Dead or Murdered? Predicting Responsibility Perception in Femicide News Reports

Gosse Minnema[a], **Sara Gemelli**[b], **Chiara Zanchi**[b], **Tommaso Caselli**[a], **Malvina Nissim**[a]

[a]University of Groningen, The Netherlands
[b]University of Pavia, Italy
g.f.minnema@rug.nl

Abstract

Different linguistic expressions can conceptualize the same event from different viewpoints by emphasizing certain participants over others. Here, we investigate a case where this has social consequences: how do linguistic expressions of gender-based violence (GBV) influence who we perceive as responsible? We build on previous psycholinguistic research in this area and conduct a large-scale perception survey of GBV descriptions automatically extracted from a corpus of Italian newspapers. We then train regression models that predict the salience of GBV participants with respect to different dimensions of perceived responsibility. Our best model (fine-tuned BERT) shows solid overall performance, with large differences between dimensions and participants: salient *focus* is more predictable than salient *blame*, and perpetrators' salience is more predictable than victims' salience. Experiments with ridge regression models using different representations show that features based on linguistic theory perform similarly to word-based features. Overall, we show that different linguistic choices do trigger different perceptions of responsibility, and that such perceptions can be modelled automatically. This work can be a core instrument to raise awareness of the consequences of different perspectivizations in the general public and in news producers alike.

1 Introduction and background

The same event can be described in many different ways, according to who reports on it, and the choices they make. They can opt for some words rather than others, for example, or they can use a passive rather than an active construction, or more widely, they can – consciously or not – provide the reader with a specific perspective over what happened.

Such choices do not just pertain to the realm of stylistic subtleties; rather, they can have substantial consequences on how we think of – or *frame* – events. Indeed, it is known that the way a piece of news is written, especially in terms of perspective-taking, heavily influences the way readers perceive *attribution of responsibility* in the events described (Iyengar, 1994). Figure 1[1] illustrates how the same event can be reported on from different viewpoints, in ways that do affect the perception of the participants' responsibilities. We are interested in unpacking *responsibility attribution* using NLP tools in the context of a socially relevant phenomenon, namely gender-based violence (GBV).

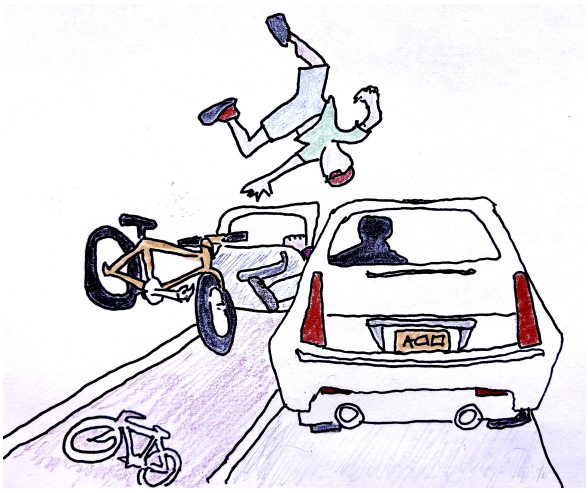


Figure 1: "Cyclist slams into car door"
Figure 1: "Car driver opens door and hits cyclist"
Figure 1: "Cyclist injured in road accident on 5th Street"
Figure 1: "Collision between bike and car"
We use alternative captions to illustrate how the same event can be described from alternative perspectives, which can evoke different perceptions in the attribution of responsibility to the actors involved.

Violence against women is worryingly common and therefore often reported in the news. A report by the European parliament (Corradi, 2021) details an estimate of 87,000 women intentionally

[1]Drawing inspired by the illustration in https://www.outsideonline.com/culture/opinion/look-you-open-your-car-door/ (accessed 2022-09-22).

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1078–1090
November 20–23, 2022.

killed in 2017. While Italy is listed in this report as one of the European countries with the lowest number of femicides, they are still too frequent and have been constant in the last 25 years (0.6 per 100,000 women in 1982 and 0.4 per 100,000 in 2017). Most discouragingly, a report from November 2018 by two Italian research institutes points out that the stereotype of a shared responsibility between the violence victim and its perpetrator is still widespread among young generations: "56.8% of boys and 38.8% of girls believe that the female is at least partly responsible for the violence she has suffered" (*Laboratorio Adolescenza* and *Istituto IARD*, 2018).

Working on Italian news, Pinelli and Zanchi (2021) observe that in descriptions of femicides, the use of syntactic constructions with varying levels of transitivity – from transitive active constructions on one side of the spectrum, via passives and anticausatives to nominalization constructions on the other side – corresponds to various degrees of responsibility attributed to the (male) perpetrator. For example, while "*he killed her*" (active/transitive) makes the involvement of an active agent fully explicit, with "*she was killed (by him)*" (passive) the event is accessed via the patient shifting attention away from the agent, and expressions such as "*the murder*" or even "*the event*" (nominal construction) moves both participants to the background. In a related contribution, Meluzzi et al. (2021) investigate the impact of argument structure constructions on responsibility attributions by means of a survey on artificially-constructed GBV reports in Italian. Their results further confirm the findings of Pinelli and Zanchi (2021) on the effects of readers' perception on the agentivity and responsibility of the perpetrators and the victims. The outcomes of both studies is in line with previous work in psycholinguistics showing that in events involving violence (at any level), the linguistic backgrounding of agents hinders their responsibility and promotes victim blaming (Huttenlocher et al., 1968; Henley et al., 1995; Bohner, 2002; Gray and Wegner, 2009; Hart and Fuoli, 2020; Zhou et al., 2021).

Based on such framing choices, how will the general reader perceive the described event? Can we model such perceptions automatically? In this paper we aim to answer these questions, still focusing on descriptions of femicides in Italian news, and exploiting *frame semantics* (Fillmore, 2006) as a theoretical and practical tool, as well as most recent NLP approaches.

Using specific pre-selected semantic frames, automatically extracted using a state-of-the-art semantic parser (Xia et al., 2021), we identify descriptions of GBV events from Italian newspapers. On these descriptions we collect human judgements through a large-scale survey where we ask participants to read the texts and ascribe a degree of *perceived responsibility* to the perpetrator, the victim, or to some more abstract concept (e.g. "jealousy", "rage"). More details are provided in §2.

