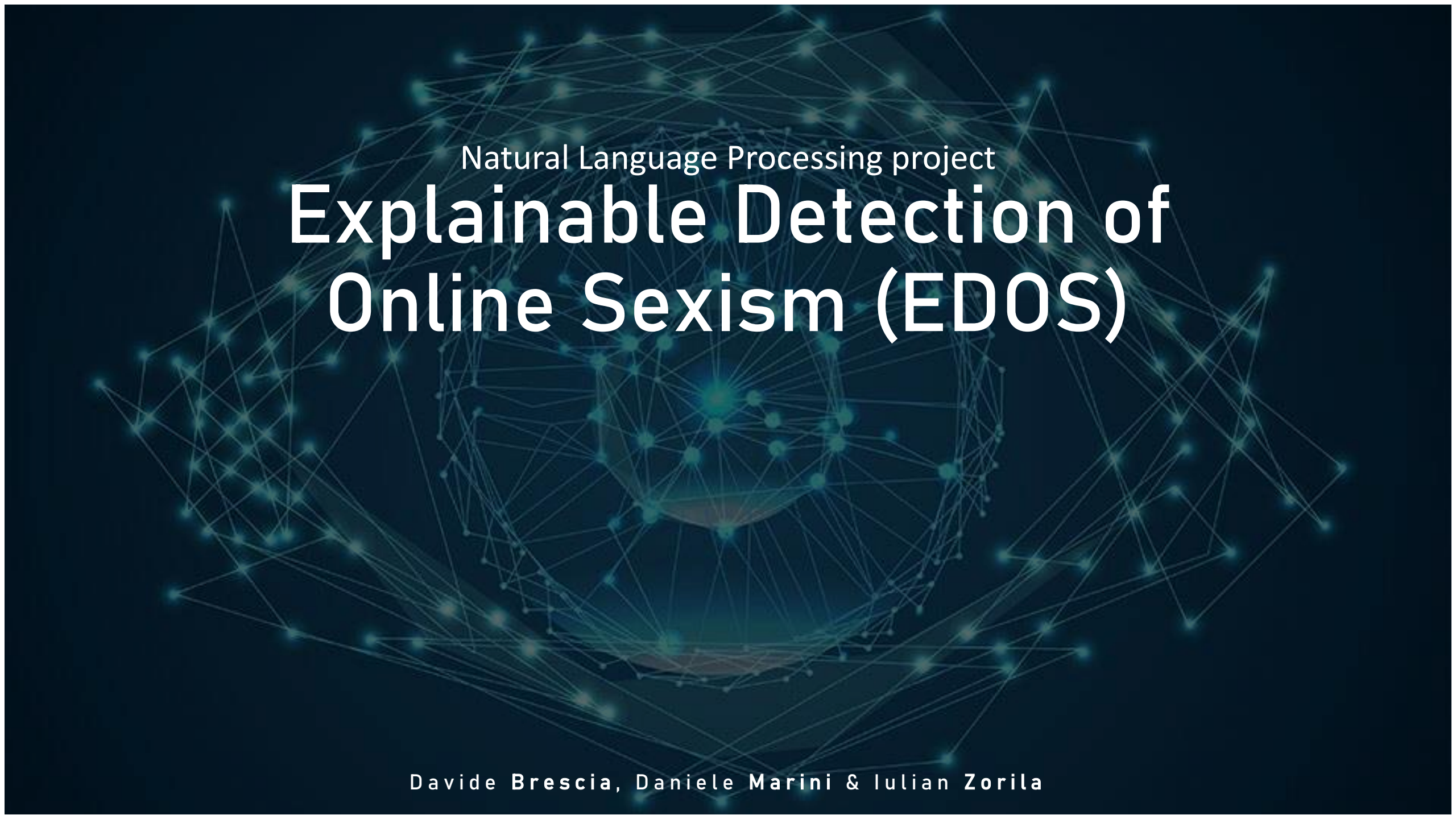
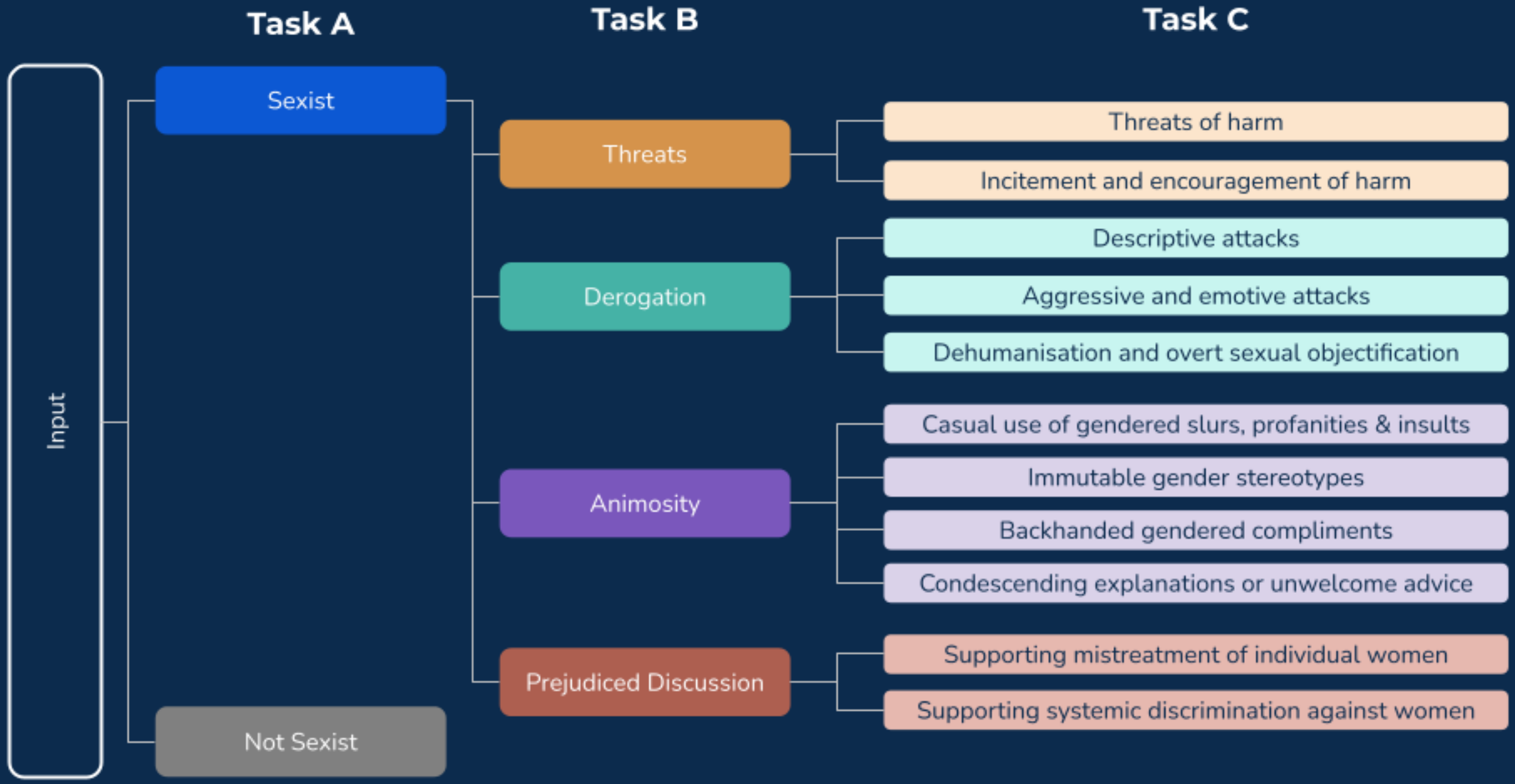


