Unveiling Stereotypes: Combining Knowledge Graphs and LLMs for Implied Stereotype Generation

Marco Cuccarini[1,2,†], Lia Draetta[3,†], Beatrice Fiumanò[4,†], Stefano Bistarelli[2], Rossana Damiano[3] and Valentina Presutti[4]

[1]*University of Naples Federico II, Department of Biology*
[2]*University of Perugia, Department of Mathematics and Computer Science*
[3]*University of Turin, Department of Computer science*
[4]*University of Bologna, Department of Modern Languages, Literatures, and Cultures*

Abstract

In recent years, hate speech detection models have achieved significantly improved results, largely due to advances in Large Language Models (LLMs). As a result, research has increasingly focused on more nuanced phenomena, such as the detection of implicit hate and stereotypes. Although the challenge of identifying implicit language has been largely explored, it remains an open issue for state-of-the-art models due to their limited ability to grasp contextual and culturally specific knowledge. In this work, we address the task of identifying stereotypes implicitly encoded in hate speech messages, and propose a method for generating them by leveraging the combined potential of LLMs and Knowledge Graphs (KGs). As a first step, we designed an ontology specifically tailored to represent implicit hate speech. We then populated the ontology using a subset of an Italian-language hate speech dataset, in which targets and implied stereotype statements were manually annotated. The remaining portion of the dataset was reserved as a test set to evaluate the impact of knowledge graph-derived information on LLM-generated stereotypes. For each input sentence, relevant knowledge was extracted from the ontology using SPARQL queries and used to enrich the prompt provided to various LLMs. We compared the results of the knowledge-enhanced approach against those of a baseline few-shot learning approach. Evaluation was conducted using BLEU, BERTScore and ROUGE metrics. Additionally, given the high subjectivity of the task, we performed a manual qualitative analysis on a subset of the model outputs to assess both the quality of the evaluation and the soundness of the generated stereotypes.
Warning: This paper contains examples of explicitly offensive content.

Keywords

Hate speech detection, Stereotype, Large Language Models, Knowledge Graph, Retrieval Augmented Generation

1. Introduction

In recent years, the detection of Hate Speech (HS) and abusive language has gathered significant attention in the field of Natural Language Processing (NLP) [1, 2, 3], becoming a crucial tool for moderating online content and limiting the spread of harmful language. While most research has focused on explicit hate speech, implicit and subtle forms of abusive language remain underexplored [4]. Scholars [5, 6] have noted that state-of-the-art hate speech detection models struggle to identify implicit hate speech and stereotypes. This challenge arises from various factors, including specific linguistic features of HS messages (e.g. irony and metaphors) and their strong dependence on sociocultural context [4]. While some studies focus solely on the classification of content as abusive or non-abusive, others aim to uncover the subtle and implicit stereotypes embedded in such content. Recognizing the complexity of the task, recent approaches leverage external knowledge to enrich prompts in zero and few-shot learning settings, aiming to provide additional context to improve detection and analysis performance. One particularly promising method is graph-based approaches, in which knowledge retrieved from external Knowledge Graphs (KGs) is integrated into LLMs prompts aiming at enhancing the model precision. Given that LLMs often suffer from limited factual accuracy, poor memorization of structured knowledge, and hallucination tendencies [7], KG-based approaches have shown promising results across a variety of tasks [8, 9]. These approaches offer an encouraging strategy to mitigate the inherent limitations of LLMs, integrating them with structured external knowledge while preserving their generative strengths [10, 11]. Building on these premises, we propose a graph-based enrichment methodology aimed at explaining subtle stereotypes embedded in hate speech sentences. First, we design a domain-specific ontology, aligning it with foundational ontologies and existing hate speech-related

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[†] These authors contributed equally to this work.

marco.cuccarini@unina.it (M. Cuccarini); lia.draetta@unito.it (L. Draetta); beatrice.fiumano@unibo.it (B. Fiumanò); stefano.bistarelli@unipg.it (S. Bistarelli); rossana.damiano@unito.it (R. Damiano); valentina.presutti@unibo.it (V. Presutti)
0009-0007-8943-8090 (M. Cuccarini); 0009-0004-6479-5882 (L. Draetta); 0009-0007-6918-0803 (B. Fiumanò); 0000-0001-7411-9678 (S. Bistarelli); 0000-0001-9866-2843 (R. Damiano); 0000-0002-9380-5160 (V. Presutti)

resources. We then populate the ontology using a subset of an Italian dataset on implicit stereotypes, which comprises manual annotations on HS targets, hateful chunks and stereotypes. Finally, starting from the target entities in each sentence, we extract relevant knowledge from the KG and integrate it into the prompt of three different LLMs. We task the models with generating the implicit stereotype that underlies each hate speech message. We compare these stereotypes with those generated by a baseline model using a non-KG-enhanced prompt. The main contributions of this work are the following:

- StereoGraph: a Knowledge Graph grounded in a dedicated ontology designed to represent implicit hate expressed in social media posts.
- A graph-based methodology to generate explicit stereotypes encoded in hateful messages.
- A fine-grained manual assessment and error analysis to evaluate the suitability of the evaluation metrics used to compare both the baseline and KG-enhanced outputs against the gold standard. This was particularly relevant given the highly subjective and culturally specific nature of task.

In the following Section (2) we present relevant related works on detection and analysis of subtle hate speech (2.1), together with graph-based approaches (2.2) to the same tasks. Section 3 describes the adopted methodology, the dataset we used for constructing the KG, and the ontology design process. The experimental setup is detailed in Section 4, while the results, including quantitative evaluation, human assessment, and error analysis are discussed in Section 5. Finally, the conclusions and limitations are presented in Sections 6 and 7, respectively. All data and code for reproducibility can be found on the following GitHub page[1].

