# A *Grice-ful* Examination of Offensive Language: Using NLP Methods to Assess the Co-operative Principle

**Katerina Korre** [a,*], **Federico Ruggeri** [b] **and Alberto Barrón-Cedeño** [a]

[a]DIT, University of Bologna
[b]DISI, University of Bologna

**Abstract.** Natural Language Processing (NLP) can provide tools for analyzing specific intricate language phenomena, such as offensiveness in language. In this study, we employ methods from pragmatics, more specifically Gricean theory, as well as NLP techniques, to analyze instances of online offensive language. We present a comparative analysis between offensive and non-offensive instances with regard to the degree to which the 4 Gricean Maxims (Quality, Quantity, Manner, and Relevance) are flouted or violated. To facilitate our analysis, we employ NLP tools to filter the instances and proceed to a more thorough qualitative analysis. Our findings reveal that offensive and non-offensive speech do not differ significantly when we evaluate with metrics that correspond to the Gricean Maxims, apart from some aspects of the Maxim of Quality and the Maxim of Manner. Through this paper, we advocate for a turn towards mixed approaches to linguistic topics by also paving the way for a modernization of discourse analysis and natural language understanding that encompasses computational methods.

***Warning:*** *This paper contains offensive language that might be triggering for some individuals.*

## 1 Introduction

Natural Language Processing (NLP) is characterized by creating applications adept at addressing real-world challenges. Among those applications, we find machine translation, text summarization, coreference resolution, and part-of-speech-tagging, to name a few [19]. These rapid technological developments, along with the recent emergence of large language models (LLMs), have brought to the fore the question of how linguistics could benefit from such advancements and contribute to the current wave. This issue is discussed in a recent *Nature* editorial, where it is illustrated that there is a distinction between NLP and Computational Linguistics, with the latter focusing more on the two aforementioned questions. More specifically, "Computational Linguistics traditionally uses computational models to address questions in linguistics and borders the field of Natural Language Processing, which in turn builds models of language for practical applications" [1]. Dupre [6] poses an opposite opinion, claiming that deep learning techniques cannot illuminate linguistic theory, as the former focuses on language performance, while the latter on language competence, which are arbitrarily different.

In this paper, we draw from the distinction between Computational Linguistics and NLP, and we use NLP methods as tools for discourse analysis. Despite the argument that, at least current deep learning techniques are pertinent to theoretical insights in linguistics [6], we believe that deep learning tools can facilitate linguistic analysis. We exemplify this in our paper, by analyzing the structure of offensive language. Offensive language detection is a popular topic in NLP, as its intricate nature, lying within the borders of linguistics, psychology, sociology, and law studies, makes it hard for current models to identify positive instances adequately [35]. Current approaches in NLP view offensive language as a detection task without delving further into the intricate dynamics of an offensive conversation. We believe that a thorough analysis of offensive language requires a pragmatic approach. By examining contextual factors, such as speech acts, perlocutionary effects, politeness strategies, and cultural norms, we can gain a deeper understanding of how and why language becomes offensive. Ignoring these pragmatic aspects would result in an incomplete and potentially flawed analysis of offensive language.

In order to perform discourse analysis on online offensive language from a pragmatic perspective, we employ part of the Gricean theory [12], which outlines four conversational principles —Quality, Quantity, Relevance, and Manner— which ensure that speakers provide truthful, informative, relevant, and clear contributions to conversations. We argue that there is a pattern in the flouting/violation of the maxims when it comes to online offensive language. The most obvious assumption is that offensive language flouts the Maxim of Manner as this type of discourse is inherently not in accordance with this Maxim. In particular, offensive language uses an inappropriate lexicon that is unsuitable for any occasion, leading to uncooperative conversations. On a similar note, Pasa et al. [27] have shown that the sarcasm of hate speech in Instagram comments flouts all four maxims. The authors hypothesize that the main factors driving these violations are the lack of concise and clear information in comments, the cultural value in Western countries that emphasizes the right to free speech, the tendency to seek excessive attention from others, and the ego that boosts self-importance while devaluing others.

Our contribution is two-fold. Inspired from previous endeavors [32, 10], we first translate the Maxims into actual metrics using NLP methods, thus bridging the gap between theoretical and computational linguistics. Secondly, we offer a computationally facilitated discourse analysis on offensive language, showing that such analyses can be semi-automated as the data can be filtered faster, allowing for a more precise examination of specific instances. To our knowledge, this is the first study that equips Gricean theory through computational methods to analyze offensive language.

Our findings indicate that violations or floutings of the Maxims do not differ when comparing offensive and non offensive online dis-

* Corresponding Author. Email: aikaterini.korre2@unibo.it

**Table 1**: The 4 Gricean Maxims and their corresponding submaxims.

| Maxim | Sub-maxims |
|---|---|
| Maxim of Quality | • Do not say what you believe to be false.<br>• Do not say that for which you lack adequate evidence. |
| Maxim of Quantity | • Make your contribution as informative as is required for the current purposes of the exchange.<br>• Do not make your contribution more informative than is required. |
| Maxim of Relevance | • Be relevant. |
| Maxim of Manner | • Avoid obscurity of expression.<br>• Avoid ambiguity.<br>• Be brief.<br>• Be orderly. |

course statistically. The only exception is the Maxim of Manner due to the intense use of profanity and possibly the Maxim of Quality, as in the 'offensive' class, untruthful comments are more frequent. To assess the effectiveness of Maxim-based metrics in discourse analysis, we also conduct a qualitative analysis.

