Bocconi

Bachelor of Science in Economics, Management and Computer Science

An Approach to the Explanation of Annotator-level Differences in Toxicity Detection Through Irony

Advisor:
Prof. Kai Zhu

Bachelor of Science thesis by:
LUCA COLACI
student ID no. 3176608

Academic Year 2023-2024

Bocconi
Università Commerciale
Luigi Bocconi

Acknowledgements

I am convinced that life, no matter how solitary a sport, is to be considered a team effort in how it is approached. This is why I believe that my academic career, and especially this final work, is no less than that. My success results from a personal effort combined with the efforts of all the people I have met inside and outside this University.

First and foremost, I want to express my deepest gratitude to my family: my parents, grandparents, uncle and aunt, and, most importantly, my sister Maria Chiara, who is the reason why I'm graduating from this University. Your support was crucial for my academic success, and I will be forever grateful for this.

I must also thank my lifelong friends, the guys from SNPL and my friends from Torre Vado, who supported me in the worst moments and rejoiced with me in the best ones with a dedication that I will never repay enough.

A special thank you also goes to my friends from Bligny, who took me in and with whom I got through these three years of study and hard work, but also many joys.

Thanks to my housemates in Atlanta, who picked up the pieces and helped me rebuild myself, with whom I shared four memorable months.

Thank you to Astra and all its members, who allowed me to give back to the student community in the best way possible. Thanks especially to the people I founded this group with, as none of this would have been possible without you.

Thanks are also due to those professors I have met over the years (on this side and the other side of the Atlantic) who have shown me the meaning of respect for the student, the desire to teach and the love for their profession. If I now have clear ideas about what I want to do in life, it is also thanks to you.

Finally, a thank you to the BEMACS guys, a community of people who wanted me as their representative for three years and whom I will never stop thanking for what we shared during this time.

Contents

1 Introduction

Hate and toxic messages have been a part of our society for most of the existence of humans on earth, with the first insults to have ever been recorded belonging to Sumerians. Many researchers in various disciplines recognize the importance of toxic language in the creation of civilisation, with the famous quote from Sigmund Freud "*The man who first flung a word of abuse at his enemy instead of a spear was the founder of civilization.*" [1]. It is, therefore, possible to understand how much this type of language influenced our society. Still, our understanding of the inner workings of such mechanisms, how they are generated, and, most of all, how different individuals perceive them is still very poor.

This problem is now more pressing than ever, as the social media era allowed for an environment that grants individuals two critical factors that pushed toxic language to its peak: first, anonymity [2], as an individual that does not fear that his identity will be revealed to the world is more likely to use harsher words and insults when commenting online; second, the possible reach of an online post, which is orders of magnitude larger than any other communication method and is also enhanced by social network's recommendation algorithms, that will show users posts that they're more likely to appreciate. This issue is not unknown to companies working in social media, as many of them apply some kind of automatic moderation algorithm, such as word blacklisting or other types of insult detection algorithms in combination with human moderation. However, social media users have become accustomed to this environment, developing slang and methods to avoid detection while still being able to express comments that may go against social media's policies.

This is, however, a problem that has ample studies spanning various fields. At the same time, an issue less addressed by scholars is the understanding of those ambiguous comments, for which it is unclear whether the comment in question should be categorised as an insult or as a harsh critique expressed through borderline words.

The inspiration for this work stems from recent news events involving some students at the Bocconi University in Milan[3], where three students were suspended after comments on Instagram regarding the introduction of gender-neutral bathrooms at the University. The suspension decision by the university generated a great debate online and off social media, even reaching the Italian Parliament, with parties supporting the university's decision on

the grounds that the comments were offensive insults towards the LGBTQ+ community. At the same time, opponents argued that the comments were criticism expressed in ironic and goliardic tones. This polarization suggested the idea that the different perceptions of these comments, and in general of online insults, could be related to how different categories of people belonging to different social and cultural groups perceive irony and how this perception influences their judgement regarding the toxicity of an online comment.
The objective of this thesis is therefore to suggest an approach to study the interaction between toxicity ratings, the community the annotator belongs to and irony and was made possible by leveraging two main resources: first, the "*Jigsaw Specialized Rater Pools Dataset*" [4], an online public dataset reporting both a toxicity score for a certain comment and information about the annotator that rated the comment, allowing for analysis that studies how raters belonging to different communities perceive toxicity on the same comment; second, a public RoBERTa-based LLM for irony detection [5], that is used to add information to the dataset by providing irony ratings for comments of the dataset, as there is no public human annotation for irony available for it.
The structure of the thesis is as follows: Section 2 will delve deeper into the various topics that are touched by the analysis and the resources used; Section 3 will explain the tools used to perform the analysis and the chosen methodologies; Section 4 will apply those methodologies to the dataset and study the results; Section 5 will deepen the problematics and limitations for the experiment, suggesting ways to solve them and, hopefully, contributing to possible future researches on the topic; finally, Section 6 will summarize the findings of this thesis.

2 Literature Review

2.1 Annotating Corpus

Corpus annotation is a critical component of computational linguistics, as it describes the act of adding metadata to text, describing linguistic features of the data such as syntax, semantics or discourse information[6] [7]. This process has been critical in the creation and development of Natural Language Processing (NLP), *i.e.* the processing of text through automated processes. As the field and research interest in it grew, various attempts at standardizing processes and methodologies were made, creating big corpora, such as the Penn Treebank [8], influencing and setting standards for subsequent works.
Initially[9], annotations were collected by hand by manual annotators, selected by the corpus creators and trained through guidelines and examples on how to execute the task correctly, usually performed on software appositely created for such tasks. On the other hand, this method introduced several issues that can be boiled down to one critical question: what are the possible sources of bias and how to limit their impact on the data?
The main source can be found within the annotators themselves: if for basic tasks, such as identifying the subject of a sentence, the possibility for multiple interpretations is close to zero, more complex tasks, such as irony or toxicity rating, which will be analyzed more thoroughly later in this section, offer various occasions where the intrinsic characteristics of the annotator might influence the annotation. To counter such issues, multiple methods have been implemented over the years:

- *Task decomposition:* Nowadays most corpus annotation efforts start with the goal of training a specific model, the final task should be divided into simpler tasks[10], to avoid making the process repetitive for the annotator, to achieve higher quality results.

- *Guidelines:* Writing effective and comprehensive guidelines for each annotation task is crucial for obtaining coherent and usable results. Usually, guidelines are contained in a document that is provided to annotators before the start of the task, and they should contain which text has to be annotated, how the annotation should be performed and how to deal with special and edge cases[7].

- *Specialized software:* The usage of specialized software for annotators, either based on open-source projects and personalized for the specific task or made ad-hoc, can improve consistency in the format of the annotations, allowing also easy compliance to internal or international formatting standards for annotations, such as the ISO 24612:2012 linguistic annotations formatting (LAF) standard.

- *Training and testing:* Nowadays, it is common practice to make annotators train on a sample corpus and evaluate them in general (*ex.*, language proficiency) and task-related tests, and then consider only annotations given by annotators that passed the minimum thresholds for these tests.

- *Rewards:* As humans perform better under an incentive, the same concept has been applied to the annotation task[11]. Incentives can be of three main types: *personal*, where the annotation task is made entertaining for the annotator (for example, by implementing a game-with-a-purpose environment); *social*, where the annotator is rewarded by feeling that he is contributing to a common effort; *financial*, where the annotator is rewarded through some form of currency, depending on the level of difficulty and time spent on each annotation.

- *Collaborative annotation:* For complex tasks, such as emotion detection, crowd-sourcing annotation can also be considered an effective method for reducing bias by making multiple annotators process the same corpora and then evaluating a proper scoring and/or exclude the piece of corpus from the final dataset based on annotator agreement.