Next, we model perception of responsibility automatically by developing a battery of regression models (both from scratch as well as atop pretrained transformer models) exploiting a variety of linguistic cues which range from surface to frame-based features. The training objective of such models is the prediction of the human perception scores. We achieve a strong correlation with a transformer-based model. The fine-grained character both of the survey and the result analysis that we conducted also allows us to observe differences in prediction complexity for the various aspects that we consider. Modeling and evaluation are discussed in §3.

The results we obtain show that **different linguistic choices do indeed trigger different perceptions of responsibility, and that such perceptions can be modelled automatically.** This finding not only confirms previous research which was conducted (manually) on a much smaller scale, but also opens up the possibility to conduct large-scale analyses of texts exposing to both producers and consumers of texts which perspectivization strategies are at play and their effects.[2]

2 Femicide perception dataset

We designed an online questionnaire study in which participants were presented with sentences extracted from the RAI Femicides Corpus (Belluati, 2021), a collection of 2,734 news articles covering 937 confirmed femicide cases perpetrated in Italy in 2015-2017, and asked to rate the level of agentivity and responsibility expressed in each sentence. The results of the questionnaire demonstrate a clear effect of semantic frames and syntactic constructions on the perception of descriptions of femicides.

[2]Our data and code are available at gitlab.com/sociofillmore/perceived-perspective-prediction.

2.1 Question formulation

The level of responsibility ascribed to event participants can be expressed in multiple ways triggering different perceptions in the readers. Since responsibility is a complex concept, we break it down into three dimensions in order to make it (i) more understandable for our participants, and (ii) to get a more nuanced picture of readers' perceptions. The three dimensions are:

1. FOCUS: does the sentence *focus* on the agent or on something else?
2. CAUSE: does the sentence describe the event as being *caused* primarily by a human or by something else?
3. BLAME: does the sentence attribute *blame* to the agent or to something else?

Example	FOCUS	CAUSE	BLAME
	ascribed to the murderer		
Her fiancé brutally murdered her	+	+	+
Blinded by jealousy, he killed her	+	+	±
Her husband's jealousy killed her	+	−	±
Her blind love for him became fatal	±	−	−
A tragic incident occurred in Rome	−	−	−

Table 1: Hypothesized perceptual ratings relative to the murderer (examples are artificial)

Table 1 shows hypothesized ratings on these dimensions for a number of artificial examples, demonstrating that the three dimensions are closely related, but do not always match: for example, the first and second sentences both focus on the role of the murderer and describe his actions as the cause, but the second sentence arguably attributes less blame to the murderer by describing him as 'blinded' by jealousy, implying that he does not bear full responsibility to his actions. Note that the ratings presented in the table merely represent a hypothesis about how the sentences are likely to be perceived; perception is inherently subjective and these examples should not be taken as a 'gold standard' of any kind.

To put the amount of responsibility attributed to the murderer in perspective, we also asked readers about the perceived level of focus, causation, and blame placed on the victim, an object (e.g. a weapon), a concept or emotion (e.g. jealousy), or on nothing at all. For a given sentence, participants were asked to give ratings on a 5-point Likert scale to each of these categories. Participants also had the option to indicate that the sentence was irrelevant and skip answering it. The full set of questions is given in Table 2. Note that, taking into account preliminary results from a pilot study, the categories have been adapted slightly to each individual question: for example, we omitted the 'none' category for the focus dimension (since there always has to be focus on something), and in the 'cause' dimension we made the descriptions of each category slightly more elaborate.

2.2 Sentence selection

Relevant sentences were extracted from the corpus following a two-step process: First, occurrences of semantic frames were automatically extracted using the LOME parser (Xia et al., 2021). This information was combined with an automatic dependency parse using SpaCy (Honnibal et al., 2020) to classify syntactic constructions. For example, *he murdered her* would be classified as "KILLING:active" (KILLING frame, expressed with active syntax), *she died* as "DEATH:intransitive", and *the tragedy* as "CATASTROPHE:nonverbal".[3] In a second step, we selected *typical frames* (Vossen et al., 2020) that encode possible ways of expressing the murder event with various degrees of emphasis on the various participants, and randomly sampled sentences containing at least one of these frames. Typical frames were selected by manually annotating the example sentences from Pinelli and Zanchi (2021) with FrameNet frames, and selecting the frames evoked by words that refer to (or imply) the event of the death of the victim ("he *killed* her" she *died*", "she was found *dead*", "a tragic *incident*"). This yielded the set of frames {KILLING, DEATH, DEAD_OR_ALIVE, EVENT, CATASTROPHE }, all of which can be used to describe exactly the same event but with different levels of dynamism (being dead vs. dying), agentivity (killing vs. dying), and generality (someone dying vs. something happening). We excluded frames that refer to events that are related to but distinct from the murder itself, such as CAUSE_HARM and USE_FIREARM ("he *stabbed* her", "he *fired* his gun" – these may refer to the cause of death, but do not include the death itself), or OFFENSES ("he was charged with *murder*" – this refers to the crime as a judicial concept, not as a real-world event). We then sampled sentences from our corpus in such a way that we created a corpus with an equal num-

[3] In this context, "nonverbal" means 'without a verb'; in this example, *tragedy* is an event expressed by a noun.

Dimension	Question	Murderer	Victim	Object	Concept	None
FOCUS	*La frase concentra l'attenzione principalmente...*	*sull'assassino*	*sulla vittima*	*su un oggetto*	*su un concetto astratto o un'emozione*	-
	'The sentence puts most attention ...'	'on the assassin'	'on the victim'	'on an object'	'on an abstract concept or emotion'	-
CAUSE	*La morte della donna è descritta come ...*	*causata da un essere umano*	-	*causata da un oggetto (es. una pistola)*	*causata da un'emozione (es. gelosia)*	*spontanea, priva di un agente scatenante*
	'The murder of the woman is described as ... '	'caused by a human being'	-	'caused by an object (e.g. a gun)'	'caused by an emotion (e.g. jealousy)'	'spontaneous, without a triggering agent'
BLAME	*La frase accusa...*	*l'assassino*	*la vittima*	*un oggetto*	*un concetto astratto o un'emozione*	*nessuno*
	'The sentence accuses ...'	'the murderer'	'the victim'	'an object'	'an abstract concept or an emotion'	'no one'

Table 2: Question dimensions and attributes

ber of examples of each frame-construction pair, and equal numbers of headlines and body-text sentences.