2. Related Works

2.1. Subtle Hate Speech Explanation

Unlike explicit hate speech, the interpretation of implicit hate speech often requires inference and integration of background knowledge [12, 13], particularly since hate expressions are usually socio-culturally dependent and rely on contextual knowledge [14]. These factors contribute to the challenge of detecting implicit hate speech and highlight the ongoing need for more sophisticated detection systems, as current state-of-the-art models still struggle to efficiently handle this task [15]. Some studies have attempted to identify subtle hate speech by leveraging different approaches

Several approaches have been explored to identify subtle hate speech, including transformer-based models [16, 17, 18], neural networks [19] or leveraging semantic information embedded in texts [19, 20]. Other approaches tried to tackle this task by incorporating the potentiality of external sources of knowledge, such as Knowledge Graphs [21].

In this context, few studies have directly addressed the challenge of unveiling or explaining subtle hate speech. Some researchers [16, 22] have focused on the role of social stereotypes, aiming to uncover their implicit meanings and to develop benchmarks for explanation-oriented tasks. Other works have specifically addressed the task of implicit hate speech explanation. Kim and colleagues [23] present a pipeline that guides transformer models' predictive decisions through the identification of key rationales. More recent studies have leveraged the generative capabilities of LLMs. For example, Huang and colleagues [24] propose a Chain-of-Explanation prompting method to generate stereotypes. Similarly, Yang et al. [25] introduce step-by-step approach that combines LLM-based chain-of-thought prompting with a human-annotated benchmark.

While several studies have focused on creating benchmarks and providing insights into implicit hate speech in English, resources for the Italian language remain limited, with only a few datasets addressing the hate speech phenomenon in depth. Notable studies [26, 27, 28, 29] have provided valuable annotated resources that distinguish between implicit and explicit hate speech and stereotypes, with the goal of detecting the more subtle and less recognizable nuances of hate. Nevertheless, research on stereotype explication remains limited. For example, Muti and colleagues [30] investigate the ability of LLMs to accurately identify implicit messages in misogynistic contexts, also exploring how prompts can reconstruct subtle meanings to make the messages explicit. However, to our knowledge, no previous work about embedded stereotypes has been carried out in the Italian cultural context. We suggest that the generation of implicit stereotypes can support the development of more comprehensive benchmarks, improving models' performance in detecting subtle forms of hate speech.

2.2. Knowledge-Enhanced Approaches

Knowledge-enhanced and Retrieval-Augmented Generation (RAG) methods [31] have emerged as a powerful paradigm to address key limitations of LLMs. More recently, this line of work has incorporated structured, graph-based knowledge, particularly KGs [8], to enhance retrieval and reasoning capabilities.

In the domain of hate speech research, knowledge-enhanced approaches have provided solutions to address the challenges posed by implicit hate speech across vari-

[1] https://github.com/marcocuccarini/StereoGraphUnveilingStereotypes

ous tasks.

Zhao et al. [21] propose MetaTox, a RAG-based approach that integrates a meta-toxic knowledge graph with LLMs for hate speech detection. First, LLMs are used to construct the KG by combining data from three English datasets. Then Qwen and LLaMA3.1 are prompted to classify tweets as toxic or non-toxic. The authors demonstrate that the MetaTox method enables to reduce false positives, leading to better generalization and reduced hallucinations from LLMs. Lin [13] combines Entity Linking techniques with summarized Wikipedia descriptions to improve performances in implicit hate speech detection and classification task. Although it does not follow a standard RAG approach, the paper proposes feeding a Multi-Layer Perceptron with embeddings of concatenated tweet and external knowledge representations, training it to perform a multi-label classification of implicit hate speech types. This approach demonstrated significant improvements when entity triggers were mentioned in text, although limitations remained for the classification of tweets requiring pragmatic understanding.

In the context of implicit hate speech, Yadav et al. [32] introduce Tox-BART, a BART-based architecture enhanced with toxicity attributes, i.e. structured meta-information on tweets, encompassing target groups, insult types, and hate intensity levels. This approach addresses limitations derived from poor quality of retrieved KG tuples, which can hinder KG-augmented approaches. Using different evaluation metrics, they demonstrate that infusion of toxicity attributes achieves performance comparable to simple KG-infusion. In the Italian context, Di Bonaventura and colleagues [33] implemented a knowledge-enhanced approach for detecting homotransphobic hate speech. The system leverages the O-Dang knowledge graph, which contains information about named entities in the Italian HT context. The approach showed promising results, outperforming baseline scores.

Compared to the reviewed literature, our approach represents a step forward, particularly in the area of Italian language hate speech detection. While most prior work has focused on the detection of implicit hate speech, our study shifts the emphasis toward the explanatory capabilities of LLMs, specifically investigating how these can be enhanced through the integration of structured knowledge. Furthermore, by focusing on stereotypes and adopting and hybrid evaluation approach (automatic and human-based), our work also provides valuable insights into the ability of LLMs to uncover sound and coherent stereotypes from implicit language, as well as into the reliability of the evaluation metrics used.

3. Methodology

In this work, we aim to perform the task of implicit stereotype generation using LLMs, comparing a baseline approach with a KG-enhanced alternative. Given a sentence and its associated hate speech target, the model is prompted to generate the subtle stereotype that contributes to the message's hateful nature. In the following sections, we briefly present the proposed pipeline (Section 3.1), describe the dataset used (Section 3.2), and outline the construction of the ontology that serves as the foundation for the knowledge graph (Section 3.3).

3.1. Pipeline Overview

Our methodology is designed to make subtle stereotypes conveyed in hateful content explicit. This is a particularly challenging task, as it requires nuanced contextual understanding and awareness of culturally specific stereotypes associated with the target. By integrating external knowledge, we investigate whether language models can effectively contextualize such messages and generate more accurate and transparent stereotypes.