The paper is organized as follows. In the next section, we introduce some essential concepts of theoretical pragmatics, as they are the basis of our metrics and the discourse analysis. In Section 3, we discuss NLP approaches that involve linguistic pragmatic aspects both in terms of human language and artificially-generated language. In Section 4, we describe our methodology, translating the Maxims into metrics and using them as discourse analytical tools. In Section 5, we present our results, discussing them in Section 6. Finally, we summarize our final remarks and potential future work in Section 7, and we close this paper with a presentation of the limitations in Section 8.

## 2 Theoretical Background

One central point in pragmatics is the work of HP Grice, who formulated several pragmatic theories applicable today to conversation and discourse analysis. These include the *cooperative principle*, according to which the contribution of the conversation "must be such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" [12]. The cooperative principle outlines the fundamental principle *guiding* communication and it is broken down into multiple sub-principles, or 'Maxims'. More specifically, it suggests that in conversation, participants generally adhere to four Maxims: the Maxim of Quantity (provide just enough information), Quality (speak truthfully), Relevance (be relevant), and Manner (be clear and concise). More information about the Maxims and their explanations, often referred to as submaxims, can be found in Table 1. These Maxims serve as implicit guidelines for effective and efficient communication.

Another essential concept in pragmatics, and upon which we touch in this study, is implicature. Implicature occurs when speakers convey meaning beyond the literal interpretation of their words, relying on context and shared understanding. Implicatures can be further divided into those that are explicitly conveyed (explicature) and those inferred by the listener. Table 2 shows an example of conversation implicature, which illustrates a type of pragmatic inference that arises when words can be arranged on a semantic scale, such as the value judgments *excellent* > *good* > *OK*. Speaker A infers from B's response because if the accommodation had been better than just OK, B could have described it as good; if it had been very good, B could have said excellent. Since B did not use good or excellent, A concludes that the accommodation was merely satisfactory. At the time of the conversation, A might also have observed signs confirming this inference, such as B's unenthusiastic tone or body language indicating discomfort. These contextual clues further help in interpreting implicatures. However, they are unavailable in online language, making the task of interpreting implicatures even harder, especially in cases of sarcasm and irony.

**Table 2**: Example of conversational implicature taken from Griffiths and Cummins [13].

| Speaker | Dialogue |
|---|---|
| A | What was the accommodation like on the work camp? |
| B | It was OK. |
| A | Not all that good, hey? |

Implicature is also related to the violation or flouting of the Gricean Maxims. Violation of the 4 Maxims occurs when speakers deviate from the expected norms of communication, potentially causing confusion or misunderstanding, most often unintentionally. In contrast, flouting involves intentionally disregarding the Maxims, giving rise to conversational implicature for rhetorical or humorous effect. For example, when someone asks, "How's the weather?" during a thunderstorm, they flout the Maxim of Relevance by intentionally ignoring the obvious context.

## 3 Related Work

This section reviews key contributions in computational pragmatics, highlighting advancements in pragmatic inference, model evaluation, and the integration of pragmatic principles into NLP systems. We conclude by providing an overview of existing pragmatic approaches for offensive language detection and analysis.

### *3.1 Pragmatics in NLP*

Jurafsky [16] defines computational pragmatics as the computational study of the relationship between utterances and their context. It examines how utterances relate to actions, discourse, and environmental factors like time and place. Inference is a key focus in computational pragmatics, addressing four main problems: reference resolution, speech act interpretation and generation, discourse structure and coherence, and abduction. Each problem involves inferring missing information from utterances. However, when it comes to the interaction between NLP and pragmatics, the focal point of research lies in the detection of pragmatic effects, either in natural language or artificially generated language. Most approaches are concerned with the capability of the models to detect and/or to understand different pragmatic phenomena (such as irony, metaphor, and sarcasm) in natural language data, including social media posts and user inputs [2, 18, 21, 34]. Another emerging area of pragmatics in NLP is concerned with the ability of the models themselves (mainly LLMs) to actually produce speech intricate enough to mimic human language, including pragmatic language functions [4, 15, 17, 29].

As a broad field of linguistics, different aspects of pragmatics have been exploited or explored in NLP. Among them are also the Gricean Maxims. For instance, Hu et al. [15] present a fine-grained analysis of the pragmatics in the language of humans and LLMs in an attempt

to answer three questions: whether models select pragmatic interpretations of speaker utterances; whether models make similar errors as humans; and whether models use similar linguistic cues to solve the task. They show that certain pragmatic phenomena, such as humor, irony, and Grice's Maxims, involve violating listeners' expectations in some way and for which the LLMs fail to choose pragmatic interpretations. On a similar note, Jwalapuram [17] evaluate computer-generated dialogues according to Grice's Maxims. They use a survey in which the user is asked to rate the system performance on a Likert scale from 1 to 5 for 4 questions that correspond to the Maxims. In this way, they are able to identify: (1) if the system provides substantive responses; (2) if the system is faithful to the factual knowledge it is provided with; (3) if the system is able to understand the user and, therefore, provide relevant replies, and, finally; (4) if the system provides awkward or ambiguous responses. While they report diverse results depending on the generated dialogues, the authors do not go into speculations as to why that could be the case. Sorower et al. [31] present a method for learning rules from natural language texts by addressing the challenge of missing data. They introduce a mention model that addresses the probability of facts being mentioned in the text based on what other facts have already been mentioned and domain knowledge in the form of Horn clause rules and by formalizing Gricean Maxims encoding them as rules in Markov Logic.