2.3 Practical implementation

Given the considerable cognitive load of analyzing (sometimes complex) sentences as well as the emotional load of reading text about a heavy and distressing topic, participants were asked to provide ratings on only one dimension, for a set of 50 sentences. Furthermore, attempting to find a balance between the *depth* (number of annotations per sentence) and *breath* (total number of annotations) of our annotations, we decided to set a target of 10 participants for each sentence and each dimension, meaning that 30 participants are needed to fully annotate each block of 50 sentences.

In order to distribute participants evenly across sentence sets and dimensions, without knowing the response rate in advance, we created 60 groups (20 sets of 50 sentences [= 1,000 in total] $\times$ three dimensions) and assigned participants to groups on a rolling basis: one group was open at a time, and once the required number of participants was reached, it was automatically closed and the next group was opened. Once a group was full, we manually inspected the responses for completeness and quality. Due to the subjective nature of the task, there are no 'wrong' responses per se, but we considered responses to be of low quality if they met at least one of the following three criteria: (i) implausibly fast completion of the questionnaire,[4] (ii) suspicious patterns of marking sentences as irrelevant and skipping them (e.g. skipping many sentences in a row), or (iii) suspicious response patterns (e.g. always giving the same ratings to each sentence).

	participant scores	*all*		*female*		*male*	
		mean	*std*	*mean*	*std*	*mean*	*std*
blame	*murderer*	2.35	1.89	2.07	1.80	2.75	2.01
	victim	0.49	0.92	0.44	0.92	0.55	0.92
	object	0.46	1.01	0.44	1.02	0.50	0.99
	concept	0.82	1.30	0.83	1.33	0.79	1.25
	no-one	1.36	1.74	1.49	1.76	1.19	1.71
cause	*human*	3.51	1.68	3.54	1.67	3.48	1.69
	object	1.37	1.85	1.36	1.84	1.40	1.91
	concept	0.86	1.32	0.88	1.31	0.76	1.34
	no-one	1.59	1.59	1.58	1.59	1.61	1.58
focus	*murderer*	2.26	1.94	2.23	1.91	2.30	1.97
	victim	2.85	1.60	2.68	1.59	3.07	1.61
	object	1.35	1.65	1.33	1.65	1.39	1.65
	concept	1.65	1.65	1.56	1.69	1.76	1.59

Table 3: Summary of perception scores per question and attribute

The link to the survey platform[5] was distributed amongst university students enrolled in bachelor's and master's degrees in different programs at several universities in Italy. Responses were collected anonymously, but participants were asked to state their gender, age, and profession.

2.4 Results

Our final dataset covers 400 sentences with ratings from 240 participants in total (153 identifying as female, 86 as male, 1 as non-binary; mean age 23.4). In Table 3, a summary of the perception scores aggregated across sentences is given. We give both the mean score (in green, on a scale from 0-5), averaged over all participants and all sentences, and the standard deviation of averaged scores across sentences. Overall, the attributes corresponding to

[4]We considered responses 'too fast' if they took less than 6 minutes (for 50 sentences, i.e. 7 sec./sentence, not including time spent reading instructions).

[5]We used Qualtrics (https://www.qualtrics.com/) to present stimuli and collect responses, alongside an in-house system for managing participants and payments.

the perpetrator tend to have higher average scores but also more variance than the other attributes (except *focus/victim*, which has a higher average but lower variance). More details about the distribution of scores per question and attribute are given in the Appendix. Due to the inherently subjective nature of the task, and in line with previous studies on perceptual norms (e.g., Brysbaert et al. 2014), we did not calculate inter-annotator agreement scores.

Table 4 (reproduced from Minnema et al. 2022) shows average scores for the *focus* question, split by typical frame and construction. This shows significant effects: sentences containing the KILLING frame tend to put higher focus on the murderer, and substantially more so when using an active construction. Meanwhile, the use of the CATASTROPHE, DEAD_OR_ALIVE, and DEATH frames, as well as the KILLING frame used in an active or passive construction increases the focus on the victim. On the other hand, there were no significant differences in focus scores for the object, and significant but smaller differences in focus on a concept or emotion. In each of these cases, the findings correspond to what we expected based on linguistic theory: if an event participant is lexically encoded in the predicate and syntactically required to be expressed, it is more likely that this participant will be perceived as being under focus. More focus on the murderer and the victim was also expected, both based on the content of the sentences, and on the fact that several frames (e.g. KILLING) lexically encode the presence of a victim and/or a killer, but not necessarily that of an inanimate concept or emotion (possibly except CATASTROPHE).

frame/construction	murderer**	victim**	object	concept / emotion*
CATASTROPHE				
nonverbal	1.319	2.713	0.760	2.190
DEAD_OR_ALIVE				
nonverbal	1.195	3.387	1.386	1.993
intransitive	1.983	3.529	1.566	1.539
DEATH				
nonverbal	0.967	3.247	1.507	1.914
intransitive	1.867	3.921	1.690	1.286
EVENT				
nonverbal	1.431	1.503	1.186	2.339
impersonal	1.169	2.201	1.309	1.949
KILLING				
nonverbal	2.007	2.387	1.032	1.673
other	2.410	2.345	1.198	1.663
active	3.897	2.659	1.570	1.651
passive	1.947	3.425	1.491	1.315

Table 4: Mean perception scores for "the main focus is on X". '*' = differences between frame-construction pairs are significant at $\alpha = 0.05$, '**' = significant at $\alpha = 0.001$ (Kruskal-Wallis non-parametric H-test). Cells with a value > 2.5 are highlighted in green.