The proposed approach is illustrated in Figure 1. Given an input sentence and its associated HS target, retrieved from the annotated dataset, we use the target to query the KG via a SPARQL query, retrieving all triples in which target is linked to its stereotypes. We then adopt a few-shot learning approach, integrating into the prompt the external knowledge retrieved from the KG in RDF format. The evaluation phase consists of a comparison between the results (i.e. generated stereotypes) obtained using the knowledge-enhanced and the baseline approach. A hybrid evaluation was performed comparing automatic metrics with human assessment.

3.2. Dataset

To address the task of subtle stereotype generation, we leveraged the Open Stereotype Corpus[2] [34] containing 3,578 Italian tweets collected between October 2018 and June 2019 from the *Contro l'Odio* dataset [35]. The dataset was annotated by five different annotators. For each message, the annotators identified the specific chunk (trigger) containing the hate content, the implicit stereotype (if present) and the stereotype cluster (a more general class aiming at creating a stereotype categorization). In the original dataset the authors automatically distinguished between agent and patient parsing each rationale, we chose to simplify this distinction aggregating the two columns under a unique class named "target". An example of the dataset structure along with a subset of annotations is presented in Figure 3. From the dataset

[2] https://github.com/SodaMaremLo/Open-Stereotype-corpus

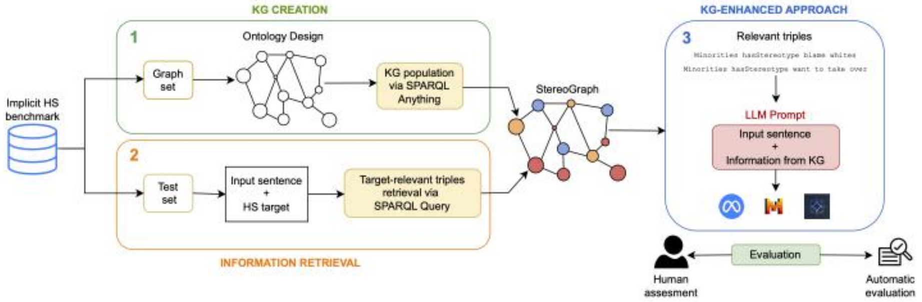
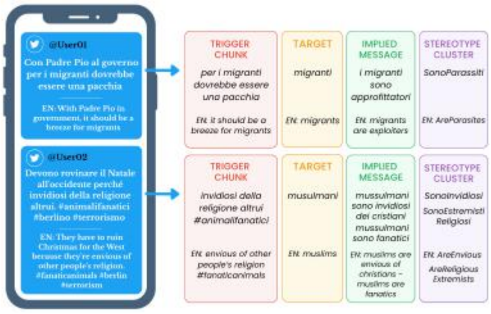

KG CREATION
1
Ontology Design
Graph set
KG population via SPARQL Anything
Implicit HS benchmark
2
Test set
Input sentence + HS target
Target-relevant triples retrieval via SPARQL Query
INFORMATION RETRIEVAL
StereoGraph
KG-ENHANCED APPROACH
3
Relevant triples
Minorities hasStereotype blame whites
Minorities hasStereotype want to take over
LLM Prompt
Input sentence + Information from KG
Evaluation
Human assesment
Automatic evaluation

Figure 1: Stereotype extraction pipeline. The dataset is split into a graph and test set. The graph set is used to populate the StereoGraph KG. Inputs from the test set are used to evaluate the approach: after identifying the HS target, SPARQL queries are used to retrieve target-relevant triples, which are incorporated in the prompt. The LLM is tasked to generate the sentence's underlying stereotype, evaluated against the gold standard using automatic metrics and human assessment.


@User01
Con Padre Pio al governo per i migranti dovrebbe essere una pacchia
EN: With Padre Pio in government, it should be a breeze for migrants

@User02
Devono rovinare il Natale all'occidente perché invidiosi della religione altrui. #animalifanatici #berlino #terrorismo
EN: They have to ruin Christmas for the West because they're envious of other people's religion. #fanaticanimals #berlin #terrorism

TRIGGER CHUNK	TARGET	IMPLIED MESSAGE	STEREOTYPE CLUSTER
per i migranti dovrebbe essere una pacchia EN: it should be a breeze for migrants	migranti EN: migrants	i migranti sono approfittatori EN: migrants are exploiters	SonoParassiti EN: AreParasites
invidiosi della religione altrui #animalifanatici EN: envious of other people's religion #fanaticanimals	musulmani EN: muslims	mussulmani sono invidiosi dei cristiani mussulmani sono fanatici EN: muslims are envious of christians - muslims are fanatics	SonoInvidiosi SonoEstremisti Religiosi EN: AreEnvious AreReligious Extremists

Figure 2: Overview of the dataset annotation structure.

we selected only the messages in which or a stereotype or hate speech was present.

