Explaining the Hardest Errors of Contextual Embedding Based Classifiers

Anonymous ACL submission

Abstract

We seek to explain potential causes for incorrect classification of the most challenging documents, namely, those that no classifier using state-of-the-art, very semantically-separable contextual embedding representations managed to accurately predict. To do so, we propose a misclassification taxonomy of incorrect predictions, which we used to perform qualitative human evaluation. We posed two (research) questions, considering three sentiment datasets in two different domains – movie and product reviews. Each evaluator (from two backgrounds) was requested to evaluate documents and compare the predominant (sentiment) model assignment to that in the gold human standard in order to decide whether misclassification was due to the model or the human and the likely reason for such misclassification. We quantified answers per category in our taxonomy across all datasets and computed their proportion, achieving a high inter-evaluator agreement of 81.7%. Differences were observed between the product and movie review domains, such as the prevalence of ambivalence in products´ and sarcasm in movie reviews. Our analysis also revealed an unexpectedly high rate of human mislabeling in the datasets and a significant number of model errors that we cannot yet explain (though can conjecture). To ensure reproducibility, our documentation, code, and datasets can be accessed on GitHub.

1 Introduction

In a scenario where the amount of user-generated content is growing exponentially, automatic text classification (ATC), one of the fundamental tasks in machine learning, plays a vital role in enabling the automatic categorization of texts into different semantic groups based on their distinctive characteristics (Li et al., 2022; Galke and Scherp, 2022).

The state-of-the-art in ATC is currently provided by Attention-Based Transformer methods (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020)), which produce contextual representations of words and documents. Indeed, in (de Andrade et al., 2023), the authors show that these contextual representations are so (semantically) separable in the embeddings space that any classifier using them achieves similar effectiveness, no matter how simple (e.g., a Nearest-Centroid classifier) or complex it may be (e.g., a Gradient Boosted Decisions Tree (GBDT) or a Support Vector Machines). Some of the results obtained in that study are the highest (state-of-the-art) ever reported in the literature for effectiveness (e.g., Macro-F1) in several experimented datasets.

With such powerful text representations and results, sometimes achieving or even exceeding human parity (Hassan et al., 2018; Yan et al., 2023), a main question that arises is: *Are we approaching the limits of what can be automatically classified by a machine learning model?*

This article delves deep into this question by analyzing the reasons for misclassification by classifiers using these powerful contextual representations. We go one step further to advance the literature and look into the **hardest cases**, i.e., documents that none of the strongest classifiers explored in the aforementioned study, using contextual embedding-based representations, were able to classify correctly.

A thorough review evidenced that such type of error or misclassification analysis is rarely performed in the literature, with a few exceptions (Martins et al., 2021). Misclassification analysis serves the purpose of revealing the how's and why's behind model (or human) failure. One of the main difficulties in performing such an analysis is the lack of standardized methodologies and methods for doing so. Accordingly, one of our contributions is the proposal of a **misclassification taxonomy** capable of categorizing incorrect classification *upon classifiers application*.

We propose and evaluate an *error taxonomy*

using a document sample for which no classifier can produce correct predictions. Aiming at simplicity when compared to more complex tasks, such as topic categorization, we focus on sentiment analysis (binary) tasks, adopting BERT to generate **contextual document representations**[1]. We evaluate the proposed taxonomy with a different sample of erroneous documents, using human evaluators with different backgrounds to assess how effective and useful the taxonomy is to explain the errors.

Unlike previous work (Martins et al., 2021) – which focuses on characterizing and assessing the impact of "hard" instances in the effectiveness of the polarity detection task using a single dataset of movie reviews and unconcerned about textual representation – we here focus on analyzing and quantifying the reasons for the misclassification of the **hardest** documents by all machine learning methods using some of the most separable representations in the literature. For this, we use datasets from two domains: movie and product reviews. We also contrast and compare the results in these two domains, gathering insights into the differences in the type of errors found in each of them.

The main questions we seek to answer are:

RQ1 *Is the proposed taxonomy effective for misclassification analysis?* To answer RQ1, we analyze evaluators' responses regarding their level of agreement – the higher the agreement, the more effective the taxonomy. We analyze inter-evaluator disagreement and correlate that with hardness in classifying.

RQ2 *Can the proposed taxonomy be used to reveal the main reasons for misclassification? Are there significant differences in the results among different domains?* In RQ2, drawing on the consensus achieved, we quantify and analyze the main reasons and causes for misclassification, outlining potential differences between domains.

Our experiments engaging eight human evaluators with two different backgrounds (Computer Science and Linguistics) and three datasets, two in the movie reviews domain and one in the product reviews domain, revealed that (i) the developed taxonomy proved effective, with an inter-evaluator agreement of over 81% for error category – this suggests that evaluators find it relatively easy to identify classification errors using the proposed taxonomy; (ii) across all assessed datasets, the predominant category of reason for error (> 50%) is "Sufficient information with model failure", a result that can be potentially leveraged for model enhancement; (iii) the evaluators found a significant amount of mislabeling by humans – i.e., evaluators considered that the model provided a correct label, but the human mislabeled the document – in around 33% of the documents in the product dataset and 16% in one of the movie datasets; (iv) in movie reviews, sarcasm[2] (> 23% of the cases) is a major reason for the model error. We believe this is a particular characteristic of this domain; (v) in product reviews, ambivalence (40% of the cases) is the main reason for model error; (vi) in at least 41% of the evaluations, evaluators reported sufficient information in the text, with the model failing in its prediction for reasons we cannot yet explain (though can conjecture).

In sum, our main contributions include:

1. The development and evaluation of a taxonomy for categorizing the main causes of misclassification of the *hardest* documents

2. A deep analysis of the results of a comprehensive qualitative experiment applying the taxonomy to three different datasets in two different domains. The results may have interesting implications for the improvement of the next generation of textual classifiers and representations

3. A release of a new dataset of challenging documents manually annotated by humans.

2 Related Work

Reasons for models' misclassification have been investigated for texts and images. In (Lee et al., 2017), five categories for misclassification of objects in images are explored (See Appendix A.1 for their Evaluation Schema). Another study, based on textual documents, aimed to categorize prediction failures (Meek, 2016) by defining four error categories (see Appendix A.1 for schema), focusing on the lack of information on the training. Unlike these works, we propose a taxonomy for ATC and investigate a more comprehensive set of reasons for test errors, focusing on the hardest cases for the classifiers using qualitative human assessment.