3 Perception score prediction

In this section, we introduce models for automatically predicting femicide perception scores, as well as a suite of evaluation measures for evaluating these models. We model our task as a multi-output regression task: given a sentence S, we want to predict a perception vector $\vec{p}$, in which every entry p_i represents the value of a particular Likert dimension from the questionnaire (e.g. *'blame on the victim'*, *'focus on an object'*).

3.1 Participant aggregation

In order to train a single model that generalizes over individual participants, we first z-score the perception values for each sentence and each participant and then take the average value across participants. Z-scores are calculated separately for each Likert dimension and participant to account for two types of variability: i) *within-dimension score intensity preference* and ii) *between-dimension preference*. Type (i) refers to different participants making different use of the score range: depending on confidence levels and other factors, participants might choose to make heavy use of the extremities of the range (e.g. very often assign '0' or '5') or concentrate most of their scores in a particular part of the range (i.e. around the center or near the high or low end). Type (ii) refers to the possibility of participants having a tendency to always assign higher or lower scores to particular dimensions. For example, some participants may always give a higher score to 'blame on the murderer' vs. 'blame on the victim'. By performing regression towards z-scored perception values, we force our models to predict *between-sentence variability*: we are most interested in predicting how each sentence is perceived relative to other sentences (e.g., does this sentence put above-average blame on the victim? below-average focus on the murderer?) and less in absolute scores since these are highly subjective and depend on many individual biases.

3.2 Metrics

We evaluate our multi-output regression problem from several angles. First, we use ***Root Mean Squared Error (RMSE)*** to measure error rates. This is complemented by $\boldsymbol{R^2}$, which estimates the proportion of variation in the perception scores

that is explained by the regression models. R^2 is defined both for each dimension and as an average over dimensions. Next, ***Cosine (COS)*** measures the cosine similarity between the gold and predicted vectors of perception values and provides an estimate of how well the relations between the dimensions are preserved in the mapping.

An alternative interpretation is the ***Most Salient Attribute (MSA)*** metric: we evaluate regression as accuracy on the classification task of predicting which Likert dimension has the highest (z-scored) perception value for each question (implemented as simply computing argmax over the output dimensions corresponding to each question). For example, if for a particular sentence, "concept" is the highest-scoring dimension for the *blame* question, this means that "blame on a concept" is more salient in this sentence compared to other sentences. Note that the fact that z-scores were computed individually for each dimension makes a major difference here: the dimension with the highest z-scored value does not necessarily also have the highest absolute value. Similarly to the risks of assigning higher or lower scores to particular dimensions, in this case participants may give more points to "murderer" on the *blame* question than to "concept", even in sentences where "concept" is very salient. In such cases, "concept" would always have a lower absolute value than "murderer", but might have have a higher z-scored value in sentences where a relatively high score was given to "concept" and a relatively low one to "murderer".

3.3 Models

We compare two types of models: *ridge regression* models (a type of linear regression with L2 regularization) trained on different types of input features, and a selection of relevant pre-trained *transformer* models, fine-tuned for multi-output regression. For reference, we also run a 'dummy' baseline model that always predicts the training set mean for each variable.[6]

Features For the ridge models, we use a series of feature representations with increasing levels of richness. By comparing models trained on different representations, we gain insights into what kind of information is useful for predicting (different aspects of) perception scores. Features are divided into three categories: *Surface* features represent the lexical content of the input sentences, either with simple (unigram) ***bag-of-words (bow)*** vectors, or with pre-trained ***FastText (ft)*** embeddings (Grave et al., 2018).[7] By contrast, *Frames* features are based on the frame semantic parses of the sentence. The first variant, ***f1***, is similar to a bag-of-words, but using counts of any frame instances (e.g. *frm:Commerce_buy*) and semantic role instances (e.g. *rol:Commerce_buy:Seller*) present in the sentence instead of unigram counts. Variant ***f2*** is similar but includes only mentions of our pre-defined frames-of-interest (KILLING, DEATH, ...). Moreover, ***f1+*** and ***f2+*** are versions of *f1* and *f2* that concatenate the bag-of-frame features to the unigram features from ***bow***. Finally, *Sentence* features are transformer-derived sentence-level representations. ***SentenceBERT (sb)*** (Reimers and Gurevych, 2019) uses representations derived from XLM-R (Conneau et al., 2020);[8] ***BERT-IT Mean (bm)*** and ***XLM-R Mean (xm)*** use last-layer representations, averaged over tokens, from Italian BERT XXL and XLM-R, respectively.

Transformers We also implement a neural regression model that consists of a simple linear layer on top of a pre-trained transformer encoder.[9] We experiment with several variants of BERT with different pretraining corpora and model sizes. ***Italian BERT XXL Base (BERT-IT)*** is a base-size monolingual BERT model trained on the Italian Wikipedia and the OPUS corpus; ***BERTino*** is a distilled version of this model. We compare these with ***Multilingual BERT Base*** (Devlin et al., 2019) and ***Multilingual DistilBERT*** (Sanh et al., 2019), trained on concatenated Wikipedia dumps for 104 language, and ***XLM-RoBERTa Base*** (Conneau et al., 2020), trained on CommonCrawl data for 100 languages. We use cased models in all cases.

Implementation Ridge regression models were implemented using *scikit-learn* (Pedregosa et al., 2011). Transformer models were implemented using *Huggingface Transformers* (Wolf et al., 2019). We split the dataset into 75% training and 25%

[6] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyRegressor.html

[7] We chose FastText over competing static embedding models because of its ability to handle out-of-vocabulary tokens. Sentence-level representations were computed by taking the mean over all unigram vectors in the sentence, weighted by occurrence count (i.e., if a word occurs several times in a sentence, it will have a higher weight).

[8] We used pre-trained SentenceBERT models available from https://www.sbert.net/.

[9] Huggingface Hub links to the exact models used are provided in the Appendix.

test data. We used 6-fold cross-validation within the training set to search for hyperparameters (i.e., six models were trained for each possible setup): α for ridge regression; initial Adam learning rate and weight decay for transformers. The parameters with best performance across folds were then used for training the final model.

