Computational Linguistics in the Netherlands Journal 12 (2022) 65-82 Submitted 12/2022; Published 12/2022

Don't Do Your Experiments Double-blind: The Importance of Checking Your Data

Nelleke Oostdijk Nelleke.Oostdijk@ru.nl
Hans van Halteren Hans.vanHalteren@ru.nl

CLS, Radboud University, Nijmegen, The Netherlands

Abstract

In this paper, we investigate what could happen if you run machine learning experiments on data found somewhere on the internet, without first examining this data. As an example, we did polarity recognition on a data set extracted from Booking.com. We found that a) the form of the data in the dataset sometimes made polarity judgements hard for humans and probably also for systems, b) naive use of the data results in a different task than polarity recognition as the content of the data fields does not always comply with the field descriptors, and c) the comments in the data set come in several, quite different, subtypes, so that recognition quality rather varies with the choice for training and test data. On the basis of these findings we conclude that our advice to inspect data before using it is indeed valuable.

1. Introduction

While data is key to any kind of research, it is appalling to see that in many studies and experiments researchers settle for data that just happens to be available or is easily come by without ever raising the question to what extent this data actually suits the research purpose at hand. This is not new. Way back, in the second half of the previous century, when text corpora constituted the prime sources that researchers would use, it happened that researchers sometimes would use data that, in the context of their specific research, was suboptimal at best. What is different is that at that time there were hardly any alternatives and researchers made do with what was available. Such corpora typically had been designed to constitute a balanced or representative collection of carefully sampled printed/edited texts. They were, certainly by today's standards, (extremely) small: from 1 million words (the Brown and LOB corpora) up to the 100MW British National Corpus that became available in 1997 (Aston & Burnard 1998). Since technological developments have unlocked the web (Kilgarriff & Grefenstette 2003) massive amounts of data can be accessed, although admittedly not for all languages and not for all tasks one may want to undertake. Many languages remain that are lesser or even under-resourced while specific tasks may require a form of annotated data that is not available. Where large data collections are available, researchers may select an existing dataset or derive a dedicated dataset including a subset of the data from the larger data set that fits their research purposes. Now we do understand that the time and effort researchers can spend on compiling and/or annotating their own dataset is limited and using an existing dataset for similar research allows for easier comparison of the research findings. However, using an existing dataset without paying due attention to where the data originates from, how and for what purpose the dataset has been compiled, and what data it comprises entails some potential risks. Thus effects (or lack thereof) found may in fact be explained through a bias in the data or the annotations, or as we will show in this paper, a classification task may appear to yield quite satisfactory results but then not actually be what you aimed for.

As the first author wanted to research the language used by reviewers with different language backgrounds and was looking for suitable data, she came across a dataset comprising some 515K reviews of 1493 luxury hotels scraped in 2017 from Booking.com by Jason Liu (www.kaggle.com/datasets/

Table 1: Field names and example content of the hotel review database.

Hotel_Address	12 Boulevard Haussmann 9th arr 75009 Paris France
Additional_Number_of_Scoring	78
Review_Date	8/16/2016
Average_Score	7.8
Hotel_Name	Millennium Hotel Paris Opera
Reviewer_Nationality	Japan
Negative_Review	No toothbrush in toilet kit Minibar fridge is not cold enough Missing sugar for coffee and no tea in the room Revolving door turn too fast and dengerus bit
Review_Total_Negative_Word_Counts	31
Total_Number_of_Reviews	515
Positive_Review	Very combinience location Good size big room with big double bed Lighting in the room is good too Cute bathroom and big thick towel
Review_Total_Positive_Word_Counts	26
Total_Number_of_Reviews_Reviewer_Has_Given	9
Reviewer_Score	7.9
Tags	[' Business trip ', ' Solo traveler ', ' Standard Double Room ', ' Stayed 3 nights ', ' Submitted from a mobile device ']
days_since_review	352 day
lat	488724588
lng	23378004

jiashenliu/515k-hotel-reviews-data-in-europe). The data is quite rich, as can be seen in Table 1. Apart from the actual review comments about the hotels, there is general information about both hotels and reviewers. With the comments split up in positive and negative, and with the nationality of the reviewer in the metadata, it appeared to be ideal for her research and she started annotating for various linguistic aspects of positive and negative comments (note that although the corresponding fields are named "reviews" in the data, we prefer "comment", reserving "review" for the whole of the comments and the score). However, during this annotation, she noticed that the content of the comments did not always correspond to the given field header. Positive comments were found in the field for negative comments and negative comments in the field for positive comments. Moreover, in a substantial number of reviews, a mix of positive and negative remarks appeared in one and the same field.

The above finding made us wonder how this would influence text modelling and classification. We decided to investigate this with an experiment in polarity recognition, namely recognizing whether a (short) comment text is positive or negative. In a scenario where researchers interested in modelling polarity would use data available from Booking.com as is, because they are assuming that it suits their purpose well, given that reviewers are prompted to place their comments in separate fields (one for positive and one for negative comments), they might run into problems as the reviewers do not always use the fields as expected. To our surprise, in various papers that use this dataset (e.g., Campos, Rocha Silva & Bernardino 2019; Forhad et al. 2021; Juliadi & Puspitarani 2022; Lal & Mishra 2020), there is no mention of the data not being what you would expect it to be.

In summary, the research question we are addressing is the following: "How and to what degree is the modelling of polarity influenced by the fact that Booking.com users have not always been

responding to the prompts asking for positive and negative comments to be entered in the respective fields." Note that what we are interested in is the polarity of the comments, and not whether the comment appeared in a positive or negative field. Of course, with reviewers filling out the review form as anticipated the two tasks would coincide but with the present data this is not the case. Even when the results for both tasks in terms of scores turn out to be similar, it is still worthwhile to see how they compare in the classification (are positive comments indeed seen as positive and negative comments as negative?).

The structure of the remainder of this paper is as follows: We will first discuss the data in more detail, as well as our selection and annotation, in Section 2. In Section 3, we discuss what happens if we use the data as is and therefore do field recognition rather than polarity recognition. Using the adjusted data, we repeat the tasks and discuss the quantitative results in Section 4, postponing a more detailed analysis of the impact of our choices to Section 5. Finally, in Section 6, we discuss our findings and what we think they imply for research with this and similar data.