[1] We run experiments in our datasets comparing BERT with other transformers such as RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020). The differences are minimal (if any) and potentially not influential in our work.

[2] Unlike (Frenda et al., 2023), we group irony and sarcasm under a single category as instances of figurative use of language intended to produce an effect on the reader.

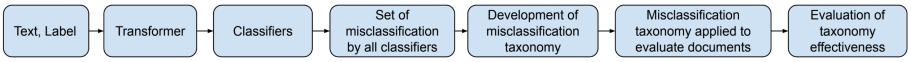


Figure 1: Flowchart of the Evaluation Methodology.

While other works (see Pandey et al. (2022) focus on human labeling by assessing the impact of (i) the number of documents; (ii) the time allocated to evaluators; and (iii) the order in the sequence of annotations in the labeling task, our study focuses on a set of (test) documents that have been misclassified by *all* classifiers using state-of-the-art, very separable (contextual) representations.

Regarding misclassification, works such as Bras et al. (2020) present a strategy to remove bias in training to reduce misclassification. Pleiss et al. (2020) proposes adapting the Area Under the Margin to identify data in a training set that precludes generalization. Both focus on the training set, on which the label of documents is known, to identify (challenging) documents that do not contribute to the learning process. Instead, we focus on misclassification at inference time, i.e., on the test set. We provide a methodology for understanding why different ATC models perform poorly in the same set of test documents, aiming to identify common characteristics of these misclassified documents.

Swayamdipta et al. (2020) and Ethayarajh et al. (2022) focus on understanding misclassified documents through using different methodologies from ours. Swayamdipta et al. (2020) present a model-based tool to characterize and diagnose datasets regarding the model's behavior on individual instances during training based on their confidence and confidence´s variability. Ethayarajh et al. (2022) seek to find challenging documents using V-usable information as a metric. Unlike the above works, we find challenging documents based on their incorrect classification by four classifiers using a very separable contextualized representation (BERT) as input. Moreover, unlike Ethayarajh et al. (2022), who do not provide qualitative experiments involving human evaluation and Swayamdipta et al. (2020), who evaluate human mislabeling only, we evaluate both automatic and human mislabeling.

Finally, in a related study, Martins et al. (2021) analyze a set of hard instances (See Appendix A.1 for evaluation schema) but, unlike ours, their study centers on characterizing and evaluating the influence of challenging cases on the classifier's effectiveness when performing polarity detection using **one single movie review dataset**. In our study, we concentrate on analyzing and quantifying the factors contributing to misclassification of the hardest documents *using multiple datasets in two domains*. Moreover, compared to Martins et al.'s, our taxonomy has more fine-grained categories, such as sarcasm (a broad category including sarcasm, irony and figurative use of language) and ambivalence (contrasting opinions in the same text). We have additional goals such as validating our taxonomy and contrasting the results in multiple different datasets and domains (Martins et al. (2021) used only one dataset), running qualitative experiments engaging human evaluators with different backgrounds.

3 Experimental Methodology

Our methodology, comprising seven steps, is summarized in Figure 1. The text and label for each document are used as input for fine-tuning a Transformer model, resulting in an encoder that produces contextual embedding vectors representing the documents using the CLS approach. We employ various classifiers with these embeddings as input, exploring different underlying techniques. From this set of classifiers, we select the set of documents for which none of the classifiers can produce correct predictions (according to the assigned labels in the datasets). Within this set, we sample documents for analysis to outline misclassification categories ("Development of the misclassification taxonomy" in Figure 1), which human evaluators use to evaluate documents ("Application of misclassification taxonomy to evaluate documents" in Figure 1) in a second sample different from the first one. Upon the application of the taxonomy, we quantify the results and evaluate its efficiency. A detailed account of the steps follows.

3.1 Datasets

Our study draws on three datasets developed for binary sentiment classification. Each dataset was constructed with text and an associated sentiment label. The first dataset comprises customers' reviews of purchased products on Amazon's website (Keung et al., 2020), which are assigned a rating from 1 to 5 stars by customers. We collected reviews containing ratings of 1 and 2 stars and labeled them as negative, while reviews containing ratings of 4 and 5 stars were labeled as positive. We discarded reviews with 3 stars (deemed neutral).

The second (PangMovie (Pang and Lee, 2004)) and the third (VaderMovie (Ribeiro et al., 2016)) datasets were used in (de Andrade et al., 2023) and we obtained the representations directly from the authors of (de Andrade et al., 2023). These datasets are made up of movie reviews, comprising text and a sentiment label (positive or negative). Table 1 presents some statistics of the datasets. As we can see, class distribution into positive and negative instances is balanced in the three datasets.

Dataset	Documents	Avg words	Positive	Negative	Classifier Macro-F1
Amazon	168000	33	84000	84000	94.2
PangMovie	10662	19	5331	5331	87
VaderMovie	10568	19	5242	5326	89.1

Table 1: Datasets Statistics

3.2 Data Representation

We fine-tuned BERT, adapting this LLM to the specific domain of sentiment classification using the texts and labels in our datasets. The aim is to improve the representation and enhance the model's effectiveness for sentiment classification. The model's fine-tuning produces an encoder, which generates CLS-based 768-dimensional embedding vectors to represent the documents. As discussed in (de Andrade et al., 2023), this fine-tuning process is fundamental to ensure the quality of the representation and the separability (into semantic classes) of the generated embedding space.

To perform fine-tuning, we used the literature´s suggested hyper-parameterization, fixing the learning rate with the value $2x10^{-5}$, the batch size with 64 documents, adjusting the model to five epochs, and setting the maximum size of each document to 256 tokens. We used differentiable heads by fine-tuning with *AutoModelForSequenceClassification*. In our experiments, we employ a five-fold stratified cross-validation procedure – fine-tuning, training, and optimizing the classifiers' parameters with the validation sets that are repeated five times. Reported results correspond to the average of the five test folds.