3.4 Results

Table 5 shows the main results on the test set for the RMSE, COS and R^2 metrics. Strongest results are obtained with the fine-tuned monolingual BERT models across all measures, with an overall R^2 scores around 0.45, meaning that these models explain almost half of the observed variance in perception scores. The multilingual BERT models (mBERT and XLM-R) perform consistently worse, with an average R^2 of 0.38 or below. Interestingly, we observe a drop in performance between the full-size and distilled models for mBERT, but not for the monolingual Italian BERT, where BERTino even performs slightly better than the original model. Drops in R^2 do not always align with drops in cosine scores: for example, XLM-R scores 0.06 R^2 points lower than BERT-IT/base, but the cosine score drops by only 0.01, while mBERT/dist loses 0.10 points on R^2 and 0.09 on COS. Thus, it appears that some models (like XLM-R) are less accurate at predicting the exact magnitude of perception scores but relatively good at capturing the overall score pattern across dimensions.

While the ridge regression models perform substantially worse than the transformer models, comparing the results between different feature representations is insightful for understanding what information is needed to predict perception: the Surface and Frames models all perform similarly with R^2 scores around 0.20 (with *f2* as a negative outlier), while the models with Neural features perform better (R^2 0.28-0.33). Simple counts of unigrams (*bow*) and frames (*f1*) give very similar overall scores; concatenating these features (*f1+*) leads to a small improvement (+0.03 R^2). This suggests that frames are useful for summarizing relevant lexical material (grouping together lexical units), but that the additional information about semantic and syntactic structure that is provided by role and construction labels does not lead to substantial gains. Using FastText embeddings instead of unigrams does not lead to gains, either. Meanwhile, comparing ridge models trained on transformer-derived features, we find best results with mean last layer representations from Italian BERT (*bm*), with slightly lower scores for the two models based on XLM-R (*sb* and *xm*); surprisingly, Sentence-BERT (*sb*) does not seem to have an advantage over averaged last-layer representations (*xm*).

Comparing R^2 scores across different questions and attributes reveals large differences in difficulty of prediction: for example, *blame on murderer* gets good scores across models, while *blame on victim* has relatively poor scores even for the strongest models (e.g. 0.24 for BERTino), and at-baseline (or worse) scores for the weaker models — notably, distilled mBERT, which performs decently on other attributes. *Caused by no-one* is even harder to predict, with no model scoring above 0.10. The *Focus* question has the overall best and most consistent performance, especially for the Italian BERT-based models, which achieve decent performance (0.46-0.66 R^2) for each of the four attributes.

This pattern is also reflected in MSA (Table 6): for *Focus*, it is substantially easier to predict the dimension with the highest score than for *Blame* and *Cause*. However, all models perform well above chance level for each of the questions, with the strongest overall scores for BERTino (56-72%).

The gain in performance achieved by the BERT-based models with respect to the surface feature models varies substantially between attributes. For example, the *bow* model has a surprisingly high score for *blame on murderer* (R^2 0.49), with only moderate gains from the BERT-IT and BERTino models (resp. +0.06 and +0.12 points). By contrast, *bow* scores poorly on *focus on concept* (R^2 0.13), whereas BERT-IT and BERTino have good scores (R^2 0.63/0.64). To get additional insight into the differences between models, we performed a feature attribution analysis. For the *bow* and *f1+* ridge regression models, we simply extracted the feature weights with the lowest and highest absolute values; for transformers, we applied the *integrated gradients* interpretation method (Sundararajan et al., 2017)[10] to obtain token-based attribution values for all sentences in the test set, and used the averaged values for tokens above a frequency threshold ($k \geq 5$, on a test set of 300 sentences) as an approximation of the overall feature importance. The results for *blame on murderer* and *focus on concept* are shown in Table 7.

[10]We used the implementation provided by the *transformers-interpret* package, see https://github.com/cdpierse/transformers-interpret

		model features	baseline	ridge								transformer				
				Surface		Frames				Neural			bert-it		mbert		xlmr
				bow	ft	f1	f1+	f2	f2+	sb	bm	xm	base	dist	base	dist	base
RMSE			0.67	0.59	0.60	0.59	0.58	0.63	0.59	0.56	0.54	0.56	0.48	0.47	0.51	0.53	0.51
COS			-0.02	0.49	0.46	0.48	0.52	0.36	0.50	0.55	0.58	0.55	0.67	0.69	0.65	0.58	0.66
R^2	*Average*		-0.01	0.20	0.19	0.20	0.23	0.08	0.18	0.28	0.33	0.28	0.44	0.45	0.38	0.34	0.38
	Blame	*murderer*	0.00	0.49	0.28	0.30	0.36	0.11	0.37	0.48	0.44	0.46	0.56	0.61	0.51	0.47	0.50
		victim	0.00	-0.05	-0.01	-0.03	-0.03	0.00	-0.08	0.09	0.13	0.09	0.17	0.24	0.15	0.01	0.10
		concept	0.00	0.05	0.11	0.08	0.07	0.02	0.09	0.22	0.27	0.23	0.37	0.33	0.26	0.12	0.25
		object	0.00	0.06	0.13	0.11	0.11	0.11	0.04	0.14	0.18	0.14	0.25	0.31	0.22	0.25	0.20
		no-one	-0.02	0.32	0.18	0.21	0.24	0.02	0.28	0.37	0.28	0.39	0.39	0.33	0.40	0.34	0.29
	Cause	*human*	-0.01	0.38	0.28	0.27	0.35	0.13	0.31	0.50	0.37	0.37	0.56	0.60	0.51	0.49	0.41
		object	0.00	0.45	0.31	0.51	0.55	0.40	0.51	0.35	0.54	0.44	0.80	0.81	0.79	0.68	0.74
		no-one	-0.01	-0.16	0.09	-0.05	-0.11	-0.18	-0.23	-0.22	0.03	-0.12	-0.07	-0.07	0.10	0.11	-0.09
		concept	-0.01	0.11	0.03	0.20	0.19	-0.02	0.11	0.07	0.19	0.00	0.39	0.31	0.04	0.18	0.31
	Focus	*murderer*	-0.01	0.51	0.33	0.34	0.43	0.15	0.42	0.48	0.43	0.51	0.66	0.65	0.61	0.61	0.58
		victim	-0.03	0.33	0.29	0.26	0.33	0.20	0.31	0.49	0.56	0.48	0.59	0.63	0.49	0.48	0.61
		concept	0.00	0.13	0.28	0.31	0.32	0.09	0.16	0.30	0.47	0.25	0.63	0.64	0.64	0.46	0.64
		object	0.00	0.06	0.21	0.14	0.13	0.08	0.07	0.32	0.36	0.41	0.46	0.46	0.19	0.21	0.37

