

**“NOME”**  
**AN EMPIRICAL STUDIES OF TEXT**  
**ANNOTATION**  
**BY MACHINES & HUMANS**

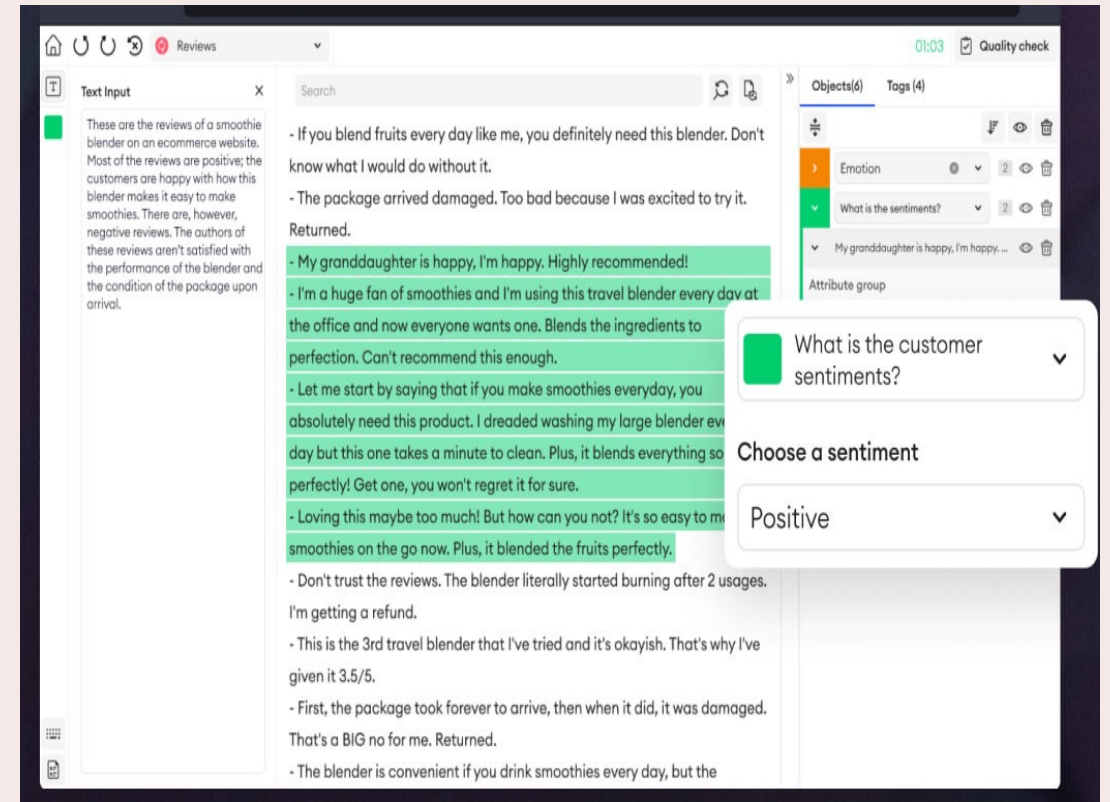
MUSA Project – Spoke No. 6 Working Group on “AI & INCLUSION”

July 2023

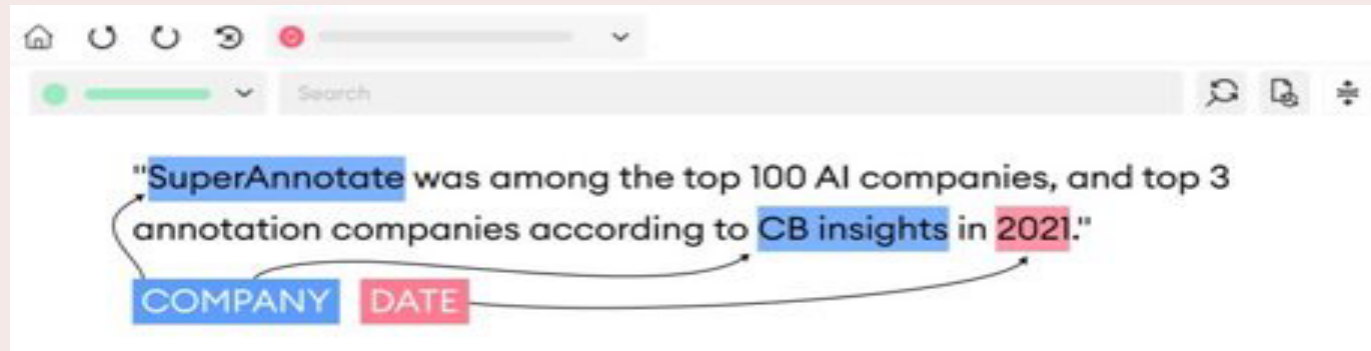
Paolo Ceravolo, Costanza Nardocci, Samira Maghool, Fatemeh  
Mohammadi