Due to the good results that the crowdsourcing method brought in the world of corpus annotation, nowadays, this is the most applied technique in manual annotation, as it exploits the "wisdom of the crowd" concept, *i.e.* the fact that the average of multiple guesses or answers to non-trivial questions usually better approach the ground truth than what a single guess would do.

2.2 Toxicity Detection

The common definition for *toxic* when talking about language is "*extremely harsh, malicious or harmful*" [12]. On the other hand, this is an umbrella term, including various categories of insults regarding, but not limited to, profanity, obscenity, sexually explicit conversation, identity-based attacks, insults, and threats [13]. This wide range of categories underlines the complexity of detecting toxicity in text, encompassing a range of negative behaviours and languages. This type of language is also not static, as it evolves over time due to natural changes in language and, especially in social media corpora, due to the pressure of moderation tools enforced by most platforms, which push users to hide insults and other words commonly related to toxic language to avoid being detected. The latter is also the main cause of obfuscation [14] [15], *i.e.* modifying words and phrases by using alternative words or commonly known terms (sometimes grammatically incorrect) to avoid moderation detection [16]. The social importance of this issue and its relevance in the social media environment for moderation purposes helped this topic gain popularity, with several papers and detection methods studied and applied for sub-categories of this problem [17] [18] [19].

One of the first works in the field [20] applied a supervised classification model to social media data to detect harassment in conversations by using n-grams, regular expressions and contextual features (namely, the "amount of harassment" detected in parent comments or replies), but it was noted that the nature of social media comments made accurate detection very difficult, due to involuntary spelling errors, shortness of sentences and voluntary user obfuscation. Nowadays, most platforms combine human moderation with predefined word blacklists. Still, these could fail due to lacking contextual information, work obfuscation and/or grammatical errors. However, [15] proposed a solution taking into account *edit distance*, *i.e.* how many edits would it take for a word to be edited into a target one. Setting a threshold to include in the word blacklist also words within a certain edit distance from popular insults can help detect obfuscation and grammatical error cases (for example, recognizing that "*@ss*" is just a voluntary obfuscation for "*ass*"). Older approaches tried to correct these errors and obfuscation during data pre-processing [21]. In contrast, more modern approaches try to leverage the presence of such features as an additional indicator of possibly offensive messages [14].

Another barrier to effective automatic toxicity detection is, in many cases, sense disambiguation, *i.e.* the ability to recognize the meaning of a word through the context it is used in. In this particular task, great work was performed in [22], specifically referring to anti-semitic hate. Here, manual annotation of corpora of both Yahoo! news comments and potentially anti-semitic websites was performed, then several possibly offensive words were identified, and Word Sense Disambiguation, as described in [23], was performed to distinguish those cases where dubious words were used in their offensive and non-offensive meaning.

2.3 Rater Identity Impact on Toxicity Detection

This analysis is considered an extension of the work made by Nitesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman "*Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation*" [4], and due to its importance for comprehending this work, a thorough explanation of their research is needed. Their goal with this experiment was to evaluate the impact of the rater's identities on toxicity annotations when annotating online comments. To do so, they built *specialized rater pools*, that is, groups of raters identifying themselves to either one of three social and cultural groups:

- U.S. Citizens identifying themselves as African American (AA);
- U.S. Citizens identifying themselves as members of the LGBTQ community;
- U.S. Citizens identifying themselves as neither African American nor members of the LGBTQ community.

After creating these rater pools, each group was assigned the same set of 25,000 comments taken from the Civil Comments dataset [24]. The dataset contains around 2 million comments from an out-of-business news website, with all of the comments labelled for toxicity and some of them also labelled by the category the toxic comment was referred to. The researchers decided to sample 8,500 identity-agnostic comments, 8,500 comments with messages referring to the African American community and 8,500 comments referring to the LGBT community. Due to self-given ethical guidelines, the samples were also controlled to limit the rater's exposure to toxic contents.

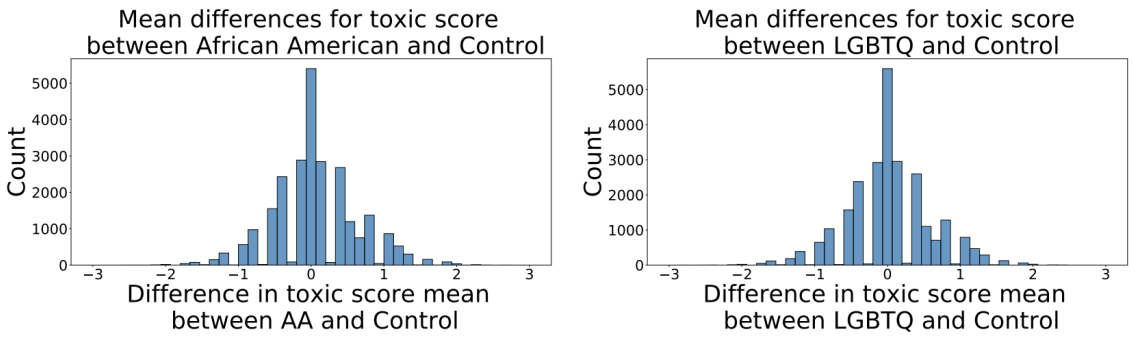

Figure 1: Histograms showing Toxicity Mean Difference distributions for both specialized rater pools against Control.

Raters were then asked to rate the comments, producing 15 ratings (5 per specialized pool) for each element in the dataset, each containing ratings on Toxicity, Identity Attack, Insult, Profanity and Threat. Ratings were given using a 4-point Likert scale[25], ranging from -2 ("Very Toxic") to 1 ("Not Toxic"), while all other labels were given on a Likert scale ranging from -1 ("Yes") to 1("No"). During annotation, additional precautions were made to safeguard the rater's well-being and limit their exposure to toxic comments, limiting the number of highly toxic comments in the sample and the amount of them that they would be exposed to per day, and all raters were given access to a peer-to-peer support platform to share opinions on the task and ask for support if needed.
After ratings were collected, the authors performed an analysis aimed at identifying if and in which direction specialized rater pools disagreed with the control one. To do that, two main approaches were taken, covering both qualitative and quantitative analysis:

- *Toxicity Mean Difference*: this value was computed for each comment in the dataset as the difference in means of the toxicity ratings given by both specialized rater pools and the control one. Plotting these differences as histograms resulted in non-skewed, centred around zero distributions (Figure 1), meaning that specialized pools disagreed with the control one in both directions (*i.e.*, they rated some comments as more toxic than the specialized pool but also the opposite).

- *Toxicity Odds Ratio*: the authors run an ordered logistic regression[26] to understand the impact of the rater category on the toxicity score. Results showed no statistically

significant difference in toxicity rating between African American raters and the control, but the odds for the control pool to rate a comment as toxic was 0.957 times that of the LGBTQ rater pool, a statistically significant difference ($p-\text{value} < 0.001$) meaning that raters from the control pool were more likely, on average, to not rate a comment as toxic with respect to the LGBTQ specialized rater pool. On top of that, authors performed the same analysis on the other four labels, finding statistical differences in ratings between specialized rater pools and control for all of them, with odds lower than 1 for all at .001 significance.