Table 5: Regression results overview: RMSE, Cosine Similarity, and R^2 scores

model features	baseline	ridge									transformer				
		Surface		Frames				Neural			bert-it		mbert		xlmr
		bow	ft	f1	f1+	f2	f2+	sb	bm	xm	base	dist	base	dist	base
Blame	0.26	0.44	0.46	0.44	0.47	0.39	0.46	0.49	0.52	0.47	0.50	0.56	0.51	0.47	0.53
Cause	0.27	0.45	0.49	0.49	0.55	0.45	0.55	0.46	0.52	0.56	0.64	0.67	0.59	0.57	0.60
Focus	0.24	0.56	0.63	0.49	0.57	0.42	0.57	0.62	0.62	0.60	0.73	0.72	0.62	0.57	0.70
mean	0.26	0.48	0.53	0.47	0.53	0.42	0.53	0.52	0.55	0.54	0.62	0.65	0.57	0.54	0.61

Table 6: Most Salient Attribute scores

	blame: murderer						focus: concept					
	ridge/bow		ridge/f1+		bertino		ridge/bow		ridge/f1+		bertino	
	feature	attr	feature	attr	feature	attr	feature	attr	feature	attr	feature	attr
+1	ex *['ex' (ex-partner)]*	0.38	rol:Killing:Killer	0.21	killer *['killer']*	0.79	che *['that' (rel.pn./comp.)]*	0.20	che *['that' (rpn./cmp.)]*	0.12	femminicidio *['femicide']*	0.49
+2	uccide *['he/she/it kills']*	0.33	ex *['ex' (ex-partner)]*	0.15	uccide *['he/she/it kills']*	0.75	pista *['course of events']*	0.19	sara *['he/she/it will be']*	0.09	figlio *['son']*	0.31
+3	moglie *['wife']*	0.31	frm:Pers_rel	0.14	assassino *['murderer']*	0.71	passionale *['out of passion']*	0.19	pista *['course of events']*	0.08	non *['not']*	0.17
+4	uccise *['killed' (ptc, f.pl.)]*	0.24	frm:Killing	0.13	ex *['ex' (ex-partner)]*	0.62	sara *['he/she/it will be']*	0.19	non *['not']*	0.08	:	0.17
+5	assassino *['murderer']*	0.22	cx:Pers_rel++nvrb	0.13	fidanzato *['boyfriend']*	0.51	femminicidio *['femicide']*	0.17	femminicidio *['femicide']*	0.08	suicidio *['suicide']*	0.15
-5	sono *['I am' / 'they are']*	-0.14	rol:Event:Event	-0.06	una *['a' (f.)]*	-0.14	omicida *['murderer']*	-0.13	nell' *['in the']*	-0.07	uccisa *['killed' (ptc, f.sg.)]*	-0.32
-4	della *['of the' (+ f.noun)]*	-0.15	sono *['I am' / 'they are']*	-0.06	.	-0.14	trovata *['found' (ptc, f.sg.)]*	-0.14	della *['of the' (+ f.noun)]*	-0.07	morta *['dead' (f.sg.)]*	-0.32
-3	-	-0.16	frm:Event	-0.08	sono *['I am' / 'they are']*	-0.15	nell' *['in the']*	-0.14	due *['two']*	-0.07	killer *['killer']*	-0.38
-2	accaduto *['happened']*	-0.17	della *['of the' (+ f.noun)]*	-0.08	trovata *['found' (ptc, f.sg.)]*	-0.20	ospedale *['hospital']*	-0.16	cx:Buildings++nvrb	-0.07	auto *['car']*	-0.41
-1	.	-0.35	.	-0.13	morta *['dead' (f.sg.)]*	-0.21	due *['two']*	-0.16	frm:Buildings	-0.09	uccide *['he/she/it kills']*	-0.42

Table 7: Comparison of most informative features for an 'easy' attribute (blame/murderer) and a 'hard' attribute (focus/concept). *[Abbreviations: rol=semantic role, frm=frame, cx=construction, nvrb=nonverbal, Pers_rel=Personal_relationship; f.=feminine [grammar], ptc.=participle, sg.=singular, pl.=plural, rel.pn.=relative pronoun, cmp.=complementizer]*

For *blame on murderer*, all three models seem to focus on similar lexical items: for example, "*uccide*" ('(he) kills') has a high positive attribution value in both the *bow* ridge regression and the fine-tuned BERTino model, and in *f1+* we find a positive score for the KILLING frame, which is an abstraction over killing-related words. We also find that personal relationships ('wife', 'ex', PERSONAL_RELATIONSHIP) get positive attributions in all three models. By contrast, we find negative attribution values for "*accaduto*" ('happened') and the corresponding EVENT frame in *bow* and *f1*, which maps neatly onto our observations discussed in §2.4. For *focus on concept*, no insightful differences between the three models are immediately obvious. We do find several intuitively relevant features in each model: "passionale" ('out of passion') and "femminicidio" ('femicide') could to examples of concepts that sentences could give focus to, whereas "omicida" ('murderer/murderous') and "killer" could be seen as emphasizing the role of a human agent rather than an abstract concept.

4 Conclusion & Future Work

This paper has presented a detailed analysis of human perceptions of responsibility in Italian news reporting on GBV. The judgments we collected confirm the findings of previous work on the impact of specific grammatical constructions and semantic frames, and the perceptions they trigger in readers.