Although we used BERT in our study, other Transformers can be easily applied within our methodology. Indeed, experiments in (de Andrade et al., 2023) showed that the contextual representations produced by different transformers (e.g., Roberta, BART) are quite similar regarding class separability, the main aspect driving our evaluations. To confirm that, we run experiments of our own in the tested datasets comparing BERT with RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020). Results are shown in Appendix A.5. As we can see, these transformers´effectiveness are very similar – BERT is statistically tied as the best method with Roberta in Amazon and marginally loses (by at most 1-2 pp) in the other two datasets. These differences, which means just a few documents in practice, is potentially not relevant in a qualitative study as ours, which uses a sample of the documents that all classifiers predicted incorrectly. We believe that the intuitions and insights we got with the current methodology, representations and models would not be substantially different if we used other Transformers.

3.3 Text Classifiers Used along with Contextual Embeddings

For document classification, we used the textual representations generated by the Transformer as input to four of the strongest classifiers used in (de Andrade et al., 2023), namely: KNN, Random Forests (RFs), Support Vector Machines (SVMs) and Logistic regression (LR). Indeed, despite using different rules and heuristics, the effectiveness of these classifiers (and of all other classifiers tested in (de Andrade et al., 2023)) is basically the same in all tested datasets when using the contextual embedding representations. This is due to the fact that these representations are already so semantically separated (by class) in the embeddings space that the employed classifier has no little effect in the classification process. For a detailed comparison among these classifiers (taken from (de Andrade et al., 2023) in two of the tested datasets check Appendix A.5.

Our decision was to explore classifiers based on different approaches – decision rules (RFs), local neighborhoods (kNN), global maximum margins (SVMs and LR) – so that if all of them misclassify the same document, this can be ascribed to the misclassified document being hard to classify. *And we do want to understand the reasons why!*.

Hence, we selected the set of documents that all classifiers misclassified in the three datasets, as presented in Table 5. A sub-sample from this set was used as a basis for devising our taxonomy, and a different (disjoint) sub-sample was used for actual evaluation, as described next. Table 2 shows the number of misclassified instances by all classifiers – there is no significant skewness in the distribution of positive and negative misclassified documents. We took a random sample of 60 misclassified documents from each dataset for evaluation, and the results are presented in Section 4.

Dataset	Misclassification	Positive	Negative
Amazon	216	115	101
PangMovie	120	54	66
VaderMovie	85	37	48

Table 2: Set of misclassifications by all classifiers.

3.4 Taxonomy Development

We conducted a preliminary round of assessment using a set of 15 randomly selected documents from PangMovie and Amazon. During this round, we convened to discuss potential sources of misclassification, aiming to better comprehend the reasons behind incorrect predictions. Through this process, we agreed upon a set of potential reasons, which represent the bulk of the categories in our taxonomy of errors. We carried out a subsequent evaluation with another set of 15 documents from each dataset, refining definitions, instructions, and the evaluation process. Upon concluding this iteration, we excluded all documents used in the preliminary stage and proceeded with a new evaluation. We randomly selected 60 samples from each dataset for manual human evaluation.

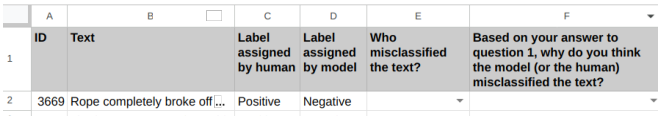
3.5 Distribution of documents

To evaluate the selected texts, we recruited eight participants with expertise in Computer Science and Linguistics, all with prior experience in annotation tasks for NLP. The participants comprised two professors holding a Ph.D. in Computer Science, one with a Ph.D. in Linguistics, and five students pursuing their bachelor's or master's degrees. All evaluators had advanced proficiency in English, which attests to their ability to perform the evaluation task.

Each participant was assigned 30 out of the 60 documents in each of the three datasets, totaling 90 documents per evaluator. Each document was assigned to be evaluated by four participants, two having a computer science background and the other two having a linguistics background. The decision to assess each document by two evaluators from each field was meant to enable quantification of (dis)agreement within the same background groups and between the two groups with different backgrounds. Section 4 presents results considering all four evaluators – impact of evaluators' background is analyzed in Appendix A.4.

3.6 Evaluation Form

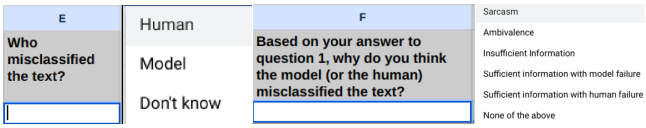
Individual forms were created for each evaluator and shared on a web cloud provider, ensuring evaluators could not access each other's forms. Our evaluation form comprised four tabs, the first containing instructions on how to evaluate the documents and the remaining ones having one sample of documents per tab, each line containing a document and the categories to be assigned to it.



	A	B	C	D	E	F
1	ID	Text	Label assigned by human	Label assigned by model	Who misclassified the text?	Based on your answer to question 1, why do you think the model (or the human) misclassified the text?
2	3669	Rope completely broke off...	Positive	Negative		

Figure 2: Screenshot of form provided to evaluators.

Figure 2 shows the form provided to evaluators, with columns for text ID, text to be evaluated, label assigned by a human, and label assigned by the machine model. Two additional columns were assigned to be filled in by evaluators with their answer to two questions: (i) "Who misclassified the text? ", for which one out of three options could be chosen: "Model", "Human", and "I don't know" (see Figure 3a); and (ii) "Based on your answer to question 1, "why do you think the text was misclassified?", for which one out of six options could be chosen, as shown in Figure 3b. Table 3 provides a description of the available options.



(a) Question 1 (b) Question 2

Figure 3: Form with questions and options for selection.