Natural Language Processing project

Explainable Detection of Online Sexism (EDOS)

Davide **Brescia**, Daniele **Marini** & Iulian **Zorila**

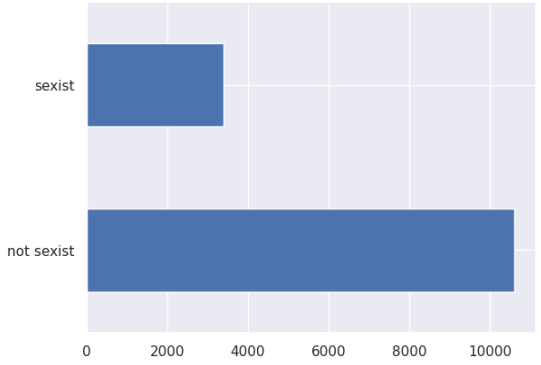

Task A
Task B
Task C
Input
Sexist
Not Sexist
Threats
Derogation
Animosity
Prejudiced Discussion
Threats of harm
Incitement and encouragement of harm
Descriptive attacks
Aggressive and emotive attacks
Dehumanisation and overt sexual objectification
Casual use of gendered slurs, profanities & insults
Immutable gender stereotypes
Backhanded gendered compliments
Condescending explanations or unwelcome advice
Supporting mistreatment of individual women
Supporting systemic discrimination against women