Apart from being used as tools for conversation analysis, Gricean Maxims are also used as metrics. Ge et al. [10] propose the task of knowledge-driven follow-up question generation in conversational surveys. They produce a human-annotated dataset and they propose new metrics based on the Gricean Maxims. Freihat et al. [9] use the Maxims for ranking community question answers as hypothesize that linguistics offers a good opportunity to predict the relevance of answers and rank them accordingly. They use different indicators for each Maxim (except quality). Although their approach did not achieve the performance of machine-learning-based approaches, it gave a linguistically motivated solution that can be improved so that it reaches the performance of machine learning methods. Tewari et al. [32] focus specifically on the Maxim of Quantity and they model it as a new metric to assess the informativeness for short texts.

Implicature has also been in the limelight of linguistically motivated NLP research. Benotti and Traum [4] investigate the pragmatic implications of comparative constructions from a computational standpoint, emphasizing the challenges in determining the superiority of one answer over another. Zheng et al. [36] introduce a dataset for recovering implicature and conversational reasoning, showing that model performance improves when a module on implicature is included during training. Similarly, Ruis et al. [29] show that fine-tuning on conversational data or benchmark-level instructions does not produce models with pragmatic understanding. However, fine-tuning on instructions at the example-level paves the way towards more useful models of human discourse.

Understanding pragmatic functions in real-life situations presents a challenge for NLP. Unlike humans, who effortlessly use context and background knowledge to deduce implicatures, NLP models find this process difficult [36]. For example, in many cases incorporating Gricean theory (i.e. the cooperative principle and the 4 Maxims) involves using survey methods, employing humans to evaluate model capabilities with regard to the understanding of pragmatic discourse [17, 32].

### *3.2 Pragmatic Approaches on Offensive Language Detection and Analysis*

Many studies on hate speech, toxic language, offensive language or any other type of harmful language detection typically focus on individual instances, neglecting its inherently conversational nature [28]. This approach might be enough for solely NLP purposes but it limits the exploration of pragmatic analysis of harmful language on a discourse analysis level. One study that takes into account the context of toxicity in online conversations is the one from Madhyastha et al. [23], where they clearly show the significance of context and the effect on annotations. Other studies, such as in the case of Gevers et al. [11], have tried to analyze the structure of hate speech or different linguistic attributes of it, such as length and lexical diversity. Saveski et al. [30] studied the structure of toxic language spread. They show that, at the individual level, toxicity is spread across many low to moderately toxic users. At a dyad level, they observe that toxic replies are more likely to come from users who do not have any social connection nor share many common friends with the poster. At the group level, they find that toxic conversations tend to have larger, wider, and deeper reply trees, but sparser follow graphs.

One of the few works that has pragmatic aspects embedded in the methodology is the work of Upadhyaya et al. [33], where they introduce a dataset for toxic language that includes annotations for speech acts that could reveal information about the stance and that could help further in the toxic language detection. More traditional approaches of discourse analysis include the work of Hidayati and Arifuddin [14] that aims to reveal the types of hate speech on social media based on the criteria developed by Austin, and the meaning of hate speech spoken by individuals to other individuals on Facebook, using qualitative descriptive methods. The results show that hate speech on social media can be classified based on illocutionary acts developed by Austin, into verdictive, behabitives, and expositive. Finally, the work of Parvaresh [26] provides a corpus-assisted analysis of hate language as found on Instagram, focusing on Afghan immigrants. The study reveals that hate speech may lack markedly hateful language and that hate language may revolve around covert ways of expressing hatred.

In this paper, we investigate the potential of using NLP methods to evaluate pragmatic discourse, using the publicly available ToxiChat dataset [3]. We build on previous research to adapt NLP techniques for assessing Gricean Maxims and the cooperative principle. These tools are employed to conduct an advanced discourse analysis of a pragmatically complex discourse type: offensive language.

## 4 Methodology

For the purposes of this study, we use metrics and NLP tools for each of the 4 Maxims. In this way, we attempt to filter different instances that will be used for a qualitative analysis in a more traditional discourse analysis manner.

### *4.1 Translating the 4 Maxims into Metrics*

The purpose of the cooperative principle and the maxims is to guide effective and efficient communication by encouraging speakers to be informative, truthful, relevant, and clear in their discourse. To quantitatively assess the success of the cooperative principle, we employ metrics and tools commonly used in NLP, aligning each one with a respective maxim. Our approach draws inspiration from prior research that has endeavored to translate these maxims into NLP met-

**Table 3**: Information about the ToxiChat dataset

| Dataset | Source | Participants | Turns | Purpose | Instances |
|---|---|---|---|---|---|
| ToxiChat | Reddit | Human + Bot | 3 | Offensiveness and Stance detection | 3,211 |

**Table 4**: ToxiChat example.

| Turn | Text | Label |
|---|---|---|
| 1 | **Title: [Question] Why do Libertarians get so much flack from the rest of reddit Like seriously I was downvoted when I said "Libertarian is a good one" on a post about third party voting.** | Safe |
| 2 | Because the rest of reddit are unironically communists. | Offensive |
| 3 | Bullshit most are democrats | Offensive |

rics [10, 32], while also introducing new methods tailored to the specific focus of our study.[1]

**Maxim of Quality** For the Maxim of Quality, we define a text classification approach aimed at detecting deceptive content.