2.4 Automatic Irony Detection

"*Irony*" derives from the Greek word $\varepsilon\iota\rho\omega\upsilon\varepsilon\iota\alpha$, a word derived from the Greek tragedy character of the *eiron*, an old man who would hide his intelligence to other characters to overcome its rival, the young and boastful *alazon*, at the end of the play.[27]. Nowadays, the term defines "the incongruity expressed between the context and statement conveyed in a piece of text"[28], *i.e.* expressing something by saying the exact opposite, but in common language, it is also used as an umbrella term including also sarcasm, satire and humour [29]. Irony can be either situational or verbal [30]:

- *Verbal irony* refers to someone conveying a message by saying the opposite, therefore closely following the definition;
- *Situational irony* refers to a situation where someone acts or expresses a message that is clearly in contrast with his environment, effectively opposing it and therefore generating the irony.

Because of its deceptive nature, even to humans, automatic detection of irony has been a widely studied problem in the field of Natural Language Processing, with different approaches being developed over the years. As the task was posed as a binary classification problem, having to determine if a piece of the corpus was to be considered ironic or not, the first approach was to implement one of the most used techniques for this kind of application, *i.e.* Support Vector Machines (SVMs). This technique requires a vector representation for each document to be classified, and then it finds a confident boundary between the categories to be told apart. For this particular task commonly-used

document vectorization methods, such as Bag-Of-Words (BOWs) didn't yield satisfying results while still outperforming previous approaches in other tasks, but more advanced feature-weighting methods allowed to achieve better results[31].

With the rise of machine learning and neural networks, these new frameworks were applied to the task, sharing with SVMs the necessity of a quality vector representation for each document. One of the earliest attempts [32] focused on detecting irony in Twitter posts and Amazon product reviews, enhancing punctuation-based feature extraction with pattern identification within the text. The approach obtained good results but lacked adaptability, as only some pre-defined patterns were extracted, making the model "blind" to all others. Other approaches to enhancing document vectorization include leveraging the presence of emoji in social media texts [33], which turned out to be a strong indicator of irony in the studied dataset.

An interesting approach was taken by [34], where word embeddings vector properties were leveraged to gain insights on irony detection. Word embedding generators like Word2Vec [35] and GloVe[36] are able to understand word similarities and translate these common characteristics into vector features, usually making similar word embedding share high cosine similarity between each other. The researchers in [34] tried to leverage this embedding to detect context incongruity, the main characteristic of irony, to correctly identify irony in a test task.

Then, with the rise of Deep Learning architectures, the task was addressed again, leveraging these new frameworks' capabilities. Many of these new applications tried to leverage context information to enhance the document vector representation: [37] used context information that could be confidently inferred from the data, while [38] implements Long-Short-Term Memory (LSTM) framework for Recurrent Neural Networks (RNNs), a critical tool for modern Natural Language Processing, to model the conversation and understand the part that triggered the ironic reply. Results show that context analysis helps identify irony in conversational data. On the other hand, researchers implementing similar frameworks [28] tested it against non-common testing datasets and found that standalone models didn't perform as well. It has to be noted, in fact, that many of the models exposed previously were trained and tested on a particular evaluation dataset of "hashtag-labelled" tweets, *i.e.* tweets collected under particular hashtags that are commonly used for ironic

posts, in particular, *#irony, #sarcasm* and *#not*, in contrast with human-labelled tweets that resulted much harder for the model to identify correctly. This limitation was conducted to hashtag-labelled tweets having a "self-enclosed" irony, meaning that contextual information and the ironic statement can all be found in the same tweet. On the other hand, human-labelled ironic tweets usually have implicit contextual information that is easily available to a human reader but not to the model (for example, an ironic comment on politics). The paper proposed an approach to solve the issue by leveraging *transfer learning*, therefore enriching knowledge of the model through outside sources (in this case, sentiment analysis of the tweets), significantly improving the model's performance.
Large Language Models (LLMs) are the latest and biggest breakthrough in Natural Language Processing, and their application for the irony detection task can be found in the work of Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves and Luis Espinosa-Anke "*TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification*" [5], where they proposed a standardized testing framework for Twitter-specific evaluation tasks, composed of 7 separate tasks: emotion recognition, emoji prediction, hate speech detection, offensive language identification, sentiment analysis, stance detection (identifying how a tweets position itself in one of 5 domains: abortion, atheism, climate change, feminism and Hillary Clinton) and irony detection. Their work showed how RoBERTa[39], a Large Language Model developed by Facebook AI department based on BERT[40], when retrained on Twitter Data, performed better than selected baselines and two of its variants (a baseline RoBERTa model with no retraining and a RoBERTa architecture trained exclusively on Twitter data).

3 Methods

3.1 Research Question

Examining previous literature, it is visible how extensive research was made in the field of toxicity rating, but very little was made to go beyond highlighting differences in offensiveness perception by trying to understand the underlying causes of such differences. With this work, the objective is to build, test, and share a possible approach to analyzing a possible component of such differences, irony. The research question of this thesis can, therefore, be formulated as follows:

- *How does the perception of irony influence the differences in toxicity ratings among annotators from diverse social and cultural backgrounds?*

3.2 Dataset

The dataset that will be used for this analysis is the one created through the work of Nitesh Goyal et al. "*Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation*" [4], published on Kaggle under the name "Jigsaw Specialized Rater Pools Dataset" at https://www.kaggle.com/datasets/google/jigsaw-specialized-rater-pools-dataset. It collects 382,500 annotations of 25,500 unique comments coming from the Civil Comments dataset [24], each one having 15 annotations performed by raters belonging to one of three different groups ("African American", "LGBTQ" or "Control"), with 5 annotations per group per comment.
This dataset was chosen for its uniqueness and extreme relevance to the research question of this thesis, as it allows to study how raters from different social and cultural groups perceive toxicity in social media comments, leveraging the benefits given by having more than one annotation per group, allowing to catch variations within them.
Following an initial analysis of the data, 1533 entries were found lacking at least one toxicity rating, as the raters had the option to opt out from rating a certain comment by checking the box "*This comment is in a foreign language or not comprehensible for another reason (e.g., gibberish, different dialect etc.)*". As understanding toxicity rating is the objective of this thesis, these comments lacking it will be discarded from the analysis.

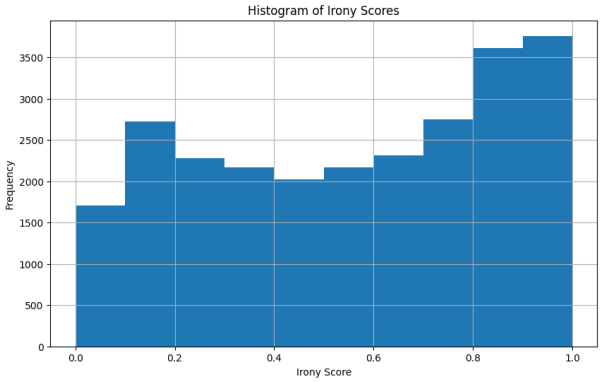

Figure 2: Histogram representing the distribution of irony scores assigned by the RoBERTa model to comments in the dataset.

3.3 Dataset Enhancement through Irony Scores

To perform the analysis irony scores need to be computed for comments in the dataset. In this case, these will be given through the RoBERTa-based LLM classifier developed by Francesco Barbieri et al. in "*TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification*" [5], publicly available through the model-sharing website Hugging Face at https://huggingface.co/cardiffnlp/twitter-roberta-base-irony?. This approach has some limitations which will be more thoroughly examined in section 5 "*Limitations & Suggestions for Future Research*", but it was deemed a good solution in terms of unbiasedness of ratings and cost efficiency.

This method produced irony scores for all comments that resulted nicely distributed in the range 0-1 (visible in Figure 2, with slight left skewing of the distribution to be expected as, due to both the comment selection method and the nature of these comments, it is expected that many of them will have high irony scores.