Introduction

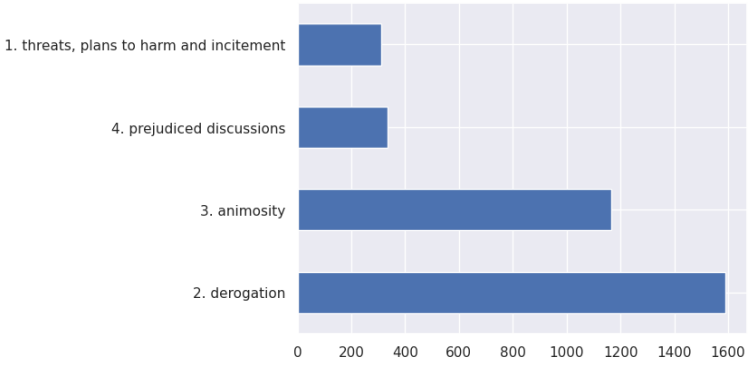
Dataset

SemEval-2023 Task 10: Explainable Detection of Online Sexism

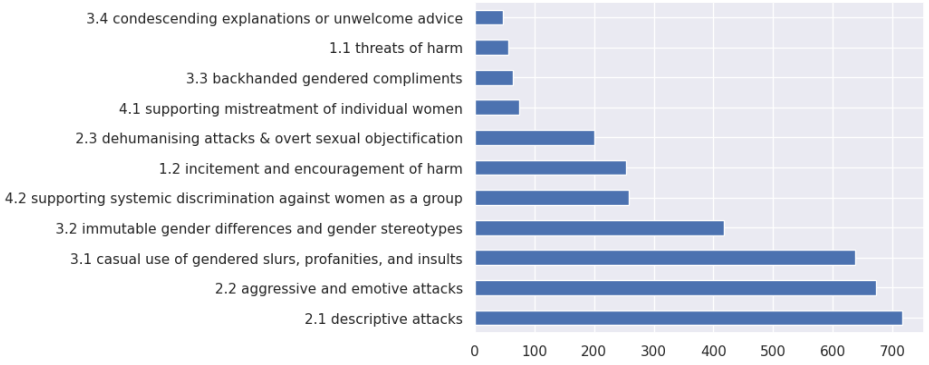
Task A



Task B

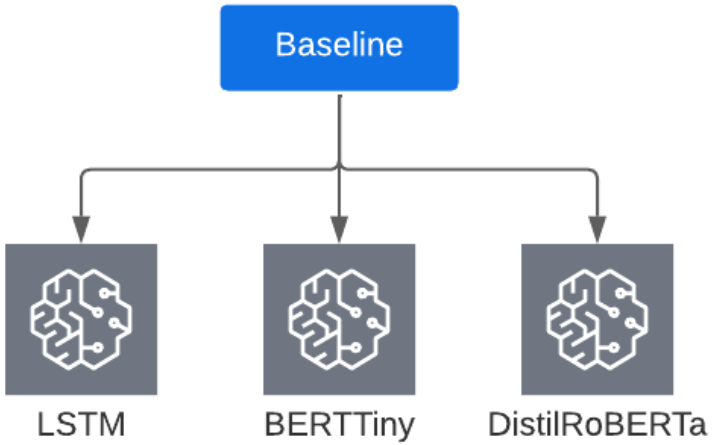


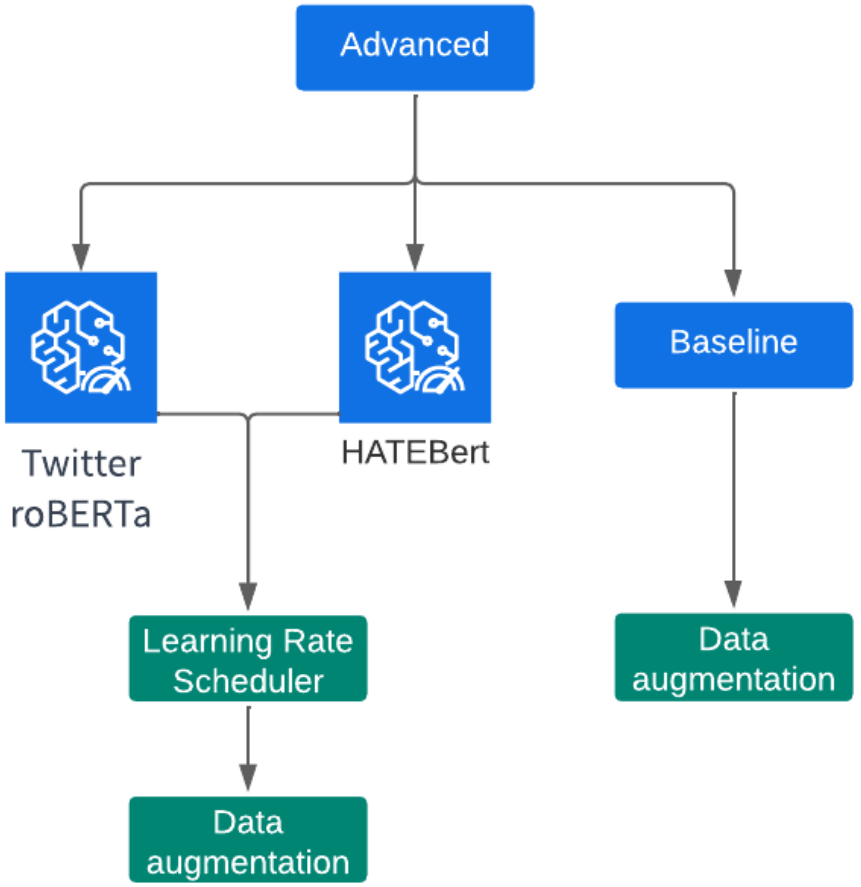
Task C



Introduction

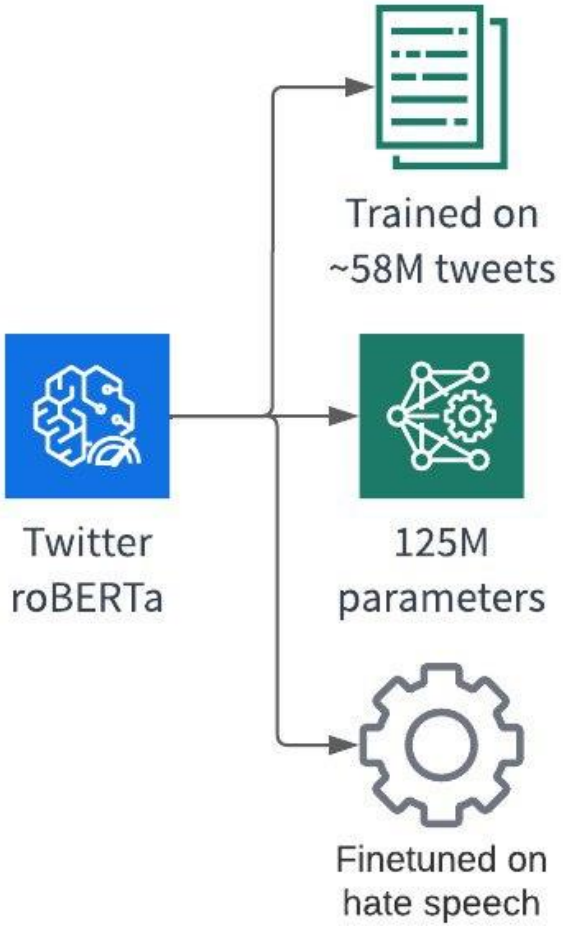
System Description

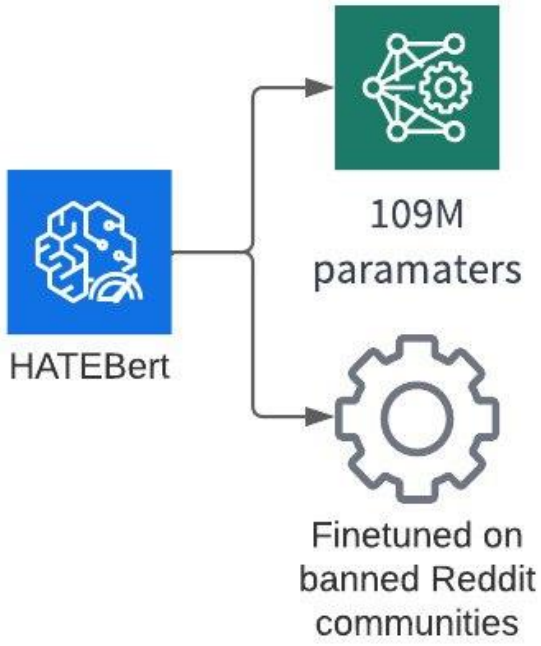




Advanced Phase

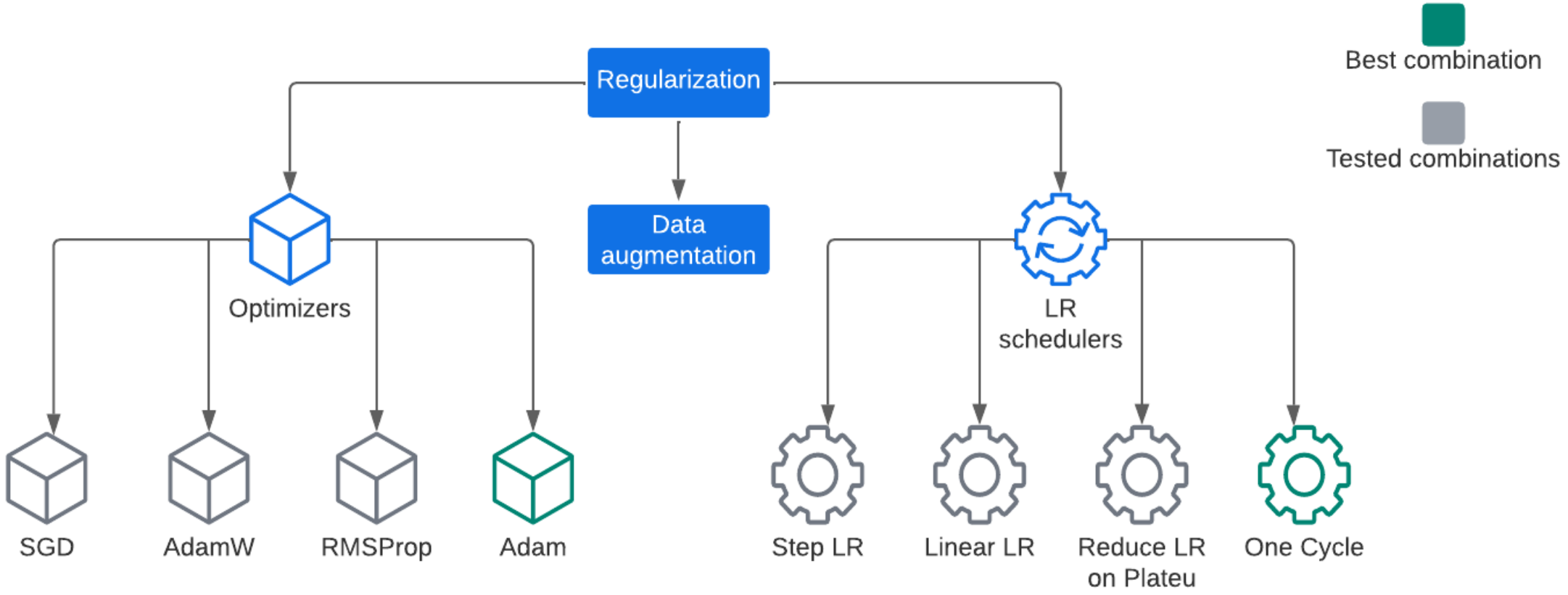
Models





Advanced phase

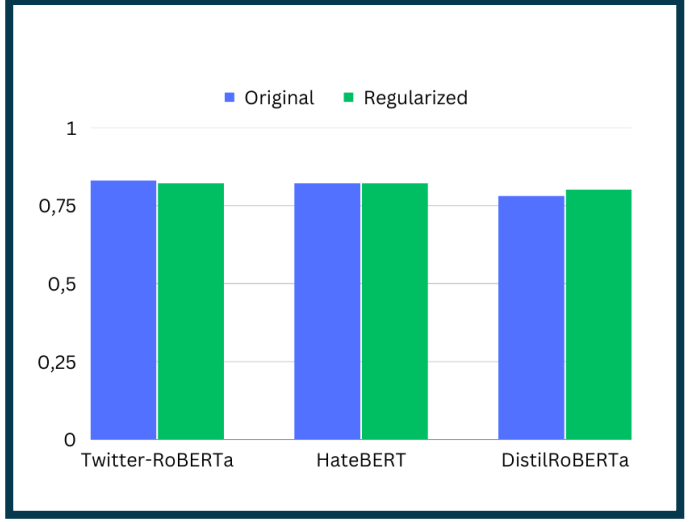
Regularization - introduction



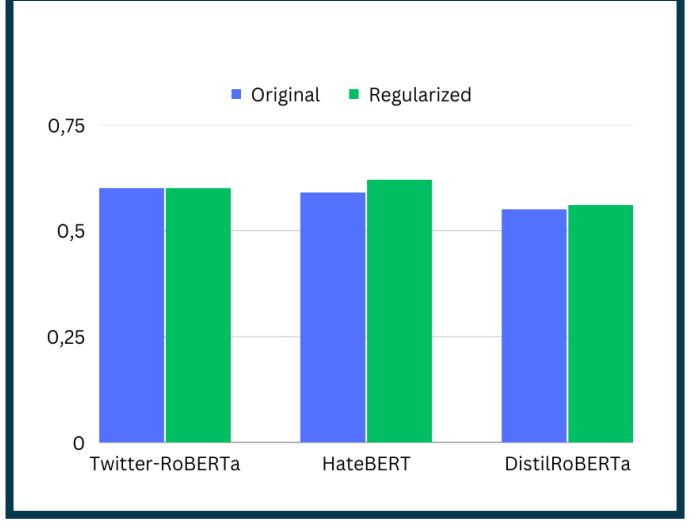
Advanced phase

Regularization - results

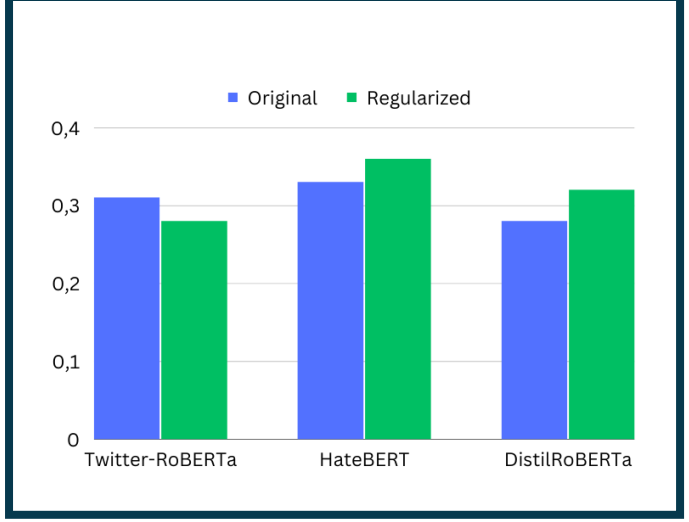
