Validating Predictive Models Of Evaluative Language For Controllable Data2Text Generation

Maurice Langner
Linguistic Data Science Lab
Ruhr-Universität Bochum, Germany
Maurice.Langner@rub.de

Ralf Klabunde
Linguistic Data Science Lab
Ruhr-Universität Bochum, Germany
Ralf.Klabunde@rub.de

Abstract

In data2text generation, tabular data is transformed into a text that expresses information from that source domain. While some text types, such as instructions, demand objective and neutral language without any expressive and evaluative content, many other text types are expected to provide expressions for these kinds of subjective meanings. In controllable, pipelined neural NLG separate learning models, notably regression models, can be used to predict whether some feature deviates sufficiently strongly from an expected value, so that evaluative language would be appropriate for verbalizing this finding. In this paper, we present an empirical study on the comprehension of evaluative adverbs and adjectival modifiers in car reviews, a text type that is characterized by a mixture of factual information with evaluations expressing positive or negative surprise. We show to what extend regression-based decision boundaries for producing evaluative content in controllable data2text NLG match the reader's expectations that are raised by those evaluative markers. Finally we show that regression values in combination with standard deviation of the technical input data constitute reasonable Boolean thresholds for both positive and negative surprise, which provide the basis for the development of more complex models that also include the scalar base of adverbs and modifiers.

1 Introduction

In controllable data-to-text natural language generation (henceforth data2text), tabular data is transformed into surface text that structures, relates and verbalizes the relevant pieces of information as close as possible to the domain-specific characteristic features and structures of the corresponding text types. The goal of adjusting the output to style, tone and structure of typical texts in the respective domain is achieved by either manually encoding template-based systems, or by fine-tuning large language models (LLMs) like BERT or GPT-X on a subset of the corpus.

An application domain that is especially challenging in regard to data2text NLG are car reviews. In these texts, technical information is seamlessly interwoven with subjective impressions of the test driver and, even more important, with the test driver's assessment of the car's features against his experience with cars of comparable quality and class.

Although a trustworthy system should not generate information that is not present in the underlying database of car features, it is possible to approximate the domain knowledge needed to reproduce the expert's assessment of the car's technical features. For example, sentence (b) in Table (1) from a driving report about the Lotus Elise Sport 2015 cannot be generated from purely technical information about the vehicle. It contains additional, subjective information on the driving experience. If a system for controllable data2text NLG shall be capable of generating sentences of this kind, the respective information must be added or the verbalisation be hard-wired into the system, including a mechanism for deciding when it is suitable. Talking about being pressed back into the seat would be inadequate, even quirky, if the car was a compact car with a 60 HP motor. LLMs may encounter such expressions during training and reproduce it when realizing the features for the target text, but controlling style and content for data2text with an LLM is a non-trivial task, which is why most NLG systems in-use do not rely on them (yet).

Accordingly, in sentence (a) in Table (1), where the strong acceleration is contrasted with a comparably low motor power, controllable data2text NLG needs a mechanism for determining when the mismatch between acceleration and power output is large enough to permit the usage of an evaluative adverb like *surprisingly* and respective modifiers like *enormous*. Again, LLMs may learn using

Proceedings of the 16th International Natural Language Generation Conference, pages 313–322
September 11–15, 2023. ©2023 Association for Computational Linguistics

(a)	Surprisingly, the sports car has an enormous acceleration from 0 to 60 in under 6.5 seconds, given the comparably low power of 136 hp.
(b)	When you hit the gas, the first law of motion kicks in and you feel that well-known feeling rise in your stomach when being pressed back into your seat.
(c)	The reason for this is that the Lotus Elise has an exceptionally low kerb weight of 1931 pounds.

Table 1: Three subsequent example sentences from a driving report

evaluative expressions from the input data, but a controllable generation, deciding where such an expression is suitable on the basis of the technical data input, should be preferred. Some examples generated by OpenAI's ChatGPT, which we report below, will clarify this point.

Sentence (c) finally gives the reason for the high acceleration given the low HP: an *exceptionally low kerb weight of only 1931 pounds*. The adverb *exceptionally* signals the impossibility to predict a value of 6.5 seconds for acceleration when only considering the power output, but when considering the joint effect with a tiny weight, the acceleration is, albeit a rare combination, technically possible.

The exemplary excerpt demonstrates the relevance of evaluative expressions on sentence and discourse level. However, although evaluative meaning often implies a subjective estimation why some features or states have a positive or negative stance, these evaluations rely on experiences and expectations that can be inferred from underlying data.