2. Analysis of the Dataset

Before we go on to describe the experiments, we first discuss the data in some more detail, describing what the data look like in their original form and also how we relabelled the data to make it more compatible with our polarity recognition task.

2.1 Characteristics of the data

As these reviews are a typical example of user-generated text, we find all the expected characteristics that sets them apart from carefully edited text that we find in more traditional, printed outlets. Thus capitalization may be non-standard and spelling is varied, while many reviewer comments are not well-formed sentences or phrases and language use may be creative. For examples, see the Appendix.

We also find that in the scraping process all punctuation has been removed and all other special characters (apostrophes, hyphens, etc.) and characters with diacritics have been replaced by blank spaces. Empty fields have been filled with the text *No Positive* or *No Negative*, but not always, so that some fields remain plain empty. Some review comments appear to have been cut off. Examples are provided in the Appendix.

Finally, and this is especially noteworthy in the context of the current research question, the shape of the data is clearly influenced by the behaviour of the reviewers. More specifically,

- The reviewer may have left one of the fields empty.
- Supposedly some reviewers want to please and therefore fill in something, even when there is nothing to fill in (exs. [1]-[2]).

  (1) sorry but nothing to say

  (2) If you want me to be very picky rooms near the elevator can get some occasional hallway noise But this rates a 10 in my book for value

- The reviewer has (partly) been ignoring the prompt and just entering comments. These may be of mixed polarity (exs. [3]-[4]) or even be in the wrong fields (examples [5]-[6]]. Example [6] is particularly interesting as here the user in first instance used the positive field to enter all comments, both positive and negative, and then completed the negative field by referring to the positive field.

  (3) The room was quite small but otherwise good

(4) It s very beautiful building Location is also not bad Stuff was trying their best but the check in was so badly organized that it s hard to get reed of the biter taste

(5) (NEG field) Hairdryer Milk that was in date A warm breakfast with a good selection of cooked food

(6) (POS field) Two double beds were ideal for teenagers The bathroom was a little dated (NEG field) As above

- The text can only be interpreted correctly if one knows what type of comment the user was prompted to enter in the field in question. This is especially true for "bare" comments, often consisting of single nouns (e.g. *location*, *staff*) that are understood to be positive comments in case they occur in the positive field and negative when encountered in the negative field (exs. [7]-[8]).

(7) It s Location Staff and Price

(8) all of it

- The reviews are supposed to be in English, but the dataset also contains reviews that are wholly or partly in some other language than English. For example,

(9) Personalen var tyv rr inte serviceminded loj och I ngsam Gick inte att best lla allt som fans p menyn g lier b de mat och dryck Fint I ge vid Hyde Park Litet rum och badrum

(10) No business center Gym closes quickly 8 30pm Mais l autre choses pas de problemes

2.2 Selection and annotation

As the amount of data precludes manual verification of all data for the current study, we decided to take a subset, consisting of every 10th review. As all the reviews of the same hotel are placed consecutively in the dataset, this means that the distribution of reviews over hotels remains (approximately) the same. Also, for each of these reviews we use both the positive and the negative comment. Note that we consider the positive and negative comments independently of each other and thus lose the link between the two comments belonging to the same review. For other tasks, it would be wise to use the two in combination as some comments can only then be correctly interpreted or the reviewer scores be explained (see, for example, [11]-[12]).

(11) (NEG field) Everything (POS field) I loved the place Staff excellent (SCORE) 10

(12) (NEG field) Same as above (POS field) When I booked through booking com they had two prices for a room one with breakfast one without I took the one with breakfast When I got to the hotel I found out the breakfast was free with any booking but I was charged 40 a day for two people I would like to know where that 40 a day went (SCORE) 2.5

In this subset, there were 16,319 empty comments (16,220 marked as being empty, either "No Positive" or "No Negative", and 99 actually empty). From the remainder, we managed to inspect and annotate 82,942 comments within the time available, so not quite the 10% of the data we originally intended to include in our experiment. As we found that the manual work could be speeded up if we sorted the comments by length and worked through them starting from the shortest to increasingly longer comments, there may be a slight bias in the resulting relabelled subset in that longer comments (and with that the mixed comments) are somewhat underrepresented.

We distinguished between:

- POS (43,692 cases). In Positive Field → Pure Positive.

(13) I loved this place nothing to complain about Wonderful Breakfast amazing

(14) Location Rooms are clean and big

- MPNEG_CONT (343). In Positive Field → Misplaced Negative, contradiction.

  (15) Bed not comfy

  (16) We had to queue up for 15 minutes every morning to get a table for breakfast Very poor experience

- MPNEG_DENY (277). In Positive Field → Misplaced Negative, denial.

  (17) Nothing really

- MPNEG_BOTH (46). In Positive Field → Misplaced Negative, both denial and contradiction.

  (18) Nothing was dirty old fashioned and just vile

  (19) Wasn t much to like about this stay roo was over priced and misrepresentation by its description on the website

- MIXEDPOS (1736). In Positive Field → Mixed Positive.

  (20) breakfast was the onlynice thing about this place cant even call it a hotel

  (21) Excellent location Awful breakfast Poor tiny room

- NEG (28,213). In Negative Field → Pure Negative.

  (22) Cold shower Very warm room No ventilation Very Noisy air con so had to leave off room next to lift so could hear lift at night mould on blind in shower dark bad lighting in shower broken tiles in shower lumpy bed cooked breakfast not piping hot

  (23) Just the rollaway bed was a bit unsteady and screechy

- MPPOS_CONT (601). In Negative Field → Misplaced Positive, contradiction.

  (24) Loved everything

  (25) for what I wanted it was perfect

- MPPOS_DENY (3,896). In Negative Field → Misplaced Positive, denial.

  (26) Not one thing

  (27) Honestly I don t have any dislikes

- MPPOS_BOTH (515). In Negative Field → Misplaced Positive, both denial and contradiction.

  (28) Nothing Everything was great

  (29) No issues would recommend visit and stay to friends and others

- MIXEDNEG (3,471). In Negative Field → Mixed Negative.

  (30) The bed was very uncomfortable Needs new mattress Everyone else found nice and helpful but receptionist not so friendly

  (31) Small but clean rooms Breakfast menu was good but was same every day

- FOREIGN (152). Non-English.