Task A



Task B



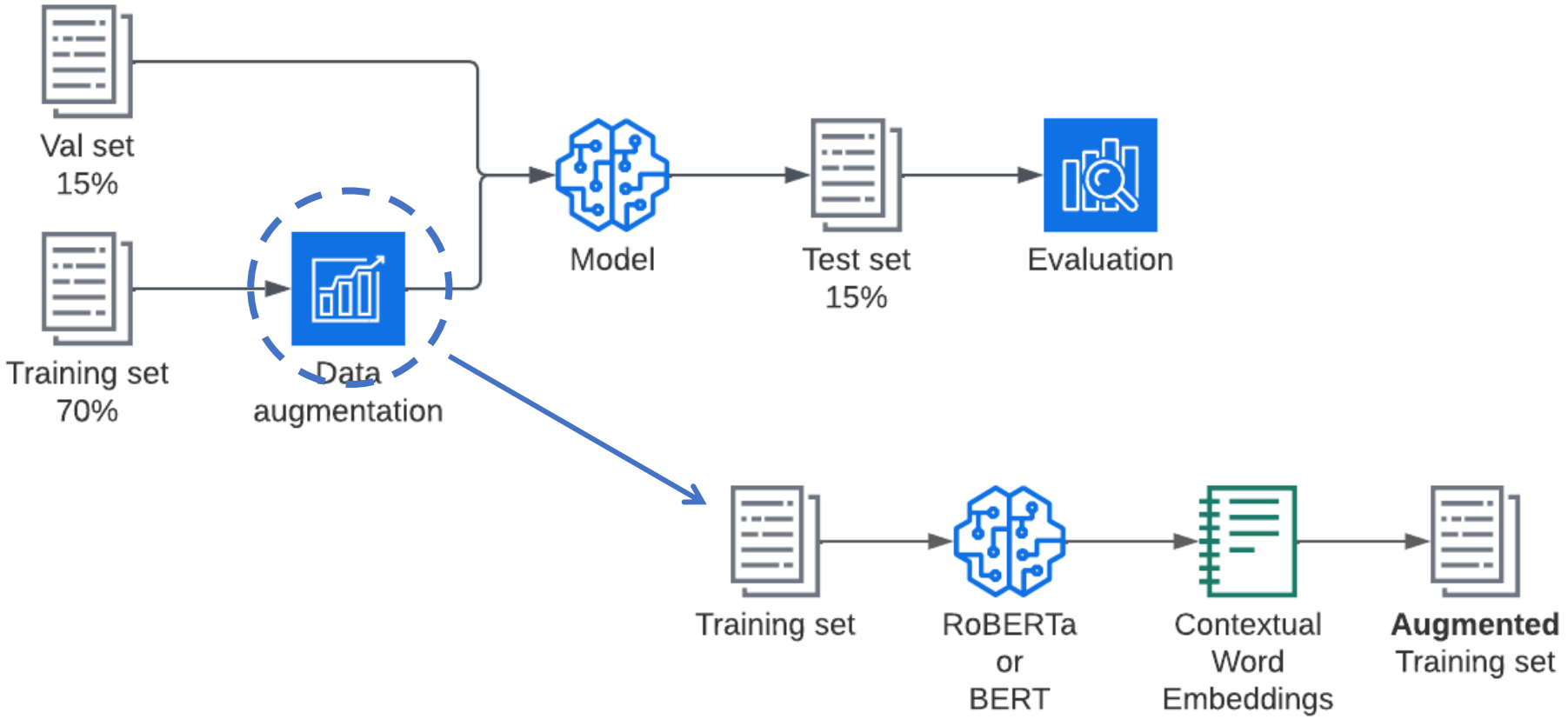
Task C



Data Augmentation

Introduction

General process for Task A, Task B and Task C



Data Augmentation

Contextual Word Embedding

Sexist samples

ORIGINAL	AUGMENTED
How about I talk shit to an idiot like you? Are you a nazi like that bitch?	How much about I talk some shit to an idiot like fucking you? — Are you a nazi fuck like them that bitch?
Husbands. Kill your piece of shit commie wives	Husbands. and Kill all your piece of baby shit commie and wives
If you want to attack culture and norms you have to go after the women themselves and call them out on their shit.	