3.7 Categories To Evaluate Misclassification

The second question in our evaluation form required the evaluator to choose a category that could account for the misclassification. The instructions tab provided evaluators with examples of each category, some of which are presented in Table 4. The first row shows an example of a text misclassified due to the model's failure despite sufficient information to make the correct decision. In this case, the model assigned a negative sentiment, though the text contains a positive opinion "precious increments artfully.....". The second row shows an example of misclassification due to human failure with sufficient information. The wording "completely broke off" indicates a negative opinion, but it is labeled as positive. The third row is a misclassification ascribed to sarcasm, where "seen it before" is a negative opinion ironically expressed. The fourth row exemplifies a misclassification due to *ambivalence*, where despite the negative word "expensive", there are two favorable opinions ("won't oxidize" and "better than soap").

Category	Description
Sarcasm	Text contains ironic expressions (words that are the opposite of what one means), humorous expressions, figurative language (metaphors, idioms)
Ambivalence	Text contains both positive and negative opinions
Insufficient Information	Text is very brief and lacks information to assign the predominant sentiment
Sufficient information with model failure	Text contains sufficient information but model was unable to correctly assign the predominant sentiment
Sufficient information with human failure	Text contains sufficient information but human was unable to correctly assign the predominant sentiment
None of the above	None of the above categories can be said to account for the misclassification

Table 3: Misclassification categories and their description

Text (Category): Examples of each category
A film of precious increments artfully camouflaged as everyday activities (**Sufficient information with model failure**)
Rope completely broke off after a couple of months (**Sufficient information with human failure**)
Final verdict: you've seen it all before. (**Sarcasm**)
Expensive but won't oxidize metal. Maybe better than soap (**Ambivalence**)

Table 4: Misclassifications included in evaluators' guidelines

The categories "Sarcasm" and "ambivalence" are designed to capture very different instances of language use. "Sarcasm" refers to instances of language use when a user makes a statement that is meant to be understood figuratively. For instance, if a movie is assessed as being "a sleeping pill that works wonders", the statement is meant to be understood as "a very boring film to the point it makes you fall asleep". "Ambivalence", on the other hand, refers to instances of language use where two contrasting sentiments are worded. Hence, if a movie is assessed as "having an excellent cast despite being very slow-paced," both a positive and negative sentiment are expressed. "Ambivalence" does not inherently implicate figurative language.

We created the taxonomy based on an extensive survey of works seeking to categorize misclassification and held discussions until we reached a consensus on the taxonomy´s categories. These categories may apply to several ATC tasks besides sentiment analysis when there is some type of opinionated comment. We believe that our methodology is robust enough to be applied to other tasks beyond sentiments, as several categories pertain to general ATC problems, regardless of the domain.

4 Results

Documents were assessed by 4 evaluators. Question 1 required selecting 1 out of 3 alternatives, whereas Question 2 had 6 alternatives. Consensus was defined as one of the alternatives having the *majority* of votes – 4, 3, or 2 [3]. If there was no majority for a document, it is classified as "No consensus".

[3] In the case of two votes, provided that the remaining two alternatives have one vote each.

4.1 Taxonomy effectiveness

To answer our first research question: "Is the proposed taxonomy for misclassification effective to be used for misclassification analysis?", we analyzed the responses from questions 1 and 2 provided by the evaluators. We consider a taxonomy effective if there is high consensus among evaluators upon the defined categories and if there is low consensus in a category that has no definition, in our case, "Don't know" for Question 1 and "None of the above" for Question 2.

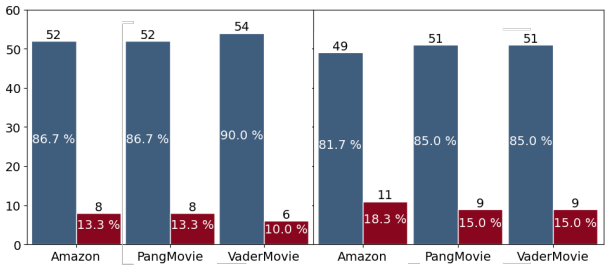
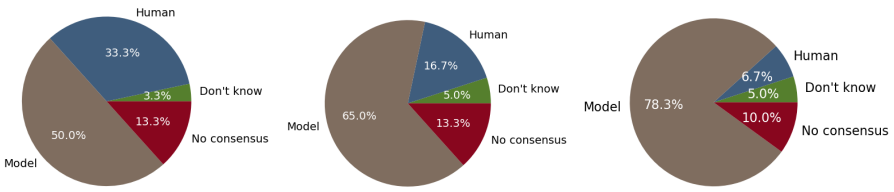


Figure 4: Consensus and No Consensus on Q1 (left) and Q2 (right).

Figure 4 shows the consensus percentages obtained for Questions 1 and 2 in the three evaluated datasets. For Question 1, out of 60 documents, 54 attained high inter-evaluator agreement in VaderMovie, and 52 in Amazon and PangMovie. In other words, in at least 86.7% of the cases (52/60), consensus was achieved in some category defined for Question 1 in the three evaluated datasets, implying low difficulty for evaluators to define a type of misclassification. We break down those numbers in Appendix A.5 to show the consensus distribution per document and A.6 per evaluator background. As shown there, the vast majority of the documents had the same categorization assigned by 4 or 3 evaluators, emphasizing high agreement and taxonomy effectiveness.

Figure 4 (b) shows the consensus percentages for Question 2. It is important to bear in mind that in Question 2, six options were available, likely leading to a higher difficulty in achieving agreement. Nonetheless, we can observe a high consensus in all datasets for this question, with the lowest value being obtained in the Amazon dataset, 49 out of 60



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 5: Percentages for answers to Question 1 in the three datasets.

documents reaching at least 81.7% consensus. As also shown in Figure 5, "No Consensus" was below 14% for Question 1 and below 19% for Question 2. In Appendix A.2, we show examples (in Table 6) of documents that posed difficulties for evaluators.