We consider evaluative expressions as a key element to generating humanlike, pragmatically rich text and understanding the underlying mechanisms for producing evaluative language in general. This is of special importance when solely being informative is not the fundamental driving force for communication. Evaluative adverbs and modifiers, which we focus on, are just one way of realizing a mentionable data mismatch that is present at the early stage of document planning. Different NLG systems, either traditional or neural models, may then proceed differently for realizing the respective evaluative content and including it in the generated output.

In this paper, we describe an empirical study on the comprehension of evaluative expressions in driving reports, performed with non-experts in order to identify the relation between their evaluative interpretation and the underlying technical data. The study inverts the typical acceptability rating approach of fixed sentences. The participants have to define lower and upper thresholds for numerical features that make the respective phrase acceptable, allowing us to empirically approximate intervals and their match with evaluative scales. We then compare the results with regression-based approaches to data analyses to show whether these models are adequate for dynamically determining decision boundaries of evaluative language in text generation.

We use the car domain with its comprehensive technical specifications, but the approach we are presenting should be transferable to any data-oriented generation model and corresponding texts with evaluative content related to these data, such as technical devices and their use, weather reports, or data and reports from sports events, such as Football games, as in the ROTOWIRE corpus (Wiseman et al., 2017).

2 Related Research

Evaluative expressions (and expressive constituents in general) have received much attention in Formal Semantics and Pragmatics, revolving around the questions how these items can be integrated into a compositional framework, and how their meanings interact with truth-conditional and inferred content (Potts, 2005; Gutzman, 2015). As meaning analyses that are focusing on conditions for the use of these expressions, they can ignore the question of what world knowledge or data their use is based on, which is an indispensable aspect of data2text NLG however (Reiter and Dale, 2000; Ferreira et al., 2020; Gatt and Krahmer, 2018).

Evaluative adverbs and modifiers, the linguistic topic of this paper, are also attributed to emotion generation or affective language generation (de Rosis and Grasso, 2000), where evaluative items are generated in order to convey information with a specific stance (Elhadad, 1991), for example communicating exam marks (Mahamood et al., 2007) or tailoring texts to specific users (Balloccu et al., 2020). The same intention is inherent to car reviews that include driving reports, where the tone of the text is coloured by the author's opinion. Experimental studies in this area of research are quite complex due to the multitude of influential factors regarding audience, personality, individual preferences and the level of knowledge or common ground. Experimental results are often non-reproducible (Mahamood, 2021), since the target group is a decisive factor. Affective language generation has been implemented in both, template-based NLG systems (de Rosis and Grasso, 2000), as well as in neural end2end language models (Singh et al., 2020; Santhanam and Shaikh, 2019). Regarding controllability, generating such evaluative markers with Transformer models and LLMs in general poses a challenge due to toxicity and fact hallucination (Ji et al., 2022; Dušek and Kasner, 2020; Shen et al., 2020). First attempts have been made to adapt GPT models to and condition on the intended evaluative stance of the text to be generated (Singh et al., 2020).

In recent approaches, end-to-end language models have been re-pipelined in order to enhance controllability of intermediate representations. Ferreira et al. (2020) propose a re-modularization of neural generation networks, chaining separately trainable and evaluable networks that are specialized for the different tasks of content selection, document planning and surface realisation. They show that these pipelined neural generation models outperform end-to-end networks, especially on unseen data, where the latter tend to produce topic-unrelated, incoherent texts and hallucinations. In pipelined data2text generation it is possible to integrate the necessary calculations for determining the adequacy of evaluative content on a polarized scale (Benamara et al., 2017) with regression and classification (Langner and Klabunde, 2022) and using large language models such as T5 (Xue et al., 2021) for transformation into surface text.

2.1 Large language models and evaluative tasks

ChatGPT shows in a clear way that trusting completely on LLMs does not necessarily result in the appropriate use of evaluative expressions. The following excerpt is ChatGPT's answer to the task of evaluating the specific values of the features 'acceleration' and 'motor power' of the car under discussion, viz. the Lotus Elise Sport 2015.[1]

> *The acceleration of under 6.5 seconds from 0 to 60 mph with a motor power of 136 HP is pretty impressive. It suggests that the vehicle has a good power-to-weight ratio, which allows it to accelerate quickly.*
> *In terms of rating, it depends on the context of the vehicle. If it's a small sports car or a high-performance sedan, then this acceleration is quite good. On the other hand, if it's a larger vehicle such as an SUV or a pickup truck, then it's still decent but not as impressive.*
> *Overall, I would rate an acceleration of under 6.5 seconds from 0 to 60 mph with a motor power of 136 HP as a very respectable performance for a car.*