# WHAT IS TEXT ANNOTATION?

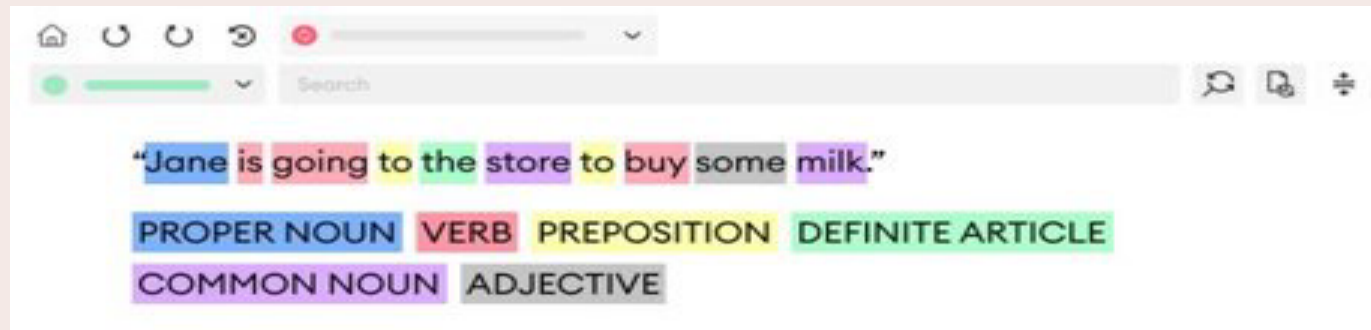
- **Assigning labels** to a text document or different elements of its content
- Use of annotated data to train AI models
- Goal: Help the machine understand the natural language of humans



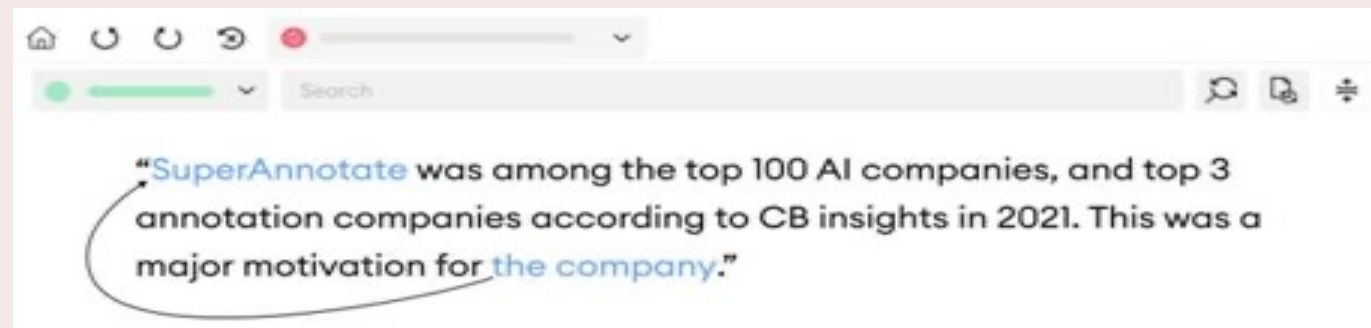
## Different Types of Automatic Text Annotation



