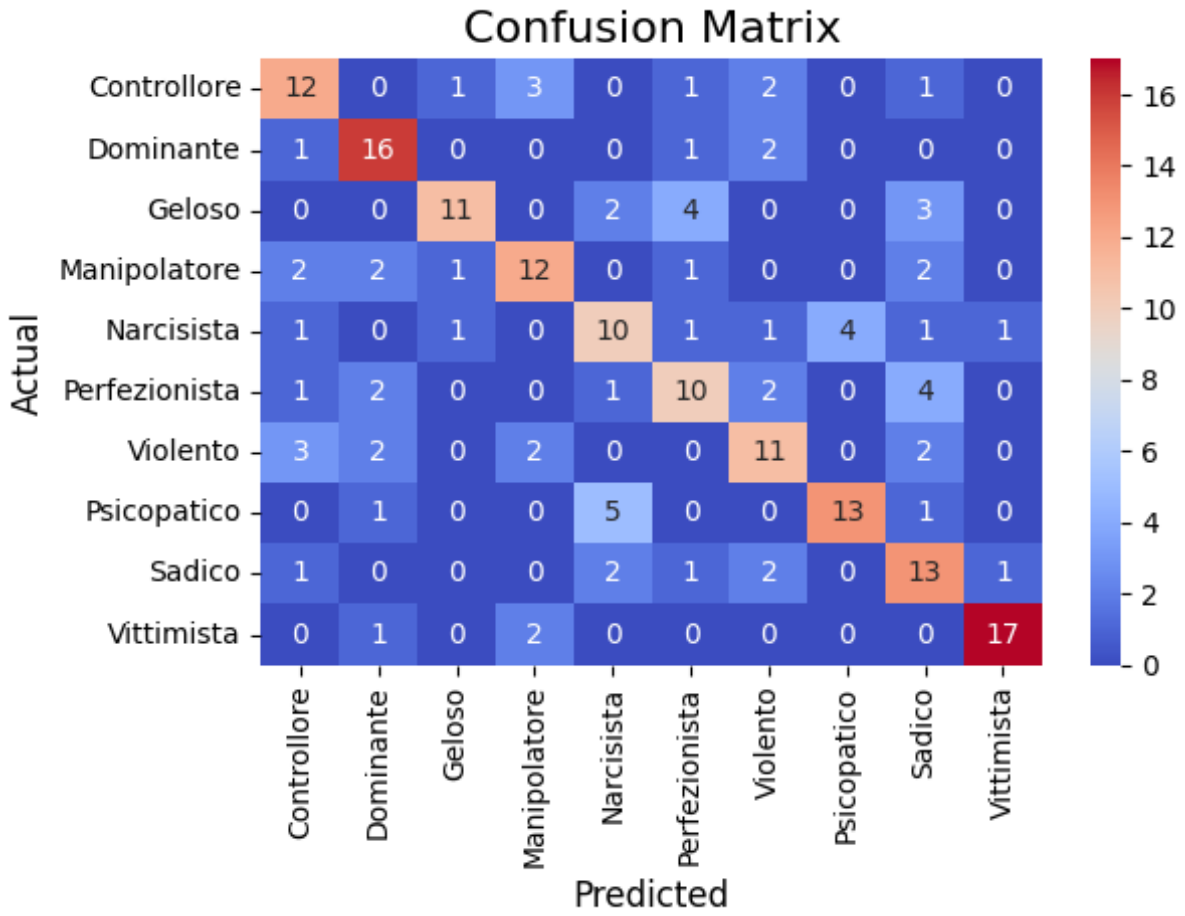
Several metrics have been evaluated:

- Accuracy
- Micro-average precision
- Macro-average precision
- Micro-average recall
- Macro-average recall
- Micro-average f1-score
- Macro-average f1-score



ROBUSTNESS OF THE RESULTS

All the metrics gave a value between 62% and 64%. The very low standard deviation (0.03) values indicates that the model is consistent and reliable in its predictions.



CONCLUSIONS AND LIMITATIONS

DATA AVAILABILITY

The limited size of the dataset and the little availability of Italian hate-speech dataset is a core problem.

MODEL AVAILABILITY

The little availability of fine-tuned model on Italian languages, and specifically to Italian hate-speech detection models, lead to initial poor results.

PROPOSED METHOD

Sequence transformers and machine learning classification models led to better results.

NEW CORPORA OF DATA

Bigger corpora of data can significantly improve the accuracy of the model

COMPUTATIONAL COMPLEXITY

The Support Vector Machine, is computionally expensive to run on large datasets.

BETTER POSSIBLE APPROACHES

With a larger corpus of data, fine-tuning becomes a powerful option.