Reports of personal experiences and stories in argumentation: datasets and analysis

Neele Falk and **Gabriella Lapesa**
Institute for Natural Language Processing, University of Stuttgart
{neele.falk,gabriella.lapesa}@ims.uni-stuttgart.de

Abstract

Reports of personal experiences or stories can play a crucial role in argumentation, as they represent an immediate and (often) relatable way to back up one's position with respect to a given topic. They are easy to understand and increase empathy: this makes them powerful in argumentation. The impact of personal reports and stories in argumentation has been studied in the Social Sciences, but it is still largely underexplored in NLP. Our work is the first step towards filling this gap: our goal is to develop robust classifiers to identify documents containing personal experiences and reports. The main challenge is the scarcity of annotated data: our solution is to leverage existing annotations to be able to scale-up the analysis. Our contribution is two-fold. First, we conduct a set of in-domain and cross-domain experiments involving three datasets (two from Argument Mining, one from the Social Sciences), modeling architectures, training setups and fine-tuning options tailored to the involved domains. We show that despite the differences among datasets and annotations, robust cross-domain classification is possible. Second, we employ linear regression for performance mining, identifying performance trends both for overall classification performance and individual classifier predictions.

1 Introduction

Although personal narratives and experiences naturally fill an important place in our everyday discussions, they still do not conform to the classic ideal of a "good" argument. A "good" argument contains facts and logical conclusions, but as soon as more personal or emotional nuances are involved, it deviates from the norm. According to the theory of Deliberative Democracy (Habermas, 1996; Fishkin, 1995), the discourse that precedes political decisions plays a central role. Here, too, until the so-called *affective turn* in Social Sciences (Hoggett and Thompson, 2012), the assumption was that an exchange of arguments that is as rational as possible can lead to better political decisions. Recent studies in Deliberative Theory, however, are increasingly concerned with the interplay between classical argumentation and *alternative forms of argumentation*, which include personal experiences, narratives and emotions, and their positive effects on discourse (Polletta and Lee, 2006; Esau, 2018; Gerber et al., 2018; Maia et al., 2020). A norm that only allows rational and logical argumentation firstly does not correspond to human realistic communication and secondly bears the danger that less educated groups or groups in which other communication standards prevail are marginalized. Arguments with personal experiences are therefore important to fulfill one of the core deliberative standards, namely inclusivity (Polletta and Gardner, 2018).

The following example is taken from a discussion about regulations to ban peanut products on airlines and illustrates further possible positive effects of arguments with personal experiences: *My daughter has been tested 4 times for her allergy to peanuts. She is in the highest category of reactivity which means if peanuts are being ingested in her vicinity, she could die. A buffer zone simply doesn't work in a confined space such as an airline.*[1] The perspective of the frightened parent and their allergy-affected daughter is likely to elicit emphatic reactions from the other participants in the discussion and illustrates the rationale of proponents of a peanut ban.

Our work represents the first step towards identifying these arguments at a large scale by developing models for automatic detection of contributions with personal experiences and stories (to which we refer with the general label of *reports* in the paper). This is the necessary first step to be able to further examine the role of such arguments in

[1] The comment was taken from the e-rulemaking platform regulationroom.org.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5530 - 5553

reasoning and their relationship with Argument Quality (Wachsmuth et al., 2017a) and Discourse Quality (Steenbergen et al., 2003),

To tackle this task we collect available datasets and investigate the best strategy to merge the available sources. In two of the datasets we employ, produced by the Argument Mining community (Regulation Room and Change My View) our targeted phenomenon is annotated as *testimony*. In the third dataset, from the Social Science community (Europolis) it is annotated as *storytelling*. The two categories are obviously not fully overlapping, but we hypothesize that they share a conceptual core which can be leveraged for robust classification.

We perform a large set of in-domain, out-domain and cross-domain experiments including different domain-adaptation strategies for the pre-trained language models used in the classification. We analyze our results using performance mining and show that the cross-domain training setup leads to the most robust results and also has the most positive effect when used with a domain-adapted LM. We also conduct regression-based error analysis and compare the most salient features of the two largest datasets that get picked up by the most robust model and show that prototypical textual properties of reports can push the model in the right direction (probability of reports close to decision boundary) but can also lead to over-generalization (higher probability of reports for false positives).

2 Related work

Argument Mining & NLP The automatic detection of arguments with personal experiences was first tackled by Park and Cardie (2014) with the goal to classify claims as verifiable or unverifiable and to be able to detect what type of evidence would be necessary as a consequence. They defined the subcategory *verifiable experiential* for verifiable claims that contain personal experiences. Their dataset contains comments from RegulationRoom,[2] an e-rulemaking platform with the goal of enabling online deliberation: governmental institutions or companies can have their proposals about new regulations discussed by citizens to get feedback. Park and Cardie (2014) conducted classification experiments with a SVM using different feature sets hypothesizing that the amount of past tense and first personal pronouns would be most predictive for verifiable experientials. They achieved a F1-score of $\sim 70\%$ on the RegulationRoom dataset. This work was further extended in Park et al. (2015b) and reformulated as a sequence classification task. Park et al. (2015a) and Park and Cardie (2018) developed a new annotation scheme targeted at elementary units in arguments. This schema introduces **testimony** as an elementary unit which corresponds to a proposition about the author's personal state or experience. A similar evidence type, called *anecdote* and defined as a personal experience of the author or a narration of a concrete example or event was classified in news editorials (Al-Khatib et al., 2016, 2017). Song et al. (2016) build a database for claims and suitable anecdotes focusing on stories with a clear narrative structure and popular main characters (e.g. the Dalai Lama). Wang et al. (2019) model different persuasion strategies in dialogues targeting social good (e.g. fund raising): personal stories which exemplify positive outcomes and benefits of a donation are one of these strategies.

Social Sciences The role of personal narratives in digital and deliberative democracy has gained more attention in the recent years. Polletta and Lee (2006) were the first who investigated the role of personal narratives in online argumentation. Their data was further analyzed in Black (2008) and Black (2013), who emphasized the importance of personal narratives in discussions for forming group identity and understanding others' perspectives. The relationship between storytelling and emotions was investigated in Esau (2018) who pointed out that these are especially useful when discussing social problems that cannot be addressed with factual information alone. Maia et al. (2020) annotated the functions of storytelling (e.g. do people tell an experience as a disclosure of harm or to propose a solution?) and examined how the different types of narratives effect the quality of a discussion. As far as available annotation is concerned, we conduct our experiments on the Europolis corpus (Gerber et al., 2018). Europolis contains spoken contributions from a transnational poll, in which citizens from different European countries got together to discuss about the EU and the topic immigration. The spoken contribution have been annotated with different aspects of deliberative quality (e.g. does the speaker show respect or value other participants?), and one of these aspects includes alternative forms of communication. Relevant to our work is the annotation category

[2] http://regulationroom.org/

storytelling, marking those contributions which contain personal experiences or concrete examples of a speaker's own country.

3 Datasets

Regulation Room (*RegRoom)* Our experiments are based on the final version of the Cornell eRulemaking Corpus (CDCP) (Park and Cardie, 2018) It contains 725 comments from Regulation Room, discussing consumer debt collection practices in the United States. The comments are annotated with different proposition types, and our category of interest is *testimony*.

Change My View (*CMV*) contains 344 comments from the subreddit *ChangeMyView*[3] (Egawa et al., 2019) and is annotated with a similar schema as introduced in Park et al. (2015a). Our reference category is *testimony*.

Europolis (Gerber et al., 2018), already introduced in section 2, contains a total of 856 transcribed speech contributions whose original language was German, French, and Polish (only available in the English translation).[4] Europolis is annotated along many deliberative quality dimensions (Steenbergen et al., 2003), and our reference category is *storytelling*.

From now on, we will use the neutral term REPORT as our positive label, of which *testimony* and *storytelling* are dataset-specific declinations. Note that, while *storytelling* is annotated at the document level in Europolis, RegRoom and CMV contain span-level annotation of *testimony*: for our experiments, a document containing a *testimony* span is considered as a positive instance of REPORT. For all datasets, reports are the minority class (RegRoom: 41%; CMV: 37%, Europolis: 35%).

Table 1 displays one example per dataset. In the example of Europolis the participant describes the general situation in their country in a more objective manner, thus reflecting a quite broad definition of a personal narrative. The RegRoom example is very personal and emotional in its tone, displaying more prototypical features of a personal experience. The CMV example is somewhere in between: it departs from a personal experience but it targets a general situation, and has a more objective tone.

3.1 Preprocessing and Feature extraction

The datasets were preprocessed by removing time stamps and URLs. With freely available tools, we extracted a total of 51 features (henceforth, *contribution-level* features) from four categories:[5]

Surface features (6 features), e.g., length in tokens; average amount of characters and syllables per word. We hypothesize that longer comments are more likely to contain reports (verbose retelling of concrete stories / examples).

Syntactic features (6), e.g., relative amount of fine-grained part-of-speech tags per comment (e.g., personal pronouns, past tense, auxiliaries, named entities). Specific categories, e.g. first-person pronouns and past tense verbs ("I had an unpleasant experience..."), are likely to be predictive of reports.

Textual complexity (19), with different measures of lexical diversity, lexical sophistication and readability. While prototypical reports are expected to exhibit lower textual complexity (character repetitions, more concrete concepts), the modulation of complexity in the reports in our datasets is an open question.

Sentiment/Polarity (20), e.g. amount of positive or negative adjectives/nouns, amount of specific emotions (joy, fear). We hypothesize that comments with reports will have more marked polarity.

4 Experiments

Task We perform binary classification at the document level: forum posts in RegRoom and CMV, spoken contributions in Europolis. We create 10 random train / dev / test splits using 15 % as development and 20 % as test data and we ensure that every document is part of the test set at least once.

Setups We experiment with the following training/test setups:

In-domain: we train and test the models on the same dataset.

Out-domain: we train the models on a single dataset and test them on the other two individually (e.g. train on Europolis and test on CMV and RegRoom). We also concatenate two datasets and test the corresponding model on the missing one, e.g. train on a joined set of CMV and RegRoom and test on Europolis (*2vs1*).

[3] On CMV, users exchange views on a variety of different topics and can reward other comments if they are convincing or have led to a change of their opinion. Research on persuasion on CMV has targeted, for example, interaction dynamics and stylistic choices (Tan et al., 2016) or the effect of social pressure (Jain and Srivastava, 2021).