3.4 Regression Methods

3.4.1 The Ordinal Linear Regression Model

To understand the effects of our controls on the toxicity ratings given to various comments ordinal regression methods will be used. These frameworks assume the existence of an underlying continuous variable Y^* that is only observable through its segmentation into

categories that have an underlying order [41]. Imagine, for example, answers from a survey where the annual income of the person responding is given into a series of categories "*Under $60k*", "*Between $60k $100k*", "*Between $100k and $120k*" and "*Over $120k*": these are an expression of a continuous variable, the income of an individual, that is flattened to be expressed as one of these categories. Ordinal regression techniques allow a researcher to leverage existing regression frameworks to gain insights on the effects of independent variables on these dependent categories[42].

The Ordinal Logistical Regression model (`ologit`) is the most used method to do such studies. This model can be written as:

$$logit(\mathbf{P}(Y \leq j)) = \log\left(\frac{\mathbf{P}(Y \leq j)}{\mathbf{P}(Y > j)}\right) = \alpha_j - \boldsymbol{\beta} X,\ j \in [1, \ldots, J-1] \tag{1}$$

where α_j is the intercept specific to the category j^{th} category, $\boldsymbol{\beta}$ indicates the vector of fitted coefficients of the regression, X the dataset and J the total number of categories of the dependent variable. The negative sign in front of the fitted coefficients is introduced to better represent how most statistical software (like R or Stata) perform this regression, with $-\boldsymbol{\beta} = \boldsymbol{\eta}$ being the definition for the "real" fitted coefficients. What this model does is to fit a series of logistic regressions where all entries belonging to a certain category j or less are mapped to 0 and all others to a 1. Then, a simple logistic regression model is fitted, with an *ad hoc* intercept for each of these regressions and common coefficients for the independent variables.

Therefore, the model assumes that independent variables act equally across all levels, with the only thing varying across the `ologit` model being the intercepts, which work similarly to thresholds to determine to which class a certain entry should belong. This assumption is called the *Parallel lines* or *Proportional odds* assumption, as employing it would also mean that the odds of belonging to a class or its successor/predecessor for a particular entry are constant among all classes. Various tests are applicable to test this hypothesis, with the most widely used being the Brant test[43], developed by Rollin Brant in 1990. This test verifies that the assumption holds for the dataset by performing a separate regression for each of the comparisons (effectively performing a Multinomial Logistical Regression, or `mlogit`, which is more difficult to interpret) and then compares

the deviations of these coefficients from the ones obtained through the `ologit` model, creating test statistic to understand if the parallel lines assumption is violated (H_0) or verified (H_A).

3.4.2 Interpreting the `ologit` Model

Interpreting such models can be difficult due to their mathematical nature. Referring to Formula 1, we are supposed to be trying to understand how a binomial independent variable $\mathbf{x}_1$ acts on the output level $Y \in [1, 2, 3]$. As the Proportional Odds assumption is assumed, then the odds will not vary across categories. Therefore:

$$logit(\mathbf{P}(Y \leq j|x_1 = 1)) - logit(\mathbf{P}(Y \leq j|x_1 = 0)) = \alpha_j + \eta_1 - \alpha_j = \eta_1 = -\beta_1 \quad (2)$$

Then, it is possible to simplify the formula further by exponentiating both sides and leveraging the property of logarithms $\log(a) - \log(b) = \log(a/b)$:

$$\frac{\mathbf{P}(Y \leq j|x_1 = 1)}{\mathbf{P}(Y > j|x_1 = 1)} / \frac{\mathbf{P}(Y \leq j|x_1 = 0)}{\mathbf{P}(Y > j|x_1 = 0)} = \exp(-\beta_1) \quad (3)$$

To simplify notations, it is possible to rewrite $\frac{\mathbf{P}(Y \leq j|x_1=1)}{\mathbf{P}(Y>j|x_1=1)} = p_1/(1-p_1)$ and $\frac{\mathbf{P}(Y \leq j|x_1=0)}{\mathbf{P}(Y>j|x_1=0)} = p_0/(1-p_0)$. Then, as $\exp(-a) = 1/\exp(a)$:

$$\exp(\beta_1) = \frac{p_0/(1-p_0)}{p_1/(1-p_1)} \quad (4)$$

This means that for individuals where $x_1 = 1$ the odds of belonging to a class greater than j are $\frac{p_0/(1-p_0)}{p_1/(1-p_1)}$ the odds of individuals where $x_1 = 0$. A similar explanation can be applied to continuous regressors.

3.4.3 An Example for `ologit` Model Interpretation

This example was built by combining examples from "*Understanding and interpreting generalized ordered logit models*" by Richard Williams [42] and the UCLA: Statistical Consulting Group FAQ website [44].
Suppose that someone is trying to understand if it is possible to predict the answers to

a question from a survey expressed on a 5-point Likert scale given the individual's age in decades and gender. Running the `ologit` command in Stata on the dataset returned the following output:

Variable	Coefficient	Std. Err.	P-value
gender	0.967	0.0263432	0.000
age	-0.426	0.2393527	0.074
/cut1	0.3768424	0.1103421	-
/cut2	2.451855	0.1825628	-

Table 1: Results from example Ordered Logistic Regression

We can interpret these coefficients by applying the rules stated above:

- For the `gender` variable a coefficient of 0.967 implies an odds ratio of $\exp(0.967) = 2.631$, meaning that the odds of answering with a higher category are 2.63 times higher for the comparison gender (for example, female) with respect to the base one (therefore male). The result is also significant at all significance thresholds.
- For the `age` variable a coefficient of -0.426 (even if significant only at 10% significance level) results in an odds ratio of 0.653, meaning that an increase of 10 years in the age of an individual makes him 0.635 times less likely (or, in other words, reduce his likeliness by 34.7%) of answering with a higher category.

3.4.4 The Multinomial Logistical Regression Model

Alternatively to the `ologit` model it is possible to implement a Multinomial Logistical Regression (`mlogit`) model. This kind of model does not take into consideration the ordering of the various categories of the outcome variable and fits a series of `logit` models between all minus one categories of the output, with this excluded category being considered as a base, effectively producing $K - 1$ sets of fitted coefficients, one for each `logit` model. Doing so effectively discards the need for the Parallel Lines assumption, as in this kind of model each `logit` regression will be able to capture best the magnitude and direction of each of the explanatory variables. However, it does require the *Independence of Irrelevant Alternatives* (IAA in short) assumption, meaning that an individual should

stick to his choice no matter which are the choices available. This assumption is taken for granted in this work, as annotators did not have a fixed number of ratings that they could allocate between the comments. The formula for a Multinomial Logit model can be written as:

$$\mathbf{P}(Y = j) = \frac{\exp(\boldsymbol{\beta}_j X)}{\sum_{i=1}^{J} \exp(\boldsymbol{\beta}_i X)}, j \in [1, 2, \ldots, J] \tag{5}$$

This effectively defines a system of models having multiple possible solutions for the fitted parameters $\boldsymbol{\beta}_j$ that yield the same probabilities for each class. A base category must be established to find a unique solution and its coefficients are set to 0. Therefore, for an imaginary regression on a 3-categories dependent variable, the system of models will look as such:

$$\begin{cases} \mathbf{P}(Y = 1) & = \frac{1}{1+\exp(\boldsymbol{\beta}_2 X)+\exp(\boldsymbol{\beta}_3 X)} \\ \mathbf{P}(Y = 2) & = \frac{\exp(\boldsymbol{\beta}_2 X)}{1+\exp(\boldsymbol{\beta}_2 X)+\exp(\boldsymbol{\beta}_3 X)} \\ \mathbf{P}(Y = 3) & = \frac{\exp(\boldsymbol{\beta}_3 X)}{1+\exp(\boldsymbol{\beta}_2 X)+\exp(\boldsymbol{\beta}_3 X)} \end{cases} \tag{6}$$