The first paragraph perfectly captures the fact that the Lotus Elise may accelerate so fast with comparably small power due to a "good power-to-weight ratio" (as also explained in the driving report, see sentence (c) in Table 1) and evaluates it as "pretty impressive". The second paragraph reveals the inconsistency in regard to the evaluation, a lack of coherence and the wrong calculation behind it. For a small sports car, the acceleration is now judged as "quite good", where initially it was "impressive", clearly a shift on the acceleration scale to the negative side. Furthermore, it names SUVs and pickup trucks as contrastive vehicles, for which the given acceleration would only be decent. This is in contradiction to the technical facts about motors and vehicle types. In the final paragraph, ChatGPT uses a third evaluative expression of the features, (*a very respectable performance*), and relates acceleration to power output in a generalizing statement.

Prompting ChatGPT with the task to produce a sentence for a driving report with the features above, using an adverb to express its opinion[2], results in the following sentences that were produced in the given order:

1 Impressively, the car with 136 hp achieved an acceleration of 6.5 seconds from 0 to 60 mph, indicating that it is remarkably well-designed for performance.

2 Honestly, the car's performance was underwhelming with a 6.5 second acceleration from 0 to 60 mph despite its 136 hp power output.

Both sentences correctly fulfilled the task, but only the first agrees to expert opinions on the Lotus Elise, and sentence number 2 completely

[1]Original prompt: *"How would you rate an acceleration of under 6.5 seconds from 0 to 60 mph with a motor power of 136 HP?"*

[2]Original prompt: *"Please produce a sentence for a road test report of a car with 136 hp and an acceleration of 6.5 seconds from 0 to 60 mph, expressing your opinion with a sentential adverb."*

contradicts the first one in its criticism. In sum, ChatGPT – as a representative example of relying on LLMs – is very well capable of enriching text with evaluations and generates sophisticated formulations, but adequacy and agreement with expert knowledge is in need of improvement.

3 Regression models for predicting expectations

The technical data we are using have been extracted from the ADAC database, the database of Europe's largest automobile association. The database contains technical information on vehicles of various types as well as independently written reports on these vehicles from ADAC's vehicle experts. We use this database due to the comprehensive technical details it provides in combination with the vehicle reports. Commercial interests are not present.

We have trained different regression models. The first is a standard linear regression model, which we expect to underfit the data since some of the features, e.g. 'power output' and 'acceleration', have exponential rather than a linear relation. The second is a polynomial regression model, which may underfit for the same reasons of mixed relation types between predictors and response. Furthermore, we implemented a deep neural network with intermediate, non-linear layers for regression. For these networks, we used Adam optimizer and mean absolute error as loss function. The best fitting model is the DNN with an MSE of 2.73 and an R2 score of 0.62 for the examples given in Figure (1).

In general, experts in automotive engineering have expectations about certain features of a car, given its technical details such as a certain acceleration given the power output and kerb weight. The images in Figure (1) show relations between each pair of the features 'acceleration', 'power output' and 'weight' as extracted from the database. The color of each "+" marker represents the feature that is not present on the axes. For example, in subfigure (c) the color encodes the weight, light-blue indicating small numbers and light-brown indicating large numbers.

The green dot in each sub-figure marks the data point for the Lotus Elise which has been described by the example sentences in Table (1). The position of these data points outside of the data cloud is a strong indicator already for using evaluative expressions. However, we require a general approach for deciding whether information is evaluative at all.

In Figure (1a), there seems to be a linear relationship between power output and weight with strong variation of the power value for instances higher than 2500 kg. Due to large variation, regression models agree ever less with each other with increasing power, the polynomial parabola turning down again, the DNN taking a steeper increase upwards.

Figure (1b) shows a rather broad distribution of the relation between weight and acceleration. There seems to be a linear decrease of seconds between 1000 and 2000 kg, which forms a baseline. But the variation range suggests that weight seems to be of less importance for predicting the acceleration value. Regression models are also less equivalent at the borders above 2800 kg where fewer data is available and variation is largest.

In Figure (1c), there is an exponential decrease of acceleration time with increasing power output. Compared with the other sub-figures, variation is small so that there is a well defined relationship that can be modeled with far better fit than the other ones. According to the distribution of data points, variation seems to grow with the decrease of seconds. The curve finally converges towards a horizontal line. Here, the DNN fits the data perfectly. The polynomial model fits as well, but for power output values higher 400, the curve rises again, which is rather unrealistic and does not fit the data - it is just the nature of a second degree polynomial. The linear regression model overestimates the acceleration for power outputs higher than 400 and due to its linear nature assumes a constant decrease which is not reflected in the data either.