3.3. Ontology Design

For the ontology design process we adopted a fully manual approach to ensure the quality of the resulting resource through several means: aligning it with foundational ontologies and related semantic resources, ensuring the conceptual correctness of the defined classes, and minimizing the potential introduction of bias. The ontology includes four top-level classes: `Situation`, `Stereotype`, `Agent`, and `Type`. The class `Situation` is aligned with the homonymous class from the foundational ontology DOLCE [36]. Its purpose is to link a given target and its associated stereotype to a specific occurrence, such as a Twitter post, in order to avoid the introduction of bias or overly generic statements about stereotypes. The class `Stereotype` captures the implicit assertions conveyed in a given sentence. The class `Agent`, aligned with the FOAF (Friend of a Friend) ontology[3], has two subclasses: `Group` and `Person`. These subclasses represent different types of targets and are connected to specific situations via the `hasTarget` relation, which links a message to its corresponding target. The class `Type` is designed to provide a taxonomy for both targets (e.g., racial target, religious target) and stereotypes (e.g., 'are dangerous', 'are unclean'). The ontology was subsequently populated using `SPARQLAnything`[4] [37] leveraging the datasets described in the previous section as data source. After this process we obtained a knowledge graph containing triples as to the followings:

```
ster:_803176483174780929
rdf:type   dul:Situation ;
rdfs:label    "Forza ragazzi, 180mila clandestini all
       anno, rom da tutte le parti, illegalita totale,
       Coop rosse e bianche che lucrano. ora sapete
      cosa votare" ;
dul:hasTarget   ster:immigrati ;
ster:hasStereoManifestation  ster:180mila-clandestini
      -allanno ;
ster:hasStereotype ster:invadendo-italia .

ster:invadendo-italia
        rdf:type      ster:Stereotype ;
        rdfs:label    "invadendo italia" ;
        ster:hasType  "SonoInvasori" .

ster:immigrati  rdf:type  foaf:Group .
```

This means that a specific post, identified by the ID `ster:_803176483174780929`, is an instance of the class `Situation`. It has a specific content, expressed trough the relation `rdfs:label`, and it is associated to a specific stereotype chunk trough the relation `ster:hasStereoManifestation`. The tweet is then associated with a particular target, `ster:immigrati`, as well as a stereotype, `ster:invadendoitalia`. The stereotype is then defined as an instance of the class

[3] http://xmlns.com/foaf/spec/

[4] https://sparql-anything.cc/

ster:Stereotype and linked to a specific cluster SonoInvasori through the relation ster:hasType.

4. Experiment Setting

In the next sections, the experimental setting is presented. The following approach consists of three main steps: Knowledge retrieval, where relevant information is retrieved from the KG (Section 4.1); Prompting, where three models are prompted using both a few-shot baseline and a few-shot KG-enriched approach 4.2; and Evaluation (4.3), where the results are assessed using both automatic metrics and manual evaluation.

4.1. Knowledge Extraction

For every sentence of the test set we extracted relevant knowledge from the Knowledge graph leveraging the following SPARQL query:

```
SELECT ?s ?stereotype
    WHERE {{
      ?s a dul:Situation ;
         dul:hasTarget <{target_uri}> ;
         ster:hasStereotype ?stereotype .
    }}
```

Using this query we were able to retrieve all the stereotype associated with a certain tweet that has the specified target. For example using "immigrati" as target we are able to extract triples like the followings, in which the first element is the ID, the second the gold stereotype and the third hateful span:

```
ster:_id sono-irregolari clandestini-musulmani
ster:_id non-rispettano-legge nn-amano-subire-le
     -nostre-leggi-sti-migranti
ster:_id spacciano immigrati-spacciatori-e-
     stupratori
```

Since our goal is to prove that this integrated information could improve implicit stereotype generation, we rely on the gold-standard targets provided in the dataset. This avoids the noise introduced by potential errors in target prediction. One limitation encountered is the over-representation of certain targets, which appear with a high number of samples. To reduce the impact of the “lost in the middle” phenomenon [38] and to balance the quantity of information, we randomly sample 20 stereotypes per target.

4.2. Prompt Construction

We decided to test three different models LLaMA-3.1-8B, gemma-2-9b-it [39] and Mistral-7B-Instruct-v0.2 [40] to explore their ability to understand the subtle stereotype embedded in the message. We selected these three distinct LLMs because they are state-of-the-art, multilingual, open-source models with comparable architecture and medium scale size.

The task is conducted in the Italian language. For the baseline, we used a few-shot learning approach and for the prompt construction we adopt a vanilla structure setup; the prompt is written in Italian. Additionally, it includes instructions on how to structure the output sentence, explicitly asking the models to generate output in the format [subject] [are/do] [predicate]. The knowledge-enhanced approach incorporates a prompt containing information about the target entity from the KG. For each target, we associate the relevant retrieved stereotypes. The full prompt is presented in Appendix A. The output produced by the LLM was preprocessed before the evaluation, removing generic elements provided by the LLM, such as the usual formulaic closing statements (e.g., asking if it can assist further).

4.3. Evaluation

For the evaluation phase, we leverage BLEU [41], BERTScore [42] and ROUGE [43]. BLEU measures how many n-grams in the generated text appear in the reference text, focusing on precision and penalizing very short outputs. ROUGE focuses on recall, checking how much of the n-grams or sequences of the reference text appear in the generated text, often used for summarization. BERTScore compares the generated and reference texts using deep contextual embeddings from BERT, capturing semantic similarity beyond exact word matches.

Since recent studies [44, 45] have highlighted the limitations of automated evaluation methods and some scholars [46, 47] are beginning to emphasize the potential of hybrid approaches and aware of the fact that stereotypes are characterized by high subjectivity and culture-specific variation, we conducted also a human-based evaluation to better understand the reliability of the metrics used in such contexts. To this end, we designed a twofold manual assessment.