[4] We translated the German and French transcriptions into English using DeepL and used the professional English translation of the Polish data. Refer to A.1 for more details.

[5] For an overview of the features and mean values, as well as extraction details, refer to appendix A.2.1.

RegRoom (testimony)	I was never informed by Bank of America that they sold my credit card and closed the card. When I realized it, I paid it off immediately. During that quarter, after long illnesses, my Father and Mother both passed (within 31 days of each other) and frankly, credit card payments were not in the forefront of my thinking. ALSO, just because a bank or credit card company has been exempted from Usury laws does not mean they do not commit the violation! THAT needs to be stopped!
CMV (testimony)	I used to work at an aquatic center that had women's only hours once a week during which only female lifeguards would cover the pool. As it was explained to me, the primary purpose of these hours was to give Muslim and Orthodox women a place to swim without violating their religion. It was common for non-religious women to swim during these times because they felt more comfortable not having to swim in front of men. I don't know what the rationale is at your gym . I would argue that yes, the women's only hours there may be sexist , but they also allow women to partake in an activity that would otherwise be prohibited to them during normal hours
Europolis (storytelling)	In Slovenia, we have a lot of immigrants from the non-EU countries, especially in the health care sector, because we need specialists in Slovenia. Slovenians do not want to work in this sector so of course people from other countries are coming to work there.

Table 1: Examples of reports (testimony or storytelling) in the three datasets

Cross-domain: we create one training set which is the concatenation of the training sets for of the in-domain experiments and test it on the each dataset-specific test set (*all*).

Classification models We experiment with the following classification models (cf. section B.1 for more details and hyper-parameters):

- *Feature-based*: we train a random-forest classifier with the features mentioned above (51 in total, appendix section 3.1).
- *BoW*: each contribution is represented as a count vector with the frequency counts for the 5000 most frequent words (the vocabulary was constructed based on the fusion of all datasets) which is fed into a random-forest classifier.
- *FeedforwardNN*: a contribution is represented as the average of the embeddings the words occuring in it and fed into a feed-forward neural network with one hidden layer of size 300 and a `ReLU`.
- *BERT*: we fine-tune BERT (Devlin et al., 2019) with a classification head on the task of predicting whether a contribution contains a report or not.
- *Domain-adapted BERT (3 models)*: we fine-tune the underlying language model (LM) with the masked language modeling objective and next sentence prediction on domains that would match the domains of our target datasets. For the Europolis dataset we sample ∼1M sentences from Europarl (Koehn, 2005) [6], for CMV and RegRoom we sample ∼1M sentences from the Webis-CMV-20 corpus (Al-Khatib et al., 2020) and ∼1M sentences from the args.me corpus (Ajjour et al., 2019). We fine-tuned one LM on each domain (*BERT-adapt-europarl*, *BERT-adapt-argue*) and one on the concatenation of the two (*BERT-adapt-mixed*).

Results Table 11 in the Appendix displays the results for all models for each training setup (averaged over all splits), with significance values. The results show that BERT outperforms the other models for all training/test setups and training on the joined dataset (*all*) works well for all test corpora (the feature-based classifiers benefit especially from it). Fine-tuning the LM on the concatenated domains (*BERT-adapt-mixed*) yields the best results when trained on *all* with a macro F1-score of 0.76 (Europolis), 0.85 (CMV) and 0.94 (RegRoom). The non-domain-adapted BERT is more robust if in-domain and out-domain experiments are also taken into account, e.g. a drop in performance from 0.72 to 0.65 F1 can be observed on Europolis when trained in-domain using *Bert-adapt-europarl*. In the following section, we will employ linear regression to build a comprehensive picture of these performance trends.

5 Analysis

To get a statistically informed understanding of the different factors influencing the behavior of our models on the different datasets, we employ linear regression. Our dependent variables are **aggregated performance** (F1 macro) in section 5.1 and **model predictions** on individual items (probability of report) in section 5.2.

For the aggregated performance analysis in Section 5.1, our independent variables (IV) are: the different experimental configurations i.e., combinations of classifier architectures (referred to as

[6] We used the EN monolingual and the English translation from the DE-EN and FR-EN parallel data.

"model" in the tables), training setup and test corpus as well as their interactions, as specified in the formula:[7]

```
F1macro ~ (model +
training setup + test corpus)^3
```

For the model predictions (section 5.2) our independent variables are: the contribution-level features in 3.1 and a subset of the experimental configurations (training setup and test corpus), as well as the interactions between the contribution-level features and the experimental configurations.

In both cases, our analysis proceeds in three steps. First, we fit incrementally complex models and assess their fit in terms of adjusted R^2 (proportion of explained variance) and significance with respect to the less complex models (i.e., models with fewer independent variables). At this step, we also check for multicollinearities. Next, once we identify the most explanatory regression model (i.e., the set of independent variables that maximizes the fit to the dependent variables), we proceed to identify its most explanatory IVs in terms of explained variance and significance (e.g., does the choice of training setup determine a strong difference in the performance of our classifiers?). Last, we identify the best values for the IVs (e.g., which of the training setups guarantees best F1?) by visualizing predicted performance with the help of effect displays (Fox, 2003), which show the partial effect of one (or more) parameters by marginalizing over all other parameters.

5.1 Aggregated Performance

With the regression analysis presented in this section, we are interested in capturing the pattern of variation exhibited by our experimental configurations, with a focus on the effects of in-domain, out-domain and cross-domain training and domain adaptation.

Data: we consider all experimental runs from the in-domain, 2vs1 out-domain, and cross-domain training setups, resulting in 630 data points.[8] We code the levels of the IV `training setup` as *in-domain* (trained/tested on the same corpus), *2vs1* (out-domain training) and *all* (cross-domain training).

[7]The formula follows the R syntax, `^3` denotes the 3-way interactions among the terms included between parenthesis, as well as their lower-order terms.

[8]This number of data points results from the multiplication of: the number of classifiers (7), test corpora (3), training setups (3), splits (10). We conducted a sanity-check analysis with split id as IV, finding no significant effect.

IV	adjusted R^2	sign.
test corpus	11.6 %	
+ training setup	23.6 %	***
+ model	44.1 %	***
+ all two-way interactions	65.3 %	***
+ all three-way interactions	77.2%	***

Table 2: Adjusted R^2 for each regression model predicting the F1 macro with step-wise addition of IVs. Significance adding more predictors is tested using the `anova` function from R.

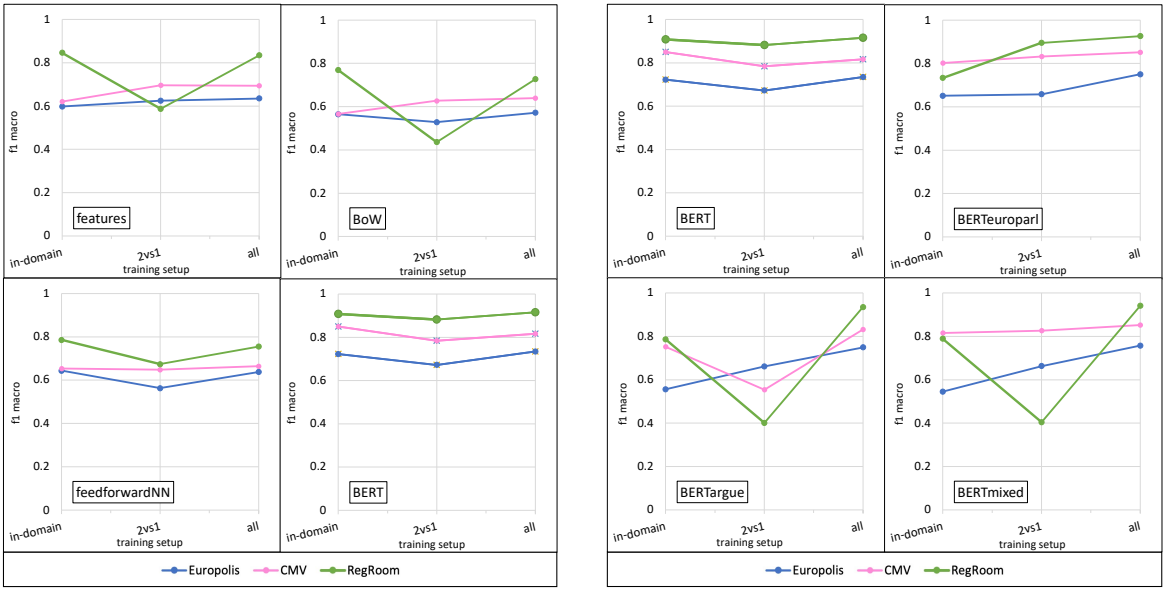
Results: Table 2 reports the fit of the simplest model, which only contains `test corpus` as IV, and of the incrementally more complex models. All IVs and interactions between explain a significant additional amount of variance[9]. The most explanatory model contains the three IVs and their two-way and three-way interactions, for a total explained variance of $R^2 = 77.2\%$.

Table 12 in the appendix displays the portion of variance accounted for by each IV in the final regression model. Most variance in performance (20% of the R^2) is determined by the classifier architecture (IV: `model`).The high amount of additional variance explained by the three-way interaction, however, (12%) indicates that the effects of the training setup and test corpus differ significantly between classifier architectures.

What is the effect of training setup and domain-adaption? To get a more detailed picture of the relationship between experimental configurations, classifier architectures, and the predicted performance we can have a look at figure 1 which displays the effect plot for corresponding the three-way interaction.

Comparing the different training setups, it becomes evident that cross-domain training setup is most robust: the difference between the predicted performance for all test corpora when trained on mixed domains (`training setup` = *all*), i.e. the difference between the colored lines at the rightmost position of the y-axis, is similar for the different classifier architectures (across all panels), while in-domain or out-domain training can lead to very different performance predictions. One can for example observe a low predicted performance when using features or BoW as classifier architecture for RegRoom in an out-domain setting (drop in the green line in the two upper panels in figure 1a). Out-domain training is rarely beneficial if no

[9]$p \leq 0.05$ '*' $p \leq 0.01$ '**' $p \leq 0.001$ '***'



(a) Non-domain-adapted classifier architectures

(b) BERT + its domain-adapted variants

Figure 1: Effect plot, 3-way interaction: predicted performance (x-axis), training setup (y-axis), classifier architectures (each panel), test corpus (colored lines).

domain adaptation is performed, small positive effects can only be obtained for CMV and Europolis with the feature-based classifier (increase in the pink and blue line in the upper left panel in figure 1a). The largest positive effects for cross-domain training in contrast to in-domain training can be observed for Europolis (increase in the blue line from *in-domain* to *all* across most panels).