  (32) Location e cortesia dello staff

  (33) Excellent bed en kussens

  (34) Extremt litet rum

Non-English (FOREIGN) and empty comments were excluded from the experiment. This left us with 50,440 positive and 32,350 negative comments.

Misplaced comments presumably stem from the reviewers' feeling that all fields have to be filled out. A reviewer who has nothing negative to say makes this clear by means of commenting in the negative field that all was positive, or by denying having any negative comments to make, or a combination of these. In effect, the negative field is filled with a comment which is positive in nature or vice versa. It is this behaviour that accounts for the main difference between classifying for the fields in the original data and classifying for our labels on the comments. The former classifies for which field a user put the comment in, while the latter classifies for the actual polarity of the comment. Both are valid classification tasks, but our stated goal was polarity classification, so we should be using the latter option.

Mixed comments bring a different problem. In such comments there are bits that are positive as well as bits that are negative. We could have left these out of the experiment altogether, but kept them in as they make the task much more interesting. However, we did not attempt to analyse all mixed comments in order to find out whether they were more positive or more negative as a whole. Instead, we assumed (optimistically) that the user placed it in the field that most characterized its overall polarity. We expect to run into this aspect again when doing our error analysis.

3. Field Recognition

In this section, we perform the task of polarity recognition while assuming that the name of the field ("positive review" versus "negative review") is a correct indicator of the polarity. In fact, we have already shown that this is not the case. This experiment, therefore, shows what would happen if we did not check the data beforehand and what conclusions would be drawn.

3.1 Data

For our experiment, we only use the fields that Booking.com provides for comments. The content of the comments is discussed in Section 2. Here, we do not investigate the data at all and ignore the selection and labelling, but just apply our classifiers to whatever is in those fields. We do exclude empty comments, though. This includes the really empty ones (where there is no text at all) as well as the ones with the comment text *No Positive* or *No Negative* that was provided in the scraping process. Furthermore, we uncased the text, as casing was of uncertain reliability, even though we realized that this might introduce new complications.

We execute the task twice, once using the full dataset and once separately for a subset of ˜10% of the reviews (˜50K reviews), which is that part of the data which we manually checked and, where appropriate, relabelled and which we use in the experiments in Section 4. Within this first experiment, using the two subsets allows us to judge the impact of the size of the training data. Between this experiment and the one in Section 4, using the 50K subset allows us to compare quality levels in handling the two tasks as training and test set sizes are comparable.

3.2 Methods

We applied two different classification methods to the data. The first is Roberta, to be exact RobertaForSequenceClassification (huggingface.co/transformers/v2.9.1/model_doc/roberta.html

#robertaforsequenceclassification), using the *roberta-base* pretrained model, cross-entropy loss, Adam optimizer, MAXLEN=256, TRAIN_BATCH_SIZE=8 and LEARNING_RATE=1e-05. A pilot experiment showed that after two epochs there was only some fluctuation but no further significant improvement. Given this observation, and the machine time constraints we were under, we decided to train for three epochs in the actual experiments. Seeing the fluctuations, it would have been better to repeat each run several times and combine the outputs, but this too was not feasible in the time available. We applied ten-fold cross-validation, each time training on 90% of the data and testing on 10%. Results were calculated once over the judgements for the data as a whole.

As we would like to analyze where things go right or wrong and how this may be influenced by specific aspects of the data, we also want to use a system which does allow us insight in its workings. A problem with Roberta is that it is rather hard to determine how it arrives at its classification. The field of Explainable AI (XAI) is dedicated to solve this problem, but explainability remains an issue for transformer systems performing NLP tasks, like sentiment analysis. As an example, Borgia et al. (2020) compare three explanation methods, namely LIME (Ribeiro, Singh & Guestrin 2016), Integrated Gradients (Sundararajan, Taly & Yan 2017) and their own attention-based method. They find that the explained classification correlates quite well with the actual system classification, but not perfectly. Furthermore, the three methods disagree quite often on the classification and the explanation. Even if we would trust that the explanations are valid for the cases we are interested in, which often happen to be the more interesting and therefore deviant cases, serious further modeling would be needed to arrive at these explanations. Not having found a ready tool for "looking into" the Roberta model, we decided to still treat it as a black box for the time being and turn to another classification method to provide us with an insight in the internal working of the classification.

We refer to an odds-based method from earlier research, henceforth referred to as "ODDS". For each uni-, bi- and trigram which occurs more than once in the training data, we calculate the relative frequency within the positive and negative cases and then determine the odds by dividing one by the other. Features with odds less than 2 are ignored, odds higher than 10 are reduced to 10. For each test case, we take the weighted average of the odds of all n-grams, using the $\log_2$ of the n-gram's overall frequency as weight. Then, we take all scores and normalize them in a nonparametric manner, by determining the odds that a given score belongs to a positive or negative case and again taking the $\log_2$. Finally, we linearly map the positive scores to the range [0,1] and the negative ones to [-1,0]. Note that all these calculations lend themselves perfectly to leave-one-out processing, so that we could use leave-one-out instead of ten-fold cross-validation for ODDS. ODDS, although with slight differences as regards the details, has been shown to yield quite adequate results, reaching second place in the VarDial 2018 shared task on distinguishing between Dutch and Flemish subtitles (van Halteren & Oostdijk 2018; van Halteren 2019). Moreover, here we are prepared to accept a slightly lower classification quality if this lets us investigate how it arrives at its result. In Section 5, we describe how we implemented such an investigation.

3.3 Classification results

The classification quality for the original dataset is shown in Table 2. From these numbers, we can conclude that a) the recognition is quite good, despite the idiosyncracies of the data (other than the field confusions) b) Roberta is better than ODDS, as expected, c) recognition is better when more training data is available, also as expected, with Roberta profiting more than ODDS.

However, we would like to look beyond these numbers. And indeed, if we do some error analysis, using the full set, we see that not all is as it should be. Among the most strongly misrecognized comments, we find examples of the types of denial discussed in Section 2.2 (exs. [35]-[36]), but also cases where it seems the reviewer was just not paying attention to the prompt (exs. [37]-[40]).