First, aiming at replicating the automatic evaluation, we compared the generated (baseline and KG-enhanced) outputs with the gold standard, assessing semantic and syntactic similarity on a scale from 1 to 5. The annotators were presented with the following questions: (Q1) To what extent, on a scale from 1 to 5, is the baseline output semantically and syntactically comparable to the gold standard? (Q2) To what extent, on a scale from 1 to 5, is the KG-enhanced output semantically and syntactically comparable to the gold standard? Then, aware that stereotypes are not universally shared, and that some are more prevalent or culturally dependent than others [48], we aimed to evaluate the extent to which the generated

Table 1
Examples the human evaluation conducted on the test set. Q1, Q2 and Q3 refer to the question that annotators answered during the evaluation phase

Gold	Baseline	KG	Annotator	Q1 [1-5] ↑	Q2 [1-5] ↑	Q3 [Y/N]
I rom sono truffatori	I rom sono falsi invalidi per commuovere.	rom_sinti sono ladri	Ann_1	3	3	NO
			Ann_2	3	3	NO
			Ann_3	4	4	SI
i migranti non sono profughi	gli avvocati pagano i migranti.	gli immigrati sono criminali	Ann_1	1	1	NO
			Ann_2	1	4	NO
			Ann_3	1	2	NO
i migranti sono criminali	gli immigrati sono violenti	gli immigrati delinquono	Ann_1	3	3	SI
			Ann_2	3	5	NO
			Ann_3	4	5	SI

stereotype might be culturally recognizable from our own perspective as white Italian researchers aged between 25 and 30. The evaluation of generated stereotypes was conducted only on content produced by the baseline model, as the KG-enhanced method provides the model with additional contextually relevant information. Annotators were asked to assess whether, in their own perspective, the generated stereotype reflects commonly held beliefs or societal biases (Q3). For example, the stereotype "gli avvocati pagano i migranti" ("Lawyers pay the migrants") was judged unrealistic by all three annotators. In contrast, "gli immigrati delinquono" ("Immigrants commit crimes") received two positive evaluations out of three, suggesting that this stereotype may reflect a commonly held bias in the Italian context. The human evaluation was conducted by three annotators on a subset of 50 sentences. An example of the conducted manual evaluation is presented in Table 1.

5. Results

In the next sections the experiment results are provided. While automated methods are efficient, they often lack precision. In contrast, human evaluation offers greater contextual understanding but is time-consuming and costly. To balance accuracy and efficiency, we applied an automatic method to the full dataset and selected a smaller subset for manual evaluation.

5.1. Computer-Based Analysis

In the Table 2 are presented the result of the generation task comparing the three models across the two approaches, i.e. baseline *versus* knowledge graph enhanced. The Results shows that adding the information from KG improves the performance of all three models, LLaMA3.1, gemma2, and Mistral7B, across BLEU, Rouge, and BERT-based scores. Gemma2 benefits the most, with its BLEU score more than doubling and a big gain in Rouge. LLaMA3.1 and Mistral7B also show consistent, though smaller, improvements. The BERT-based scores indicate better semantic relevance with KG. Overall, the KG helps the models produce more accurate and meaningful results.

Table 2
Baseline vs KG-enhanced evaluation scores

Model	Method	BLEU ↑	ROUGE ↑	BERT-based ↑
Gemma 2	Baseline	0.029	0.142	0.521
	KG	0.061	0.253	0.596
LLaMA 3.1	Baseline	0.071	0.264	0.571
	KG	0.076	0.298	0.618
Mistral 7B	Baseline	0.077	0.301	0.573
	KG	0.080	0.302	0.608

5.2. Human-based Analysis

The annotators were provided with answers from both the baseline and the KG-enhanced method. Each answer was evaluated on the basis of its similarity to the gold standard, the normalized results are presented in Table 3. Furthermore, for the baseline generation only, annotators were asked to assess whether the stereotypes reflect commonly held beliefs or communal biases. LLaMA 3.1 the highest average scores for both baseline and KG-enhanced outputs, demonstrating strong overall performances. Gemma 2 shows lower results across all metrics, while Mistral7B performs the lowest on both baseline and KG averages. Human evaluation further confirms that incorporating knowledge from the graph improves model performance across all models and annotators. In addition, the variation in annotators' scores highlights the subjective nature of the task and the challenge of achieving consistent judgments. Annotator 2, for example, generally rates outputs higher, particularly for KG-enhanced responses, while Annotator 3 is more critical. Human-evaluated results confirm the trends observed in computer-based scores (for all the models and the annotators the score are higher in the case of the KG-enhanced approach), demonstrating how our method improves the model's ability to explicitly address implicit hate speech and suggesting that automatic measures can be informative for this type of task.

Regarding the assessment of the generated stereotype the human evaluation reveals divergence tendency: LLaMA shows the average highest scores across the three annotators, and the value appears to be high especially according to Annotators 1 and 3. Gemma2 shows a similar tendency, especially regarding the annotators 2 and 3. Finally, Mistral tends to have an overall lower score about the stereotypes soundness, suggesting that it may produce less biased or not realistic content.

5.3. Human-based vs Computer-based metric

To better understand the relationship between automatic metrics and human judgment, we compared the results of BLEU, ROUGE and BERT Score with human evaluation over a sample of 50 sentences, as seen in Figure 3. The three plots help identify which metric aligns more closely with human evaluation.

From the plots, it is evident that the BERT Score metric (shown in the third plot) correlates more consistently with the annotators' evaluation, suggesting it is a more reliable indicator of quality for this task. This is due to the nature of BERT score, which leverages contextual embeddings to measure similarity on a semantic level. Conversely, BLEU and ROUGE metrics (depicted in the first and second plots, respectively), which operate more on the lexical-syntactic level, show more variability and several limitations in accurately matching human judgment.

Understanding the relationship between automatic and manual assessment is crucial for contextualizing the values obtained from each metric and evaluating model performance in a meaningful way. The comparison also helps to understand which metrics are more robust and reliable, especially for tasks requiring deep contextual and pragmatic understanding.