3.4.5 Interpreting the `mlogit` Model

The `mlogit` model is not interpreted through odds ratios like the `ologit` model, but rather through ratios of relative risk (RRR), *i.e.* the ratio of risk in a group relative to a control group (that is, the base category of the regression) given a one-unit change in a predictor variable. A formal definition can be given through relative probabilities:

$$\text{RRR}_{ij} = \frac{\text{P}(Y = j \mid x_i + 1)}{\text{P}(Y = k \mid x_i + 1)} \Bigg/ \frac{\text{P}(Y = j \mid x_i)}{\text{P}(Y = k \mid x_i)} = e^{\beta_{ij}} \tag{7}$$

where $\text{P}(Y = j \mid x_i)$ is the probability of the outcome being in category j given the predictor x_i, $\text{P}(Y = k \mid x_i)$ is the probability of the outcome being in the reference category k given the predictor x_i and β_{ij} is the estimated coefficient for predictor x_i for outcome category j. Then, this ratio can be interpreted as:

- $RRR > 1$ points towards an increase in risk of belonging to category j with respect to base category k for a one-unit increase in variable x_1;
- $RRR < 1$ points towards a decrease in risk of belonging to category j with respect

to base category k for a one-unit increase in variable x_1;

- $RRR = 1$ indicates no increase nor decrease of risk of belonging to category j with respect to base category k for a one-unit increase in variable x_1.

3.4.6 An Example for `mlogit` Model Interpretation

This example is based on the official documentation for the `mlogit` Stata command [45]. Suppose someone wants to understand the relationship between the preferred means of transport of some individuals and some general demographics (gender and age). Performing the `mlogit` command on the survey's data yields the following results:

	Foot		Bus	
Variable	Coefficient	P-value	Coefficient	P-value
age	-0.011745	0.038	-0.007961	0.046
female	0.5616934	0.006	0.4518496	0.219

Table 2: Results from example Multinomial Logistic Regression

Assuming a baseline category "Car" used as a comparison, it is possible to interpret the results of the regression as follows:

- Variable `age`, which is statistically significant at 5% level for both comparisons, yields RRRs of 0.9883 and 0.9921, meaning that individuals are less likely to prefer walking or taking the bus with respect to taking the car as they get older;

- Variable `female`, which is statistically significant for the Foot vs. Car comparison but not for the Bus vs Car comparison, yields, respectively, RRRs of 1.7536 and 1.5710, meaning that an individual is statistically more likely to prefer walking over taking the car if it is a woman, while nothing statistically significant can be said on the Bus vs Car comparison, even if data suggests that women are more likely to prefer taking the bus than the car.

3.4.7 The Generalized Ordinal Logistical Regression Model

Between the two models described above lies the Generalized Ordinal Logistical Regression (`gologit`) model[42], also called the Partial Proportional Odds model. This framework is based on the standard `ologit` model but is able to relax the proportional odds assumptions only for some independent variables, effectively reducing the model's complexity and facilitating its interpretability with respect to the full `mlogit` model. The formula for such model, though being very similar to the `ologit` model one (Equation 1), can be written as

$$logit(\mathbf{P}(Y \leq j)) = \alpha_j - \boldsymbol{\beta}_j X, j \in [1, \ldots, J-1] \tag{8}$$

where the main change is the addition of an index to vector $\boldsymbol{\beta}_j$, as in this case, there will be some of the coefficients (the ones referring to independent variables that do not satisfy the Parallel Lines assumption) that will differ depending on the category, while others will be shared.

3.4.8 Interpreting the `gologit` Model

Interpretations of the results of this regression model can be performed identically to the `ologit` model (refer to Section 3.4.2 for a more accurate explanation).

3.4.9 An Example for `gologit` Model Interpretation

This example is based on "*Understanding and interpreting generalized ordered logit models*" by Richard Williams [42]. Suppose that someone wants to understand the relation between an answer to a political question, to be expressed on a 3-point Likert scale (Yes, Neutral or No), and the general demographics of the respondent (gender and age, expressed in decades). By running the `gologit` Stata command on the dataset of the survey, these are the results: For variables where the Parallel Lines assumption seems to be

Explanatory Variables	P-Value	Coef	Y vs Ne, No	Y, Ne vs No
Female	.843	0.037	-	-
Age	.001	-	-0.172	–0.071

Table 3: Results from Partial Proportional Odds Model

respected, the coefficient will be shared across all categories, while for variables where this assumption needs to be relaxed, coefficients for single comparisons are reported, together with a p-value representing a test of joint significance across all estimated coefficients for that variable. Results can be, therefore, interpreted as follows:

- The `gender` variable, although being not statistically relevant, reports a coefficient common across comparisons of 0.037, therefore with an odds ratio of $\exp(0.037) = 1,037$ that would indicate that women are slightly more likely to answer more negatively than men's;
- The `age` variable, statistically significant at under the 1% level, does not respect the Parallel Lines assumption. Nonetheless, its coefficients report RRRs of, respectively, 0.842 and 0.932, meaning that older individuals are more likely to answer negatively, but the effect of age is not of the same magnitude across comparisons.

4 Results and Analysis

4.1 The Model

The proposed relation for the following regressions is as follows:

$$toxic\_score = irony\_score + rater\_category + irony\_score \times rater\_category \quad (9)$$

This is a simplification of the model to allow for a simple understanding of the terms, while more specific model formulas can be found at Equations 1, 5 and 8. The rationale behind this setup is to allow for the rater categories to directly influence the irony score (following the findings of previous research on this dataset[4]) while adding information through the irony score given by the RoBERTa model and creating a way for rater categories and irony scores to interact as well through the interaction term. All the following regression will be performed on Stata 18.

4.2 Ordinal Logistic Regression and Interpretation

First, an `ologit` regression is performed on Stata through the namesake command, setting "Control" as the base category for *rater_cat*. Results are shown in Table 4. The model

Variable	**Coefficient**	**Std. Err.**	**z**	**P-value**
Irony	-0.6863	0.0196	-34.99	0.000
Rater Category				
African American	0.0667	0.0180	3.71	0.000
LGBTQ	0.0372	0.0179	2.08	0.037
Interaction Terms				
African American × Irony	0.0990	0.0280	3.54	0.000
LGBTQ × Irony	0.0310	0.0277	1.12	0.264
/cut1	-3.0116	0.0139		
/cut2	-1.3472	0.0127		
/cut3	-0.9331	0.0126		

Table 4: Results from Ordered Logistic Regression

was found statistically significant above the intercept-only model at all significance levels,

and results can be interpreted as follows:

- Irony acts significantly at all significance levels on the toxicity score, reporting an odds ratio of $\exp(-0.6863) = 0.5034$, meaning that, for a 1-unit increase in the irony score, individuals are approximately half as likely to rate the comment as less toxic.

- Coefficients for the rater categories, 0.0667 for the African American category and 0.0372 for the LGBTQ one, are found to be both significant respectively at all significance levels and at the 5% significance level, a result that is slightly inconsistent with previous works on this dataset[4] where only the LGBTQ coefficient was found significant. The two coefficients can be interpreted as follows:

  - The coefficient for the African American pool yields an odds ratio of $\exp(0.0667) = 1.069$, meaning that a rater in the AA rater pool was slightly more likely to rate a comment as less toxic with respect to a rater in the control pool.
  - The coefficient for the LGBTQ pool yields an odds ratio of $\exp 0(.0372) = 1.038$, meaning that a rater in this pool was also slightly more likely to rate the same comment as less toxic with respect to a rater in the control pool.