(35) (NEG) All was very good Rooms clean and staff very friendly (POS opposite: Rooms and breakfast were excellent)

Table 2: Classification results for the experimental subset, evaluated against the original labels (fields).

Train set	ODDS			Roberta		
	Prec	Rec	F1	Prec	Rec	F1
50K subset	94.60	94.68	94.63	95.75	95.84	95.79
Full 500K	95.01	95.10	95.05	96.40	96.47	96.43

(36) (POS) I can not think of one (NEG opposite: Elevator No room service Tiny rooms Name on Booking com is different than on hotel signs)

(37) (NEG) The hotel is close to picadily and ver convenient (POS opposite: Great French atmosphere)

(38) (NEG) Very friendly staff (POS opposite: It s comfort and location)

(39) (POS) Rooms are very small staff very rude refused to refund money only upgraded to another room after several phone calls with booking agent then it was still no better Only difference was the bed size still couldn t open my suitcase unless I put it on the bed Photos are of somewhere else or they have been altered (NEG opposite: The whole place)

(40) (POS) The front desk staff was unfriendly We got a room facing all the construction Asked for a room not facing the street construction Not sure if the lady at the front desk was having a bad day but she gave a tiny room where the TV was tucked in a corner So only one side of th room could see the TV The bed was horribly uncomfortable (NEG opposite: The bed was uncomfortable The TV was tucked in the corner of the room Front desk staff was unpleasant)

What an error analysis should also pay attention to is the set of most common errors. Here, 1731 (5.6%) of the errors are caused by comments consisting simply of the word *nothing* and being classified as NEG instead of POS. The reason is that this specific comment occurs 12 times more often under NEG than under POS. Other examples of comments that might occur on both sides are *location* (502/1.6% of the errors), *breakfast* (498/1.6%), *everything* (420/1.4%) and *staff* (181/0.6%). Also high in the list are further contradictions and denials, such as *all good* (135/0.4%) and *nothing at all* (66/0.2%). Strangely, some of these are found as errors both ways. *everything was great* causes 135 errors when in the positive field and 51 when in the negative field.

Now, we see that the fact that reviewers do not always fill out the review form as expected impacts on the results and will become manifest when doing an error analysis after modelling. However, that means that all the work in modelling has already been done and that it is often no longer feasible to redo the modelling in reaction to the findings. In other words, time has been wasted. At least a fast check beforehand (and rather a thorough one) can prevent this.

4. Polarity recognition with the adjusted data

In this section, we use the knowledge gained during the annotation of (the 50K subset of) the data. We now try to classify for the positive or negative content of each comment text and take various subsets of the data for training, depending on the detailed content label.

4.1 Data and Methods

For purposes of testing, we keep the subdivisions described in Section 2.2, but will present results with the corresponding positive and negative classes combined, e.g., MIXEDPOS and MIXEDNEG will be listed together as MIXED. Similarly, we have PURE, MPDENY, MPCONT and MPBOTH. The label with which the prediction needs to correspond is only the polarity of the detailed label,

Table 3: Classification results for the experimental subset, evaluated against our annotation.

		PURE	MPCONT	MIXED	MPDENY	MPBOTH	total
		71,905	944	5,207	4,173	561	82,790
ALL	ODDS	97.59	96.61	89.28	81.83	71.30	96.09
82,790	Roberta	97.69	97.78	88.19	93.17	94.12	96.84
PURE*+MIXED	ODDS	97.51	96.72	87.96	70.24	86.99	95.46
78,056	Roberta	97.62	97.99	88.96	35.68	51.69	93.65
PURE*+MPDENY	ODDS	97.70	97.35	86.90	86.41	93.76	96.41
77,022	Roberta	97.56	98.31	80.05	93.27	67.91	96.05
PURE*	ODDS	97.60	96.93	85.83	72.15	87.70	95.50
72,849	Roberta	97.65	97.56	83.27	36.78	62.57	93.44
ORIG	ODDS	96.33	92.06	82.81	7.93	56.51	90.71
82,790	Roberta	97.34	60.27	88.27	6.88	8.20	91.18

so POS or NEG. For purposes of training, we want to see the impact of various choices as regards the inclusion of problematic data. We use five different training sets

- ALL. All selected comments.
- PURE*+MIXED. PURE* plus the mixed comments.
- PURE*+MPDENY. PURE* plus the denying misplaced class.
- PURE*. All purely positive and purely negative comments. These consist of the PURE class and the contradicting misplaced class (MPCONT).
- ORIG. The field of the comments in the original data. This simulates that we think we are doing the actual polarity recognition without noticing the problems in the data. However, we are not evaluating against the columns but against the actual polarity, for which we would have had to notice...

The methods are exactly the same as described in Section 3.2 for the experiment with classifying the data on the basis of the field indicator. Furthermore, the experiment uses the same setup, being 10-fold cross-validation for Roberta and leave-one-out for ODDS.

4.2 Classification results

The average F1 scores for all combinations of training and test sets are shown in Table 3. Each pair of rows lists the results for a specific training set for the two systems. The number under the name of the training set is the number of training items. Each column lists the results for a specific test set (or the complete test data). The number under the name of the test set is the number of test items. Within each column, blue indicates the best ODDS result and red the best Roberta result.

Taking the data set as a whole (column "total") again confirms Roberta being the better classifier, although with a smaller margin than above, and also shows that this task too seems to be handled very well. Roberta does best when seeing all types of comments as training data but, interestingly, ODDS prefers not to be bothered with mixed comments in its training, as it scores highest with PURE*+MPDENY.

Looking at the individual types of comments yields some more insights. To classify PURE comments, Roberta prefers to train on all comments, but the gain over PURE* alone is minimal. ODDS again does better without the MIXED comments added in. Both systems arrive at the same

quality. The ORIG train set is also able to keep up here, as PURE in the correct field is the normal situation for the bulk of the comments. Still, it does score lower as it is hampered by the non-conforming cases. The MPCONT test set, which are also in the PURE* training set, are classified (almost) equally well as the PURE test set. For ODDS, the scores tend to be a bit lower overall, with the best score again at PURE*+MPDENY. For Roberta, scores are even a bit to a lot better, but surprisingly not when using only the PURE* (i.e. PURE+MPCONT) training set. Apparently, MPCONT is so different, and has so few examples, that Roberta gets confused. Here, Roberta agrees with ODDS that MIXED training material is confusing rather than useful. For the ORIG training set, scores are quite a bit lower, as in that training set these cases end up in the opposite class. ODDS appears to be able to keep up, but Roberta drops to 60%.