Figure 1b compares the performance of the non-adapted BERT model (top-left panel) to its three domain-adapted variants. This comparison helps us further characterize the domain-specificity of the annotations. First of all, we can observe that BERT is more robust across different training setups and test-corpora. If we look at the course of the lines in the upper left panel in figure 1b we can see that it is relatively stable, almost parallel: the panel shows only a small drop in predicted performance for out-domain training and a small improvement with cross-domain training for RegRoom and Europolis. In contrast, the domain-adapted variants are subject to a much higher variance (e.g., a strong drop in predicted performance for BERTargue and BERTmixed, trained out-domain, tested on RegRoom).

Domain-adaptation can be useful in a cross-domain training setup, as we can see slight improvements for the predicted performance comparing the colored lines at the right-most position of the y-axis of the upper left panel with the others, and this observation is similar for all domain-adapted variants.

When comparing the different domain-adapted variants, the best model is BERTeuroparl which works well for all test corpora for out-domain and cross-domain training, whereas BERTargue leads to a low predicted performance when used with Europolis or in an out-domain training setup. This indicates that the LM adapted to a deliberative context (since Europarl contains parliamentary debates) is compatible with all test corpora, whereas BERTargue may be too domain-specific.

5.2 Item-based predictions: error analysis

In this section, we employ regression for error analysis, with the goal of finding out which linguistic properties of a contribution drive the model prediction away or towards the gold label. In this analysis, we focus on the predictions of the most robust classifier if both in-domain and out/cross-domain are taken into account, namely the non-adapted BERT (see section 5.1).

We extract two subsets predictions from the `BERT` and perform regression analysis on each of them. We examine the predictions for the false positives (FPs) in order to find out in which cases the model overgeneralizes (which properties cause the model to predict a high probability for reports?), but also to find out what puts the model on the right track (probability close to the decision boundary). Similarly, we can examine the predictions for FNs,

where typical features for reports may ensure that the model's probabilities go in the right direction, while particularly atypical features or a disadvantageous training/test combination may cause the model to incorrectly predict lower probabilities.

Dependent variable In this analysis, our dependent variable is the **probability that a comment contains a report**. The distribution of the probabilities, however, is heavily skewed towards the upper and lower bound. We therefore transform the individual probabilities with a log transformation to reduce the skewness. In order to gain the same advantage for both error types and to be able to better compare the two in the effect plots, we invert the probabilities of the FPs and map them to the same range as the false negatives ($1 - p(reports)$). For both error types, the resulting values range between -10 and -0.5 and in both cases the predicted classification label would change to the gold label when the probability exceeds the upper threshold.[10] The distribution of the dependent variables is displayed in the histograms in section D.1. Thanks to this transformation, a positive effect or an increase in the dependent variable can be interpreted as beneficial for both error types.

Independent variables and model selection In our analysis, we focus on the difference between the two largest datasets and include *Europolis* and *RegRoom* as `test corpus`, excluding *CMV*. Qualitatively, Europolis and RegRoom share the deliberative focus, while CMV is persuasion-driven and more likely to exhibit idiosyncratic properties that we would not be able to sufficiently discuss here for reasons of space. The `training setup` variable contains *RegRoom*, *Europolis*, and *all* so that we can examine the effects of in-domain, out-domain (this time in a 1vs1 version), and a cross-domain training setup. The final subsets for this analysis contain 776 datapoints for the FPs and 1,212 data points for the FNs.

Our analysis builds on the assumption that specific linguistic properties of the input drive the predictions towards or away from the gold label. For this reason, the first core of IVs is represented by the contribution-level features used to train the feature-based random forest classifier (cf. section 3.1). To avoid multicollinearities and for simplification purposes we applied a correlation-based feature reduction methodology whose criteria are described in section D.2 in the Appendix. All features were further centered and scaled.

[10] We trained the regression models without a log transformation. This led to significantly worse results for the FNs, no large difference for the FPs. We therefore report the models with the log transformation.

IV	false positives		false negatives	
	adjusted R^2	sign.	adjusted R^2	sign
features	5.2 %		18.9 %	
+ test corpus	5.1%	-	23.2 %	***
+ training setup	5.3%	-	31.0 %	***
+ two-way interactions	8.7%	***	36.0 %	***
+ three-way interactions	9.7%	-	40.0 %	***

Table 3: Adjusted R^2 and significance for each regression model (FPs, FNs) predicting the probability of reports with step-wise addition of IV. Significance between the richer model and its nested counterpart is tested using anova.

We incrementally added the experimental configuration features (`training setup` and `test corpus`), as well as two- and three-way interactions, on top of the contribution-level ones. Similar to section 5.1 we compare nested models starting from the one containing only contribution-level features as IVs, effectively assessing whether the variation in performance is only due to linguistic properties or whether certain linguistic properties affect specific experimental configurations.

Results: The results for both subsets (FPs and FNs) are shown in table 3. It is noticeable that the most explanatory regression model for the FNs can explain significantly more variance (40%) than the one for the FPs (10%); reasons for this could be the smaller amount of data and the poorer distribution of the dependent variable. In both cases, the contribution-level features alone explain roughly half of the variance accounted for by the most complex models. For the FPs, `training setup` and `test corpus` significantly contribute to the fit only when the two-way interactions are taken into account. On the other hand, for the FNs, all incremental steps significantly improve the fit.[11]

Which feature types have the greatest impact on the errors? Table 4 summarizes the relative contribution of different feature groups to the total amount of explained variance.[12] The surface

[11] To identify the most explanatory regression models (set of IVs) for FPs and FNs we employed step-wise model selection. A detailed discussion of this process, along with the list of selected IVs and their explained variance and significance can be found in appendix section D.2, in tables 13 and 14.

[12] The sum of the explained variance (R^2) of a feature group is the sum of the R^2 of the individual features in that group plus their interactions with `training setup` and/or `test corpus`.) The feature type "experimental configurations" contains only the amount explained by `training setup`,

group of IV	false positives	false negatives
experimental configurations	23 %	27 %
surface features	0 %	3 %
syntactic features	27 %	32 %
sentiment/polarity	23 %	21 %
textual complexity	27 %	17 %

Table 4: Effect sizes (relative amount of R^2) for different groups of IVs in the most explanatory regression models for the FPs and FNs.

based features have an extremely low impact on the prediction of the probability of reports: this is surprising given that previous work has shown that the length of a contribution has a great impact on model predictions (*length bias*, cf. Wachsmuth and Werner (2020)). The other feature groups are all involved relatively equally in explaining performance variance, with textual complexity and syntactic features dominating FPs, and syntactic features FNs.

What is the impact of contribution-level features? If we look at the effect plots for the most explanatory IVs, we can see which features are particularly helpful for the errors: positive effects drive the model towards the gold label. The effects for FNs highlight the dominant role of syntactic features. For example, we see positive effects for past tense verbs, personal pronouns, and post length (cf. figure 5, section D.2 in the appendix) which also confirm the hypotheses that these features are prototypical for reports. Some features are more discriminative when training on specific data: for example, the effect for personal pronouns is positive when training on RegRoom but slightly negative when training on Europolis (cf. figure 6, section D.2 in the appendix).

While some of the features can lead to overgeneralization of the model we can identify a feature of textual complexity that proves to be useful for FPs. The effect plot in figure 7 (section D.2 in the appendix) displays the interaction between mean average type token ratio (*mattr50*) and `training setup`. This feature puts the model on the right track when trained on RegRoom but moves the probability further away from the decision boundary when trained on Europolis or a mix. This example once again emphasizes that RegRoom, in combination with specific features, can take on an advantageous role as a training corpus.

`test corpus`, and their interaction.

Storytelling vs. testimony: discussion The reports of personal experiences in RegRoom resemble prototypical narratives; they contain very personal and individual experiences, exhibiting many of the expected characteristics of a typical report. Our regression-informed error analysis shows that training on RegRoom positively impacts performance, while the opposite is often true for Europolis. Indeed, the reports in Europolis can describe more general experiences or a concrete situation in a country (e.g. in the example from Europolis, table 1). As a result, they are less prototypical and difficult to detect based on structural and linguistic features. Ideally, through training on Europolis or on mix, it is possible to recognize reports that are not only about an individual experience, but about a collective one, an aspect that is especially important in a discussion with a deliberative focus. An initial empirical investigation of the argument quality of the different types of reports gives evidence that reports exhibit a higher quality than contributions without reports (see section E in the appendix a detailed description of a pilot case-study). An interesting research question in this context is to what extent the individuality of a report influences quality or to what the commonality of the reported experience positively impacts deliberation.

6 Conclusion

This work targeted the automatic identification of personal experiences and stories in argumentation. Leveraging available annotation in three argumentative datasets (two, Regulation Room and Change My View, from the Argument Mining community; one, Europolis, from the Deliberative Theory community), we evaluated different classifier architectures and training setups. Mixing training data from different domains leads to robust results across all corpora and models and boosts the performance of the classifiers based on the domain-adapted LMs.[13]

Our experiments established an empirical foundation that will allow us to investigate the target phenomenon at a larger scale and will lead to a better understanding of the compatibility of the underlying annotations (storytelling from Deliberative Theory and testimony in Argument Mining) as well as of the impact of storytelling/testimony on the quality of an argument.

[13]The experimental code and the dataset splits are available at https://github.com/Blubberli/storytestimony

Acknowledgments

The research reported in this paper has been funded by Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB (Powering up e-deliberation: towards AI-supported moderation). We thank the anonymous reviewers for their valuable feedback. We would also like to thank Eva Maria Vecchi, Enrica Troiano, Sebastian Padó, and Jonas Kuhn for their feedback on different versions of this work. We also thank Marlène Gerber for giving us access to the Europolis dataset.

Ethics statement

Two of the three datasets employed in this work, *Regulation Room* and `Change My View`, are publicly available along with the annotation used in our experiments. *Europolis* has been kindly provided to us and the splits will be relased at the terms and conditions of its owners.

Our experiments and analysis do not employ any author-specific feature. We cannot exclude, however, that the reporting of very specific personal experiences may facilitate the identification of the author of a post.