**Named entity recognition (NER)**



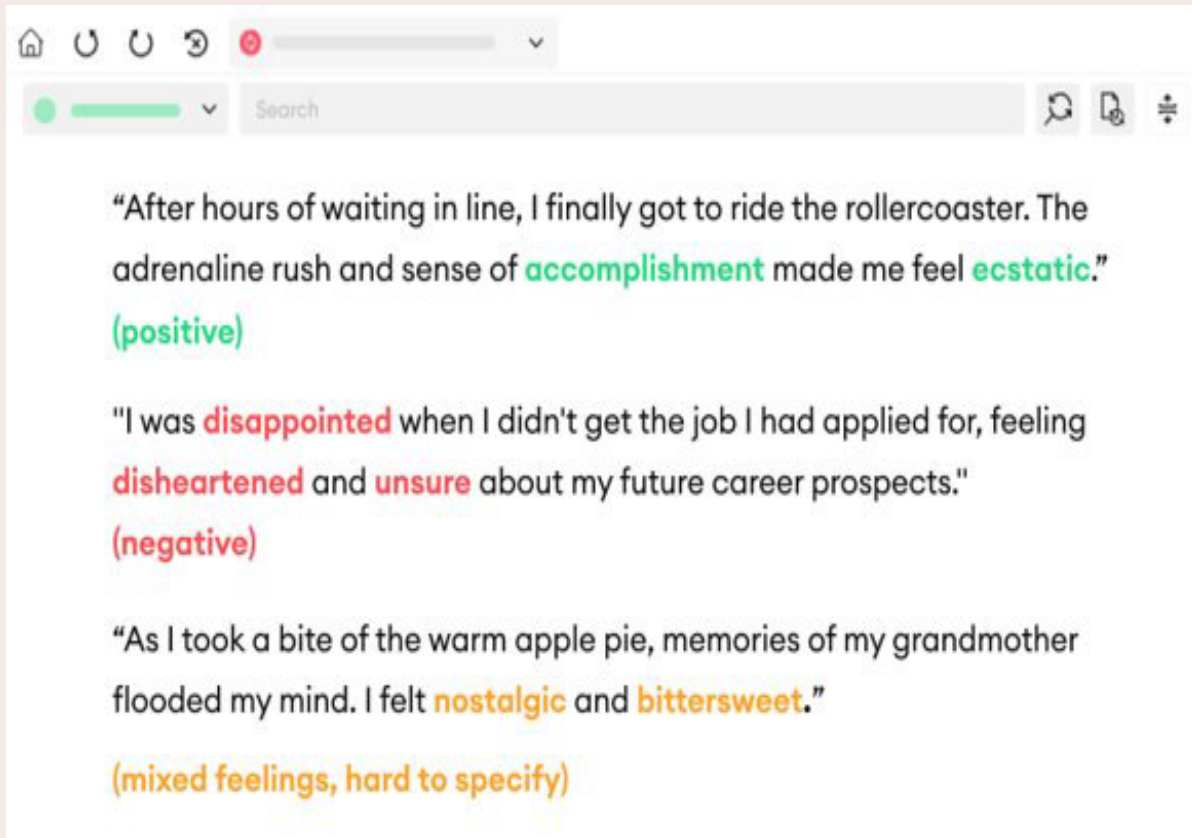
**Part-of-speech tagging**



**Coreference resolution (relationship annotation)**

# Different Types of Automatic Text Annotation

## Sentiment annotation



A screenshot of a web browser interface. The address bar shows a search bar with the text "Search". Below the address bar, there are three paragraphs of text with sentiment annotations. The first paragraph is highlighted in green and labeled "(positive)". The second paragraph is highlighted in red and labeled "(negative)". The third paragraph is highlighted in orange and labeled "(mixed feelings, hard to specify)".

"After hours of waiting in line, I finally got to ride the rollercoaster. The adrenaline rush and sense of **accomplishment** made me feel **ecstatic**."  
(positive)

"I was **disappointed** when I didn't get the job I had applied for, feeling **disheartened** and **unsure** about my future career prospects."  
(negative)

"As I took a bite of the warm apple pie, memories of my grandmother flooded my mind. I felt **nostalgic** and **bittersweet**."  
(mixed feelings, hard to specify)

## named entity linking (NEL)

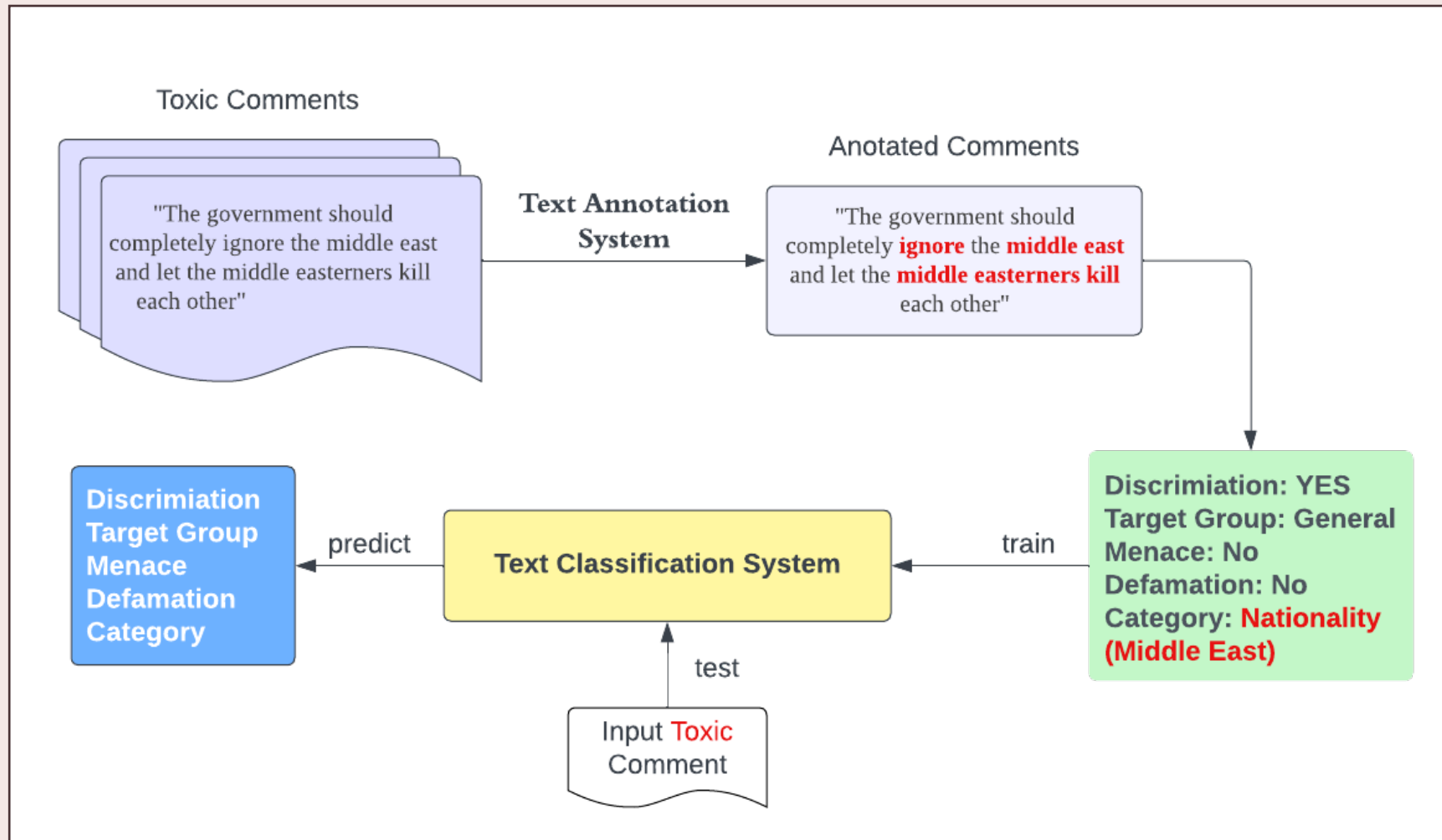
"Paris is the capital of France"