4.2 Response Analysis

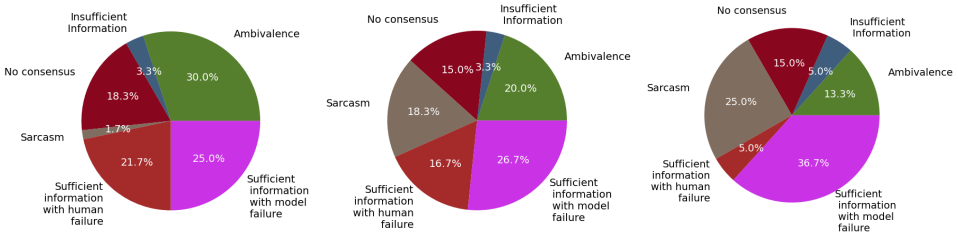
Upon concluding our consensus analysis, we explored the responses given by evaluators. Figure 5 shows that 50% of the misclassifications in the Amazon dataset were ascribed to the model. This is even higher in the movie datasets, emerging as the main misclassification reason in 65% of the cases in PangMovie and almost 80% in VaderMovie. Percentages for the option "Don't know" are very low in all datasets. Together with the option "No consensus", they achieve at most 18.3% in PangMovie (and 16.3% and 15% in Amazon and VaderMovie, respectively) of all analyzed documents in all datasets.

Though lower than Model's errors, the percentage of errors ascribed to the "Human" category is significant, mainly in the Amazon dataset (33%) (See Appendix A.3 for more details). This means that in 33% of the misclassifications, 3 or 4 evaluators (in the majority of the cases) considered that the model classified the document correctly and the human-provided a wrong label. Though lower in the movie domain, human mislabeling is not negligible in the two datasets – 16.7% in PangMovie and 6.7% in VaderMovie. This relatively high percentage of human mislabeling merits further investigation in future studies, though manual labeling has been acknowledged as a complex task, very much prone to errors (Zhu et al., 2023).

Figure 6 presents the results for Question 2. Consensus cases show clear differences between the two domains. The main reason for misclassification in Amazon was "Ambivalence", with 30% of the cases, whereas "Sarcasm" is almost non-existent. This can be accounted for by the fact that in product reviews, texts tend to be more focused on features of a product, so-called *aspects*, there being less irony or sarcasm in the reviews. Most misclassifications occurred when the text concomitantly expressed both positive and negative opinions about product aspects ("Ambivalence"). This is a challenge both for the model and the human to predict the "correct polarity" for the document. This raises the question as to whether there is a single correct polarity label for these documents or whether different product aspects should be given different polarities (Brauwers and Frasincar, 2022).

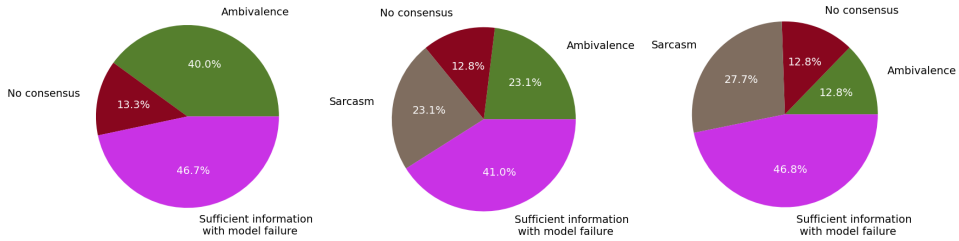
In the movie domain, we see a different result, with "Sarcasm" as the main reason for misclassification in VaderMovie and the second main one in PangMovie, almost tied with "Ambivalence". We believe sarcasm is a particular characteristic of the movie review domain, possibly due to the fact that reviewers assess artistic productions and feel the need to use figurative language to express their opinions about them. As in the Amazon dataset, "Ambivalence" is a major reason for misclassifications, especially in PangMovie. This suggests that in the movie domain, reviewers also tend to point out both positive and negative aspects, bringing a challenge both for models and humans to ascribe polarity to the texts. In this sense, *sarcasm detection* (Verma et al., 2021) and *aspect analysis* (Brauwers and Frasincar, 2022) are both interesting lines of investigation worth pursuing.

Finally, "Sufficient Information with Model Failure" is the major reason for errors in both movie datasets (36.7% in Vader and 26.7% in PangMovie) and the 2nd most frequent for products (25% of the cases) for Question 2. If we delve into the results by exploring the answers for which evaluators selected Model failure in Question 1, we can see (Figure 7) that almost half of the errors lie in this category for the three datasets. Evaluators considered there was sufficient information in the text for the model to produce the correct classification, but for some unknown reason (neither "Ambivalence" nor "Sarcasm") they could not point out (or which our



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 6: Percentages for answers to Question 2 in the three datasets.



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 7: Response analysis for Question 2 in cases where "Model Failure" was selected for Question 1.

taxonomy failed to elicit), the model got it wrong.

This can be due to several reasons, such as lack of sufficient training data and borderline cases. Although it is possible to perform this analysis in open LLMs, such as BERT, which is not the case for closed-source solutions such as GPT, it is still hard due to Transformer complexity. We will devote our attention to this challenging issue in future work. Nevertheless, to give initial insights for analyses, Figure 12 in Appendix A.7 presents the TSNE visualization of the misclassifications using the BERT-based vectors.We can observe that many misclassified documents lie on class borders, but there re cases demanding further investigation.

4.3 Final Remarks

Our misclassification assessment by qualified human evaluators prioritizes fine-grained analysis of a small representative sample of documents. Several rounds of discussions were held till a taxonomy was reached. Our study is exploratory and involves human evaluation, demanding careful manual data analysis. Evaluators had to answer 2 questions for each document in 3 datasets in 2 domains and were requested to comment on dubious cases. Each of the 8 evaluators was requested to evaluate 90 documents and compare the predominant (sentiment) model assignment to that in the gold human standard in order to decide whether misclassification was due to the model or the human and the likely reason for such misclassification. We believe this is a very complex task, which constraints sample size and number of participants, a not uncommon issue in qualitative experiments (Sharp et al., 2019).