As far as the societal impact of the use of reports of personal experiences and stories is concerned, we are fully aware that they can be a two-edged sword. On the one hand, as pointed out in the paper, the use of personal experiences triggers empathy and promotes inclusivity. On the other hand, however, the non verifiable nature of such reports makes them an easy vehicle to generate and spread fake news, and this, coupled with the emotional loading reports tend to have, makes them a useful tool highlight certain episodes for propagandistic purposes. Robust computational models to identify human-generated reports of personal experiences and a deeper understanding of the linguistic features of these contributions can also serve as a first step into the identification of false stories generated by automatic systems and of those human-generated ones that are best candidates to serve manipulation purposes.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59, Berlin Heidelberg New York. Springer.

Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7067–7072. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Laura W. Black. 2008. Deliberation, storytelling, and dialogic moments. *Communication Theory*, 18(1):93–116.

Laura W. Black. 2013. Framing democracy and conflict through storytelling in deliberative groups. *Journal of Public Deliberation*, 9(1).

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821.

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.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Katharina Esau. 2018. Capturing citizens' values: On the role of narratives and emotions in digital participation. *Analyse und Kritik*, 40(1):55–72.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key?

In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Fishkin. 1995. *The Voice of the People: Public Opinion and Democracy*. Yale University Press.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):p221 – 233.

John Fox. 2003. Effect displays in r for generalised linear models. *Journal of Statistical Software*, 8(15):1–27.

Marlène Gerber, André Bächtiger, Susumu Shikano, Simon Reber, and Samuel Rohr. 2018. Deliberative abilities and influence in a transnational deliberative poll (europolis). *British Journal of Political Science*, 48(4):1093–1118.

Jürgen Habermas. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. The MIT Press.

P. Hoggett and S. Thompson. 2012. *Politics and the Emotions: The Affective Turn in Contemporary Political Studies*. Bloomsbury Publishing.

Ayush Jain and Shashank Srivastava. 2021. Does social pressure drive persuasion in online fora? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9201–9208, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior research methods*, 50(3):1030–1046.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rousiley C. M. Maia, Danila Cal, Janine Bargas, and Neylson J. B. Crepalde. 2020. Which types of reason-giving and storytelling are good for deliberation? assessing the discussion dynamics in legislative and citizen forums. *European Political Science Review*, 12(2):113–132.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates, Inc.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015a. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, page 206–210, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joonsuk Park, Arzoo Katiyar, and Bishan Yang. 2015b. Conditional random fields for identifying appropriate types of support for propositions in online user comments. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 39–44, Denver, CO. Association for Computational Linguistics.

Francesca Polletta and Beth Gharrity Gardner. 2018. The forms of deliberative communication. *The Oxford Handbook of deliberative democracy*, pages 69–85.

Francesca Polletta and John Lee. 2006. Is telling stories good for democracy? rhetoric in public deliberation after 9/ii. *American Sociological Review*, 71(5):699–723.

Wei Song, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Anecdote recognition and recommendation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2592–2602, Osaka, Japan. The COLING 2016 Organizing Committee.

M. Steenbergen, Andre Baechtiger, Markus Spörndli, and J. Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017c. Dagstuhl-15512-argquality. Technical report.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

A Preprocessing

A.1 The Europolis dataset and its automatic translation

The discussions collected in the Europolis dataset were held in 2009 in order to investigate whether and how deliberation can take place in a multilingual/transnational setting, and what effects deliberation can have for citizens (e.g. increasing political engagement or interest). Participants from 27 countries participated to simultaneously translated small-groups discussions about the topics of climate change and immigration. A sample of the discussions of 13 groups, whose original language was French, German, and Polish, has been annotated by Gerber et al. (2018) on different dimensions of the Discourse Quality Index (DQI) by Steenbergen et al. (2003).

The dataset contains the professional translation of the Polish speeches into English, as well as the transcription of the speeches French and German, which we translated into English with DeepL (`https://www.deepl.com/translator`). The quality of the English translation of the French and German texts has been checked by one native speaker each. Native speakers were instructed to check for the semantic integrity of the conveyed message and the grammaticality of the output.

We acknowledge that this quality check does not rule out the possibility that the automatic translation has distorted the linguistic features we employ for the feature-based classifiers and in the regression analysis. We do believe, however, that the pattern of results in table 11 is still relatively stable despite the potential distortion. More specifically, the F1 macro of the feature-based classifier trained and tested on Europolis is more than acceptable: had the translation affected the features dramatically, the performance would have dropped much more with respect to the feature-based representations trained/tested on in-domain native English models (F1 macro: Europolis=0.60; CMV=0.62; RegRoom=0.85). Moreover, had the features been distorted inconsistently, the generalization learnt from the out-domain (2vs1) feature-based classifier tested on Europolis and trained on CMV+RegRoom would have exhibited a drop in performance, not a gain (test: Europolis, 2vs1 = 0.63; test: Europolis, in domain = 0.60).

As regards the fact that translated features are also employed in the regression analysis at the item level, we believe that the validity of this analysis is not affected, because the translated text was the input of the BERT classifiers which in turn produced the probability values we analyse with the features extracted from that text.

A.2 Features

This section provides the details regarding the features briefly introduced in section 3.1 and employed in the experiments. Tables 5, 6, 7, 8 and 9 list all features names grouped by type, along with a short description and information on the values. For each feature, table 10 displays the mean value per corpus, separately for documents with a positive (report) vs. negative label (no report).

A.2.1 Extraction details

Syntactic features (Table 5)

- Quantify the relative amount of a certain part-of-speech tag (e.g., adverbs, adjectives, named entities) in a document.

These features have been computed using `spacy` (`https://spacy.io`) for tagging, with a model trained on English blogs, news and online comments (en_core_web_md).

Surface features (Table 6)

- Measure the length of the sentences / words, the number of complex words and a combination of these information in the form of readability metrics (flesch reading ease and gunning fog index)

The two readability metrics (*Flesch Reading Ease* (Flesch, 1948) and *Gunning Fog Index*) have been computed with the python package `readability` [14].

Lexical diversity (Table 7)

- These metrics are different variants of the type/token ratio, designed to be less sensitive to text length.

These features has been extracted with `TAALED` [15]. For more details refer to Kyle et al. (2021).

[14] `https://github.com/andreasvc/readability/`

[15] `https://www.linguisticanalysistools.org/taaled.html`

feature name	value
adverbs	relative amount of adverbs in a contribution
auxiliary	relative amount of auxiliary verbs in a contribution
named_entities	relative amount of named entities in a contribution
past_tense	relative amount of past tense verbs in a contribution
personal_pronouns	relative amount of first personal pronouns in a contribution
subordinate_conj	relative amount of subordinate conjunctions in a contribution

Table 5: Syntactic features: overview. Total number: 6.

feature name	description	value
postlength	the number of words of a comment	raw frequency
chars_per_word	number of characters per word	mean of all scores
syllables_per_word	number of syllables per word	mean of all scores
long_words	number of words with more than 7 characters	mean of all scores
flesch	flesch score based on average length of a sentence and average number of syllables per word	mean of all scores
gunning fog	weighted average of the number of words per sentence and number of long words (words with more than three syllables)	mean of all scores

Table 6: Surface features: overview. Total number: 6.

Lexical sophistication (Table 8) The metrics of lexical sophistication are computed based on word / co-occurrence information taken from existing reference corpora and word lists, e.g. the Corpus of Contemporary American English (COCA) or the (Averil Coxhead's) High-Incidence Academic Word List (AWL).

- *Word Frequency*: given a text, its word frequency value is calculated as the average of the frequencies of the words occurring in it, based on frequency estimates from different reference corpora (see above).
- *Range indices*: given a text, its range indices are calculated as the average of document frequencies of the words occurring in it, estimated on reference corpora.
- *Mutual information*: uses the mutual information scores of academic bigrams, computed based on reference corpora.
- *Academic list indices* relative amount of academic words and n-grams using word lists as reference.
- *(Psycholinguistic) Word Information*: average of different psycholinguistic scores (e.g. concreteness, familiarity, imageability).
- *Semantic network*s: measures indicate how word forms are semantically related. More sophisticated texts contain words with fewer senses and words with more hypernyms (more subordinate terms).
- *Contextual distinctiveness* measures the diversity of contexts in which a word is encountered, e.g. "love" occurs in many different contexts, while the number of contexts where the word "bride" occurs is more restricted.

This set of features has been extracted with TAALES [16], see Kyle et al. (2018) for details.

Sentiment features (Table 9) The sentiment features rely on a number of pre-existing sentiment, social-positioning and cognition dictionaries (e.g. EmoLex) which serve as a look-up table.

- The features correspond to macro-feature component scores produced by PCA

To extract the sentiment features, we use SEANCE [17]. The metrics and the retrieval of the feature components are described in Crossley et al. (2017).

[16] https://www.linguisticanalysistools.org/taales.html

[17] https://www.linguisticanalysistools.org/seance.html

feature name	description	value
mtld_original_aw	computes type token ratio of increased word windows / segments	mean of all scores
mattr50_aw	Moving average type token ratio (50-word window)	mean of all scores
hdd42_aw	for each word type, compute the probability of encountering one of it's tokens in a random sample of 42 tokens, same range as type token ratio	mean of all scores

Table 7: Lexical diversity features: overview. Total number: 3.

feature name	feature type	description	value
COCA_spoken_Bigram_Frequency	N-gram	academic bigram frequency scores	mean of all scores
COCA_spoken_Frequency_AW	Word Frequency	frequency scores of words in spoken language	mean of all scores
COCA_spoken_Range_AW	Range indices	number of documents that the words occurs, domain: spoken language	mean of all scores
COCA_spoken_bi_MI2	mutual information	bigram association strength (mutual information squared), academic bigrams	mean of all scores
All_AWL_Normed	Academic list indices	number of academic words	relative amount of academic words
WN_Mean_Accuracy	Word Information	Average naming accuracy	mean of all scores
LD_Mean_Accuracy	Word Information	Average lexical decision accuracy	mean of all scores
LD_Mean_RT	Word Information	Average lexical decision accuracy	mean of all scores
MRC_Familiarity_AW	Word Information	unigram familiarity scores, MRC database	mean of all scores
MRC_Imageability_AW	Word Information	unigram imageability scores, MRC database	mean of all scores
Brysbaert_Concreteness_Combined_AW	Word Information	concreteness norms by Brysbaert et. al. (2013)	mean of all scores
McD_CD_AW	Contextual Distinctiveness	Co-occurrence probability of word with 500 highly frequent context lemmas (within 5 unigrams to the left and right of the target lemma)	Kullback-Leibler divergence relative entropy
Sem_D_AW	Contextual Distinctiveness	Semantic variability of contexts (1,000-word chunks of text) in which word occurs	Natural log of mean LSA cosine of similarity between contexts containing target words; reverses sign
content_poly	semantic networks	number of senses of content words	mean of all scores
hyper_verb_noun_Sav_Pav	semantic networks	hypernymy score for nouns and verbs, all senses and paths	mean of all scores