If you do want to attack culture and norms you have got to fucking go up after women the working women themselves and call upon them straight out on their shit.
just the inside of your wallet: cash, bank cards, credit cards, debit cards. That is the inside that matters to femoids.	just calculate the random inside contents of your wallet: cash, plastic bank report cards, credit cards, electric debit cards. That is the math inside that matters all to some femoids.

Data Augmentation

Results

	MODEL NAME	TASK B F1-DIFFERENCE	TASK C F1-DIFFERENCE
BASELINE	LSTM	16% +19%	3% +7%
BASELINE	BERTTiny	35% +8%	21% -3%
BASELINE	DistilRoBERTa	55% -1%	28% +7%
ADVANCED	Twitter-RoBERTa	60% +1%	**28% +10%**
ADVANCED	HateBERT	**62% -3%**	34% -2%

Final Results

Models Comparison

	DistilRoBERTa w/ LR Scheduler	RoBERTa pretrained on Hate Speech	RoBERTa pretrained on Hate Speech w/ LR Scheduler	HateBERT	HateBERT w/ LR Scheduler	LSTM w/ Augmentation	BERTTiny w/ Augmentation	DistilRoBERTa w/ Augmentation	RoBERTa pretrained on Hate Speech w/ LR Scheduler and Augmentation	HateBERT w/ LR Scheduler and Augmentation
Task A	0.8	**0.83**	0.82	0.82	0.82	0.74	0.73	0.78	0.82	0.82
Task B	0.56	0.6	0.6	0.59	**0.62**	0.35	0.43	0.54	0.61	0.59
Task C	0.32	0.31	0.28	0.33	0.36	0.1	0.19	0.35	**0.38**	0.34

Discussion

Part One

Smaller models performed overly well in binary classification but lower in more complex tasks.

	LSTM	BERTTiny	DistilRoBERTa
Task A	0.73	0.74	**0.78**
Task B	0.16	0.35	**0.55**
Task C	0.03	0.21	**0.28**

Moreover, due to its low complexity, it was challenging to enhance performance in Task A.

Discussion

Part Two

In Task A the usage of slang challenges the model, as these words are extremely rare and the model is not always able to associate the right meaning.

> *True. **Landwhales** have the pick of normal guys because Chad gets all the Top Stacies*

Discussion

Part Three

In Task A the model usually classifies offensive phrases as sexist, and conversely, sexist phrases that perhaps do not contain insults or negative terms are more likely to be misclassified as non-sexist phrases.