On the basis of the results of our survey, we have investigated to what extent different NLP architectures can predict the human perception judgements. The results of our experiments indicate that fine-tuning monolingual transformers leads to the best results across multiple evaluation measures. This opens up the possibility of integrating systems able to identify potential perception effects as support tools for media professionals.

In the future, we plan to run a more detailed analysis of the data considering differences along individual and demographic dimensions of the respondents. In addition to this, natural follow-up experiments will focus on the application of the approach to other languages and cultural contexts both targeting GBV as well as other socially relevant topics, e.g. car crashes (Te Brömmelstroet, 2020).

Ethics Statement

Limitations This work has a strong connection with multiple theoretical frameworks: Frame Semantics, Construction Grammar, and Critical Discourse Analysis. The way we have structured the questionnaire aimed at collecting data from human participants with respect to different sentences - which in different ways contained variations in syntactic structures and semantic frames that could be linked to findings and claims about the "*perception*" and its effects in the interpretation of sentences. The use of state-of-the-art NLP tools to identify these properties in a large collection of data represents both an advantage (i.e., allows to deal with a large number of data, reducing human subjective interpretation) and a limit (i.e., errors from the systems may result in non optimal examples for human judgements).

While representing an *unicum* in the language resource panorama, since there are no previous comparable and available corpora, the number of available sentences used to train the models is somewhat limited. The final corpus, however, represents an optimal compromise between number of judgements needed to obtain a solid representations of perceptions by users and number of data points that could be used by stochastic NLP architectures to learn from the data.

Finally, the outcome of the perception judgements can be generalized to the population of Italian young adults attending universities (i.e., undergrad students). This is a limitation of the data collection process. We tried to minimize this by reaching out to students in multiple universities (i.e., geographical variation) and at different faculties (from Arts/Humanities, to Computer Science and Physics) and disciplines (from Linguistics, to Media and Communication Studies, Computer Science, and Physics).

Data collection The questionnaire was conducted using the Qualtrics XM platform. Participation to the questionnaire was on a voluntarily basis. Participants could interrupt their participation in any moment. Only fully completed questionnaires have been retained. Participants received compensation (5 euros) - upon completion of the questionnaire. Participants have been recruited mainly among undergraduate students at different universities in Italy.

Participation was fully anonymous: 1) partici-

pants could access the questionnaire via a unique special access token that could be obtained by filling in a form; 2) no personal information other than the participants' email address was stored; 3) IP addresses were not stored or tracked; 4) the special access token and the participants' email were decoupled. Participants could receive their compensation only by providing the unique access token.

Dual use The experiments we have run investigate to what extent models are able to predict human perceptions along three dimensions with respect to GBV. The very nature of the task limits the potential misuse by malevolent agents. At the same time, malevolent agents can purposefully misrepresent the results to minimize the negative aspects associated to the reporting of the phenomenon by media. By making the models and the data publicly available, together with a detailed explanation of how the models work and how results should be interpreted in a correct way, we mitigate these risks.

Intended use As it is the case for supervised models, sensitivity to the training material is high. At the moment, we have not tested the portability of the models to other topics. We do recommend to use these models only on data compatible with the phenomenon we have taken into account, i.e., GBV against women. Although the application of the models to any other type of texts reporting violence and killing against other targets may still give some valid results, we discourage its use since risks of unforeseen behaviors are high, with potential harmful consequences for the victims of violence.

Acknowledgements

The research reported in this article was funded by the Dutch National Science organisation (NWO) through the project *Framing situations in the Dutch language*, VC.GW17.083/6215. We would like to thank the CRITS research center at the Italian public broadcaster (RAI) for providing access to their femicide dataset. We also thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

M. Belluati. 2021. *Femminicidio. Una lettura tra realtà e interpretazione*. Biblioteca di testi e studi. Carocci.

Gerd Bohner. 2002. Writing about rape: Use of the passive voice and other distancing features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, 40:515–529.

M. Brysbaert, A.B. Warriner, and V. Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Consuelo Corradi. 2021. Femicide, its causes and recent trends: What do we know? Briefing requested by the DROI Subcommittee of the European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/653655/EXPO_BRI(2021)653655_EN.pdf, accessed 2022-08-24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

Charles J. Fillmore. 2006. Frame semantics. In D. Geeraerts, editor, *Cognitive Linguistics: Basic Readings*, pages 373–400. De Gruyter Mouton, Berlin, Boston. Originally published in 1982.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Kurt Gray and Daniel M. Wegner. 2009. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96:505–520.

Christopher Hart and Matteo Fuoli. 2020. Objectification strategies outperform subjectification strategies in military interventionist discourses. *Journal of Pragmatics*, 162:17–28.

Nancy M Henley, Michelle Miller, and Jo Anne Beazley. 1995. Syntax, semantics, and sexual violence: Agency and the passive voice. *Journal of Language and Social Psychology*, 14(1-2):60–84.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Janellen Huttenlocher, Karen Eisenberg, and Susan Strauss. 1968. Comprehension: Relation between perceived actor and logical subject. *Journal of Verbal Learning and Verbal Behavior*, 7:527–530.

Shanto Iyengar. 1994. *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Laboratorio Adolescenza and *Istituto IARD*. 2018. Adolescenti e stili di vita: Sintesi risultati. https://www.istitutoiard.org/wp-content/uploads/2018/12/Indagine-Adolescenti-2018_sintesi-risultati.pdf, accessed 2022-08-24.

Chiara Meluzzi, Erica Pinelli, Elena Valvason, and Chiara Zanchi. 2021. Responsibility attribution in gender-based domestic violence: A study bridging corpus-assisted discourse analysis and readers' perception. *Journal of pragmatics*, 185:73–92.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Erica Pinelli and Chiara Zanchi. 2021. Gender-based violence in italian local newspapers: How argument structure constructions can diminish a perpetrator's responsibility. *Discourse Processes between Reason and Emotion: A Post-disciplinary Perspective*, page 117.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.

Marco Te Brömmelstroet. 2020. Framing systemic traffic violence: Media coverage of dutch traffic crashes. *Transportation research interdisciplinary perspectives*, 5.