↓  
wikipedia.org/wiki/**Paris**

↓  
wikipedia.org/wiki/**France**

# TEXT CLASSIFICATION

- Using Text Annotation for HS and Discrimination Detection



# WHY IS IT HATE SPEECH? MASKED RATIONALE PREDICTION FOR EXPLAINABLE HATE SPEECH DETECTION

No.	Model	Example sentence and Rationale	Label
<b>Case 1 Explicit hate speech</b>			
1	Human	imagine unironically believing the high iq kike meme	HS
	BERT-MRP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-RP	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
	BERT-HX	imagine un ##iro ##nical ##ly believing the high iq ki ##ke me ##me	HS
2	Human	<user> ffs these ghetto blacks are ignorant like cracker mountain hillbillies	HS
	BERT-MRP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-RP	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
	BERT-HX	<user> ff ##s these ghetto blacks are ignorant like crack ##er mountain hill ##bill ##ies	HS
<b>Case 2 Implicit hate speech</b>			
3	Human	yah christian island republic shall not have queers or witches on it if we find any death	HS
	BERT-MRP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-RP	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	HS
	BERT-HX	ya ##h christian island republic shall not have queer ##s or witches on it if we find any death	NO
4	Human	you can not culturally enrich a moslem	HS
	BERT-MRP	you can not culturally en ##rich a mo ##sle ##m	HS
	BERT-RP	you can not culturally en ##rich a mo ##sle ##m	NO
	BERT-HX	you can not culturally en ##rich a mo ##sle ##m	NO
5	Human	<user> he is infected with jihadi virus he will spread it to others	HS
	BERT-MRP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-RP	<user> he is infected with jihad ##i virus he will spread it to others	HS
	BERT-HX	<user> he is infected with jihad ##i virus he will spread it to others	OF



# A DEEP NEURAL NETWORK BASED MULTI TASK LEARNING APPROACH TO HATE SPEECH DETECTION

- Hate
- Offensive
- Racism
- Sexism
- Aggression
- Overtly aggressive(OAG)
- Covertly aggressive (CAG)
- Harassment

- (i). Explicit attack towards individual (EI)
- (ii) Explicit attack towards group (EG)
- (iii). Implicit attack towards individual (II)
- (iv). Implicit attack towards groups (IG)