We train a BERT-based text classifier [5] on the Deceptive Opinion Spam Corpus (DOSC) [24, 25].[2] The corpus contains 1,600 customer reviews (both positive and negative) about 20 hotels. Half of the reviews are labeled as deceptive, while the remaining half are labeled as truthful. We group reviews based on the target hotel and build train (reviews of 16 hotels), validation (reviews of 2 hotels), and test (reviews of 2 hotels) splits such that all reviews belonging to a hotel are in the same split. We follow standard practice [5] and fine-tune the BERT-based text classifier for up to five epochs. We consider five different seed runs to ensure a sound evaluation. The classifier achieves an average macro F1-score of $0.926 \pm 0.021$ on the DOSC corpus test set.

**Maxim of Quantity** The Maxim of Quantity has been first studied in Tewari et al. [32]. The authors propose informativeness as a metric of the Maxim of Quantity based on syntactic cohesion. They use a dependency parser to transform segments into graphs of syntactic relations, defining syntactic cohesion as the sum of these relations. Syntactic cohesion is computed by comparing two sets of heads and their dependents, with normalized values falling between -1 and 1, indicating optimal, slightly cohesive, or fragmented cohesion. They normalize cohesion, dividing the score by the total number of words in the segment. Informativeness in an instruction sequence is the sum of syntactic cohesion values across all segments, with a normalized score ranging from 0 to 1, indicating under-informative, optimally informative, or over-informative sequences. In this study, we employ the same methodology.

**Maxim of Relevance** For the Maxim of Relevance, we implement a methodology to assess relevance within conversations using BERT embeddings and cosine similarity. Beginning with data preprocessing, we apply a custom binary relevance calculation function, which uses BERT embeddings to measure the similarity between conversation titles and their two subsequent responses. This process computes relevance scores by capturing the coherence of each response with respect to both the conversation title and the preceding response.

**Maxim of Manner** Our approach for the Maxim of Manner is inspired by Kiyavitskaya et al. [20] and focuses on assessing the instances in accordance to two main aspects of the Maxim: ambiguity and orderliness. There are many types of ambiguity, such as lexical, syntactic, and pragmatic. However, language models are not sensitive enough to successfully capture such delicate linguistic nuances yet [20, 22]. For that reason, we focus only on lexical ambiguity. We formulate our approach as follows:

Let $S$ be a sentence consisting of words $w_1, w_2, \ldots, w_n$. We define the ambiguity $\text{amb}(w_i)$ of word $w_i$ as the number of senses (synsets) that $w_i$ has in WordNet [7]. The ambiguity of $S$ is computed as

$$\text{amb}_{\text{total}}(S) = \sum_{i=1}^{n} \text{amb}(w_i) \tag{1}$$

Let $D$ be a dataset of sentences, where $S_j$ is a sentence in $D$. The maximum total ambiguity value is defined as:

$$\max(\text{amb}_{\text{total}}(S)) = \max_{S_j \in D} \text{amb}_{\text{total}}(S_j) \tag{2}$$

The Normalized Ambiguity of a sentence $S$ is then defined as:

$$\text{amb}_{\text{norm\_total}}(S) = \frac{\text{amb}_{\text{total}}(S)}{\max(\text{amb}_{\text{total}}(S))} \tag{3}$$

We also apply a readability metric as a proxy for text obscurity. We use the Flesch readability metric [8], which evaluates the ease of reading a text based on sentence length and word syllable count, providing a score from 0 to 100 (higher scores indicate easier readability and lower scores suggest more complex texts).

Regarding profanity, we use the *better profanity* library,[3] which enables us to identify instances of profanity within the data, thereby automatically violating the Maxim of Manner. The library includes a word list and returns True if any word in the provided string matches a word in the list. By systematically analyzing instances for ambiguity, obscurity, and orderliness, the methodology ensures adherence to principles of clarity and coherence in online discourse.

### *4.2 Data*

To conduct a discourse analysis based on the cooperative principle and the four maxims, we require that the data consist not only of isolated comments but also of dialogues with conversational turns. To our knowledge, there are few datasets containing instances of dialogues with offensive language, and those that do typically offer no more than two turns. Therefore, the data used for this study are sourced from the ToxiChat dataset [3], primarily constructed for stance analysis in online offensive contexts. Details about the data are presented in Table 3, with an illustrative example available in Table 4. In our study, we use only the train set of the dataset, and since we are interested in a pragmatic analysis of natural language, we are only concerned with the turns in the thread that are produced by humans and not the turn produced by the bot.

## 5 Results

**Table 5**: Truthfulness of ToxiChat instances with respect to the offensiveness of each instance. The bars are annotated with the percentages per class (safe/offensive).

| Predictions | True | Untrue |
|---|---|---|
| Offensive | 558 (60%) | 371 (40%) |
| Safe | 1609 (70%) | 673 (30%) |

**Quality** Table 5 shows the results of the BERT based deception classifier, assessing the truthfulness of the instances. Predominantly

[1] Code available at: https://github.com/katkorre/A-Griceful-Examination-of-Offensive-Language.git

[2] We use the `bert-base-uncased` model card from HuggingFace.