Using these regression models allows us to take features x_0^c,...,x_{n-1}^c of car c in order to predict an expected value for feature x_n^c that shall be verbalised in the text. A deviation of the real feature value in the tabular data from this expected value both qualifies and quantifies the generation of evaluative adverbs or modifiers. Other options of lexicalizing the intended affective tone exist as well, but they are not subject of the empirical study described in this paper.

Applying regression models to the Lotus Elise, we see that these models predict the car to be slower, to weigh more and to have nearly double the power. All regression values therefore would justify the usage of evaluative adjectives with positive stance like *surprisingly* or *incredibly*.

There is one caveat to this setup. How do we quantify the divergence from the real value that triggers evaluative content? Its empirical counterpart is: How do evaluative adverbs influence the reader's expectation about a certain feature? Due to data sparseness, we cannot choose the threshold in such a way that the distribution of evaluative expressions in the empirical data best matches the distribution predicted by the models. For this reason, we decided to perform an empirical study for determining the intervals of values that license the usage of evaluative content on the basis of the reader's expectations.

4 Empirical study on expectation values raised by evaluative adverbs

In order to evaluate the adequacy of the regression models, we need to consider how evaluative adverbs and modifiers are interpreted and whether the regression models correctly capture expectations of readers. For this reason, we conducted a study on the interpretation of these evaluative items and their influence on the expectations about numerical features in driving reports. The study is designed as a webserver application, participants from Germany and Austria as well as the USA were acquired via Prolific. Participants were selected w.r.t. their first language, highest education level (at least an academic bachelor degree), possession of a driving licence and ownership of a car. For each study in German and English, 50 participants took part. Additionally, 50 further participants took part in a shorter ablation study on modifiers in German. The participants were paid 13 pounds per hour. 20 minutes were scheduled for each participant, but most were significantly faster (8-13 minutes median time). For the ablation study, 12 minutes were scheduled. For the German studies we analyse in this paper, no outliers had to be excluded. We took into account the features 'acceleration', 'mileage', 'maximum speed', 'power output' and 'displacement', the former four because they are the most well-known and intuitive features, the latter representing features that are less intuitively accessible. We used the German counterparts to the evaluative adverbs *surprisingly*, *disappointingly*, *amazingly* and *unfortunately*, as well as the modifiers *good*, *bad*, *low*, *high*, *slow*, *fast*, *average* and *enormous*.

polarity	item
-2	Disappointingly, the car goes slowly from 0 to 60 mph in [...] seconds with a power output of 200 hp.
0	With a power output of 200 hp, the car goes from 0 to 60 mph in [...] seconds.
2	Amazingly, the car goes from 0 to 60 mph in only [...] seconds with a power output of 200 hp.

Table 2: Example items and their polarity of group e (estimating acceleration given power output of 200 hp)

4.1 Methods

The study comprises two tasks. The first task type is selecting an option from a menu as answer to a question in the car domain. These questions function as distractors, while also being a means for collecting information on the participants experience with cars that can be used for further research in regard to text production. For some items the participants were instructed to select an evaluative adverb that agrees with their judgement of the given features, as in *How would you rate the acceleration of the following car in relation to the power (hp)? "The sports car goes from 0 to 60 mph in 7.5 seconds with an output of 560 hp."*, with answer options *{surprisingly fast, normal, disappointingly slow}*.

The main task type presents a sentence from a road test report (henceforth *item*), where two features are named. These sentences are grouped by their degree of neutrality or polarity towards a positive or negative evaluation. Seven categories are possible, with polarity ratings between -2 and 2, being the most negative and most positive expressions. Sentences of category 0 are neutral, 0.5 and -0.5 contain only modifiers, whereas categories -1 and 1 only contain an evaluative adverb with respective polarity. Categories -2 and 2 contain both an adverb and one or two additional modifiers in the grammatical phrases that contain the features. Examples for three polarities are given in Table (2). In the respective group, participants were asked to estimate acceleration for a car with 200 hp given the differently polar expressions.