Class	Type	Sentence
Offensive	II	"@JasminePore: If you dressed up as a cat for Halloween you are basic." ...or a pussy
Offensive	II	"@chieflothaohes: In D.C., we call them rollers, in Chiraq we call them thots." All across America they have been known as a h*e.
Offensive	EI	"@YoungRicooo: Happy birthday b**ch n**ga @WexicanAfro702" Thanks pal
Offensive	EI	@SimplyHype: @kobe Bryant shut the f**k up, you f**king trash ni**a, work on another ring; <a href="http://t.co/6RleNKR7EE">http://t.co/6RleNKR7EE</a>
Offensive	EG	RT @AthleteChamp: Destroying a team that talked trash before the game.
Offensive	EG	RT @BadAdviceDog: r u african because ur african b***h
Offensive	IG	;@RudePost: A new breed of white girl... <a href="https://t.co/0Zsu3zQixD">https://t.co/0Zsu3zQixD</a> ; she slammed that b**ch;
Offensive	IG	@Weakass_Drex: If Jennifer Hudson can lose weight anybody can you fat a*s h*es!!!
Harassment	II	@benshapiro You're on a roll today Ben! Obama can't even act like a good Christian let alone lecture the Jews!
Harassment	II	SOMEONE F****NG HIRE @lindsaylohan AS THEIR SPIN DOCTOR PLZ
Harassment	EI	RT @IrishDude10: Go back to prison you f***ing n**ger #Mayweather
Harassment	EI	@Bambi_cop Untag you f*****g ni**er
Harassment	EG	RT @TrapHouseMario: White people literally Be like " I'm not racist you f***ing Ni**er"
Harassment	EG	@GORBtheOTHER The Jews are the masterminds idiot
Harassment	IG	RT @janimine: Diversity is a code word for #WhiteGenocide <a href="https://t.co/A28ASFGR4">https://t.co/A28ASFGR4</a>
Harassment	IG	@AmancioHitler I said "a glass of juice" not "gas the Jews"!!
CAG	II	Anna hazare is agent
CAG	II	Anirban Kundu So you are the pesticide
CAG	EI	Anna is a "Natak Raja" !!!
CAG	EI	So that Modi doesnt have to campaign separately... Feku and his never ending gimmicks
CAG	EG	Also try to stop violence from communists
CAG	EG	Judiciary system of India.....?????????
CAG	IG	The Great Powers should avoid such dangerous inhuman dramas to happen in the society.. it creates bad results on one's health
CAG	IG	Well first of all municipal corporation need to keep cows off the streets. Its cows owner's responsibility.

# ETHOS: A MULTI LABEL HATE SPEECH DETECTION DATASET

**Does this comment contains hate speech? (required)**

Yes  
 No

**Does this comment incites violence? (required)**

Yes  
 No

**Is this comment targeting a specific individual (directed) or a group/class of people (generalized)? (required)**

Directed  
 Generalized

**Which category of hate speech is it? (required)**

Gender  
 Race  
 National Origin  
 Disability  
 Religion  
 Sexual Orientation

Hate speech detection system with binary information

Wish you cut your veins. Don't shout out you have mental problems. Act. Cut them;

Labels: Hate Speech 87%

Ban  
Allow

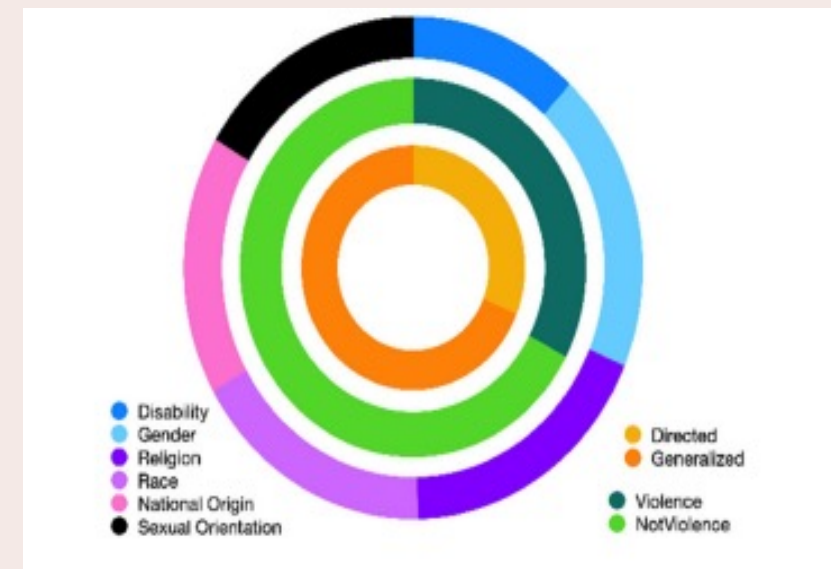
---

Hate speech detection system with multilabel information

Wish you cut your veins. Don't shout out you have mental problems. Act. Cut them;