5 Conclusion

We addressed the hard task of unveiling the reasons why models misclassified the hardest documents, those which no classifier could correctly classify, despite using the best, most separable state-of-the-art text representations. For this, we devised a taxonomy of errors and ran a qualitative experiment that requested eight evaluators with two distinct backgrounds to use our taxonomy to qualify the errors using three datasets in two domains – most prior work has been limited to a *single domain or dataset*. The high consensus among the evaluators emerged as a strong finding in our analysis. We have found significant differences regarding reasons for misclassification in the product and movie review domains. Sarcasm is very pronounced in movie reviews, while Ambivalence is more prevalent in product reviews. There is also a high proportion of human labeling errors and a noteworthy number of Model errors, which we cannot yet account for. We believe that our methodology serves to identify challenging documents based on evaluator disagreement in other domains and scenarios.

In future work, we will focus on "unexplained" cases, seeking a better understanding of the high human mislabeling rate. We will investigate sarcasm detection and aspect analysis as preprocessing steps as well as apply multi-perspective analysis to our study (Frenda et al., 2023).

Limitations

Despite relevant contributions, our study has some limitations. Our evaluation targeted two domains, three datasets and the task of sentiment analysis. Increasing the number of dataset domains and expanding our analysis to the task of Topic Classification will provide new valuable insights. The size of our evaluation group is relatively small, although this is common in qualitative studies (Silverman, 2004). We will increase the number of evaluators in future studies. Our work uses BERT's contextual representations. Although (de Andrade et al., 2023) shows BERT produces representations that are as (semantically) separable in the embedding space as representations produced by other Transformers (e.g., RoBERTa, BART), we intend to test our methodology with different Transformers in the future.

References

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. 55(4).

Claudio M.V. de Andrade, Fabiano M. Belém, Washington Cunha, Celso França, Felipe Viegas, Leonardo Rocha, and Marcos André Gonçalves. 2023. On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *Information Processing & Management*, 60(4):103336.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13844–13857. Association for Computational Linguistics.

Lukas Galke and Ansgar Scherp. 2022. Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4038–4051, Dublin, Ireland. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William D. Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Han S. Lee, Alex A. Agarwal, and Junmo Kim. 2017. Why do deep neural networks still not recognize these images?: A qualitative analysis on failure cases of imagenet classification. *CoRR*, abs/1709.03439.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. 13(2).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Karen Martins, Pedro O.S Vaz-de Melo, and Rodrygo Santos. 2021. Why do document-level polarity classifiers fail? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1782–1794, Online. Association for Computational Linguistics.

Christopher Meek. 2016. A characterization of prediction errors. *CoRR*, abs/1611.05955.

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. page 271–es.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, pages 17044–17056. Curran Associates, Inc.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5:1–29.

H. Sharp, J. Preece, and Y. Rogers. 2019. *Interaction Design: Beyond Human-Computer Interaction*. Wiley.

D. Silverman. 2004. *Qualitative Research: Theory, Method and Practice*. SAGE Publications.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Palak Verma, Neha Shukla, and A.P. Shukla. 2021. Techniques of sarcasm detection: A review. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 968–972.

Ming Yan, Haiyang Xu, Chenliang Li, Junfeng Tian, Bin Bi, Wei Wang, Xianzhe Xu, Ji Zhang, Songfang Huang, Fei Huang, et al. 2023. Achieving human parity on visual question answering. *ACM Transactions on Information Systems*, 41(3):1–40.

Yu Zhu, Yingchun Ye, Mengyang Li, Ji Zhang, and Ou Wu. 2023. Investigating annotation noise for named entity recognition. *Neural Comput. Appl.*, 35(1):993–1007.

A Appendix

A.1 Summary of Evaluation Schemas reported in Related Work

Table 5 shows a summary of the evaluation schemas reported in related work. Compared to them, our schema is much more robust and comprehensive.

A.2 Consensus Distribution

This subsection presents evaluator consensus distribution for Questions 1 and 2, analyzed in Section 4.1. Regarding Question 1, as can be seen in Figure 8a, out of the 52 documents that achieved evaluator consensus in the Amazon dataset, 33 reached full agreement among all four evaluators, 16 documents reached full agreement among three, and 3 documents reached full agreement between two evaluators. This points to documents with full agreement among three or four evaluators representing a significant portion of the total number of documents with consensus, in turn demonstrating the robustness of our final results. Similar results were obtained for VaderMovie and PangMovie regarding the joint proportion (i.e., the sum of the proportions) of evaluations with 4 and 3 agreements.

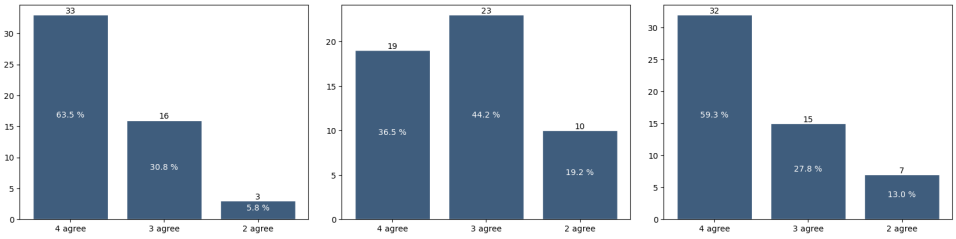
Regarding Question 2, results show less consensus among the evaluators, which may be due to the number of categories they had to choose from. This is reflected in the graphs in Figure 9. The Amazon dataset showed higher consensus among a higher number of evaluators, possibly accounted for by the type of review - product review. As movie reviews assess artistic productions and implicate more sarcasm and figurative language, the full consensus is harder to achieve, though still attainable.

Regarding documents for which there was no consensus among the evaluators (Figure 4 - Left), there are 8 for the Amazon dataset, 8 for the PangMovie dataset and 6 for the VaderMovie dataset. As for question 2 (Figure 4 - Right), there are 11, 9, and 9 documents without consensus for Amazon, Pang Movie, and Vader Movie datasets, respectively. To exemplify challenging documents, we provide three examples from each dataset in the "No consensus" category for Question 2, as shown in Table 6.