5.4. Error Analysis

To gain deeper insight into the functionality and limitations of our approach, and to identify areas for potential future improvements, we conducted an error analysis on the tweets where the KG-enhanced method showed the lowest performance. Overall we observed that errors frequently occurred when the input contained named entities or subjects that differed from the primary target. For example, in the tweet:

> Finanzia l'invasione degli immigrati: ecco la prova. La vergogna di George Soros, "padrone" d'Italia.
> English: "He funds the immigrant invasion: here is the proof. The shame of George Soros, the 'master' of Italy."

the KG-enhanced output was: "George Soros finanzia l'invasione degli immigrati" (English: "George Soros funds the immigrant invasion"), while not conceptually incorrect, this differs from the gold standard:"i migranti vogliono invadere l'Italia" (English : "The migrants want to invade Italy."). A similar issue occurred in the tweet:

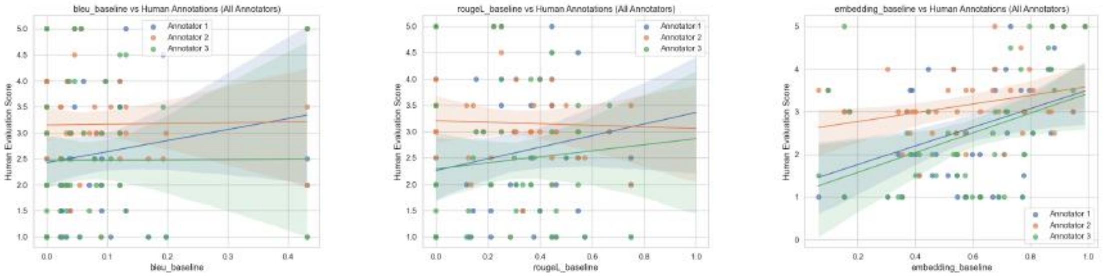
> Che senso ha ministro Trenta rispettare chi non rispetta noi? Che senso ha difendere la loro cultura o presunta cultura quando essi disprezzano la nostra? La ministra Trenta contro Salvini: sbagliato dire che l'Islam è terrorismo
> English: What's the point, minister Trenta, of respecting thos who don't respect us? What's the point of defending their culture or so-called culture when they despise ours? Minister Trenta against Salvini: it's wrong to say that Islam is terrorism"

The KG-enhanced output was "la ministra Trenta disprezza la cultura italiana." (English: Minister Trenta despises Italian culture.) whereas the gold standard was: "i musulmani vanno contro i valori dell'Occidente" (English: Muslims go against Western values). In other cases, when the model encounters a target associated with a high number of stereotypes, it tends to concatenate many of them into a generic and incoherent output.

In some cases, both the baseline and the KG-enhanced approaches struggle to recognize irony and fail to produce a reliable underlying stereotype. For example, consider the following sentence:

> #Dimartedi Stasera indottrinamento pro Europa. Alla bisogna sono benvenuti anche gli stranieri. Bravo #Floris, vai a cager English: #dimartedi tonight: pro-Europe indoctrination. If needed, even foreigners are welcome. Well done #Floris, go to hell.

Both the baseline and the KG-enhanced approaches generate the "gli stranieri sono benvenuti" (English: Immigrants are welcome), failing to detect the subtle irony in the original message.



(a) BLEU scores compared to all annotators

(b) ROUGE-L scores compared to all annotators

(c) BERTScore compared to all annotators

Figure 3: Overview of the Italian dataset annotation structure with comparisons of three metrics—BLEU, ROUGE, and BERT Score—against human annotators for the Llame model and the baseline.

Table 3
Baseline vs proposed method human-base score. Q1, Q2 and Q3 refer to the three questions presented to the annotators.

Model	Metric	Annotator 1	Annotator 2	Annotator 3
Gemma 2	(Q1) Baseline Average	0.291	0.444	0.327
	(Q2) KG Average	0.378	0.561	0.362
	(Q3) Stereotype %	0.396	0.449	0.673
LLaMA 3.1	(Q1) Baseline Average	0.332	0.434	0.332
	(Q2) KG Average	0.469	0.648	0.411
	(Q3) Stereotype %	0.588	0.469	0.673
Mistral 7B	(Q1) Baseline Average	0.270	0.316	0.321
	(Q2) KG Average	0.357	0.622	0.449
	(Q3) Stereotype %	0.367	0.286	0.449

Finally, we observed challenges in tweets with complex hypotactic structures and multiple subjects. In such cases, models often fail to correctly identify the primary target and to produce relevant output. Furthermore, the KG-enhanced method tends to generate overly long responses in these situations, which can reduce the coherence and precision of the generated content. In summary, the worst-performing examples often occur because the model misidentifies the target of the hate tweet, leading to reduced accuracy. However, in many cases, the model still manages to extract a correct implicit message, which, while different from the gold standard, is present in the tweet. In such cases, the prediction is valid, but the reference annotation fails to recognize it as correct.

6. Conclusion

In this work, we aim to investigate whether large language models are able to uncover implicit stereotypes embedded in hate speech messages. This task is important as it helps uncover the subtle content of hate speech messages and supports hate speech detection models in identifying abusive language. Specifically, we explore the role that additional information from a knowledge graph may play in the understanding and generation of underlying stereotypes. We compare a baseline few-shot approach with a knowledge-enhanced method, leveraging different LLMs. We observed that prompts enhanced with additional information outperformed the baseline approach. To better assess the reliability of the automatic evaluation metrics, we also conducted a manual evaluation, replicating the task performed by the automatic metrics. The human evaluation confirmed the results, showing higher scores for the knowledge graph-enhanced approach. While the manual assessment was aligned with the automated results, we observed a high degree of variability in the scores. This suggests that evaluating such generated content is inherently subjective and can vary based on the annotators' culture, age, or beliefs. These findings highlight the importance of contextualizing evaluation metrics and recognizing that they may carry biases or oversimplify complex phenomena. From the error analysis, we observed that the KG-enhanced approach occasionally struggles to manage the quantity of information provided, suggesting that further studies are needed to better understand the extent to which such models can effectively integrate additional knowledge.