These items were collected by automatically extracting sentence adverbs from the ADAC corpus and manually selecting evaluative ones. These sentences containing the adverbs were randomly assigned to a predefined group of polarities. The missing polar items in each group were then manually constructed by modifying the corpus extracted

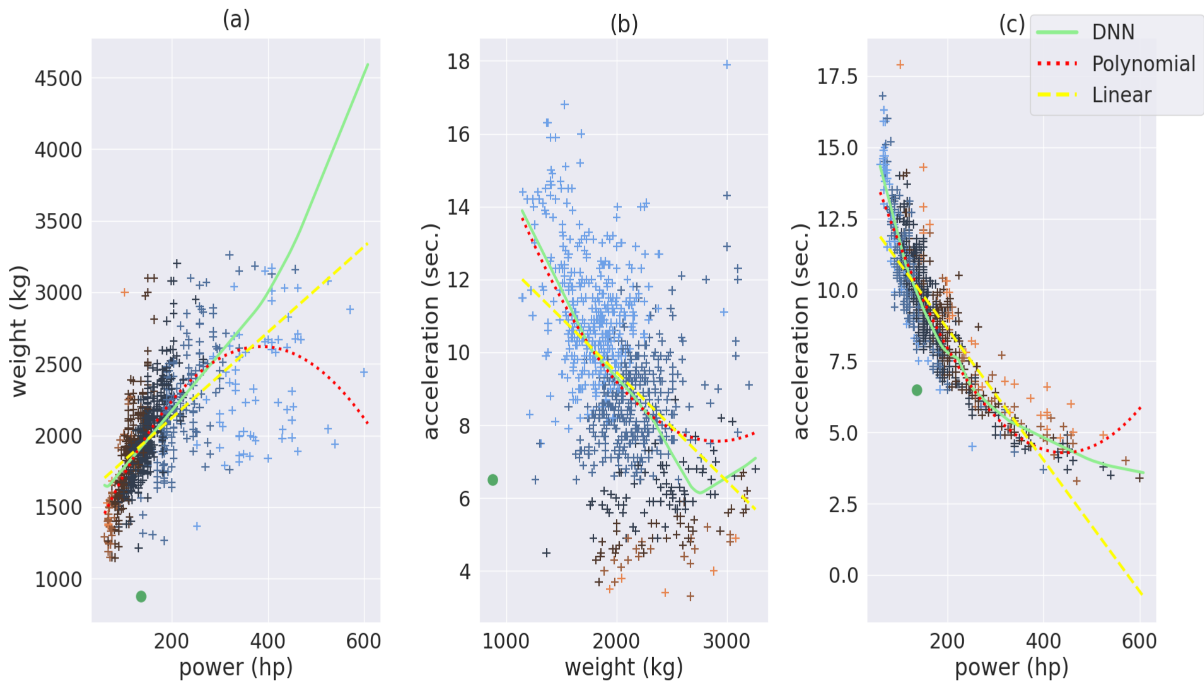

Figure 1: Bivariate plots for weight, acceleration and power output. Each "+" represents a data point. The larger green dot marks the Lotus Elise, which is an outlier in all graphs. The line plots represent three types of regression models and their fit to the data, the linear (yellow dashed), the polynomial (red dotted) and the deep neural network for regression (lightgreen solid).

item, so that the effect of different polarities can be tested within each group.

Instead of letting the participants rate the acceptability of fixed sentences, we decided to remove one of the numerical statements from the items and let the participants determine the lower and upper threshold such that the resulting interval of values agrees to their expectations the expression raises. The participants could either use a slider or two text fields to enter the thresholds. Minimal and maximal values had been determined on the basis of our database on technical features.

Each sentence provides a single session item the participants have to deal with; the order of the 15 items is randomized in order to prevent bias. Towards the end of the study 3 expressions have been shown simultaneously that express the same features, but with polarity categories -2, 0 and 2. These 3 items are not randomized and agree to 3 items all participants have seen before. The task is to adjust the thresholds also in comparison to the choices they make for the other items on the page. We integrated this final page in order to assess whether results vary when the participant's expectations for one item are delineated more clearly by the expectations raised by alternative evaluations.

The study is based on the following hypotheses: We assume that the expectation values differ significantly between neutral sentences and those with evaluative adverbs in agreement with polarity. For example, if the sentence states with positive stance that a sports car has a surprisingly high maximum speed given a certain power output, we assume the participants' expectation interval to be higher than for the statement without evaluative adverb. Likewise, we assume that negative polar expressions vary significantly from positive polar questions in regard to the expected values, since it should differ from the neutral one, but in the opposite direction to the positive expression.