- The interaction terms between *rater_cat* and *irony_score* were not found both significant, with the one referring to the AA pool significant at all significance levels and the one referring to the LGBTQ pool having a p-value of 0.264. The first one, having a coefficient of 0.0990, yields an odds ratio of $\exp(0.0990) = 1.104$, meaning that a rater in the African American rater pool, given a one-unit change in the irony score, is more likely to give a lower toxicity rating to a comment than a rater in the control pool.

Overall, irony seems to have a statistically significant predicting power over the dependent variable. However, performing a test for the Parallel Lines assumption on the `ologit` regression yielded test statistic of $\chi^2(10) = 328.74, p\text{-value} < 0.001$, meaning that the null hypothesis (that is, the Parallel Lines assumption holds) is rejected at all significance levels. This means that the results for this model are not statistically relevant, and a Generalized Linear Model would be a better choice for our regression.

4.3 Multinomial Logistic Regression and Interpretation

Initially, an attempt was made to fit a Generalized Ordinal Logistic model due to its better performance and ease of interpretability, as explained in Section 3.4.7. The search for the best optimal combination of constraints (*i.e.*, to understand for which independent variables the Parallel Lines assumption needed to be relaxed and for which it was not a necessity), on the other hand, yielded a Multinomial Logistic Regression model, as none of the variables respected the necessary assumption. Results of this regression can be found in Table 5.

Variable	**Threshold**	**Coefficient**	**Std. Err.**	**z**	**P-value**
Irony	-2	-0.5437	0.0380	-14.32	0.000
African American	-2	0.2325	0.0372	6.25	0.000
LGBTQ	-2	0.1428	0.0365	3.91	0.000
African American × Irony	-2	-0.0557	0.0563	-0.99	0.322
LGBTQ × Irony	-2	-0.0080	0.0553	-0.14	0.886
Constant	-2	2.8760	0.0250	115.13	0.000
Irony	-1	-0.7143	0.0216	-33.03	0.000
African American	-1	0.1238	0.0202	6.12	0.000
LGBTQ	-1	0.1104	0.0201	5.50	0.000
African American × Irony	-1	0.0553	0.0311	1.78	0.076
LGBTQ × Irony	-1	-0.0203	0.0309	-0.66	0.512
Constant	-1	1.3393	0.0140	95.62	0.000
Irony	0	-0.6801	0.0201	-33.84	0.000
African American	0	0.0416	0.0183	2.27	0.023
LGBTQ	0	0.0126	0.0182	0.69	0.490
African American × Irony	0	0.1121	0.0286	3.92	0.000
LGBTQ × Irony	0	0.0336	0.0284	1.18	0.238
Constant	0	0.9437	0.0128	73.48	0.000

Table 5: Multinomial Logistic Regression Estimates

These results should be interpreted as ratios of relative risk, therefore:

- The `irony` variable yields three coefficients statistically significant at all significance levels, with ratios of relative risk being $RRR_{irony,-2} = 0.5807$, $RRR_{irony,-1} = 0.4894$, $RRR_{irony,0} = 0.5062$, indicating that an increase in irony score yields, across

all levels, a halving of the probability that the comment is assigned to the comparison category 0, *i.e.* "Not Toxic".

- The coefficients of the variable indicating if a rater belongs to the African American rater pool were found significant for all comparisons at the 5% significance level. These coefficients yielded $RRR_{AA,-2} = 1.2618, RRR_{AA,-1} = 1.1317, RRR_{AA,0} = 1.0425$, meaning that a rater in the African American pool was overall more likely to assign a comment the "Not Toxic" rating.

- The coefficients for the LGBTQ rater pool were not found significant in all comparisons, as only -2 vs. 1 and -1 vs. 1 comparisons were statistically significant at all significance levels. This is not unexpected, as category 0 being labelled as "Unsure" could bring less statistically relevant results (more on this in Section 5.1.1). The two statistically relevant coefficients produced ratios of relative risk $RRR_{LGBTQ,-2} = 1.1535, RRR_{LGBTQ,-1} = 1.1168$, indicating that raters belonging to this specialized pool were overall more likely to assign a comment the "Not Toxic" rating.

- Interaction terms were overall found not significant, with the only one significant at all levels being the interaction between African American pool members and the irony score in the 0 vs. 1 comparison, yielding $RRR_{AA\times irony,0} = 1.1186$, suggesting that members of this specialized rater pool felt less the effect of irony in assigning the "Unsure" rating with respect to the "Not Toxic" one. However, this effect can be considered insignificant due to critiques exposed in Section 5.1.1.

Overall, this dataset did not show significantly different findings from the Ordinal Logistic Regression ones, only offering greater insights into the magnitude of these effects between comparisons.

4.4 Generalized Ordinal Logistic Regression and Interpretation

Following the result of the original paper [4], a `gologit` model was run, assuming that the Parallel Lines assumption holds for the impact of rater categories. Results can be found in Table 6.

Variable	Threshold	Coefficient	Std. Err.	z	P-value
Rater Category	-2				
African American		0.0688	0.0180	3.83	0.000
LGBTQ		0.0403	0.0179	2.25	0.024
Irony	-2	-0.6571	0.0299	-22.00	0.000
African American × Irony	-2	0.1659	0.0347	4.79	0.000
LGBTQ × Irony	-2	0.1302	0.0343	3.80	0.000
Constant	-2	2.9600	0.0183	162.08	0.000
Rater Category	-1				
African American		0.0688	0.0180	3.83	0.000
LGBTQ		0.0403	0.0179	2.25	0.024
Irony	-1	-0.7678	0.0206	-37.33	0.000
African American × Irony	-1	0.1287	0.0285	4.52	0.000
LGBTQ × Irony	-1	0.0740	0.0283	2.62	0.009
Constant	-1	1.3794	0.0132	104.77	0.000
Rater Category	0				
African American		0.0688	0.0180	3.83	0.000
LGBTQ		0.0403	0.0179	2.25	0.024
Irony	0	-0.6542	0.0199	-32.90	0.000
African American × Irony	0	0.0733	0.0282	2.60	0.009
LGBTQ × Irony	0	-0.0058	0.0280	-0.21	0.835
Constant	0	0.9255	0.0127	72.91	0.000

Table 6: Generalized Ordered Logit Estimates with Partial Proportional Odds

It is possible to notice that this new regression did not indicate any result diverging from previous analyses; however, it helped to improve the overall significance of the coefficients: in this last regression, all coefficients resulted statistically significant at the 5% significance level, with directions and magnitude coherent with previous results. These summarized findings can be found in Table 7.

4.5 Logistic Regression and Interpretation

Finally, it was decided to run a simple Logistic Regression (`logit`) model to validate our findings further. In order to do so, the outcome variable needed to be encoded into a binary one, and this was done by grouping together the "Toxic" and "Very Toxic"

Variable	-2 vs. 1	-1 vs. 1	0 vs. 1
African American	↑ (7.13%)	↑ (7.13%)	↑ (7.13%)
LGBTQ	↑ (4.11%)	↑ (4.11%)	↑ (4.11%)
Irony	↓ (48.08%)	↓ (46.36%)	↓ (48.01%)
African American × Irony	↑ (18.06%)	↑ (13.74%)	↑ (7.62%)
LGBTQ × Irony	↑ (13.90%)	↑ (7.68%)	↓ (0.58%)
Constant	↑ (96.32%)	↑ (98.56%)	↑ (96.32%)

Table 7: Summary of Changes in Probability for Generalized Ordered Logit Model

categories (encoding them as 1) and the "Unsure" and "Not Toxic" categories (encoding them as 0). The independent variables were left untouched. Therefore, the reference model still follows Equation 9. Results of this regression can be found in Table 8.