To sum up, the findings of this research suggest that knowledge graph-based approaches are highly promising, even in the hate speech domain, where they remain largely underexplored.

7. Limitation and Future Work

In this work we focused on the integration of stereotypes, retrieving targets from the gold standard. This allows us to concentrate the analysis on the knowledge insertion process within the LLM, minimizing the introduction of noise. As future work, we intend to test the approach using a state-of-the-art target detection model. Although this may introduce errors due to target misclassifications, it would enable full autonomy for the proposed method and enhance its applicability in real-world scenarios. Target detection methods can also return multiple potential targets in cases of uncertainty, providing a fuller stereotype context for posts that may involve more than one target. While we noticed that different stereotypes are associated to the same target, as a future work we may consider an approach based on semantic similarity to select the most contextually relevant stereotypes. This approach could offer a more focused context for the prompt and reduce the likelihood of model misunderstandings. During the error analysis phase, we identified errors potentially caused by the 'lost-in-the-middle' phenomenon. Future work should explore in greater depth how models manage different quantities of input information. Finally, it is important to highlight that the manual evaluation we conducted—particularly regarding the cultural shareability of the generated stereotypes, is inherently biased and reflects the perspectives of the researchers involved in this study. As future work, it would be interesting to carry out a large-scale, prospectivist survey to explore the diversity of opinions on stereotypes and to investigate the dominant worldview conveyed by different large language models.

Ethical Considerations

We acknowledge that when dealing with hate speech, particularly stereotypes targeting minorities, it is essential to be mindful of the potential of introducing bias or unintentionally amplifying hateful content. We made efforts to control and reduce the presence of bias and to remain aware of its potential introduction. During the experimental phase, we prompted LLMs to generate implied stereotypes, which in some cases resulted in the generation of hateful or offensive content. The generated hateful content is intended solely to remain within the context of this experimental research. Its occurrence also provides additional insights into how LLMs can produce harmful language despite safety filters.

References

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[2] J. S. Malik, H. Qiao, G. Pang, A. van den Hengel, Deep learning for hate speech detection: a comparative study, International Journal of Data Science and Analytics (2024) 1–16.

[3] D. Nozza, F. Bianchi, G. Attanasio, Hate-ita: Hate speech detection in italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 2022, pp. 252–260.

[4] N. B. Ocampo, E. Sviridova, E. Cabrio, S. Villata, An in-depth analysis of implicit and subtle hate speech messages, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1997–2013. URL: https://aclanthology.org/2023.eacl-main.147/. doi:10.18653/v1/2023.eacl-main.147.

[5] J. Mun, E. Allaway, A. Yerukola, L. Vianna, S.-J. Leslie, M. Sap, Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9759–9777. URL: https://aclanthology.org/2023.findings-emnlp.653/. doi:10.18653/v1/2023.findings-emnlp.653.

[6] Y. Zhang, S. Nanduri, L. Jiang, T. Wu, M. Sap, BiasX: "thinking slow" in toxic content moderation with explanations of implied social biases, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4920–4932. URL: https://aclanthology.org/2023.emnlp-main.300/. doi:10.18653/v1/2023.emnlp-main.300.

[7] M. Bombieri, P. Fiorini, S. P. Ponzetto, M. Rospocher, Do llms dream of ontologies?, ACM Trans. Intell. Syst. Technol. (2025). URL: https://doi.org/10.1145/3725852. doi:10.1145/3725852.

[8] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag ap-

proach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).

[9] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-augmented generation with knowledge graphs for customer service question answering, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2905–2909.

[10] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, T. Luong, FreshLLMs: Refreshing large language models with search engine augmentation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13697–13720. URL: https://aclanthology.org/2024.findings-acl.813/. doi:10.18653/v1/2024.findings-acl.813.

[11] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).

[12] M. Dadvar, D. Trieschnigg, R. Ordelman, F. De Jong, Improving cyberbullying detection with user context, in: European conference on information retrieval, Springer, 2013, pp. 693–696.

[13] J. Lin, Leveraging world knowledge in implicit hate speech detection, in: L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, J. Zhao (Eds.), Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 31–39. URL: https://aclanthology.org/2022.nlp4pi-1.4/. doi:10.18653/v1/2022.nlp4pi-1.4.

[14] N. Lee, C. Jung, J. Myung, J. Jin, J. Camacho-Collados, J. Kim, A. Oh, Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis, 2024. URL: https://arxiv.org/abs/2308.16705. arXiv:2308.16705.

[15] A. Albladi, M. Islam, A. Das, M. Bigonah, Z. Zhang, F. Jamshidi, M. Rahgouy, N. Raychawdhary, D. Marghitu, C. Seals, Hate speech detection using large language models: A comprehensive review, IEEE Access (2025).

[16] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 345–363. URL: https://aclanthology.org/2021.emnlp-main.29.

[17] M. S. Jahan, M. Oussalah, D. R. Beddia, N. Arhab, et al., A comprehensive study on nlp data augmentation for hate speech detection: Legacy methods, bert, and llms, arXiv preprint arXiv:2404.00303 (2024).

[18] M. Zhang, J. He, T. Ji, C.-T. Lu, Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12073–12086. URL: https://aclanthology.org/2024.acl-long.652/. doi:10.18653/v1/2024.acl-long.652.

[19] S. Ghosh, M. Suri, P. Chiniya, U. Tyagi, S. Kumar, D. Manocha, Cosyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6159–6173.

[20] H. Ahn, Y. Kim, J. Kim, Y.-S. Han, SharedCon: Implicit hate speech detection using shared semantics, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 10444–10455. URL: https://aclanthology.org/2024.findings-acl.622/. doi:10.18653/v1/2024.findings-acl.622.