[3] https://pypi.org/project/better-profanity/

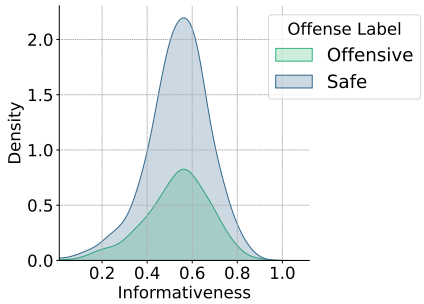


**Figure 1**: Histogram of Informativeness of ToxiChat instances with respect to the offensiveness of each instance.

in both classes, there are more instances labeled as True rather than Untrue. However, we notice that in the 'safe' class, around 70% of the instances are marked as True, while this percentage drops in the 'offensive' class, with 60% of the cases classified as True. Proportionately, untruthful statements are more likely to appear in offensive language. Therefore, it is more likely that the Maxim of Quality is flouted or violated in offensive contexts.

**Quantity** The evaluation results of the Maxim of Quantity, based on informativeness, are presented in Figure 1. We observe that, in most cases, both offensive and non-offensive instances consistently achieve a reasonable level of informativeness throughout the thread excerpts, with the majority of the values falling close to 0.5 which indicates an optimally informative instance.

We analyze this further with Table 6 as we proceed to form three thresholds that correspond to three classes of informativeness to compare the offensive against the safe class. A threshold of 0.25 was set to delineate instances deemed 'Under-Informative', indicating low levels of informativeness. A threshold of 0.75 was designated for data instances categorized as 'over-informative', denoting instances that contain redundant information. Finally, the values in-between denote the optimum level of informativeness. Most of the instances are optimally informative. Comparing the 'safe' and 'offensive' classes, there are more under- or over-informative instances in the 'safe' class. The difference in the number of under- or over-informative instances between the 'safe' and 'offensive' classes might be influenced by the uneven distribution of instances across these classes. Since the dataset contains significantly more 'safe' instances than 'offensive' ones, this imbalance can skew the analysis. Instead of normalizing or stratifying in this study, we maintain the raw data characteristics to interpret with a focus on real-world relevance.

**Relevance** In terms of relevance, our results are shown in Figure 2. Most instances, offensive or not, are deemed relevant by our model. Similar to previous Maxims, relevance is rarely flouted or violated, and when it is, it most frequently occurs in the 'safe' class. An exception is seen in responses to the title, where violations also occur in the 'offensive' class. This suggests that there is likely no correlation between the offensiveness of an instance and its violation of the maxim of relevance. However, it is important to consider the domain of the data, which is sourced from Reddit. Given that Reddit revolves around specific questions and answers, the room for irrelevant responses is limited.

**Manner** We evaluate the Maxim of Manner in terms of ambiguity, readability and profanity. Figure 3 displays a boxplot of our ambiguity detection results. The two boxes are similar in size, with the 'safe' class showing a slightly larger range of values. The general pictures accounts to the fact that, in terms of ambiguity, 'offensive' and 'safe' dialogue instances do not differ to a significant degree. The picture is similar when calculating readability, with both classes presenting high scores in the Flesch readability metric, with the lower quartile being close to 50 in both cases. The 'safe' class, however, also presents a higher number of outliers that tend to have lower readability. Among those scores there are also negative ones which indicates a very short sentence or an extremely complex one. This could also be due to internet language and formatting. We initially hypothesized that the Maxim of Manner is typically flouted in the context of toxic or offensive language, and our results confirm this through the high frequency of profanity. Offensive language often relies on strong, explicit terms to convey hostility or aggression, which naturally includes a higher frequency of profanities. This is obvious in Figure 5 which shows that instances labeled as offensive contain more profanities compared to those that are labeled as safe.

**Table 6**: Informativeness thresholds of ToxiChat instances with respect to the offensiveness of each instance.

| Category | Optimally Informative | Over-Informative | Under-Informative |
|---|---|---|---|
| Offensive | 886 (95.37%) | 21 (2.26%) | 22 (2.37%) |
| Safe | 2186 (95.79%) | 50 (2.19%) | 46 (2.02%) |

## 6 Exemplary Discourse Analysis

Quantifying the maxims and examining the results in Section 5 have allowed us to form a more concrete idea and hypothesis, while it also allows us to filter results that would be of discourse analysis interest. To perform a discourse analysis, we first proceed to select instances according to the results of the previous section. For that reason, we look only at the offensive class and we randomly choose one example that violates each maxim, and one example that does not and proceed to compare the instances. The selected examples can be found in Table 7.

Comparing the two examples for the Maxim of Quality, the one that does not violate the maxim, does not contain any information that could potentially be untrue. The use of hedging with 'seem' and the simile introduced with 'like' mitigate the certainty of the author of the comment, despite the fact that it is an offensive comment. The example that violates the maxim, however, is full of potentially false assumptions, such as "he used all the sexual energy into fighting". This information is misleading and does not contribute to an effective cooperative (online) conversation.

Looking at the examples for the Maxim of Quantity, both responses generally adhere to the maxim. Response 1, which looks as an additional comment from the author of the title thread, provides enough information to support its point without overwhelming details, justifying the reasoning to their initial question. Response 2 offers a concise and direct answer. However, it is possible it could be considered slightly under-informative as it does not precisely reply to the initial question. About the second example, that according to the used algorithm violates the maxim of quantity, Response 1 provides an abundance of specific criticisms, making it slightly verbose and less clear due to its structure. That could lead us to the conclusion that it violates the Maxim of Quantity. The second response expresses an opposite opinion from the one presented in the title. It does not answer directly the question. However, with that response 'I like them', we are to lead to the implicature that the author of the response does not support banning 'GenderCritical', contrary to the suggestion in the title.