Table 8: Lexical sophistication features: overview. Total number: 16.

feature name	description
action_component	ought verbs, try verbs, travel verbs, descriptive action verbs
affect_friends_and_family_component	affect nouns, participant affect, kin noun, affiliation nouns
certainty_component	sureness nouns, quantity
economy_component	economy words
failure_component	power loss verbs, failure verbs
fear_and_digust_component	fear- / disgust- / negative nouns
joy_component	joy adjectives
negative_adjectives_component	negative adjectives
objects_component	objects
polarity_nouns_component	polarity nouns, aptitude nouns, pleasantness nouns
polarity_verbs_component	polarity verbs, aptitude verbs, pleasantness verbs
politeness_component	politeness nouns
positive_adjectives_component	positive adjectives
positive_nouns_component	positive nouns
positive_verbs_component	positive verbs
respect_component	respect nouns
social_order_component	ethic verbs, need verbs, rectitude words
trust_verbs_component	trust verbs, joy verbs, positive verbs
virtue_adverbs_component	hostility adverbs, rectitude gain adverbs, sureness adverbs
well_being_component	well-being words

Table 9: Sentiment features: overview. Total number: 20

features	Europolis		CMV		RegRoom	
	reports	no reports	reports	no reports	reports	no reports
action_component	0.51	0.51	0.53	0.52	0.54	0.55
adverbs	0.07	0.08	0.06	0.06	0.05	0.05
affect_friends_and_family_component	0.16	0.16	0.27	0.24	0.24	0.23
All_AWL_Normed	0.05	0.05	0.05	0.05	0.07	0.07
auxliliary	0.01	0.03	0.02	0.02	0.02	0.03
Brysbaert_Concreteness_Combined_AW	2.37	2.38	2.45	2.44	2.44	2.45
certainty_component	0.20	0.21	0.20	0.20	0.18	0.19
chars_per_word	4.23	4.22	4.32	4.30	4.52	4.50
COCA_spoken_bi_MI2	9.58	9.59	9.21	9.20	9.15	9.05
COCA_spoken_Bigram_Frequency	232	232	191	192	205	188
COCA_spoken_Frequency_AW	7906	7759	6962	7133	7642	7592
COCA_spoken_Range_AW	0.65	0.65	0.58	0.58	0.56	0.56
content_poly	9.42	9.63	9.72	9.61	9.64	9.66
economy_component	0.29	0.29	0.14	0.17	0.16	0.184
failure_component	0.04	0.04	0.05	0.04	0.05	0.06
fear_and_digust_component	0.12	0.14	0.18	0.20	0.24	0.25
flesch	76.90	76.56	72.86	74.37	72.32	73.46
gunningFog	12.61	12.74	12.80	12.46	12.38	12.33
hdd42_aw	0.62	0.63	0.80	0.80	0.62	0.65
hyper_verb_noun_Sav_Pav	3.95	3.87	4.12	4.14	4.40	4.38
joy_component	0.49	0.58	0.79	0.71	0.54	0.40
LD_Mean_Accuracy	0.96	0.97	0.96	0.96	0.96	0.96
LD_Mean_RT	625	624	629	629	634	634
long_words	0.16	0.157	0.17	0.17	0.19	0.19
lsa_average_top_three_cosine	0.15	0.151	0.17	0.17	0.17	0.17
mattr50_aw	0.75	0.754	0.78	0.77	0.78	0.77
McD_CD	0.85	0.830	0.87	0.88	0.90	0.88
MRC_Familiarity_AW	596	594	592	592	589	589
MRC_Imageability_AW	308	309	319	319	312	313
mtld_original_aw	49.57	50.58	70.02	70.00	62.91	62.88
named_entities	0.04	0.03	0.02	0.02	0.02	0.02
negative_adjectives_component	0.19	0.31	0.26	0.27	0.53	0.58
objects_component	0.09	0.09	0.11	0.13	0.17	0.17
past_tense	0.03	0.02	0.03	0.03	0.06	0.03
personal_pronouns	0.04	0.05	0.04	0.02	0.05	0.01
polarity_nouns_component	0.27	0.30	0.42	0.44	0.36	0.36
polarity_verbs_component	0.51	0.52	0.32	0.37	0.38	0.35
politeness_component	0.37	0.37	0.18	0.19	0.24	0.23
positive_adjectives_component	-0.02	0.01	0.10	0.08	0.02	-0.01
positive_nouns_component	-0.22	-0.20	-0.22	-0.26	-0.52	-0.56
positive_verbs_component	0.21	0.24	-0.13	-0.01	-0.11	-0.09
postlength	203	132	341	259	162	106
respect_component	0.17	0.17	0.09	0.09	0.06	0.06
Sem_D	2.15	2.15	2.11	2.11	2.10	2.10
social_order_component	0.50	0.57	0.48	0.44	0.49	0.49
subordinate_conj	0.02	0.03	0.02	0.02	0.02	0.02
syllables_per_word	1.28	1.28	1.34	1.33	1.38	1.38
trust_verbs_component	0.15	0.15	0.19	0.19	0.21	0.21
virtue_adverbs_component	0.13	0.14	0.18	0.17	0.19	0.15
well_being_component	0.05	0.05	0.08	0.10	0.03	0.04
WN_Mean_Accuracy	0.99	0.99	0.99	0.99	0.99	0.99

Table 10: Mean values for the features per dataset: positive (report) vs. negative (no report) subsets. Features are sorted in alphabetical order.