Labels: Hate Speech 87% Incites Violence 92% Directed 100% Disability 100%

Ban  
Allow

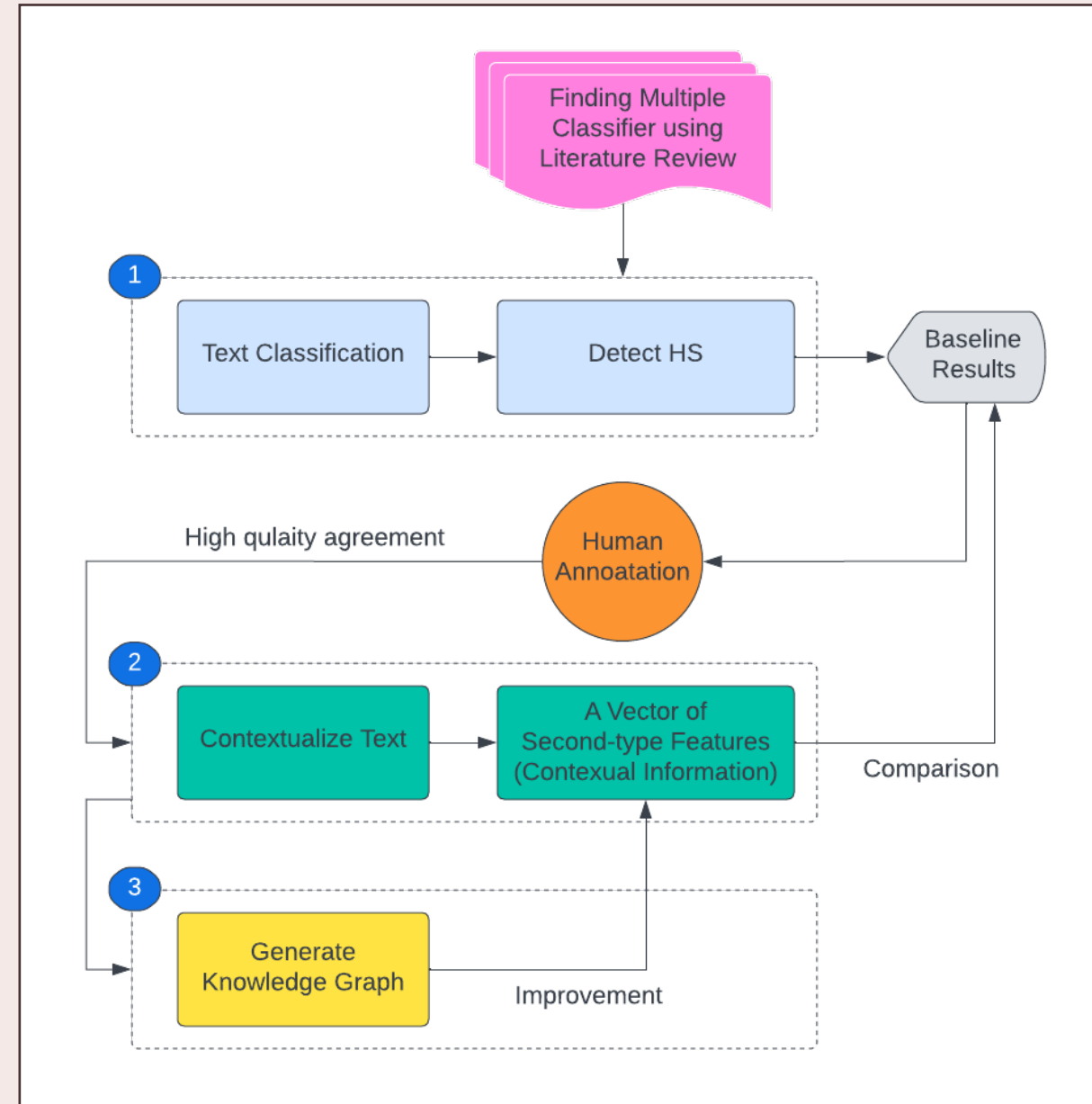