Variable	Coefficient	Std. Err.	z	P-value
Irony	0.7125	0.0216	33.01	0.000
Rater Category				
African American	-0.1241	0.0202	-6.14	0.000
LGBTQ	-0.1115	0.0201	-5.55	0.000
Interaction Terms				
African American × Irony	-0.0548	0.0311	-1.76	0.078
LGBTQ × Irony	0.0222	0.0309	0.72	0.472
Constant	-1.3382	0.0140	-95.71	0.000

Table 8: Logistic Regression Results for Binary Toxic Score

Interpreting the results of this model closely matches the interpretation method of the `ologit` model (refer to Section 3.4.2), but the formula of the `logit` model has only positive signs, so coefficient signs in this regression will be swapped but maintain the same meaning as in the other regressions.

Overall, results closely match previous findings, with coefficients referring to irony scores and rater categories statistically significant at all significance levels; interaction terms, however, were not found statistically significant. The direction and magnitudes of the effects of these variables match the results of previous regressions (refer to the sections above for more detailed insights).

5 Limitatons & Suggestions for Future Research

5.1 Limitations

In this section, the limitations of the chosen approach will be analyzed to understand which steps can be taken to improve the significance of the analysis results.

5.1.1 Dataset

The chosen dataset for the analysis is unique in the panorama of publicly available data sources on the topic, as rater anonymization performed by most companies before publishing the dataset and/or delivering it to the researchers does not allow to gain insights on the rater's identity. However, some downsides must be noted:

- The dataset does not provide specifications on the object of the comment, *i.e.* to whom the comment is referred, even if this information was used to select the comments from the Civil Comments dataset. Having this additional information could allow for better insights regarding how the rater's identity and irony interact by controlling for the object of the insult (a possible hypothesis could be that individuals will rate a comment as more toxic if the insult is directed towards their community).

- Specialized rater pools for this experiment are not mutually excluding[4], and this decision caused the dataset creators to consider the intersection between the two classes. A possible solution could have been to have multiple levels of classification per rater (*e.g.*, ethnicity and gender) and then perform the analysis within these classifications, confronting results within each categorization.

- The fact that option 0 appeared with the "Unsure" label to raters during the annotation work could have led to misannotations by the latter, as better labelling for this option would have been "Neither Toxic Nor Not Toxic".

5.1.2 Irony Scoring Methodology

As described in Section 2.4, automatic irony detection is not an exact science, depending heavily on the sensibility of the annotators creating the dataset to train the model and

the model specification themselves. The RoBERTa-based irony detection model chosen to support this analysis was selected due to its online open-source availability, ease of use and the absence of reported biases specific to the model, facilitating the re-creation of the analysis for those who wanted it. However, two main downsides of this approach must be noted:

- First, this model is created by retraining a general-purpose RoBERTa model on Twitter data, an approach that granted it good performance on the task it was built for. However, the comments of the dataset used for the analysis do not come from Twitter and, though being similar in format and also coming from an online source, might reduce the performance of the model based on the fact that these kinds of LLMs seem to perform better if, after a general training, they are retrained on domain-specific comments[5].

- Second, as the objective of the analysis is to understand the role of irony in toxicity detection, but also how irony influences the toxicity score given to a certain comment depending on the community the rater belongs to, a better approach could be to have the comments rated for irony by either the same annotator that provides the toxicity score or by a new set of specialized rater pools matching the characteristics of the ones that performed the toxicity scoring. Both of the alternative approaches have their pros but also cons (for example, the first alternative method could result in overloading the workload of annotators[10], reducing the overall quality of the annotations). Still, they are both valid examples of how this work could be improved.

Both downsides were thoroughly analyzed before the start of the analysis. While the proposed approaches for their solution were considered unfeasible due to logistical or resource constraints, they hold significant potential. Overall, the LLM-based scoring method proposed in this work is deemed the best approach possible, but alternative methods could be able to yield even better results.

5.1.3 Model Specifications

It is difficult to find quantitative works on the effect that irony has on an individual's perception, as most of the current work on it focuses on its detection (refer to Section

2.4). This causes a lack of reference points for statistical models, making it difficult to assess the quality of the model's specification. The decision to go with a simple model with interaction terms arises from the technical difficulties caused by defining, justifying and running more complex models. However, there are ample possibilities for higher complexity models to be tested by adding higher grade terms and/or more control variables. While logistic models were deemed suitable for this research due to their mathematical frameworks and capabilities, it's worth considering more complex models for future research. Despite their complexity, alternative formulations for the Generalized Ordinal Logistic Regression model [46] could potentially provide more nuanced insights into the relationship between irony and toxicity.

5.2 Future Research Directions

This section serves as a catalyst for possible improvements in future works on this topic and an inspiration for other researchers to delve into related fields and issues. With Section 5.1 highlighting the pain points of the approach taken in this work, researchers seeking deeper insights into this topic are urged to start by enhancing the dataset with additional features. This includes, but is not limited to, a higher number of categories for the specialized rater pools, additional information on the object of the comment rated (not to be shown to raters but to be added as a control during regressions), and human-labelled irony scores for the comments in the dataset. Furthermore, exploring different regression formulations and utilising different regression methods is strongly encouraged. Finally, this dataset provides additional labels for each comment, inquiring about the specific type of toxic language perceived by the rater: these labels, not analyzed in this thesis, could provide additional findings on how irony and the rater's category interact in explaining differences in those ratings as well.

6 Conclusions

This work aimed at understanding if differences in toxicity ratings given by annotators belonging to different social and cultural groups could be explained through the introduction of irony scores in the statistical model, further enhancing previous works on the topic[4]. This analysis was performed by leveraging the "Jigsaw Specialized Rater Pools Dataset", enriched through a RoBERTa-based irony detector[5], and by performing several types of logistical regressions to validate our findings.

Results align with previous findings on this dataset regarding the impact of the rater identity on the toxicity rating, showing that ratings given by a specialized rater pool differ statistically significantly from ratings given by a control and, therefore, more general, rater pool. Irony was found to be a significant explanatory variable for the rating of the comment, reducing its likeliness to be assigned to lower toxicity ratings. Interaction terms between these two explanatory variables were not found significant in all models, but allowed to gain further insights into how these two terms interact with each other, with irony reducing its effect on the toxicity rating in most of the cases, and especially for African American raters.

The chosen approach has limitations, as constraints in available resources determined many choices between dataset selection, irony scoring and model definition. However, these limitations do not reduce the significance of the findings and can be used to further research on the topic, trying different datasets and model specifications to understand more and better what plays into action when studying differences in toxicity perception between individuals.

These findings can greatly impact how people consider toxicity, especially in the social media context, and improve individuals' understanding of how others perceive toxicity. Practical applications of these results can be found in training and fine-tuning of automatic moderation systems used by social media networks, training of moderators for online communities and, in general, spreading awareness regarding the existing differences in perception of toxicity so that these can be taken into consideration both in academical and real-world applications.

References

[1] Josef Breuer and Sigmund Freud. *Ueber den psychischen Mechanismus hysterischer Phänomene:(Vorläufige Mittheilung)*. Veit & Company, 1893.

[2] Eli Omernick and Sara Owsley Sood. "The Impact of Anonymity in Online Communities". In: *2013 International Conference on Social Computing*. 2013, pp. 526–535. DOI: 10.1109/SocialCom.2013.80.

[3] Tiziana De Giorgio. "Alla Bocconi sospesi tre studenti per i commenti sui bagni gender neutral. Sui social le offese: "Li puoi usare per andare a trans"". In: *La Repubblica* (Feb. 2024).

[4] Nitesh Goyal et al. *Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation*. 2022. arXiv: 2205.00501 [cs.HC].