		test data											
		Europolis				**CMV**				**RegRoom**			
training data	model	prec.	recall	F1	macro	prec.	recall	F1	macro	prec.	recall	F1	macro
europolis	features	0.70±0.11	0.28±0.06	0.39±0.09	0.60±0.05	0.39±0.18	0.12±0.07	0.18±0.09	0.45±0.06	0.75±0.12	0.18±0.05	0.29±0.05	0.52±0.03
	BoW	0.67±0.10	0.23±0.05	0.34±0.06	0.57±0.04	0.50±0.13	0.40±0.10	0.44±0.09	0.58±0.07	0.77±0.17	0.1±0.03	0.19±0.06	0.47±0.03
	feedforwardNN	0.53±0.08	0.63±0.13	0.57±0.07	0.64±0.05	0.40±0.17	0.37±0.13	0.38±0.07	0.53±0.05	0.59±0.25	0.35±0.14	0.43±0.17	0.59±0.10
	BERT	0.69±0.07	0.64±0.22	0.63±0.18	0.72±0.10	0.64±0.10	0.71±0.07	0.67±0.07	0.73±0.05	0.90±0.09	0.24 ±0.04	0.37±0.06	0.58±0.03
	BERT-adapt-europarl	0.62±0.11	0.48±0.25	0.5±0.18	0.65±0.11	0.48±0.12	0.6±0.13	0.53±0.11	0.59±0.09	0.72±0.13	0.23±0.05	0.35±0.06	0.56±0.04
	BERT-adapt-argue	0.72±0.19	0.28±0.25	0.33±0.21	0.56±0.11	0.60±0.13	0.69±0.12	0.63±0.11	0.69±0.09	0.93±0.04	0.5±0.06	0.65±0.06	0.74±0.03
	BERT-adapt-mixed	0.75±0.14	0.24±0.24	0.30±0.20	0.55±0.11	0.39±0.06	0.64±0.06	0.49±0.05	0.49±0.04	0.59±0.09	0.4±0.05	0.48±0.06	0.60 ±0.04
CMV	features	0.50±0.09	0.31±0.07	0.38±0.07	0.57±0.05	0.71±0.09	0.34±0.12	0.45±0.13	0.62±0.08	0.93±0.06	0.31±0.07	0.47±0.07	0.63±0.04
	BoW	0.61±0.28	0.04±0.03	0.08±0.05	0.43±0.03	0.67±0.17	0.24±0.09	0.35±0.10	0.56±0.06	0.60±0.52	0.01 ±0.01	0.02±0.02	0.38±0.01
	feedforwardNN	0.37 ±0.07	0.43 ±0.20	0.38 ±0.10	0.50±0.05	0.59 ±0.15	0.56 ±0.19	0.56 ±0.15	0.65 ±0.10	0.69 ±0.15	0.37 ±0.16	0.45 ±0.13	0.59±0.08
	BERT	0.58±0.09	0.39±0.05	0.47±0.06	0.62±0.04	0.81±0.08	0.81±0.06	0.81±0.06	0.85±0.04	0.96±0.02	0.78±0.05	0.86±0.03	0.89±0.02
	BERT-adapt-europarl	0.62 ±0.06	0.45±0.05	0.52±0.04	0.65 ±0.02	0.77±0.11	0.74±0.16	0.75±0.12	0.80±0.08	0.93±0.03	0.90±0.02	0.9±0.02	0.92±0.02
	BERT-adapt-argue	0.59±0.11	0.32±0.07	0.4±0.08	0.59±0.05	0.70±0.13	0.66±0.19	0.68±0.15	0.75±0.10	0.92±0.03	0.85±0.06	0.88±0.04	0.90±0.03
	BERT-adapt-mixed	0.63±0.09	0.30±0.04	0.42±0.05	0.60±0.04	0.76±0.12	0.83±0.11	0.78±0.06	0.82±0.06	0.94±0.03	0.8±0.06	0.87±0.04	0.90±0.03
RegRoom	features	0.46±0.05	0.85±0.05	0.59±0.06	0.59±0.02	0.58±0.06	0.80±0.07	0.68±0.05	0.71±0.04	0.86±0.06	0.78±0.07	0.82±0.05	0.85±0.04
	BoW	0.49±0.05	0.57±0.05	0.53±0.04	0.61±0.04	0.49±0.07	0.85±0.07	0.62±0.06	0.61±0.06	0.81±0.05	0.63 ±0.07	0.71±0.05	0.77±0.04
	feedforwardNN	0.44±0.05	0.77±0.13	0.55±0.04	0.55±0.05	0.52±0.10	0.75±0.10	0.61±0.07	0.63±0.08	0.76±0.04	0.74±0.08	0.74±0.04	0.78±0.02
	BERT	0.62±0.06	0.64±0.04	0.63±0.04	0.71±0.03	0.69±0.10	0.91±0.05	0.78±0.08	0.81±0.07	0.88±0.05	0.91 ±0.06	0.89±0.04	0.91±0.03
	BERT-adapt-europarl	0.69±0.12	0.1±0.03	0.19±0.04	0.49±0.03	0.52±0.15	0.30±0.08	0.37±0.08	0.55±0.06	0.70±0.31	0.6±0.37	0.63±0.33	0.73±0.21
	BERT-adapt-argue	0.59±0.06	0.62±0.05	0.60±0.04	0.69±0.03	0.10±0.91	0.9±0.06	0.77±0.08	0.80±0.07	0.8±0.16	0.72±0.28	0.74±0.24	0.79±0.18
	BERT-adapt-mixed	0.56±0.05	0.58±0.04	0.57±0.03	0.66±0.03	0.72±0.08	0.90±0.06	0.80±0.06	0.82±0.05	0.84±0.15	0.70±0.40	0.70±0.36	0.79±0.22
2vs1	features	0.49±0.05	0.69±0.04	0.57±0.04	0.63±0.02	0.67±0.11	0.56±0.10	0.60±0.07	0.70±0.05	0.99±0.02	0.24±0.05	0.39±0.07	0.59±0.04
	BoW	0.63±0.10	0.17±0.04	0.27±0.05	0.53±0.03	0.53±0.10	0.60±0.08	0.56±0.07	0.63±0.06	0.87±0.16	0.06±0.02	0.12±0.04	0.43±0.02
	feedforwardNN	0.44±0.05	0.74±0.09	0.54±0.04	0.56±0.03	0.56±0.08	0.60±0.10	0.57±0.06	0.65±0.03	0.77±0.05	0.45±0.11	0.56±0.09	0.67±0.06
	BERT	0.61±0.06	0.53±0.05	0.56±0.05	0.67±0.03	0.73±0.08	0.74±0.08	0.73±0.06	0.78±0.05	0.95±0.02	0.77±0.04	0.85±0.02	0.88±0.02
	BERT-adapt-europarl	0.66±0.08	0.42±0.05	0.5±0.05	0.66±0.04	0.76±0.07	0.83±0.07	0.79±0.06	0.83±0.04	0.87±0.05	0.89±0.03	0.88±0.03	0.89±0.03
	BERT-adapt-argue	0.59±0.05	0.5±0.06	0.55±0.05	0.66±0.04	0.5±0.14	0.32±0.09	0.38±0.09	0.55±0.06	0.8±0.33	0.03±0.02	0.06±0.03	0.40±0.02
	BERT-adapt-mixed	0.62±0.08	0.49±0.05	0.54±0.05	0.66±0.04	0.70±0.11	0.95±0.04	0.8±0.08	0.83±0.07	0.90±0.32	0.03±0.02	0.06±0.03	0.40±0.02
all	features	0.64±0.08	0.38±0.06	0.48±0.05	0.63±0.04	0.67±0.09	0.55±0.12	0.59±0.09	0.69±0.06	0.92±0.05	0.69±0.06	0.79±0.05	0.83±0.03
	BoW	0.68±0.09	0.24±0.05	0.35±0.06	0.57±0.04	0.59±0.13	0.52±0.15	0.53±0.09	0.64±0.06	0.86±0.07	0.50±0.06	0.63±0.05	0.73±0.03
	feedforwardNN	0.53±0.05	0.60±0.12	0.55±0.04	0.64±0.03	0.58±0.10	0.63±0.15	0.59±0.09	0.66±0.06	0.74±0.05	0.67±0.09	0.70±0.04	0.76±0.02
	BERT	0.66±0.08	0.66±0.14	0.65±0.10	0.73±0.07	0.74±0.13‡	0.83±0.08	0.78±0.10	0.82±0.08	0.90±0.04	0.90±0.09	0.90±0.06	0.92±0.05
	BERT-adapt-europarl	0.68±0.04	0.69±0.14	0.68±0.08	0.75±0.05	0.82±0.09*	0.82±0.08*	**0.82±0.07***	0.85±0.05*	0.92±0.03	0.90±0.05	0.9±0.03	0.93±0.02
	BERT-adapt-argue	0.68±0.06	0.69±0.11	0.68 ±0.05	0.75±0.03	0.80±0.10	0.79±0.08	0.79±0.05	0.83±0.04	0.93±0.03*	0.92±0.02*	0.92±0.02*	0.93±0.02*
	BERT-adapt-mixed	0.70±0.06*	0.67±0.07*	**0.68±0.04***	0.76±0.03*	0.80±0.08*	0.85±0.08*	**0.82±0.06***	0.85±0.04 *	0.93±0.04*	0.93±0.02*	**0.93±0.03***	0.94±0.03*

Table 11: Precision, recall and F1 score for the positive class (reports) and F1 macro across training / test setups.∗ denotes statistical significance between the domain-adapted BERT and the non-adapted BERT model.‡ denotes statistical significance between BERT trained in-domain vs. *all*

.

B Classification experiments

B.1 Classifiers

In what follows, we provide the implementation details for the classification models employed in our experiments.

- Random forest classifier: we use sklearn `https://scikit-learn.org` with the `n_estimators` parameter set to 1000. The other parameters are set to the default.
- Feed-forward neural network: we use pretrained word embeddings with subwords ($d = 300$), provided by finalfusion (`https://finalfusion.github.io/pretrained`), pretrained with skipgram (Mikolov et al., 2013). The English word embeddings were trained on the CoNLL 2017 corpus.
- BERT: we use `BERTForSequenceClassification` from the huggingface library `https://huggingface.co/docs/transformers/model_doc/bert`. We use the sequence of the first 512 tokens and train for a maximum of 20 epochs. We pick the model that achieves the best macro F1 score on the validation set. Parameters: batchsize = 16, lr=2e-5, optim=Adam, model=bert-base-uncased.

B.2 Results

Table 11 reports the full set of experimental results for the automatic recognition of contributions containing reports. We report the precision, recall and F1 score for the positive class (report) and the F1 macro score. Each column shows the results for one test corpus: Europolis, CMV and RegRoom. The scores represent the average of the 10 scores obtained for each test split. We therefore also report the standard deviation.

C Analysis

C.1 Aggregated performance analysis

Table 12 provides the details of the fit of the regression model predicting aggregated performance (F1 macro): relative importance of each IV (measured by the relative amount of explained R^2), the GVIF, and significance.

IV	f1 macro		
	expl.var	**GVIF**	**sign.**
model	20.9	3.0	***
test corpus	11.9	4.6	***
training setup	12.2	4.6	***
training setup : test corpus	8.8	4.3	***
test corpus: model	5.0	2.6	***
training setup : model	8.6	2.6	***
test corpus : training setup : model	12.1	2.0	***

Table 12: Effect sizes (relative amount of R^2), significance and GVIF for the most explanatory regression model with `training setup` = *in-domain*, *2vs1*, *all*.

D Item-based performance analysis

The following tables and paragraphs contain more details about the regression analysis to predict model performance at the item-level (probability of report).

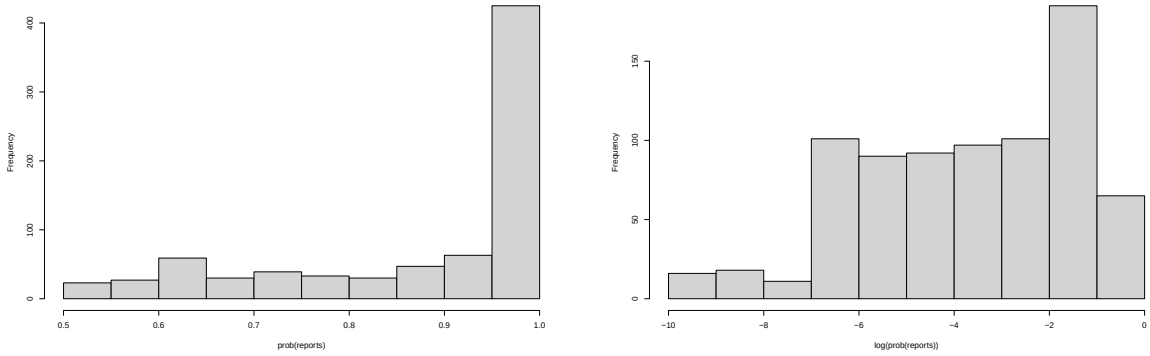
D.1 Dependent Variable

Table 2 and 3 display the distribution of predicted probabilities by the best classification model (BERT) for each error type before and after transformation.

D.2 Feature Reduction and Model Selection

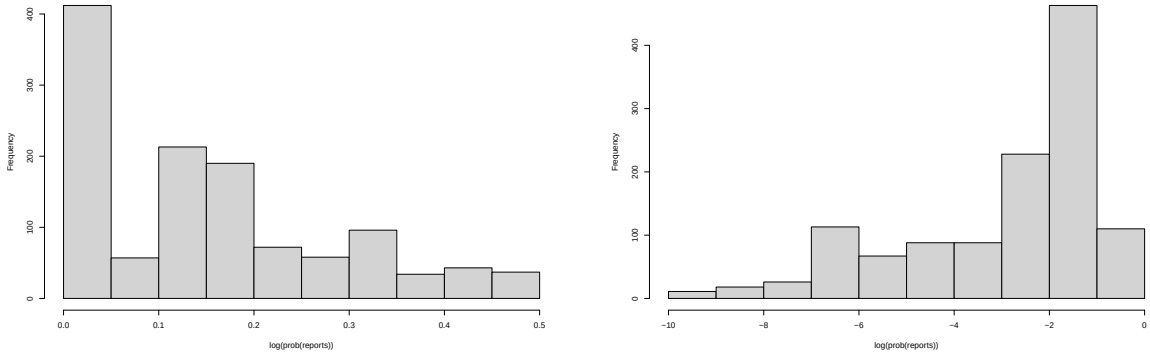
While tackling a regression analysis task with very many IVs as it is in our case, there is not just one strategy for model selection (i.e., which IVs and interactions to include in the regression model). Given the large pool of (potentially correlated) 51 contribution-level features which we wanted to combine with the experimental configuration features (`training setup` and `test corpus`), and being interested in potential interactions as well, we decided to pre-select the contribution-level features based on their correlation.