# RESEARCH OUTLINES

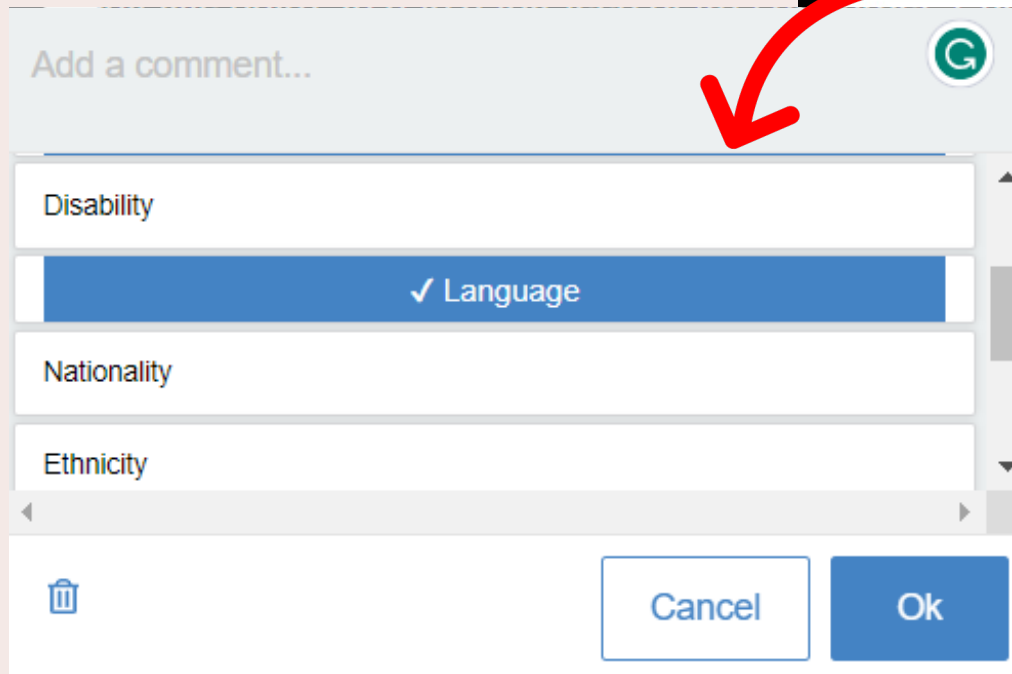
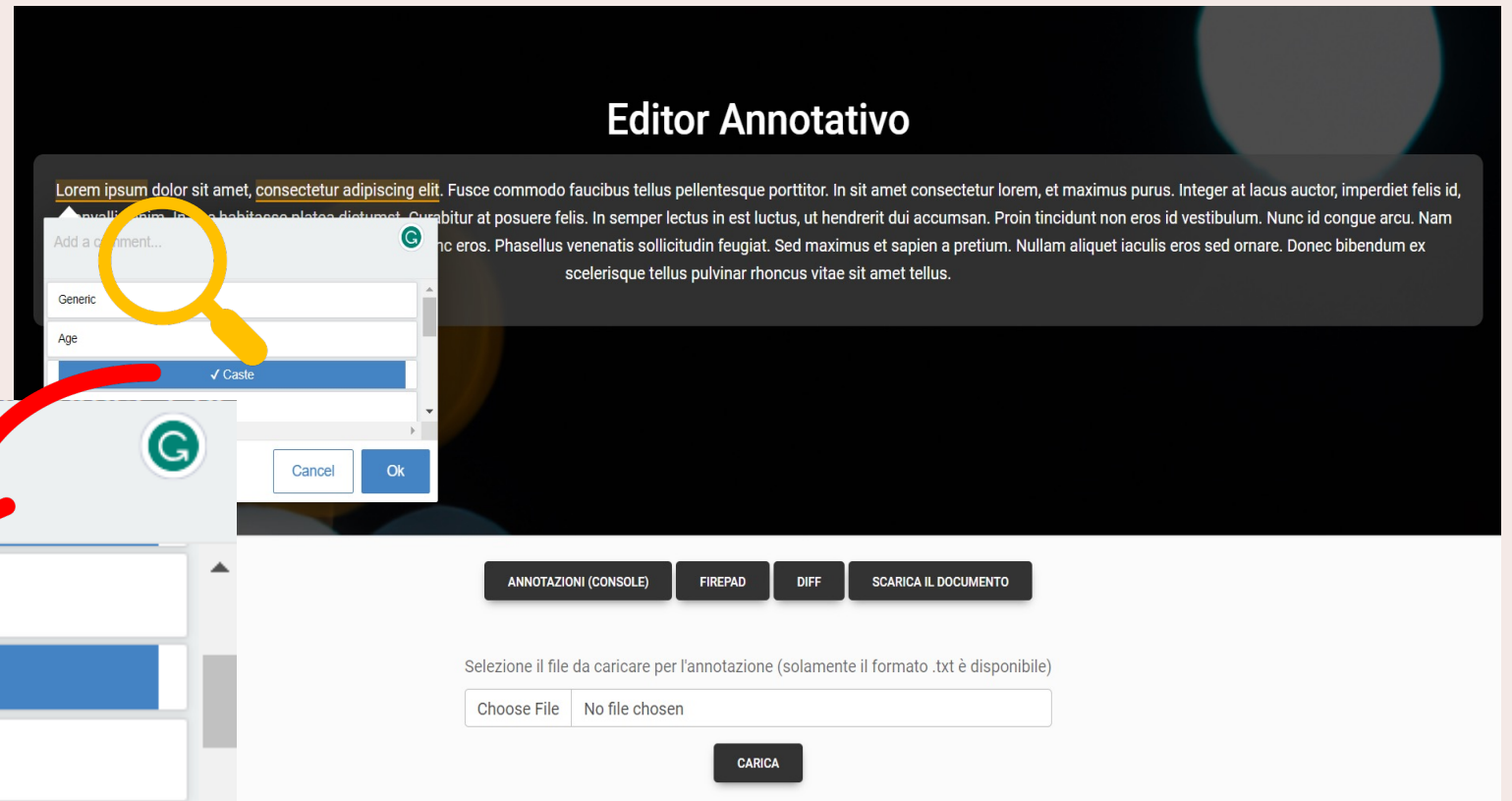
- Aim of the Project
- Main Areas of Applications
- Double Directions