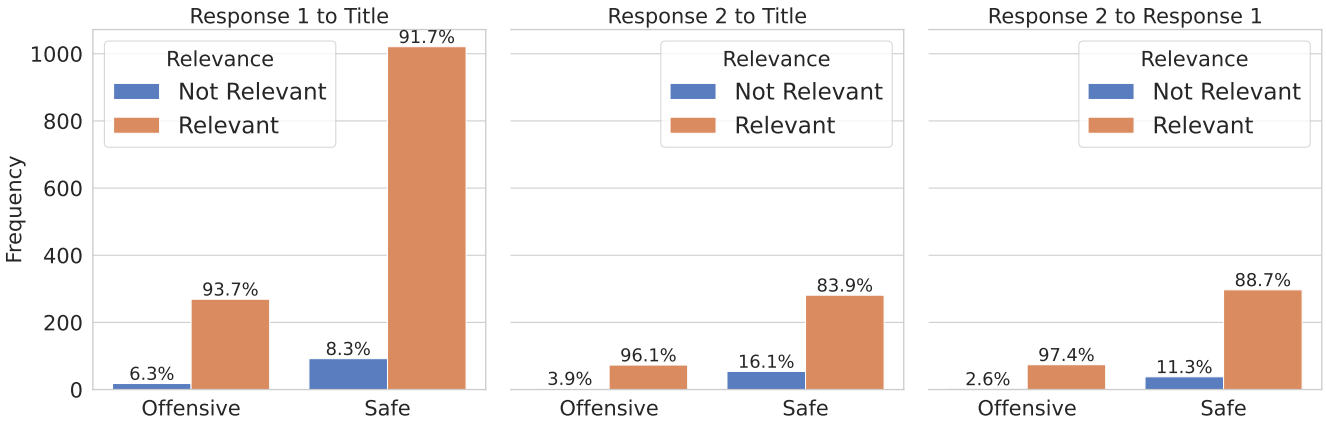


**Figure 2**: Distribution of relevance scores concerning offensiveness. The first plot shows the distribution of the first replies to the title, the second plot shows the distribution of the second replies to the title, and the third plot shows the distribution of relevance for the second reply to the first reply.

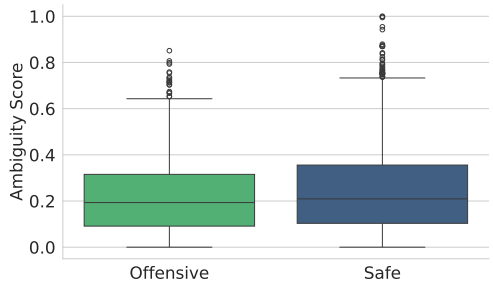


**Figure 3**: Ambiguity scores of ToxiChat instances with respect to the offensiveness of each instance.

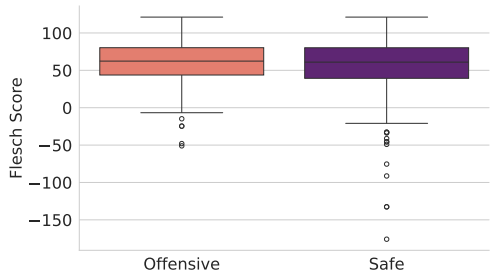


**Figure 4**: Flesh Readability. Higher scores indicate easier readability, with lower scores suggesting more complex texts.

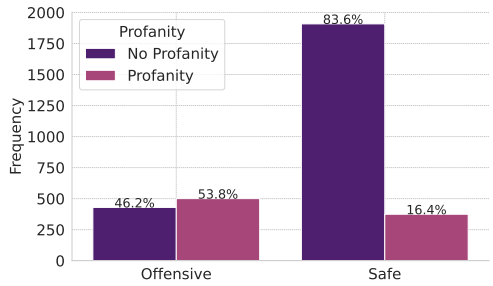


**Figure 5**: Distribution of profanity in ToxiChat instances with respect to the offensiveness of each instance. The bars are annotated with the percentages per class (safe/offensive).

**Table 7**: Randomly selected offensive instances that violate the Gricean Maxims or not. For Quantity and Relevance we report 2 or 3 conversation turns, as the metrics take into account the whole conversation. For Quality and Manner, we report the individual instances.

| Maxim | Text | Violates |
|---|---|---|
| Quality | a lot of you seem like genuinely shitty people | No |
| | Mike tyson didnt have sex for 5 years Boxing over Bitches He used all the sexual energy into Fighting also not wasting excess Testestrone and using it as fuel I guess he is champ and sometimes sacrifice like that what makes sucessful people different over normal people What are your thoughts | Yes |
| Quantity | **Title**: Do you get along with the NONcrazy Liberals Leftists; **Response 1**: And before you say there are no noncrazy Democrats read the comment section of this NY post opinion piece URL They despise the fascist farleft takeover of everything as much as we do that comment section alone was a real eyeopener; **Response 2**: I get along with everyone until they start being a dick | No |
| | **Title**: Im curious can we get GenderCritical banned; **Response 1**: I mean it hates autistic men male babies and encourages misandry and hatred against men as a whole It even surpassed incels in terms of idiocy and hatred because at least incelsfrom rMGTOW like their daughters to some extent GenderCriticial even hates sons and male children and anything to do with men; **Response 2**: I like them They show the true nature of feminism to everyone I wish they had 1mm members so they cant be brushed off as a minority | Yes |
| Relevance | **Title**: British women abroad are a FUCKING DISGRACE Morgoths Review morgoths review; **Response**: Tourists are the worst kind of twatthot trash | No |
| | **Title**: Have an amazing daynight No updoot required You have no choice You must be happy right now or else I will come over to you and give you a hug No homo lt3 you can accomplish your days; **Response**: yes homo | Yes |
| Manner | Could someone make hitlers easy bake oven it might still exist but I cant find it | No |
| | Stoning the whores is the one thing Muslims got right | Yes |