The first step in our selection of contribution-level features is a correlation analysis conducted on the full dataframe (false-positives and false negatives). We clustered the 51 features based on their pairwise Spearman correlation. The output of the clustering is displayed in the dendrogram in figure 4). Based on the assumption that correlated features are likely to distort the performance of the regression model, we established a conservative threshold of Spearman ≥ 0.2 and, for each subcluster with a correlation higher than this threshold, we manually selected only one feature and discared the others. The manual selection was based on qualitative consideration (e.g., the more general feature, or the more interpretable). For example, for the subcluster that contains the *MRC imageability* score



(a) Probabilities without transformation. (b) Probabilities with transformation ($log(1 - prob(reports))$.

Figure 2: Histogram of the probability of reports for the **false positives**, with and without transformation.



(a) Probabilities without transformation. (b) Probabilities with transformation ($log(prob(reports))$.

Figure 3: Histogram of the probability of reports for the **false negatives**, with and without transformation.

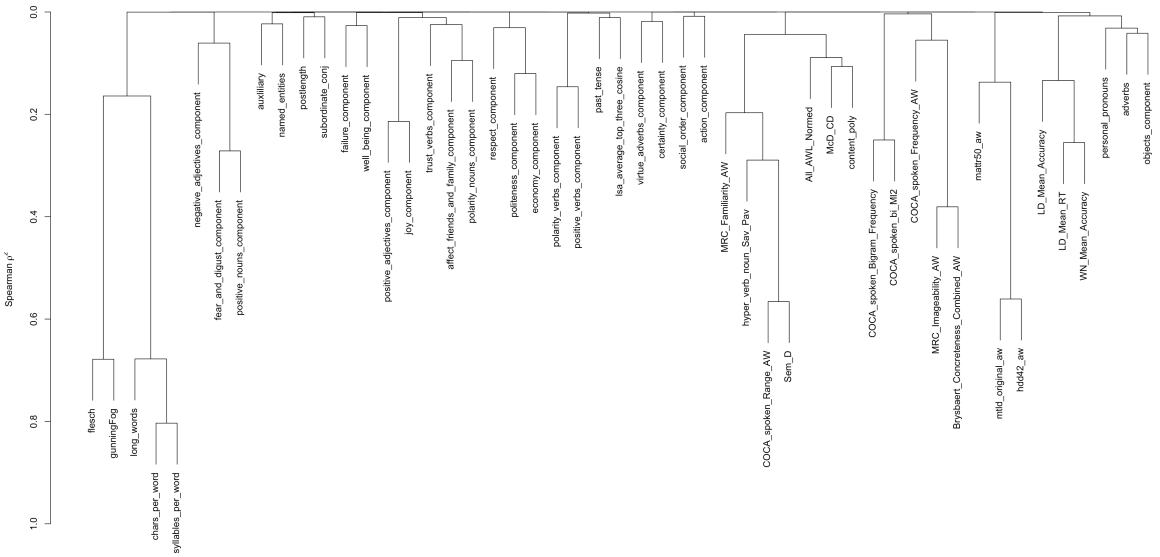


Figure 4: Dendrogram (result of hierachical clustering with Spearman correlation) of the item-based features.

term	df	GVIF	p.value	explvar
testCorpus:trainSetup	2	1.890	0.000	2.095
economy_component	1	1.175	0.001	1.377
personal_pronouns	1	1.127	0.002	1.186
mattr50_aw:trainSetup	2	1.441	0.002	1.493
auxiliary	1	1.045	0.005	0.947
All_AWL_Normed	1	1.051	0.015	0.703
politeness_component	1	1.489	0.033	0.539
economy_component:testCorpus	1	1.228	0.033	0.534
adverbs	1	1.098	0.059	0.421
COCA_spoken_Bigram_Frequency:testCorpus	1	1.289	0.060	0.417
subordinate_conj	1	1.035	0.063	0.407
trainCorpus	2	1.136	0.147	0.452
mattr50_aw	1	2.066	0.195	0.198
COCA_spoken_Bigram_Frequency	1	1.209	0.254	0.153
respect_component	1	1.446	0.311	0.121
lsa_average_top_three_cosine	1	1.140	0.402	0.083
testCorpus	1	3.173	0.492	0.056
lsa_average_top_three_cosine:testCorpus	1	1.313	0.560	0.040
sum(R^2)				11.221

Table 13: Terms of the final regression model for the **false positives**, with degrees of freedom, variance inflation factor, statistical significance and explained variance.

and the *Brysbeart concreteness* we keep the *Brysbeart concreteness* score because concreteness is a more general notion. Concreteness quantifies the extent to which the word's referent can be perceived and imageability, the extent to which the word's referent can be perceived visually.

The output of correlation-based qualitative feature selection is a set of 37 features. The features we discarded features are: flesch, gunning Fog Index, long words, characters per word, syllables per word, fear and disgust component, joy component, COCA spoken range norms, Sem_D, COCA spoken bigram mutual information, MRC Imageability, hdd42_aw and LD_Mean_Accuracy.

At this point the analysis proceeds per subset (FPs vs. FNs). We first run a regression model with the 37 contribution-level features on the FP and FN subset, respectively, without interactions. Next, we perform step-wise model selection on the regression model [18]. Unsurprisingly, we find no collinearities. The output of the stepAIC selection are two feature-based regression models, one for the FPs and one for the FNs. The feature-based regression models contain 11 contribution-level features for FPs and 20 contribution-level features for FNs respectively.

The next step in our analysis is to incrementally add the experimental configuration features, `training setup` (3 levels: *Europolis, RegRoom, all*) and `test corpus` (2 levels: *Europolis, RegRoom*). First, we add them IVs first on top of the contribution level features. Then, we add the following two-way interactions: those between contribution-level features and experimental configuration features (training setup and test corpus) as well as the two-way interaction between training setup and test corpus. We simplify the final models again using stepAIC and checked for multicollinearities. At each step, we test the significance between a richer model and its nested counterpart using the ANOVA function from R. (e.g., the model with contribution-level features in nested in the model with contribution-level + experimental configuration features).

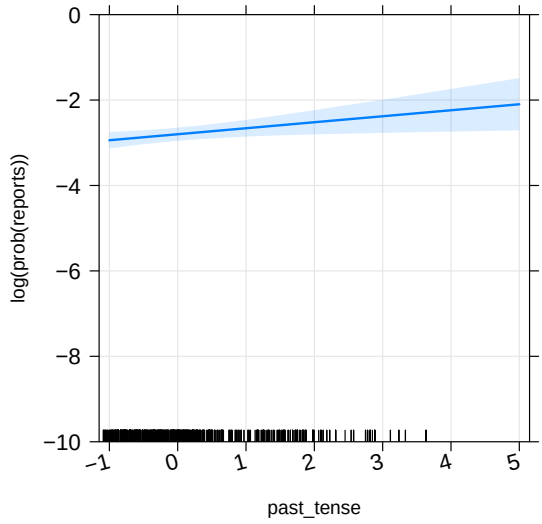
As the output of this further process of selection of the IVs, we have two final regression models, one for the FPs and one for the FNs, which unsurprisingly differ in terms of the selected predictors and explained variance. In the next section we provide the details for the fit of the models. Tables 13 and 14 provide the details of the fit of the selected model regression model for FPs and FNs, respectively. For each selected IV (or interaction), the tables display: degrees of freedom, GVIF variance inflation factor, significance, as well as explained variance in $R2$.

Effect plots Figures 7 to 5 display the effect plots referred to in the discussion in section 5.2.

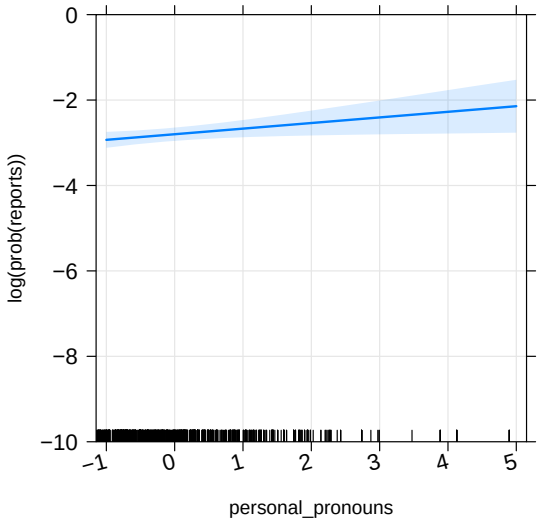
[18] We used the stepAIC function by the *MASS* package in R with standard settings