The first row in Table 6 shows an Amazon product review where the text begins positively but then brings in an issue with the product. Row 4 shows a movie review from the Pang Movie dataset, where the reviewer uses the words "distended" and "dragging", creating uncertainty for categorization. Row 6 shows a series of references to other movies and directors, which requires previous knowledge of those movies and their evaluations. Therefore, we believe that the methodology of this study serves to identify challenging documents based on evaluator disagreement.

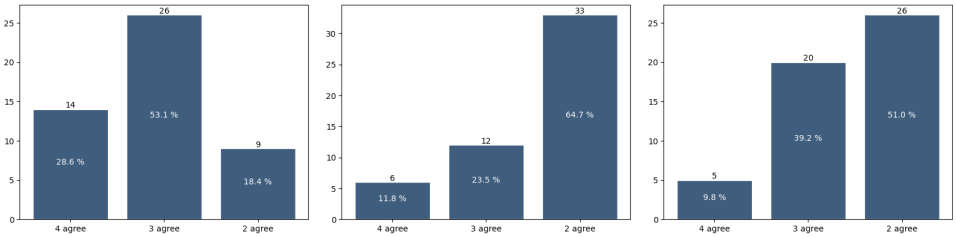
Related Work	Taxonomy categories
(Meek, 2016)	"**Mislabeling errors**": human labeling errors; "**Representation errors**": limitations in the feature set used for evaluation; "**Learner errors''**": prediction errors when there is sufficient information for accurate classification; "**Boundary errors**": correct predictions could be achieved by adding more examples, indicating absence of labeled examples for a specific class in the training set.
(Lee et al., 2017)	"**Similar Labels**": the term representing the predicted object in the image is not in the ground truth (GT) but is semantically similar to the GT. The set of true labels is the set of terms that textually describe the objects in the image. "**Not Salient**": the predicted object exists in the image but is not present in the GT; "**Challenging Images**": the GT is challenging even for a human being; "**Incorrect GT**": incorrect annotation by humans; and 5) "incorrect prediction class": machine prediction is incorrect but with sufficient information in the image for humans to detect.
(Martins et al., 2021)	"**Neutral**": when polarity is not clearly defined "**Discrepant**": when polarity differs from its associated labeling

Table 5: Summary of Evaluation Schemas Reported in Related Work.



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 8: Consensus for Question 1.



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 9: Consensus for Question 2.

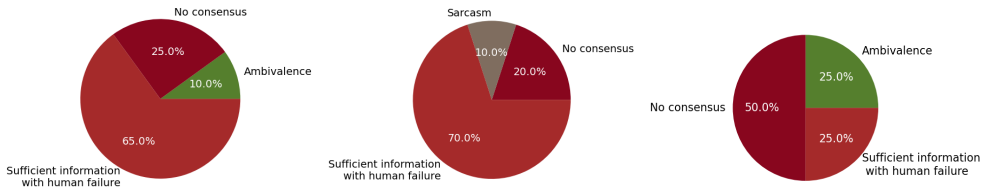
Text	Dataset
They are ok except. The fitted pops off.	Amazon
I tried a few LED harnesses and none were bright enough to see my black dog at night running through the woods. This vest, as long as not directly in front of head/tail, is super visable.	Amazon
Very short on the sides. Overall, good fit but I do not like to show my belly. Too bad lad got that. Fabric very soft.	Amazon
The script kicks in, and mr. hartley's distended pace and foot-dragging rhythms follow.	Pang Movie
Eastwood winces, clutches his chest and gasps for breath. it's a spectacular performance - ahem, we hope it's only acting.	Pang Movie
Parts seem like they were lifted from terry gilliam's subconscious , pressed through kafka's meat grinder and into buñuel's casings	Pang Movie
The recording session is the only part of the film that is enlightening and how appreciative you are of this depends on your level of fandom.	Vader Movie
It shows that some studios firmly believe that people have lost the ability to think and will forgive any shoddy product as long as there's a little girl on girl action.	Vader Movie
A light, engaging comedy that fumbles away almost all of its accumulated enjoyment with a crucial third act miscalculation.	Vader Movie

Table 6: Texts illustrating the "No consensus" category

A.3 Inter-evaluator agreement for Question 2 in cases of "Human Mislabeling"

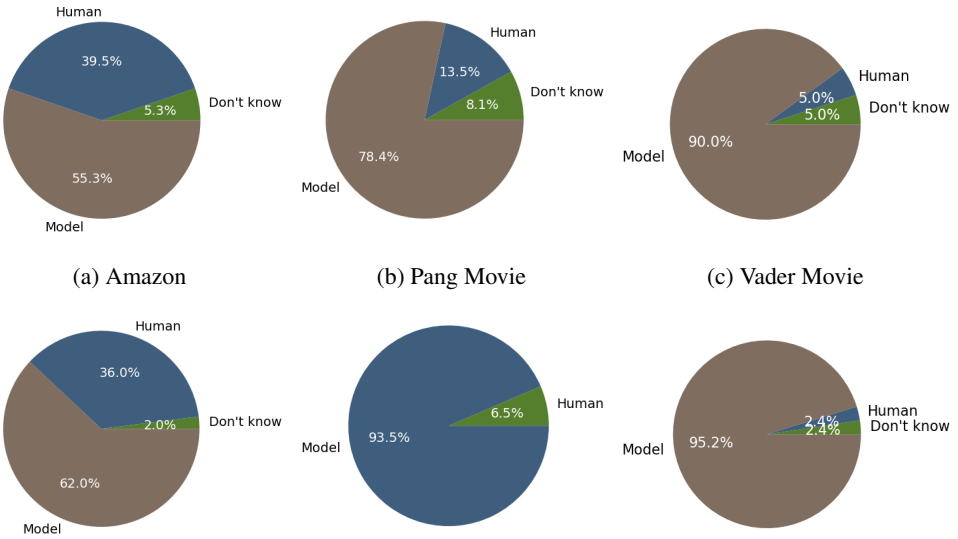
In Figure 10, similar to Figure 7, we have the quantification of Question 2, but now restricted to the documents that were evaluated as human mislabeling in Question 1. In other words, documents the evaluator considered to have been correctly classified by the Model but which had been incorrectly labeled by the human (positive or negative). We can observe that, in general, the number is lower; for instance, in the Amazon dataset, we have 20 documents evaluated as mislabeled by humans.