Examining the examples for the Maxim of Relevance, the response in the first one does not directly address the specific claim about British women abroad being a disgrace. Instead, it broadens the criticism to include all tourists, which dilutes the specificity of the original statement. This shift can be seen as a deviation from the maxim of relevance because it does not maintain the specific focus introduced by the title. While the response is contextually related (both the title and the response criticize behavior abroad), the failure to address the specific group mentioned (British women) makes it less relevant. Thus, it can be argued that the response violates the maxim of relevance by not directly engaging with the specific claim made in the title. This could therefore be considered a potential error of the model. About the second example, the initial message and the response both adhere to the Maxim of Quantity. The initial message provides enough context and information to be encouraging and humorous, and the response is brief but relevant and clear. The playful nature of the exchange is maintained through both contributions, and neither is too sparse nor too detailed for the context. The response directly addresses the 'No homo' part of the initial message, playfully contradicting it. It provides a relevant and humorous counterpoint to the initial message without adding unnecessary information. Therefore, this could be a false positive error for the algorithm.

Finally, about the Maxim of Manner, the first example is offensive probably towards Jews. It could be considered a slightly ambiguous statement, as it is unclear whether the speaker is making a dark joke, referring to a specific object or concept, or whether they misunderstand the implications of the words they are using. About the last example, the statement is highly offensive and lacks clarity. It uses derogatory language and promotes violence without any regard for decency or ethical considerations.

Despite occasional algorithmic errors in detecting the maxim floutings or violations, the pragmatic discourse analysis facilitated by our approach effectively highlights the nuances in which offensive language interacts with conversational norms. By applying computational algorithms, we can systematically filter and analyze large datasets, revealing patterns in offensive language.

## 7 Conclusions

In this paper, we adopt a pragmatics-based approach to hate speech analysis, reinforced by NLP methods. We draw from the linguistic theory of the 4 Gricean Maxims and the Co-operative principle, and we employ NLP methods as tools to assess whether the maxims are flouted or violated in offensive online contexts. Our approach provides an essential step before any type of discourse analysis, that allows a better understanding of the data and consequent filtering of instances for qualitative discourse analysis. Our experimental results showed some patterns in the flouting/violations of the maxims in offensive language settings, such as the flouting/violation of the maxim of manner due to ambiguity and profanity. With this paper, we advocate for more mixed approaches that will encompass both computational and traditional linguistics, and which will contribute to better data analysis.

In future work, we aim to dive deeper into the potential of the metrics, particularly when coupled with advanced LLMs. This exploration will contribute to further automate the assessment process of the cooperative principle, potentially enhancing its accuracy. Additionally, we intend to investigate other discourse domains posing challenges to NLP, such as sentiment analysis, humor, and sarcasm detection, which represent pragmatically charged categories. Furthermore, we are interested in examining intersections among the maxims to gain a comprehensive understanding.

## 8 Limitations

Our work is not without limitations. First of all, the metrics and NLP techniques that we used for the assessment of the maxims is not perfected and should be tested in other settings, as well as evaluated in more contexts, ideally from human experts. Even with NLP models with very high performance, there is always the possibility of error. Therefore, manual examination for discourse analysis is essential. Another limitation relates to the fact that, during the qualitative analysis, we examined each example for only one maxim flouting or violation each time, though it is possible that more than one floutings or violations could co-occur.

## Acknowledgements

This research is carried out under the project "RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete", in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118.

## References

[1] Language models and linguistic theories beyond words. *Nature Machine Intelligence*, 5(7):677–678, 07 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00703-8. URL https://doi.org/10.1038/s42256-023-00703-8.

[2] A. A, S. G, S. H R, M. Upadhyaya, A. P. Ray, and M. T C. Sarcasm detection in natural language processing. *Materials Today: Proceedings*, 37:3324–3331, 2021. ISSN 2214-7853. doi: https://doi.org/10.1016/j.matpr.2020.09.124. URL https://www.sciencedirect.com/science/article/pii/S2214785320368164. International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science.

[3] A. Baheti, M. Sap, A. Ritter, and M. Riedl. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.397. URL https://aclanthology.org/2021.emnlp-main.397.

[4] L. Benotti and D. Traum. A computational account of comparative implicatures for a spoken dialogue agent. In H. Bunt, editor, *Proceedings of the Eight International Conference on Computational Semantics*, pages 4–17, Tilburg, The Netherlands, Jan. 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-3704.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[6] G. Dupre. (what) can deep learning contribute to theoretical linguistics? *Minds and Machines*, 31(4):617–635, December 2021. ISSN 1572-8641. doi: 10.1007/s11023-021-09571-w. URL https://doi.org/10.1007/s11023-021-09571-w.

[7] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[8] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532. URL https://doi.org/10.1037/h0057532.