4.2 Evaluation

27 out of 32 binary comparisons between two differently polar items and the distributions of estimated values are significant. For each item, participants estimated a lower and an upper threshold that matches their expectation raised by the item's polarity and thus its evaluative stance. Across participants, this results in two normal distributions, one for each threshold. 15 items are grouped by the two contained features, the source feature as orientation and the target feature, whose numeric value

is masked and shall be estimated. Groups consist of 2 to 5 items, which means that not all possible polarities are tested for each feature pair. In each group, we tested all possible pairs of expressions for significance using ANOVA.

For illustration reasons, we concentrate on an example where participants should estimate acceleration on the basis of a given power output (compare Table 2). We have tested a multitude of feature combinations across all studies. Results are generalisations made from all these items.

As shown in Table (2), participants estimated acceleration from 0 to 100 km/h for a car with 200 hp. In Figure (2), there are two graphs containing the distribution plots for the lower threshold (left) and the upper threshold (right) for each of the three items in Table (2). The two maxima of the neutral expression's curve for the lower threshold are located between 3 and 7.5 seconds for acceleration, the single maximum for the upper threshold is positioned at about 11 seconds. We judge these curves as approximations of the lower and upper bound of acceleration values that are normal for a car with 200 hp. Now, the distributions for both, the positive item and the negative one, vary from the neutral item, both in opposite directions of each other but in agreement with their evaluative pole. Distributions for both thresholds of the positive item are shifted to the left towards better acceleration, with maxima of 2.5 seconds and 7.5 seconds respectively. The standard deviation is much smaller for the positive item, which means there was more agreement on the estimated values. On the other side, the distributions for the negative item are shifted to the right towards worse acceleration, the maximum of the lower threshold located around 10 seconds and the maxima of the upper threshold at 13 and 18 seconds. The differences between all curves are strongly significant (compare Table 3).

As Figure (1) shows, acceleration values for cars with 200 hp vary between 6.5 and 10.5 seconds, which agrees surprisingly well with the maxima of the kernel density estimates for the neutral item. Therefore, the interval between 0 and 7.5 may permit the usage of a positive evaluative adverb, whereas values between 10 and 20 seconds may license the usage of the negative evaluation.

As for the three threshold pairs for the simultaneously shown items on the final page of the study, we can compare their distributions with separately shown items in order to assess the influence of presenting alternative evaluations at the same time. Again, we compare the distributions of the corresponding pairs of items with the same polarity in regard to both thresholds. Only two thresholds are significantly different, namely the neutral upper bound and the positive upper bound. All other thresholds are more or less equivalent. These results indicate that participants do have a good intuition about the intervals that agree with a certain evaluative stance even without presence of alternatives as an orientation point.

One item that is rather special contains the feature *displacement* in a group with a neutral and a strongly positive item with adverb and modifiers. This feature is rather unknown among non-experts of the domain and therefore it is also harder to estimate reasonable values for it. Variance in the data is also really strong, since the relation between *displacement* and *power* depends on many other factors. This is also mirrored in the distributions for lower and upper threshold between the neutral and the positive item. Standard deviation is much larger, there is no significant difference and regression models perform less precise and less confident.

Another hypothesis concerns the effect of modifiers, which we expected to intensify the influential effect of the evaluative adverb. There is no empirical proof of this in our data, however.

The ablation study on the effect of modifiers confirms the findings for evaluative adverbs. Although the bell curves show that for both thresholds, the expected values differ between neutral and modifier-enriched sentence into the same direction as the neutral and adverb-enriched sentences, the effect-size of modifiers seems to be smaller than for evaluative adverbs. One interesting thing to be mentioned here is the influence of dual modifiers (one for each noun phrase of the two opposing car features), when one modifier puts its feature into perspective, such as in one group of items in this ablation study. Participants are supposed to guess acceleration by maximum speed. The modifier-enriched sentences ask for a good acceleration given a "rather average" maximum speed of 200 km/h. The positive sentence with adverb only elicits expectations with best acceleration values, whereas sentences with only modifiers and both adverb and modifiers are nearly indistinguishable and lie in between the neutral and the adverb-enriched