- ✓ Knowledge Graph embedding techniques
- ✓ Developing a Demonstrator



# A COLLABORATIVE EDITOR

- Collaboration
- Text Annotation
- Text Editing
- Intuitive and Easy to Use



# OUR IMPLEMENTATION SO FAR!

## CLASSICAL APPROACH

Model	Accuracy	F1_score
SVM	64.00%	0.52
DT	63.20%	0.62
LR	64.80%	0.54
MNB	64.80%	0.56
<b>KNN</b>	<b>66.40%</b>	<b>0.62</b>
GB	65.20%	0.61

## DL APPROACH

```
example_text = "i am so glad i was born in the west "  
preprocessed_text = clean(example_text) # Apply preprocessing steps  
  
# Convert the preprocessed text to padded sequence  
sequence = tokenizer.texts_to_sequences([preprocessed_text])  
padded_sequence = pad_sequences(sequence, maxlen=max_sequence_length)  
  
# Load the trained model  
# Load the model architecture and weights  
model = load_model('C:/Users/Occhipinti/Desktop/PhD unimi/codes/workspace/models/modelGN+CNN+ATT.h5')  
  
# Make prediction  
prediction = model.predict(padded_sequence)  
  
predicted_label = "Hate speech" if prediction > 0.5 else "Not Hate Speech"  
  
print("Predicted Label:", predicted_label)  
  
1/1 [=====] - 0s 159ms/step  
Predicted Label: Hate speech
```

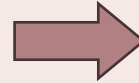
```
example_text = "Nietzsche was a fraud who couldn't account for the rate of a receding hairline"  
preprocessed_text = clean(example_text) # Apply preprocessing steps  
  
# Convert the preprocessed text to padded sequence  
sequence = tokenizer.texts_to_sequences([preprocessed_text])  
padded_sequence = pad_sequences(sequence, maxlen=max_sequence_length)  
  
# Load the trained model  
# Load the model architecture and weights  
model = load_model('C:/Users/Occhipinti/Desktop/PhD unimi/codes/workspace/models/modelGN+CNN+ATT.h5')  
  
# Make prediction  
prediction = model.predict(padded_sequence)  
predicted_label = "Hate speech" if prediction > 0.5 else "Not Hate Speech"  
  
print("Predicted Label:", predicted_label)  
  
1/1 [=====] - 0s 86ms/step  
Predicted Label: Hate speech
```

# RESEARCH QUESTIONS AND FURTHER STEPS

## What annotators do:

1. Does the text contain any form of discrimination?
2. Does the text show menace?
3. Does the text show defamation?
4. Is the text targeting a specific individual or social groups?
5. How does the text categorize individuals and social groups) by resorting to factors of discrimination?

- |                                      |   |
|--------------------------------------|---|
| <input type="checkbox"/> Age         | <input type="checkbox"/> Ethnicity          |
| <input type="checkbox"/> Caste       | <input type="checkbox"/> Gender identity    |
| <input type="checkbox"/> Sex         | <input type="checkbox"/> Religious beliefs  |
| <input type="checkbox"/> Disability  | <input type="checkbox"/> Sexual orientation |
| <input type="checkbox"/> Language    | <input type="checkbox"/> Social Classes     |
| <input type="checkbox"/> Nationality |   |



## What Machine learns:

"I am so glad I was born in the **west**."

1. Discrimination, 2. No menace, 3. No defamation, 4. General, 5. Nationality

"the **mentally ill** should be **kept away from the general population** to prevent **violence** against them"

1. Discrimination, 2. Menace, 3. No defamation, 4. General, 5. Disability

"**women** shouldn't be on screen more than 30 minutes a day"

1. Discrimination, 2. No menace, 3. No defamation, 4. General, 5. Gender identity

## Moving Forward... INTERSECTIONALITY and the Proxy

- Does the text simultaneously interlace two or more factors of discrimination?
- Which are the additional elements or individual qualities that might have a role in categorize individuals and social group?
- How the text negatively impact on individuals and social groups?
- Beyond words, what is the role of stereotypes and how machines make use of them?

**intersectionality**

[in-tər-,sek-shə-'na-lə-tē]



# WHAT DO WE AIM TO ACHIEVE?

- Data on the severity of non-inclusive language in written texts by making use of “NOME” in light of discriminatory words and stereotypical expressions
- Definition of strategies to cope with non-inclusiveness in written texts developing additional functions of NOME
- Three levels of analysis:
  - ❖ detection of discriminatory words
  - ❖ detection of stereotypical expressions
  - ❖ detection of hate speech and language forms of incitement to violence and hatred



1. How Machines  
Discriminates

BUT (also)

2. How Can AI be used to  
Tackle discrimination





THANK YOU FOR  
YOUR ATTENTION!