Additionally, we can observe a high prevalence of the category human mislabeling with sufficient



(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 10: Results for Question 2 in cases where Response to Question 1 was "Human' Failure'.



(a) Amazon (b) Pang Movie (c) Vader Movie

(d) Amazon (e) Pang Movie (f) Vader Movie

Figure 11: Percentages for answers to Question 1 by evaluators with a Computer Science background (a, b and c) and a Linguistics background (d, e and f).

information in the text, which corresponds to 65% in the Amazon and 70% in the Pang Movie datasets. This means that the evaluator considered the document to have been mislabeled by the human, despite there being sufficient information in the text for the human to choose the "right" label according to the evaluator's assessment.

Regarding the VaderMovie dataset, numbers are low, which may bias some proportions – there are only four mislabeled documents evaluated as human mislabeling, and only 1 sample was considered to have been mislabeled by the human, despite there being sufficient information in the text for the human to choose the 'right" label.

A.4 Differences in Evaluation carried out by Computer Scientists and Linguists

We carried out an additional analysis focusing on the evaluators' backgrounds. Since each document was rated by two evaluators with a Linguistics background and two with a Computer Science one, we examined our data to investigate differences ascribable to evaluators' backgrounds. Figure 11 represents the quantification of the responses to question 1 by evaluators having a Computer Science background (11a, 11b, 11c) and a Linguistics one (11d, 11e, 11f), in which there was inter-evaluator agreement of the two evaluators. We can notice that evaluators' backgrounds had little impact on the results for all datasets.

A.5 Comparison between BERT and the Classifiers using the Contextual Representations (from de Andrade et al. (2023))

For the sake of self-containedness, in Table 7, we show the results reported by (de Andrade et al., 2023) for the comparison between BERT and classifiers that used the textual representations generated by the Transformer as input. Here, we consider the results of four of the strongest classifiers used in (de Andrade et al., 2023), namely: KNN, Random

Forests (RFs), Support Vector Machines (SVMs), and Logistic regression (LR) applied to two of the datasets we exploit – PangMovie and VaderMovie. Indeed, despite using different rules and heuristics, the effectiveness of these classifiers (and of all other classifiers tested in (de Andrade et al., 2023)) is basically the same in all tested datasets when using the contextual embedding representations. This is due to the fact that these representations are already so semantically separated (by class) in the embedding space that the employed classifier has little effect on the classification process.

Dataset	BERT	RF	SVM	KNN	LR
Amazon	**94.2(0.7)**	**94.2(0.1)**	**94.1(0.1)**	**94.3(0.1)**	**94.2(0.1)**
PangMovie	**87.0(0.6)**	**86.8(0.8)**	**87.2(1.0)**	**87.1(0.6)**	**87.1(0.8)**
VaderMovie	**89.1(0.7)**	**89.4(0.6)**	**89.4(0.5)**	**89.3(0.7)**	**89.5(0.6)**

Table 7: Macro-F1 (%) and confidence interval of 95%. Best results (including statistical ties) are marked in **bold**. BERT is the original method while the other columns correspond to the respective classifiers run using the contextual embeddings produced by BERT.

A.6 Comparison Among Transformers

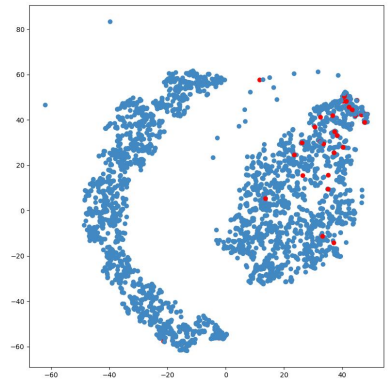
We run experiments in the tested datasets comparing BERT with RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020). Results are shown in Table 8. As we can see, these transformers' effectiveness are very similar – BERT is statistically tied as the best method with Roberta in Amazon and marginally loses (by at most 1-2 pp) in the other two datasets.. These differences, which means just a few documents in practice, are potentially irrelevant in a qualitative study as ours, which uses a sample of the documents that all classifiers predicted incorrectly. We believe that the intuitions and insights we got with the current methodology, representations, and models would not be substantially different if we used other Transformers.

Dataset	BART	BERT	RoBERTa
Amazon	93.0 (0.2)	94.2 (0.7)	94.5 (0.3)
PangMovie	88.1(0.5)	87.0(0.6)	89.0(0.4)
VaderMovie	90.4(0.6)	89.1(0.7)	91.3(0.5)

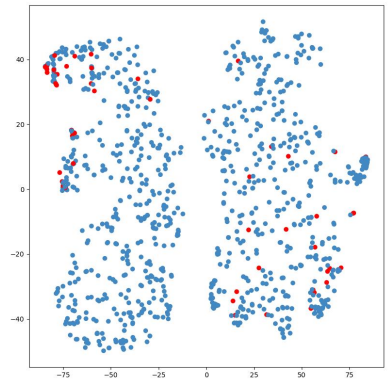
Table 8: Results regarding the evaluation metric Macro-F1.

A.7 TSNE Visualization of the Errors in the Dataset

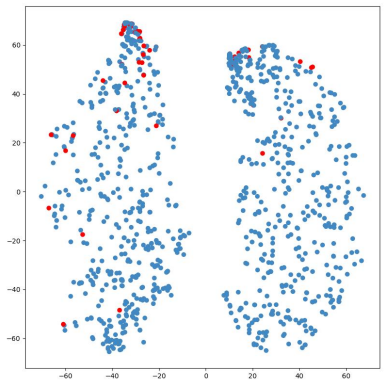
Figure 12 presents the TSNE visualization of the documents in the analyzed datasets using the BERT-based vectors. We marked in red the misclassified documents. We can see that many misclassified documents lie on the class borders, but there are other cases demanding further investigation.



(a) Amazon



(b) Pang Movie



(c) Vader Movie

Figure 12: TSNE three datasets. In red, it is the set of documents misclassification by all classifiers used in this study.