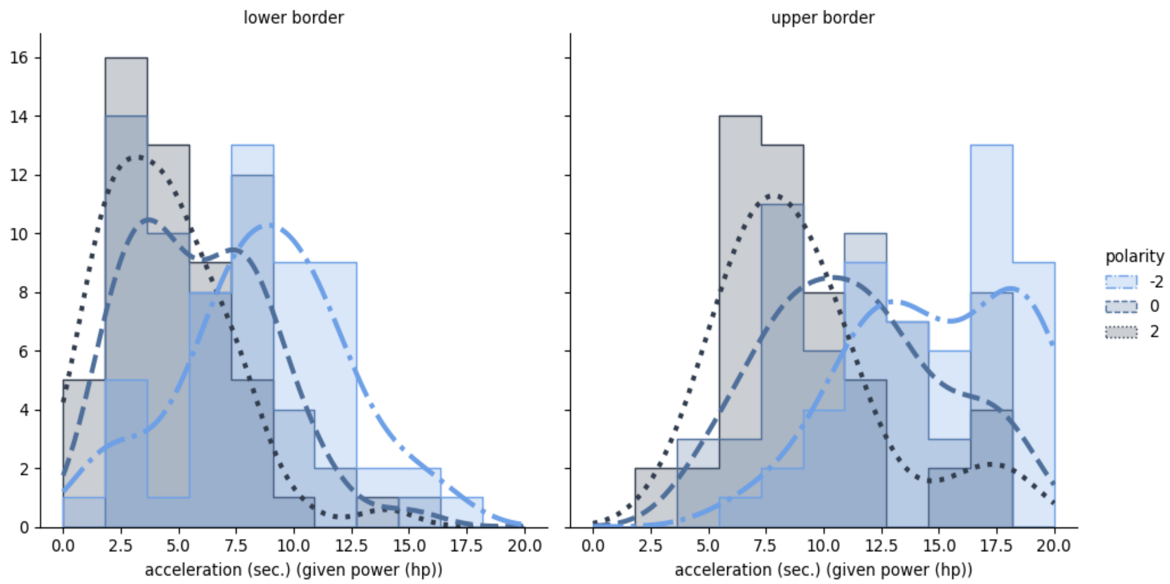

Figure 2: Distribution plots for lower border and upper border estimations on acceleration, given a specific power output value. Lower values on x-axis are better. The dotted line is the positive expression, the dashed line represents the neutral expression and the dash-dot line is the negative polar expression.

comp	lower border	upper border
-2 vs. 0	4.2 (p=5.65e-5)	4.7 (p=7.58e-6)
2 vs: 0	2.79 (p=0.006)	3.11 (p=0.002)
2 vs. -2	6.90 (p=4.91e-10)	7.97 (p=2.68e-12)

Table 3: Item pairs and their significance scores (residual probability) for curves in Figure 2

sentence. This indicates that modifiers may also delimit the effect size of evaluative adverbs in case they give additional information that narrows down where expectations should be centered.

Once again, for the ablation study, the maxima of the distributions of participants estimations correlate surprisingly well with the real observed data for each item.

Overall, when comparing the maxima of the distribution plots to the subset of the technical database that agrees with the feature that is named in the items (e.g. 200 hp for the items in Table 2), the majority of participants, who are no experts on car technology, have a surprisingly good intuition. Nonetheless data for less known features such as displacement shows that the participants must have enough knowledge about the source feature, otherwise the match between real data and maxima of distribution curves deteriorates.

4.3 Bringing empirical data and regression together

The empirical study supports a transparent mapping from regression to the use of evaluative adverbs and modifiers.

Figure (3) depicts a tripartite plot. The upmost part contains the regression model predictions of acceleration given a power output of 200 hp. The light-blue asterisks mark the three model predictions, whereas the dark-blue triangles pointing upwards and downwards to both sides result from adding or subtracting the standard deviation to or from the regression values. The close vicinity of the marks show how close regression models are in prediction in this case. The central part of the plot is a bar plot of cars with 200 hp (+/- 2.5%) and their respective acceleration value. The undermost part contains the kernel density estimation curves for the upper threshold of an item that asks for maximum speed given power output. The left, light-brown curve displays the distribution for the negative item, the central curve represents the neutral item and the rightmost curve displays estimations for the positive item. Non-neutral items contain both adverb and modifiers in this group. As the figure shows, the maximum of the neutral item's curve is neatly aligned with the maximum of the already narrow distribution of real data points as well as the regression values, which visualizes that regression