Piek Vossen, Filip Ilievski, Marten Postma, Antske Fokkens, Gosse Minnema, and Levi Remijnse. 2020. Large-scale cross-lingual language resources for referencing and framing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3162–3171, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Karen Zhou, Ana Smith, and Lillian Lee. 2021. Assessing cognitive linguistic influences in the assignment of blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 61–69, Online. Association for Computational Linguistics.

A Appendix

A.1 Questionnaire Results

Figures A.2 through A.4 show the distribution of z-scored perception scores per question and attribute.

A.2 Transformer models

Below are details about the exact versions of the pre-trained transformer models that we used:

- **Italian BERT XXL (BERT-IT)**: published by the Bavarian State Library at https://huggingface.co/dbmdz/bert-base-italian-xxl-cased. N.B.: 'XXL' refers to the corpus size, not the size of the model itself.
- **BERTino**: https://huggingface.co/indigo-ai/BERTino; this is a DistilBERT model, using Italian BERT XXL as its teacher but trained on a different corpus.
- **Multilingual BERT (mBERT)**: https://huggingface.co/bert-base-multilingual-cased
- **Multilingual DistilBERT**: https://huggingface.co/distilbert-base-multilingual-cased
- **XLM-RoBERTa**: https://huggingface.co/xlm-roberta-base

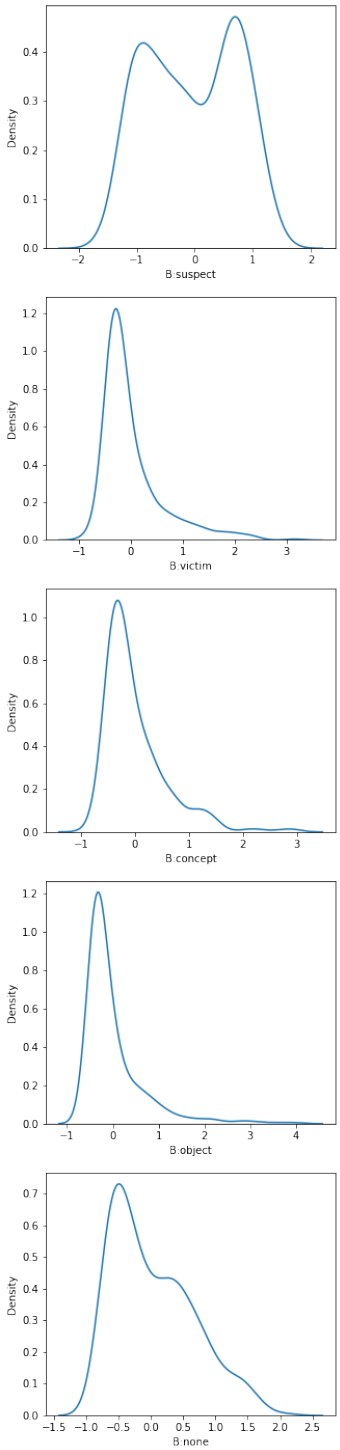
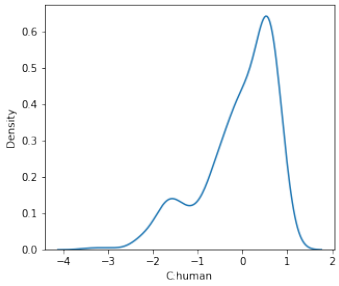
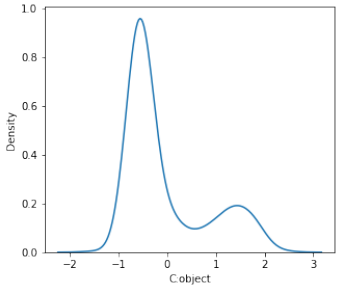
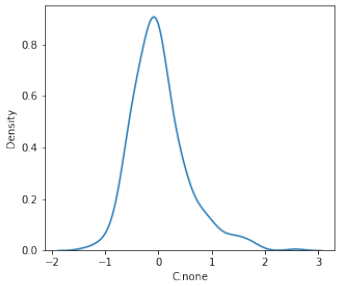


Figure A.2: Density plot of aggregated z-scores for *blame*







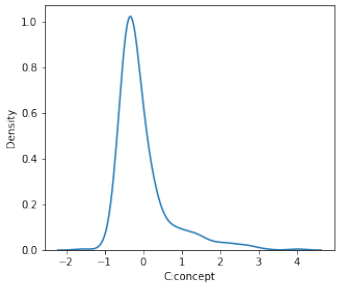
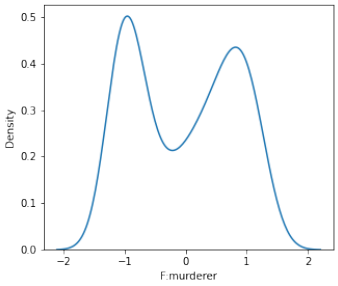
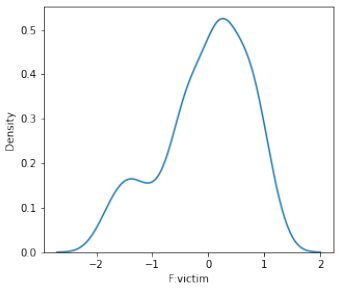
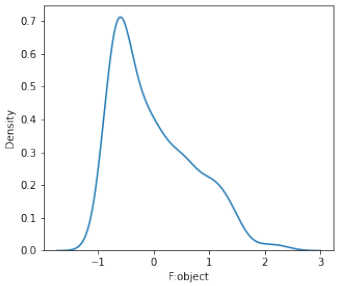


Figure A.3: Density plot of aggregated z-scores for *cause*







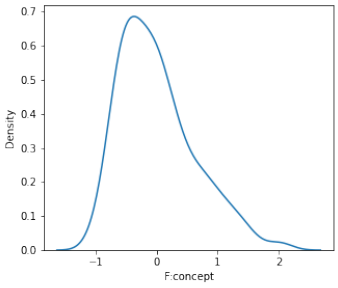


Figure A.4: Density plot of aggregated z-scores for *blame*