term	**df**	GVIF	**p.value**	**explvar**
trainSetup	2	1.719	0.000	7.698
testCorpus	1	2.063	0.000	4.518
past_tense	1	1.467	0.000	3.330
subordinate_conj	1	2.540	0.000	3.048
personal_pronouns	1	1.326	0.000	2.043
economy_component	1	2.311	0.000	1.869
auxiliary	1	2.219	0.000	1.634
positive_nouns_component	1	2.283	0.000	1.545
adverbs	1	1.138	0.000	1.302
mtld_original_aw	1	3.105	0.000	0.886
respect_component	1	1.839	0.000	0.836
All_AWL_Normed	1	2.688	0.000	0.755
auxiliary:testCorpus	1	1.709	0.000	0.690
McD_CD:testCorpus:trainSetup	2	1.680	0.000	0.812
LD_Mean_Accuracy:trainSetup	2	1.707	0.000	0.806
All_AWL_Normed:trainSetup	2	1.557	0.000	0.765
well_being_component	1	1.504	0.001	0.594
COCA_spoken_Bigram_Frequency	1	2.561	0.001	0.560
personal_pronouns:trainSetup	2	1.412	0.001	0.706
postlength	1	2.630	0.003	0.429
subordinate_conj:testCorpus:trainSetup	2	1.417	0.004	0.548
failure_component:testCorpus:trainSetup	2	1.290	0.005	0.528
Brysbaert_Concreteness_Combined_AW	1	2.692	0.005	0.389
postlength:trainSetup	2	1.582	0.006	0.515
Brysbaert_Concreteness_Combined_AW:testCorpus:trainSetup	2	1.645	0.007	0.499
economy_component:testCorpus:trainSetup	2	1.510	0.009	0.471
certainty_component:testCorpus:trainSetup	2	1.579	0.009	0.470
mtld_original_aw:testCorpus:trainSetup	2	1.495	0.011	0.447
content_poly:trainSetup	2	1.896	0.022	0.380
failure_component	1	1.739	0.026	0.247
certainty_component	1	2.453	0.026	0.246
auxiliary:trainSetup	2	1.427	0.030	0.347
economy_component:trainSetup	2	1.548	0.031	0.346
past_tense:trainSetup	2	1.437	0.038	0.324
All_AWL_Normed:testCorpus	1	2.176	0.042	0.205
failure_component:testCorpus	1	1.485	0.042	0.204
McD_CD:testCorpus	1	2.164	0.044	0.202
LD_Mean_Accuracy:testCorpus:trainSetup	2	1.647	0.045	0.308
COCA_spoken_Bigram_Frequency:trainSetup	2	1.629	0.050	0.297
certainty_component:testCorpus	1	2.139	0.066	0.167
content_poly:testCorpus	1	2.610	0.067	0.167
positive_nouns_component:trainSetup	2	1.455	0.071	0.263
LD_Mean_Accuracy	1	2.761	0.081	0.151
postlength:testCorpus:trainSetup	2	1.418	0.084	0.246
LD_Mean_Accuracy:testCorpus	1	2.447	0.115	0.123
subordinate_conj:trainSetup	2	1.657	0.119	0.211
respect_component:testCorpus	1	1.410	0.137	0.110
respect_component:trainSetup	2	1.369	0.155	0.185
COCA_spoken_Bigram_Frequency:testCorpus:trainSetup	2	1.561	0.156	0.184
certainty_component:trainSetup	2	1.576	0.158	0.183
mtld_original_aw:testCorpus	1	2.612	0.175	0.091
COCA_spoken_Bigram_Frequency:testCorpus	1	2.186	0.191	0.085
well_being_component:trainSetup	2	1.267	0.216	0.152
failure_component:trainSetup	2	1.341	0.222	0.149
content_poly	1	3.062	0.258	0.063
McD_CD	1	2.523	0.263	0.062
respect_component:testCorpus:trainSetup	2	1.491	0.306	0.117
postlength:testCorpus	1	2.056	0.323	0.048
subordinate_conj:testCorpus	1	1.912	0.349	0.043
content_poly:testCorpus:trainSetup	2	1.718	0.354	0.103
testCorpus:trainSetup	2	1.969	0.413	0.088
mtld_original_aw:trainSetup	2	1.649	0.445	0.080
McD_CD:trainSetup	2	1.671	0.775	0.025
Brysbaert_Concreteness_Combined_AW:testCorpus	1	2.236	0.777	0.004
positive_nouns_component:testCorpus	1	1.663	0.860	0.002
economy_component:testCorpus	1	1.563	0.876	0.001
Brysbaert_Concreteness_Combined_AW:trainSetup	2	1.726	0.905	0.010
sum(R^2)				44.914

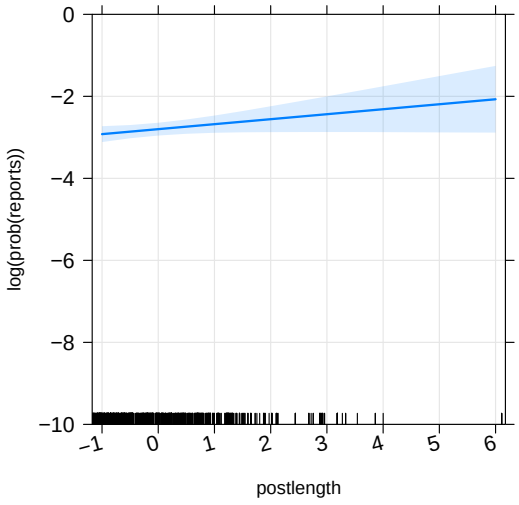
Table 14: Terms of the final regression model for the **false negatives**, with degrees of freedom, variance inflation factor, statistical significance and explained variance.



(a) Effect of *past tense verbs*



(b) Effect of *personal pronouns*



(c) Effect of *contribution length*

Figure 5: Single effects of the most explanatory syntactic features for the final regression model for the subset of **false negatives**. A positive effect (increase in the line) indicates that the feature drives the model in the right direction.

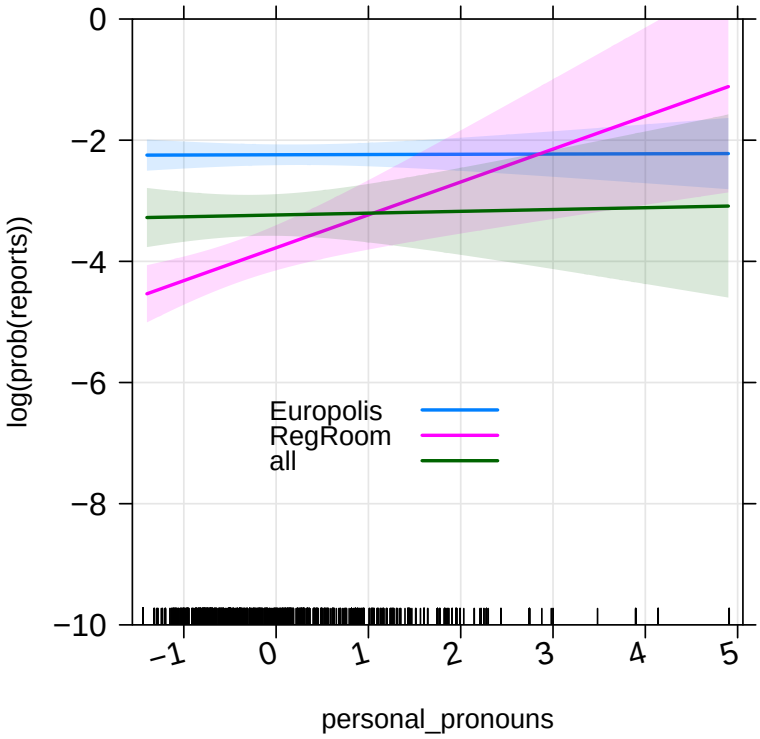


Figure 6: Effect of the interaction between the amount of personal pronouns and `training setup` for **false negatives**. A positive effect (increase in a line) indicates that the feature pushes the model in the right direction.

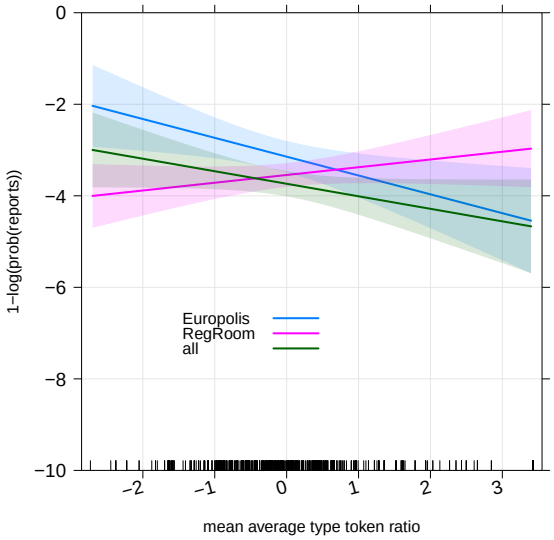


Figure 7: Effect of the interaction between mean average type token ratio (`mattr50`) and `training setup`, **false positives**. A positive effect (increase in a line) indicates that the feature pushes the model in the right direction.

E Case-study: Reports and Argument Quality

To investigate the impact of reports on the quality of a contribution, we conducted a pilot study on available Argument Quality datasets.

The first Argument Quality dataset we employed is the Dagstuhl-15512-ArgQuality corpus (Wachsmuth et al., 2017c): it is small but contains quality annotations for a very fine-grained taxonomy of argument quality dimensions (Wachsmuth et al., 2017b). We employed the best classifier from our experiments to predict whether a comment contained or not a report and found 59 comments containing report (out of a total of 320). We use t-tests to carry out a pairwise comparison of the means of the quality scores for each argument quality dimension (arguments with reports vs. argument without reports) and found the values of the documents containing reports scoring significantly higher in appropriateness, emotional appeal and sufficiency than the ones not containing reports (table 15). In particular, the higher score for emotional appeal is in line with the expectation that contributions with reports are more effective on the affective dimensions of argument quality.

Next, we conducted the same analysis on the grammarly Argument Quality corpus (GAQ) (Ng et al., 2020) which contains 3,373 comments from online fora with gold annotations for each of the 3 core dimensions of the taxonomy of (Wachsmuth et al., 2017b) (cogency, effectiveness, reasonableness) and overall quality. Our classifier detected reports in 1,288 comments. We conducted the same type of analysis as above, and found comments containing reports scoring significantly higher than the non-reports ones in all dimensions (table 16).

Last, we used a state-of-the art multi-task regression classifier (Lauscher et al., 2020) (trained on the GAQ corpus) to automatically predict argument quality scores for our three datasets. The performance of this classifier on RegulationRoom has already been validated in a manual annotation study by (Falk et al., 2021). For each dataset, we compared the means of contributions with and without reports (based on the gold standard for each corpus). While we found no significant difference in CMV or RegRoom, we did find that for Europolis (Table 17) contributions containing reports have significantly higher means for all dimensions of Argument Quality. This is in line with the findings by (Gerber et al., 2018) who state that people who score high on the deliberative quality dimensions also use reports to back up their claim.

Quality dimension	t-value	p-value
appropriateness	2.04	≤ 0.05
emotional appeal	2.29	≤ 0.05
sufficiency	2.02	≤ 0.05

Table 15: T-tests for the comparison of quality dimensions: arguments with reports vs. arguments without reports. Corpus: Dagstuhl-15512-ArgQuality. Argument with reports have significantly higher appropriateness, emotional appeal, and sufficiency

Quality dimension	t-value	p-value
cogency	5.00	≤ 0.001
effectiveness	5.91	≤ 0.001
reasonableness	5.60	≤ 0.001
overall	5.84	≤ 0.001

Table 16: T-tests for the comparison of argument quality dimensions: arguments with reports vs. arguments without reports. Corpus: GAQ corpus. Argument with reports have significantly higher cogency, effectiveness, reasonableness, and overall quality.

Quality dimension	t-value	p-value
cogency	3.17	≤ 0.05
effectiveness	3.25	≤ 0.05
reasonableness	3.19	≤ 0.05
overall	3.27	≤ 0.05

Table 17: T-tests for the comparison of argument quality dimensions: arguments with reports vs. arguments without reports (gold standard annotation). Corpus: Europolis. Argument with reports have significantly higher cogency, effectiveness, reasonableness, and overall quality.