> *Because they're afraid of confrontation, a woman will always escape*

Model Prediction: not sexist

True Value: sexist

> *If my daughter got pregnant from this shit, That judge has a lawsuit coming!*

Model Prediction: sexist

True Value: not sexist

Discussion

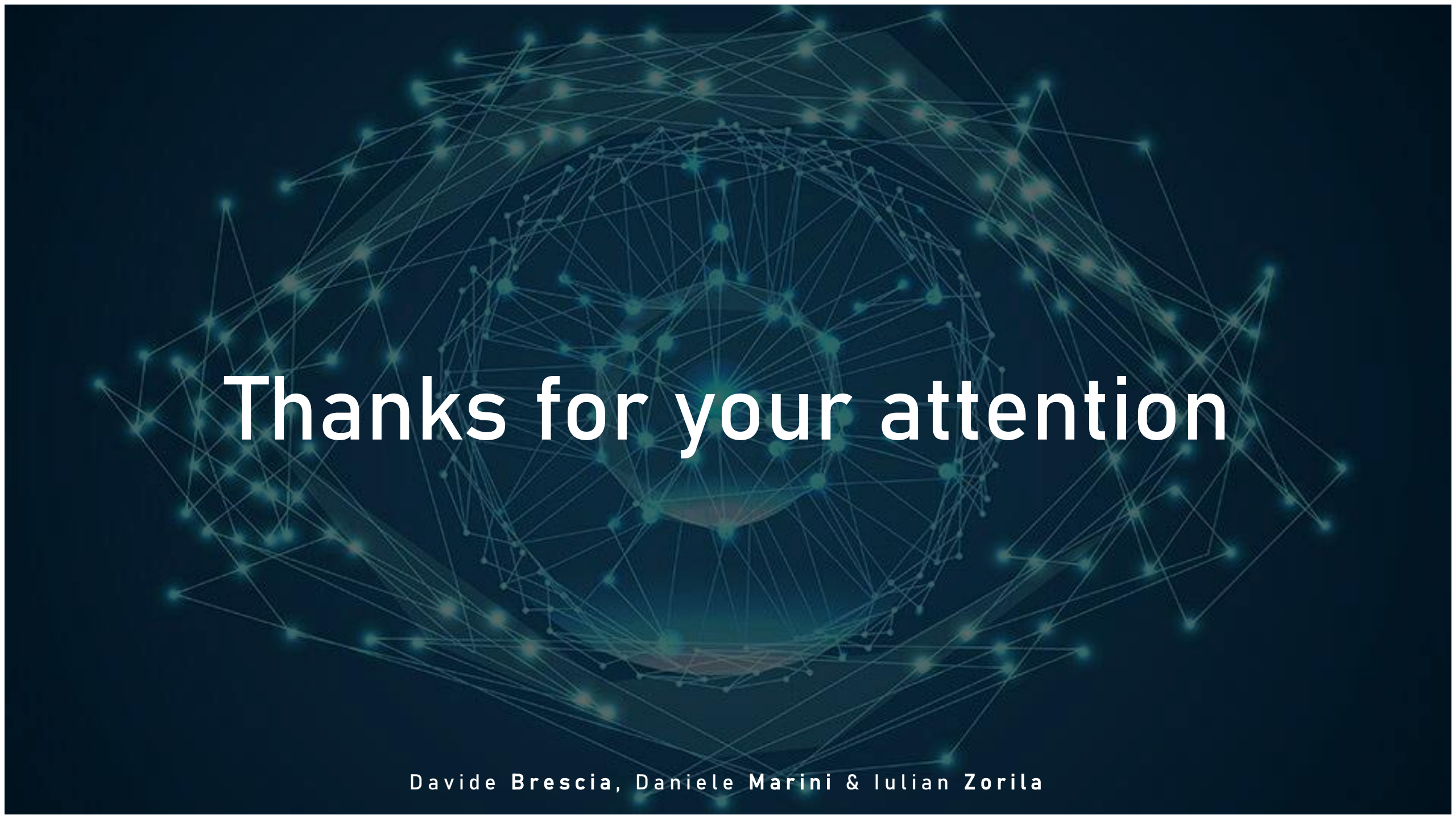
Part Four

Task type and model used for classification strongly impacted the effectiveness of data augmentation.

	DistilRoBERTa w/ LR Scheduler	RoBERTa pretrained on Hate Speech	RoBERTa pretrained on Hate Speech w/ LR Scheduler	HateBERT	HateBERT w/ LR Scheduler	LSTM w/ Augmentation	BERTTiny w/ Augmentation	DistilRoBERTa w/ Augmentation	RoBERTa pretrained on Hate Speech w/ LR Scheduler and Augmentation	HateBERT w/ LR Scheduler and Augmentation
Task A	0.8	**0.83**	0.82	0.82	0.82	0.74	0.73	0.78	0.82	0.82
Task B	0.56	0.6	0.6	0.59	**0.62**	0.35	0.43	0.54	0.61	0.59
Task C	0.32	0.31	0.28	0.33	0.36	0.1	0.19	0.35	0.38	0.34

Without Data Augmentation

With Data Augmentation


Thanks for your attention
Davide Brescia, Daniele Marini & Iulian Zorila