As could be expected, the classification of MIXED comments fares considerable worse. In fact, the scores are higher than expected, given that it is often hard to determine the overall sentiment on the basis of the mixed subcomments. Apparently, the users did tend to put the comment in the field most appropriate to their stated opinion. Roberta works best when training on PURE*+MIXED, as expected. However, ODDS does better if the training material includes MPDENY and MPBOTH, which maybe then contain (parts of) mixed content too, and manages to outperform Roberta, although only slightly. When MIXED examples are excluded from the training set, both systems lose quality, Roberta much more than ODDS. With ORIG, Roberta does not much worse than with the manually verified material, unsurprising as we are using the original field labels. ODDS, on the other hand, does not do so well here.

For MPDENY, both systems clearly need MPDENY training examples to work properly. Roberta then scores around 93% whether or not MIXED is included in the training. ODDS does not get further than 86% and even 81% if MIXED is added. However, without MPDENY training, ODDS scores around 70%, whereas Roberta scores only half that. For both systems, ORIG expectedly scores very bad, lower than 10%. MPBOTH, finally, offers an extremely mixed bag of results. Both systems are able to reach about 94%, but only with one specific training set, and a different one at that. With other training sets, Roberta is stuck at 50 to 60%. ODDS is doing better then, with 60 to 70%. ORIG here performs as expected for Roberta with 8.2%, but ODDS seems to stick around chance behaviour, which means it is doing better by accident.

If we compare the experiments in Sections 3 and 4, we can say that both tasks can be executed about equally well, with a slight edge for the polarity task over the user-selected field task. However, trying to do polarity recognition on the basis of field assignment performs considerably worse, especially for the non-standard cases. This does not come as a surprise, as it is well-known that machine learning degrades when training and test are not completely compatible. Looking at the overall comparison of ODDS and Roberta, even though this is not the focus of this paper, we can conclude that Roberta is best in general but ODDS does manage to do better in niche situations. Furthermore, what is the best combination of training data also depends very much on the set up and the system. Preclassifying into types and subsequently split processing could be a viable strategy in tasks like these.

5. Detailed analysis of the impact of data adjustment

As already mentioned, we have no way of finding out why Roberta makes the choices it does but we can see exactly what ODDS is doing. At each word position, we can see which n-grams make which contribution to the score. Trying to look at all n-grams is a bit overwhelming, so that we chose to merge all contributions of the n-grams which are active at specific word positions.

5.1 Word contributions

The first visualization is at the word level. For each word, we total its contribution during testing over the whole dataset. We then plot these scores for two different training set selections against each other.

Figure 1: Total contributions in ODDS of n-grams containing a specific word, comparing ORIG to ALL.

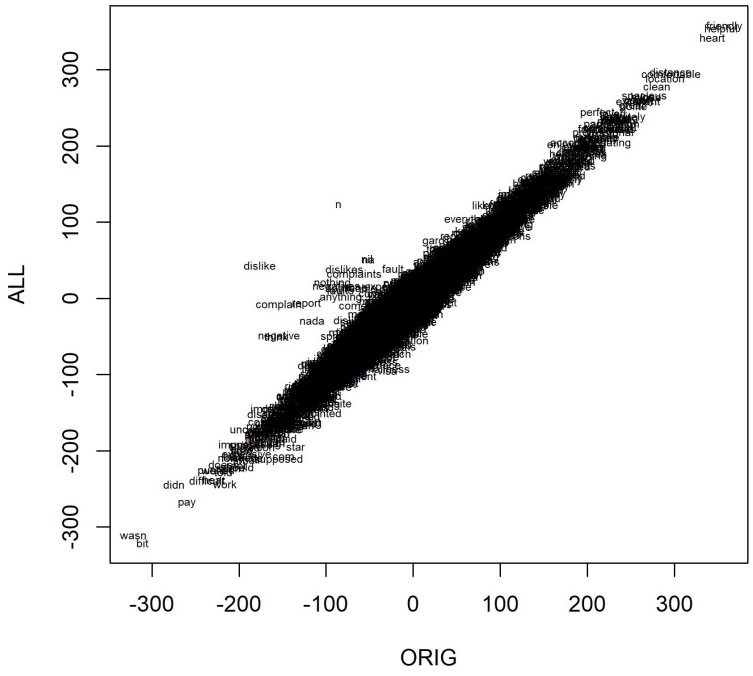

Figure 1 shows this plot for the original field labeling versus our polarity-of-content labeling. Overall, there are no enormous changes (Pearson's r = 0.959) but there are quite large changes for individual words. However, remember that these numbers are not linked to the unigrams but to all n-grams containing them. Most of these words are involved in denying that anything is wrong, e.g. *nothing to dislike* or *no complaints*. After *dislike* with a jump of +336.20 in the lead, there are six more words which climb more than 150 points, namely *complain*, *n*, *complaints*, *nothing*, *negatives*, and *think*. *no* is hardly touched (+9.14), probably because it is so common also outside the misplaced comments. *n* stems from the comment *n a*, in which *a* obviously is not touched either. *think* is seen in comments like *cannot think of any negative comments* or *i m struggling to think of anything*. With 110 examples of reviewers having trouble to think of anything negative and only 2 of anything positive, *think* is pushed strongly towards a positive opinion. We saw already above that this trend of having far more misplaced positives than misplaced negatives, also exists in general, making the changes in odds mostly positive. The largest negative change is to *supposed* with -81.67, going from negative to very negative.

Given this prominence of denials, we can expect the same differences between adding and leaving out misplaced denying comments in our training with relabeled data (Figure 2). Indeed, we see the same words outside the diagonal, but the changes are much smaller (r= 0.997). Here, *dislike* only gains +123.13

Figure 2: Total contributions in ODDS of n-grams containing a specific word, comparing PURE* to PURE*+MPDENY.