[9] A. Freihat, M. Qwaider, and F. Giunchiglia. Using grice maxims in ranking community question answers. In *Proceedings of The Tenth International Conference on Information, Process, and Knowledge Management (eKNOW 2018)*, Rome, Italy, 03 2018.

[10] Y. Ge, Z. Xiao, J. Diesner, H. Ji, K. Karahalios, and H. Sundaram. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li,

and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 113–124, Hong Kong, China, Dec. 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.paclic-1.12.

[11] I. Gevers, I. Markov, and W. Daelemans. Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands Journal*, 12:33–48, Dec. 2022. URL https://www.clinjournal.org/clinj/article/view/146.

[12] P. Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics*, volume 3 of *Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[13] P. Griffiths and C. Cummins. *An Introduction to English Semantics and Pragmatics*. Edinburgh Textbooks on the English Language. Edinburgh University Press, United Kingdom, 2 edition, Dec. 2016. ISBN 9781474412810.

[14] A. Hidayati and Arifuddin. Hate speech on social media: A pragmatic approach. *KnE Social Sciences*, 5(4):308–317, Mar. 2021. doi: 10.18502/kss.v5i4.8690. URL https://knepublishing.com/index.php/KnE-Social/article/view/8690.

[15] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230.

[16] D. Jurafsky. Pragmatics and computational linguistics. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*, page 578. Blackwell Publishing Ltd, Malden, MA, USA, 2006. ISBN 978-0-631-22547-8.

[17] P. Jwalapuram. Evaluating dialogs based on Grice's maxims. In V. Kovatchev, I. Temnikova, P. Gencheva, Y. Kiprov, and I. Nikolova, editors, *Proceedings of the Student Research Workshop Associated with RANLP 2017*, pages 17–24, Varna, Sept. 2017. INCOMA Ltd. doi: 10.26615/issn.1314-9156.2017_003. URL https://doi.org/10.26615/issn.1314-9156.2017_003.

[18] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, and N. Aussenac-Gilles. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1025.

[19] D. Khurana, A. Koli, K. Khatter, et al. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*, 82: 3713–3744, 2023. doi: 10.1007/s11042-022-13428-4.

[20] N. Kiyavitskaya, N. Zeni, L. Mich, and D. M. Berry. Requirements for tools for ambiguity identification and measurement in natural language requirements specifications. *Requirements Engineering*, 13(3): 207–239, 2008. doi: 10.1007/s00766-008-0063-7. URL https://doi.org/10.1007/s00766-008-0063-7.

[21] Y. Li, S. Wang, C. Lin, and F. Guerin. Metaphor detection via explicit basic meanings modelling. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.9. URL https://aclanthology.org/2023.acl-short.9.

[22] A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, S. Swayamdipta, N. Smith, and Y. Choi. We're afraid language models aren't modeling ambiguity. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.51. URL https://aclanthology.org/2023.emnlp-main.51.

[23] P. Madhyastha, A. Founta, and L. Specia. A study towards contextual understanding of toxicity in online conversations. *Natural Language Engineering*, 29(6):1538–1560, 2023. doi: 10.1017/S1351324923000414.

[24] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In D. Lin, Y. Matsumoto, and R. Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1032.

[25] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In L. Vanderwende, H. Daumé III, and K. Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 497–501, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1053.

[26] V. Parvaresh. Covertly communicated hate speech: A corpus-assisted pragmatic study. *Journal of Pragmatics*, 205:63–77, 2023. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2022.12.009. URL https://www.sciencedirect.com/science/article/pii/S037821662200296X.

[27] T. A. Pasa, Nuriadi, and H. Lail. An analysis of sarcasm on hate speech utterances on just jared instagram account. *Journal of English Education Forum (JEEF)*, 1(1):10–19, Jun. 2021. URL https://jeef.unram.ac.id/index.php/jeef/article/view/94.

[28] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity detection: Does context really matter? In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.396. URL https://aclanthology.org/2020.acl-main.396.

[29] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms, 2023.

[30] M. Saveski, B. Roy, and D. Roy. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449861. URL https://doi.org/10.1145/3442381.3449861.

[31] M. Sorower, J. Doppa, W. Orr, P. Tadepalli, T. Dietterich, and X. Fern. Inverting grice's maxims to learn rules from natural language extractions. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/8c7bbbba95c1025975e548cee86dfadc-Paper.pdf.

[32] M. Tewari, S. Bensch, T. Hellström, and K.-F. Richter. Modelling grice's maxim of quantity as informativeness for short text. In *Proceedings of the 10th International Conference in Languages, Literature, and Linguistics (ICLLL 2020)*, pages 1–7, Japan, 2020. URL https://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-176269.

[33] A. Upadhyaya, M. Fisichella, and W. Nejdl. Toxicity, morality, and speech act guided stance detection. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4464–4478, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.295. URL https://aclanthology.org/2023.findings-emnlp.295.

[34] C. Van Hee, E. Lefever, and V. Hoste. Exploring the fine-grained analysis and automatic detection of irony on twitter. *Language Resources and Evaluation*, 52(3):707–731, 2018. ISSN 1574-0218. doi: 10.1007/s10579-018-9414-2. URL https://doi.org/10.1007/s10579-018-9414-2.

[35] W. Yin and A. Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7, 2021. URL https://api.semanticscholar.org/CorpusID:231942329.

[36] Z. Zheng, S. Qiu, L. Fan, Y. Zhu, and S.-C. Zhu. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.182. URL https://aclanthology.org/2021.findings-acl.182.