[5] Francesco Barbieri et al. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. DOI: 10.18653/v1/2020.findings-emnlp.148. URL: https://aclanthology.org/2020.findings-emnlp.148.

[6] Tony McEnery. *Corpus linguistics*. Edinburgh University Press, 2019.

[7] Matthew Petrillo and Jessica Baycroft. "Introduction to manual annotation". In: *Fairview research* (2010), pp. 1–7.

[8] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Computational Linguistics* 19.2 (1993). Ed. by Julia Hirschberg, pp. 313–330. URL: https://aclanthology.org/J93-2004.

[9] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. Accessed on April 15, 2024. 2012. URL: https://learning.oreilly.com/library/view/-/9781449332693/.

[10] Marta Sabou et al. “Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 859–866. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.

[11] Massimo Poesio et al. “Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation”. In: *ACM Trans. Interact. Intell. Syst.* 3.1 (Apr. 2013). ISSN: 2160-6455. DOI: 10.1145/2448116.2448119. URL: https://doi.org/10.1145/2448116.2448119.

[12] *Definition of TOXIC — merriam-webster.com*. https://www.merriam-webster.com/dictionary/toxic. [Accessed 30-04-2024].

[13] *Annotation instructions for Toxicity with sub-attributes*. https://github.com/conversationai/conversationai.github.io/blob/main/crowdsourcing_annotation_schemes/toxicity_with_subattributes.md. [Accessed 30-04-2024].

[14] Chikashi Nobata et al. “Abusive Language Detection in Online User Content”. In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 145–153. ISBN: 9781450341431. DOI: 10.1145/2872427.2883062. URL: https://doi.org/10.1145/2872427.2883062.

[15] Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. “Using crowdsourcing to improve profanity detection”. In: *2012 AAAI Spring Symposium Series*. 2012.

[16] Sanaz Jabbari, Ben Allison, and Louise Guthrie. “Using a Probabilistic Model of Context to Detect Word Obfuscation.” In: *LREC*. Citeseer. 2008.

[17] Marcos Zampieri et al. “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)”. In: *CoRR* abs/1903.08983 (2019). arXiv: 1903.08983. URL: http://arxiv.org/abs/1903.08983.

[18] John Pavlopoulos et al. “Improved Abusive Comment Moderation with User Embeddings”. In: *CoRR* abs/1708.03699 (2017). arXiv: 1708.03699. URL: http://arxiv.org/abs/1708.03699.

[19] Marcos Zampieri et al. “Predicting the Type and Target of Offensive Posts in Social Media”. In: *CoRR* abs/1902.09666 (2019). arXiv: 1902.09666. URL: http://arxiv.org/abs/1902.09666.

[20] Dawei Yin et al. “Detection of harassment on web 2.0”. In: *Proceedings of the Content Analysis in the WEB* 2.0 (2009), pp. 1–7.

[21] Ying Chen et al. “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”. In: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 2012, pp. 71–80. DOI: 10.1109/SocialCom-PASSAT.2012.55.

[22] William Warner and Julia Hirschberg. “Detecting hate speech on the world wide web”. In: *Proceedings of the second workshop on language in social media*. 2012, pp. 19–26.

[23] David Yarowsky. “Unsupervised word sense disambiguation rivaling supervised methods”. In: *33rd annual meeting of the association for computational linguistics*. 1995, pp. 189–196.

[24] Daniel Borkan et al. “Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW ’19. San Francisco, USA: Association for Computing Machinery, 2019, pp. 491–500. ISBN: 9781450366755. DOI: 10.1145/3308560.3317593. URL: https://doi.org/10.1145/3308560.3317593.

[25] Rensis Likert. “A technique for the measurement of attitudes.” In: *Archives of Psychology* 22.140 (1932), p. 55.

[26] Alan Agresti. *Categorical data analysis*. Vol. 792. John Wiley & Sons, 2012.

[27] Meyer Howard Abrams. *A glossary of literary terms*. Cengage, 2018.

[28] Shiwei Zhang et al. “Irony detection via sentiment-based transfer learning”. In: *Information Processing & Management* 56.5 (2019), pp. 1633–1644. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.04.006. URL: https://www.sciencedirect.com/science/article/pii/S0306457318307428.

[29] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. "From humor recognition to irony detection: The figurative language of social media". In: *Data & Knowledge Engineering* 74 (2012). Applications of Natural Language to Information Systems, pp. 1–12. ISSN: 0169-023X. DOI: https://doi.org/10.1016/j.datak.2012.02.005. URL: https://www.sciencedirect.com/science/article/pii/S0169023X12000237.

[30] Salvatore Attardo. "Irony as relevant inappropriateness". In: *Journal of Pragmatics* 32.6 (2000), pp. 793–826. ISSN: 0378-2166. DOI: https://doi.org/10.1016/S0378-2166(99)00070-3. URL: https://www.sciencedirect.com/science/article/pii/S0378216699000703.

[31] Clint Burfoot and Timothy Baldwin. "Automatic satire detection: Are you having a laugh?" In: *Proceedings of the ACL-IJCNLP 2009 conference short papers*. 2009, pp. 161–164.

[32] Dmitry Davidov, Oren Tsur, and Ari Rappoport. "Semi-supervised recognition of sarcasm in Twitter and Amazon". In: *Proceedings of the fourteenth conference on computational natural language learning*. 2010, pp. 107–116.

[33] Paula Carvalho et al. "Clues for detecting irony in user-generated contents: oh...!! it's "so easy";-)". In: *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*. TSA '09. Hong Kong, China: Association for Computing Machinery, 2009, pp. 53–56. ISBN: 9781605588056. DOI: 10.1145/1651461.1651471. URL: https://doi.org/10.1145/1651461.1651471.

[34] Aditya Joshi et al. "Are Word Embedding-based Features Useful for Sarcasm Detection?" In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1006–1011. DOI: 10.18653/v1/D16-1104. URL: https://aclanthology.org/D16-1104.

[35] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[37] Aniruddha Ghosh and Tony Veale. "Magnets for Sarcasm: Making Sarcasm Detection Timely, Contextual and Very Personal". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 482–491. DOI: 10.18653/v1/D17-1050. URL: https://aclanthology.org/D17-1050.

[38] Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. "The Role of Conversation Context for Sarcasm Detection in Online Interactions". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Ed. by Kristiina Jokinen et al. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 186–196. DOI: 10.18653/v1/W17-5523. URL: https://aclanthology.org/W17-5523.

[39] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[40] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[41] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Vol. 7. Stata press, 2006.

[42] Richard Williams. "Understanding and interpreting generalized ordered logit models". In: *The Journal of Mathematical Sociology* 40.1 (2016), pp. 7–20. DOI: 10.1080/0022250X.2015.1112384. eprint: https://doi.org/10.1080/0022250X.2015.1112384. URL: https://doi.org/10.1080/0022250X.2015.1112384.

[43] Rollin Brant. "Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression". In: *Biometrics* 46.4 (1990), pp. 1171–1178. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2532457 (visited on 05/22/2024).

[44] UCLA: Statistical Consulting Group. *FAQ: How Do I Interpret The Coefficients In An Ordinal Logistic Regression?* 2021. URL: https://stats.oarc.ucla.edu/other/mult-pkg/faq/ologit/ (visited on 05/23/2024).

[45] StataCorp. 2023. *Stata 18 Base Reference Manual.* College Station, TX: Stata Press.

[46] Bercedis Peterson and Frank E. Harrell Jr. "Partial Proportional Odds Models for Ordinal Response Variables". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 39.2 (1990), pp. 205–217. DOI: https://doi.org/10.2307/2347760. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2347760. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2347760.