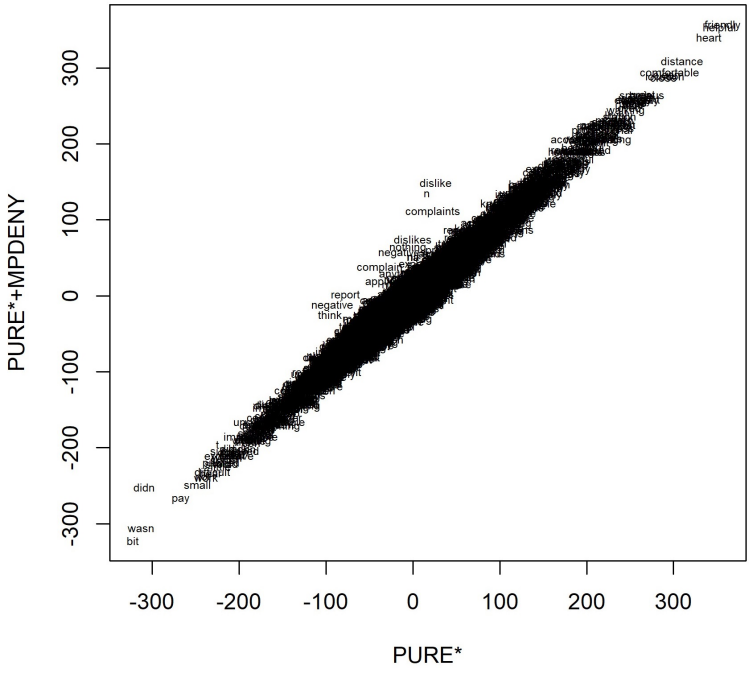

Figure 3: Total contributions in ODDS of n-grams containing a specific word, comparing PURE* to PURE*+MIXED.

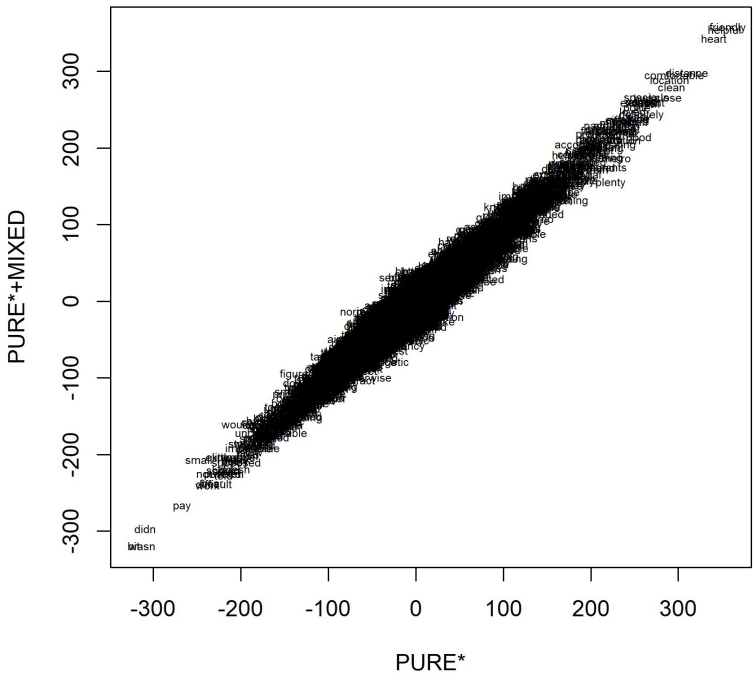

Figure 3 shows the plot for PURE* versus PURE*+MIXED. Here we do not see any words jumping outside the cloud, but only very many small movements that together bring more change (r=0.990) than the visually more remarkable ones in Figure 2. Again, we see that including the mixed cases significantly affects the task. The most prominent examples on each side are *norm* (+57.39), where people are unhappy but realize that what they are unhappy with what is apparently the norm in that country (example [41]), and *plenty* (-68.03), which is by itself generally positive but also used often in mixed negative comments offering alternatives for something negative (example [42]).

(41) The bathroom was good The thing is that compared to USA it was small but small bathrooms seems to be the norm here in Europe

(42) There is no bar area however there are plenty of good cafes and bars nearby

5.2 Full comment visualisation

The plots in Section 5.1 show us large effects and allow us to discover types of problems and solutions. However, they still do not give all that much insight in what exactly happens in each individual reviewer's comment. For that, we use another visualization. As stated, we have available the sum of all n-gram contributions at each word position. If we translate these numbers into colours, e.g. blue for positive, red for negative, and darker representing higher numbers, then we can see exactly how the system arrives at its final decision. Again, we have to stress that these numbers/colours are not linked to the unigrams but to all n-grams containing these words.

Table 4: Shifting opinions in ODDS for a MIXEDPOS example.

TRAIN	ODDS SCORE	ODDS	TEXT	ROBERTA
ORIG	-0.00193	NEG	there is nothing to dislike here i stay in the hotel on a regular basis last two years or so and i am so much pleased to see that the level of service has not become any worse i see staff members are being changed but the not the service	NEG
PURE*	-0.03495	NEG	there is nothing to dislike here i stay in the hotel on a regular basis last two years or so and i am so much pleased to see that the level of service has not become any worse i see staff members are being changed but the not the service	NEG
PURE*+MPDENY	0.01169	POS	there is nothing to dislike here i stay in the hotel on a regular basis last two years or so and i am so much pleased to see that the level of service has not become any worse i see staff members are being changed but the not the service	POS
PURE*+MIXED	-0.01664	NEG	there is nothing to dislike here i stay in the hotel on a regular basis last two years or so and i am so much pleased to see that the level of service has not become any worse i see staff members are being changed but the not the service	NEG
ALL	-0.02646	NEG	there is nothing to dislike here i stay in the hotel on a regular basis last two years or so and i am so much pleased to see that the level of service has not become any worse i see staff members are being changed but the not the service	POS