[21] Y. Zhao, J. Zhu, C. Xu, X. Li, Enhancing llm-based hatred and toxicity detection with meta-toxic knowledge graph, 2024. URL: https://arxiv.org/abs/2412.15268. arXiv:2412.15268.

[22] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: https://aclanthology.org/2020.acl-main.486/. doi:10.18653/v1/2020.acl-main.486.

[23] J. Kim, B. Lee, K.-A. Sohn, Why is it hate speech? masked rationale prediction for explainable hate speech detection, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6644–6655. URL: https://aclanthology.org/2022.coling-1.577/.

[24] F. Huang, H. Kwak, J. An, Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23, ACM, 2023, p. 90–93. URL: http://dx.doi.org/10.1145/3543873.3587320. doi:10.1145/3543873.3587320.

[25] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, S.-Y. Yun, HARE: Explainable hate speech detection with step-by-step reasoning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5490–5505. URL: https://aclanthology.org/2023.findings-emnlp.365/. doi:10.18653/v1/2023.findings-emnlp.365.

[26] V. Tonini, S. Frenda, M. A. Stranisci, V. Patti, How do we counter dangerous speech in italy?, in: CEUR Workshop Proceedings, volume 3878, CEUR-WS, 2024, p. 103.

[27] W. W. Schmeisser-Nieto, G. Ricci, S. Frenda, M. Taulé, C. Bosco, Implicit stereotypes: A corpus-based study for italian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024, pp. 997–1004.

[28] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, C. Bosco, et al., Hate speech annotation: Analysis of an italian twitter corpus, in: Ceur workshop proceedings, volume 2006, CEUR-WS, 2017, pp. 1–6.

[29] B. Cristina, P. Marinella, F. Benamara, C. P. Giovanni, P. Viviana, M. Véronique, T. Mariona, et al., Sterheotypes project. detecting and countering ethnic stereotypes emerging from italian, spanish and french racial hoaxes, in: Proceedings of the Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (SEPLN-CEDI-PD 2024), 2024.

[30] A. Muti, F. Ruggeri, K. A. Khatib, A. Barrón-Cedeño, T. Caselli, Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21091–21107. URL: https://aclanthology.org/2024.emnlp-main.1174/. doi:10.18653/v1/2024.emnlp-main.1174.

[31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[32] N. Yadav, S. Masud, V. Goyal, M. S. Akhtar, T. Chakraborty, Tox-BART: Leveraging toxicity attributes for explanation generation of implicit hate speech, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13967–13983. URL: https://aclanthology.org/2024.findings-acl.831/. doi:10.18653/v1/2024.findings-acl.831.

[33] C. Di Bonaventura, A. Muti, M. A. Stranisci, O-dang at hodi and haspeede3: A knowledge-enhanced approach to homotransphobia and hate speech detection in italian, in: CEUR Workshop Proceedings, volume 3473, CEUR-WS, 2023.

[34] S. M. Lo, M. A. Stranisci, A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, E. Jezek, V. Patti, Subjectivity in stereotypes against migrants in italian: An experimental annotation procedure, in: Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025.

[35] A. Capozzi, M. LAI, V. Basile, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. F. RUFFO, C. Musto, M. Polignano, et al., Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project, in: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, 2019.

[36] S. Borgo, R. Ferrario, A. Gangemi, N. Guarino, C. Masolo, D. Porello, E. M. Sanfilippo, L. Vieu, Dolce: A descriptive ontology for linguistic and cognitive engineering, Applied ontology 17 (2022) 45–69.

[37] L. Asprino, E. Daga, A. Gangemi, P. Mulholland, Knowledge graph construction with a façade: a unified method to access heterogeneous data sources on the web, ACM Transactions on Internet Technology 23 (2023) 1–31.

[38] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. URL: https://aclanthology.org/2024.tacl-1.9/. doi:10.1162/tacl_a_00638.

[39] Gemma Team, Gemma (2024). URL: https://www.kaggle.com/m/3301. doi:10.34740/KAGGLE/M/3301.

[40] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[41] K. Papineni, S. Roukos, T. Ward, W. jing Zhu, Bleu: a method for automatic evaluation of machine trans-

lation, 2002, pp. 311–318.

[42] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[43] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[44] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, arXiv preprint arXiv:2408.08921 (2024).

[45] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, Z. Liu, Evaluation of retrieval-augmented generation: A survey, in: CCF Conference on Big Data, Springer, 2024, pp. 102–120.

[46] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Knowledge graph validation by integrating llms and human-in-the-loop, Information Processing & Management 62 (2025) 104145.

[47] G. Faggioli, L. Dietz, C. L. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al., Perspectives on large language models for relevance judgment, in: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, 2023, pp. 39–50.

[48] S. T. Fiske, Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion), Perspectives on psychological science 12 (2017) 791–799.

A. Appendix

(**baseline:** context = ∅, **KG-method:** context = [[*target*, *hasStereotype*, *stereotype*],...])

```
// Task:
Given a social media post, extract the implicit message.
Output ONLY the implicit message using the format:
[subject] [are/do] [predicate].

// Rules:
- do not add explanations
- do not use quotation marks
- do not mention the author of the post
- extract only ONE implicit message per post

// Examples: post = "a roma nuovi campi nomadi e gli italiani abbandonano le
abitazioni. e noi abbiamo gente come la boldrini che difende tutto questo."
implicit_message = i rom sono pericolosi

post = "Papampero a San Vittore e festa del Pd? Te credo! I detenuti votano
tutti a sinistra, pure rom e islamici."
implicit_message = la sinistra protegge i criminali stranieri

// Task: Extract from the following post

post = "Prima gli italiani: gli immigrati non devono avere più diritti di
noi"
context = {}
implicit_message = {}
```