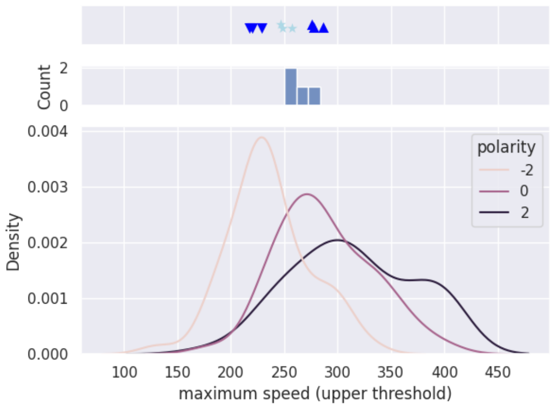

Figure 3: Predicted values (+/- std),
real data and estimations for speed given power output

is very precise and participants have a surprisingly good intuition matching real data. Even more intriguing is the fact that the regression values after modification by either adding or subtracting standard deviation also neatly match the maxima of the curves for the positively and negatively polar items respectively. Across items, one can recognize the pattern that depending on the semantics of the response feature (higher is better or lower is better), non-neutral sentences correspond to the maxima of the respective distributions of estimated values. In regard to thresholds for binary decisions on generating evaluative content, this would mean that if the regression value deviates more than standard deviation from the real value, evaluative content is licensed according to the polarity of the deviation.

This study makes no statement about which specific adverb and/or modifier should be used. It shows that regression values combined with standard deviation are a good starting point for deciding when to become negatively or positively evaluative in a description, but the subtle meaning differences between, e.g., *surprisingly*, *very surprising*, and *totally surprising* were not a topic of this paper. Such a fine-grained lexicalization process, which is a downstreamed module in a pipelined NLG system, requires access to grammatical, semantic and discourse-related constraints, which is outside the scope of the model presented in this paper.

5 Conclusion

For controllable generation of evaluative adverbs in data2text NLG, we devised a learning-based, generalizable approach to approximate thresholds for binary decisions on the presence of evaluative language and validated our findings with an empirical study on the expectations raised by comprehending evaluative adverbs. The elicited data supports the adequacy of the model and shows a surprisingly good match between regression predictions, real data and human estimations. In a pipelined neural generation system, these learning-based models permit determining generation of evaluative language at an early stage in document planning and therefore improve controllability of evaluative content before applying transformers for surface realisation.

Supplementary Materials Availability Statement: Source code of regression models, source code of the web application for the empirical study, empirical data collected during the study as well as source code for analysis of the data including functions for reproducing all graphs and figures in this paper are available online on Github: `https://github.com/MMLangner/evalAdvInData2TextNLG/`. The database containing technical data of cars is proprietary, which is why we are not allowed to distribute it. Please contact the authors for information on how to obtain the technical database in the required format from the ADAC.

References

Simone Balloccu, Steffen Pauws, and Ehud Reiter. 2020. A NLG framework for user tailoring and profiling in healthcare. In *Proceedings of the First Workshop on Smart Personal Health Interfaces co-located with 25th International Conference on Intelligent User Interfaces, SmartPhil@IUI 2020, Cagliari, Italy, March 17, 2020*, volume 2596 of *CEUR Workshop Proceedings*, pages 13–32. CEUR-WS.org.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Fiorella de Rosis and Floriana Grasso. 2000. *Affective Natural Language Generation*, pages 204–218. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. *arXiv*, pages 131–137.

Michael Elhadad. 1991. Generating adjectives to express the speaker's argumentative intent. In *AAAI Conference on Artificial Intelligence*.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2020. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 552–562.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(c):1–64.

Daniel Gutzman. 2015. *Use-Conditional Meaning: Studies in Multidimensional Semantics*. Oxford Studies in Semantics and Pragmatics. Oxford University Press, Oxford.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Maurice Langner and Ralf Klabunde. 2022. Realizing a denial of expectation in pipelined neural data-to-text generation. In *Proceedings of the 6th Workshop on Advances in Argumentation in Artificial Intelligence 2022 co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), Udine, Italy, November 28, 2022*, volume 3354 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Saad Mahamood. 2021. Reproducing a comparison of hedged and non-hedged NLG texts. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 282–285, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Saad Mahamood, Ehud Reiter, and Chris Mellish. 2007. A comparison of hedged and non-hedged nlg texts. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, page 155–158, USA. Association for Computational Linguistics.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge.

Sashank Santhanam and Samira Shaikh. 2019. Emotional neural language generation grounded in situational contexts. *CoRR*, abs/1911.11161.

Xiaoyu Shen, Ernie Chang, Hui Su, Jie Zhou, and Dietrich Klakow. 2020. Neural data-To-Text generation via jointly learning the segmentation and correspondence. *arXiv*, (2019):7155–7165.

Ishika Singh, Ahsan Barkati, Tushar Goswamy, and Ashutosh Modi. 2020. Adapting a language model for controlled affective text generation. *CoRR*, abs/2011.04000.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.