Let us start again with the effect of denials. Table 4 shows the word position contributions for each of our five training set combinations. For most of the comment, scores remain constant. For some words, such as *regular* and *basis*, there are minor changes. But the most serious impact is visible in *nothing to dislike*. In PURE*, *to dislike* is neutral and *nothing* is negative. To understand why, we would have to investigate all occurrences of these words in the data and this understanding is not important enough at this point to undertake that task. What is important here, is that without knowing the value of *nothing to dislike*, there are more negative than positive connotations in this comment, leading to an overall erroneous opinion that this is a negative comment. In ORIG, *nothing to dislike* is extremely negative, as it occurs many times in the negative field. However, shifts in the other words, and the fact that scores are judged in relation to the scores in other comments, still lead to a less negative overall judgement. In PURE*+MPDENY, *nothing to dislike* is extremely positive, as it occurs the same number of times marked as MPDENY, and therefore is taken to be POS. This now allows the system to recognize that the comment is overall positive

(0.01169). In PURE*+MIXED, the system misses the denials but sees all the mixed comments, which apparently make *to dislike* seem negative. Still, other shifts make the overall less negative than with PURE*, but still negative in the end (-0.01664). Finally, ALL regains the information from the denials, but it looks like the information from the mixed comments keeps the final opinion at negative (-0.02646), even more negative than with PURE*+MIXED, which cannot really be seen in the individual positions in the visualization. As for Roberta, we cannot see the details, but see that it agrees with the odds-based system in the first four rows. In the last one, with all information available, it chooses POS rather than ODDS' NEG. This may appear consistent with Table 3, where Roberta outperforms ODDS with the full training set (96.84% versus 96.09%), but in fact is not, as ODDS was better when considering the mixed test set (89.28% versus 88.19%).

Table 5: Shifting opinions in ODDS for a MIXEDPOS example.

TRAIN	ODDS SCORE	ODDS	CONTRIBUTIONS	ROBERTA
ORIG	-0.10505	NEG	everything else was ok	NEG
PURE*	0.22758	POS	everything else was ok	POS
PURE*+MPDENY	0.14755	POS	everything else was ok	POS
PURE*+MIXED	-0.11631	NEG	everything else was ok	POS
ALL	-0.19251	NEG	everything else was ok	POS

A similar strong change can be seen in Table 5, now connected to denials like *everything was fine*. Once these misclassifications are corrected, *everything* goes from strong negative to strong positive. However, the positive load is lost again when we are adding MIXED training data. Furthermore, *else*, *was* and *ok* get a negative load from the MIXED material, so that PURE*+MIXED and ALL fail on their final classification. Roberta was also mistaken with ORIG and put right when given the relabeled data. However, unlike ODDS, it is not confused by the MIXED training data. This time, the results are consistent with Table 3.

Table 6: Shifting opinions in ODDS for a PURE POS example.

TRAIN	ODDS SCORE	ODDS	CONTRIBUTIONS	ROBERTA
ORIG	-0.04294	NEG	rooms are fine	POS
PURE*	0.13594	POS	rooms are fine	POS
PURE*+MPDENY	0.04022	POS	rooms are fine	POS
PURE*+MIXED	-0.09918	NEG	rooms are fine	POS
ALL	-0.15252	NEG	rooms are fine	POS

The local effects do not always have to be extreme, as we can see in Table 6. In principle, we see the same pattern as in Table 5, with PURE* and PURE*+MPDENY being right, but MIXED material confusing the matter again, even more so than with ORIG. However, the individual words change much less. In fact, in this context, they are all more or less negative even for PURE* and PURE*+MPDENY. Still, since an overall negative score is more common and the threshold is therefore negative, they are classified correctly there. Here, Roberta is never confused, not even by the ORIG data.

6. Conclusion

In the previous sections, we tried to answer the research question "How and to what degree is the modelling of polarity influenced by some Booking.com users not (strictly) following the prompts

asking for positive and negative comments for specific fields." We did this on the basis of a dataset of about 500K reviews with both a positive and a negative comment field, which we used to study the task of polarity recognition, i.e., of determining whether a short comment expresses a positive or a negative opinion.

The data was extremely varied, due to various influences, described in Section 2.1. Part of the variation was caused by the typical informal and uncareful text production known from other user-generated text. This, however, did not appear to hinder the classification overly much, as scores of over 95% F-score proved to be possible. Still, some of the remaining 5% may well be caused by this type of variation. We did not examine this in detail, as this was not our main research question.

We rather focused on a different type of variation, namely that caused by the frequently occurring discrepancy between the field label, positive or negative, and the actual content of the comment. On top of the occasional confusion of the fields, we found two major types of discrepancy. More or less expected were the comments in which there was a mix of positive and negative statements. Unexpected, and related to the fact that the reviewers were asked explicitly for positive and negative aspects in separate fields, were the statements in which the reviewers went against the intent of the field (cf. Section 2.2). This could be either by denying that anything was bad/good or by contradicting by way of stating something good/bad, or even both. Both the mixed and the contrary comments did lead to (sometimes serious) degradation of the classifier performance. Only after we used relabeled comments during training as well, did performance reach a good level again.

Whether this problem is noticed depends on whether there is an inspection of the data, either beforehand in preparation for experiments or afterwards in error analysis. If neither is done, the problem will go undetected, as the classifier performs quite well. It is just performing another task than the stated one, namely determining in which field the reviewer put the comment, instead of which polarity the comment has. When taking into account the various types of comments, real polarity recognition becomes possible again at the stated 95% F-score. Only the mixed comments remain problematic, with an F-score of just under 90%. Comment-internal annotation would be needed to solve this problem.

What this study illustrates is that you need to examine your experimental data carefully. If you do not do this, there may be consequences. At best, you may get suboptimal results. In the worst case, you may be modeling the wrong task. It is tempting to assume that there just happens to be a problem with this specific dataset. And obviously, we cannot pose that similar problems will exist with any other internet-based dataset. However, we do think that these kinds of problem occur more often. It seems unwise not to inspect your data before you invest serious amounts of work in them. Unfortunately, we cannot predict what might be present in there. You will just have to see for yourself.

Acknowledgements

We thank all the contributors to Huggingface for making the tools available so that us non-experts can also perform experiments like the ones in this paper. Furthermore, we thank the anonymous reviewers for pointing us in ways that let us improve this paper.

References

Aston, Guy and Lou Burnard (1998), *The BNC handbook exploring the British National Corpus with SARA*, Edinburgh University Press.

Bodria, Francesco, André Panisson, Alan Perotti, and Simone Piaggesi (2020), Explainability methods for natural language processing: Applications to sentiment analysis, *SEBD, June 21-24, 2020, Villiasimius, Italy.*

Campos, Diogo, Rodrigo Rocha Silva, and Jorge Bernardino (2019), Text mining in hotel reviews: Impact of words restriction in text classification, *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019)*, pp. 442–449.

Forhad, Md. Shafiul Alam, Mohammad Shamsul Arefin, A. S. M. Kayes, Khandakar Ahmed, Mohammad Jabed Morshed Chowdhury, and Indika Kumara (2021), An effective hotel recommendation system through processing heterogeneous data, *Electronics*. https://www.mdpi.com/2079-9292/10/16/1920.

Juliadi, Revin Novian and Yan Puspitarani (2022), Supervised model for sentiment analysis based on hotel review clusters using RapidMiner, *Sinkron : jurnal dan penelitian teknik informatika* **7** (3), pp. 1059–1066.

Kilgarriff, Adam and Gregory Grefenstette (2003), Introduction to the special issue on the web as corpus, *Computational Linguistics* **29**, pp. 333–347.

Lal, Kavita and Nidhi Mishra (2020), Feature based opinion mining on hotel reviews using deep learning, *in* Raj, Jennifer S., Abul Bashar, and S. R. Jino Ramson, editors, *Innovative Data Communication Technologies and Application*, Springer International Publishing, Cham, pp. 616–625.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016), "Why should I trust you?": Explaining the predictions of any classifier, *Proc. of the 22nd ACM SIGKDD Int.l Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Shortis, Tim (2007), Gr8 Txtpectations. The creativity of text spelling, *English Drama Media* **8**, pp. 21–26.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017), Axiomatic attribution for deep networks, *Proc. of the 34th Int.l Conf. on Machine Learning, volume 70*, pp. 3319–3328.

van Halteren, Hans (2019), Domain bias in distinguishing Flemish and Dutch subtitles, *Natural Language Engineering* **26**, pp. 493 – 510.

van Halteren, Hans and Nelleke Oostdijk (2018), Identification of differences between Dutch language varieties with the VarDial2018 Dutch-Flemish subtitle data, *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 199–209. https://aclanthology.org/W18-3923.

Appendix A. Examples

Here we provide examples illustrating the typical characteristics of the dataset at hand.

A.1 Reviews as an example of user-generated text

- Capitalization may be non-standard. In many cases, there is no capitalization as one would expect to find in written text (e.g. the use of a capital letter to mark the beginning of a sentence or proper name). Capitals are not used or are used with words in seemingly random positions. In some cases, capitalization is used for alternative purposes, such as stressing particular parts of comments. For example,

  (43) My colleagues stayed in the same hotel on the same floor booked after me using the same system and paid HALF THE PRICE I DID I paid 340 and they paid 171 and 200 respectively When I complained at the hotel they simply said Oh I m sorry that s how it works

(44) THOUGHT the pillows were a little old and worn

- Spelling is varied. Apart from standard British and US spelling, and spelling errors that reviewers (native and non-native speakers) make (see [45]-46]), we find influences of 'Txt spelling' (Shortis 2007), i.e. spelling as is used in informal online communication (e.g. clipped forms like *fab* and *comfy*, but also *gr8*, *vg* and *allllllllll*).

(45) Without contac me They have to put me in diffrent cause they had problem w water

(46) Norhing

(47) Location was excellent The staff extremely helpful Room was fab Wine and chocolates in room a lovely gesture for hubbys birthday Highly recommend this hotel

(48) The room is good with right amount of aminities and location was great as well Staff super helpful and friendly Gr8 value for the money

(49) Allllllllll good breakfast was amazing and the staff too

- Many reviewer comments are not well-formed sentences or phrases. They sometimes contain grammatical errors, but more often reviewers choose to leave out parts they consider to be obvious. Thus the grammatically well-formed version of example [50] is something like *The room was on the street side. It was very noisy.*.

(50) Street side very noisy

- Language use may be creative. See, for instance, examples [51]-[52].

(51) Every thing was p e r f e c t

(52) Nothing at all I am a frequent traveller and this stay was definitely my worst nightmare Disgusting by all means Kindly avoid avoid avoid things cannot be worse

A.2 Effects of the scraping process on the shape of the data

- All punctuation has been removed and all special characters (apostrophes, hyphens, etc.) and characters with diacritics have been replaced by blank spaces. For example,

(53) great breakfast great staff super views from the main bar

(54) Our room had a really strong smokey odor I think because of the guests before us it wasn t nice to sleep in as non smokers

(55) did not rate the breakfast more than 4 10

- Empty fields have been replaced by *No Positive* or *No Negative*, but not always. Some fields are found to be plain empty.

- Some review comments appear to have been cut off. This happens with some longer comments but also occasionally with short comments. For example,

(56) This hotel has opened almost 3 months ago and here it is on top of all hotels in Milan and there is more than enough reasons for it actually from our stay here we decided to lookup Room Mate as the first option wherever we go hoping this will be a chain standard This is still growing chain hailing from Spain and is currently expanding internationally best of luck guys and keep it up I even looked up some of their other hotels and they are towards the top in each city I looked up The reason for the raving reviews is very simple they put their selves in the travellers shoes and acted accordingly

whatever pains you when staying at a hotel is not here like Spacious rooms we had a triple room with double king size bed and single bed and still the room felt spacious at par with some 5 tar hotels electrical outlets conveniently placed very well lit spacious bathroom with two basins and both a bathtub and a big walk in shower with rain forest head toiletries that you are happy to use breakfast till noon free wifi is not only in premises but you can ask for the free wifi portable router to take it with you on the go with free 100 mbs everyday throughout your stay a storage room if you need to leave your luggage or as we did keep the baby stroller there instead of taking it up to the room everyday staff are remarkable in every aspect special thanks go to Piero Francesco Andrea Hany Hassan and Sara for making every encounter more pleasant than the one before and taking all measures to make us feel at home if you look up the address via Google Earth you will not find the hotel as the street was photographed while Giulia was under construction so the hotel is the construction work you find when zooming in to street level till google updates the photo this is an amazing location at the door of the hotel you re a couple of steps to the left to Duomo and another couple of steps to the right to Vittorio Emmanuelle gallery Duomo metro is

(57) Stairs only as no lift is available and