Explaining the Hardest Errors of Contextual Embedding Based Classifiers

Claudio M. V. de Andrade[1], **Washington Cunha**[1], **Guilherme Fonseca**[2]
Ana Clara S. Pagano[1], **Luana de C. Santos**[1], **Adriana S. Pagano**[1]
Leonardo Rocha[2], **Marcos André Gonçalves**[1]
claudio.valiense@dcc.ufmg.br, washingtoncunha@dcc.ufmg.br,
guilhermefonseca8426@aluno.ufsj.edu.br, anapagano@ufmg.br, lcs2017@ufmg.br
apagano@ufmg.br, lcrocha@ufsj.edu.br, mgoncalv@dcc.ufmg.br
[1] Federal University of Minas Gerais, Brazil
[2] Federal University of São João Del-Rei, Brazil

Abstract

We seek to explain the causes for the misclassification of the most challenging documents, namely those that no classifier using state-of-the-art, very semantically-separable contextual embedding representations managed to predict accurately. To do so, we propose a 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. Evaluators with two different backgrounds evaluated documents by comparing the predominant sentiment assigned by the model and the label in the gold dataset in order to decide on a likely misclassification reason. Based on a high inter-evaluator agreement (81.7%.), we observed significant differences between domains, such as the prevalence of ambivalence in product reviews and sarcasm in movie reviews. Our analysis also revealed an unexpectedly high rate of incorrect labeling in the gold dataset (up to 33%) and a significant amount of incorrect prediction by the model due to a series of linguistic phenomena (including amplified words, contrastive markers, comparative sentences, and references to world knowledge). Overall, our taxonomy and methodology allow us to explain between 80%-85% of the errors with high confidence (agreement) – enabling us to point out where future efforts to improve models should be concentrated.

1 Introduction

In a scenario where the amount of user-generated content is growing exponentially, automatic text classification (ATC) plays a vital role in enabling 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 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, was able to classify correctly.

A thorough review evidenced that this 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 predictions *upon classifiers application*.

We propose and evaluate an *error taxonomy* using a document sample for which no classifier

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 419–434

was able to produce correct predictions. Due to the very complex nature of the error analysis task, we adopt BERT to generate **contextual document representations**[1]. We evaluated 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 in explaining the errors.

Unlike previous work (Martins et al., 2021) – which focuses on assessing the impact of "hard" instances on the effectiveness of polarity detection using a single dataset (movie reviews) and not concerned with textual representation – here we 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 used datasets from two domains: movie and product reviews. We also compare and contrast 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 agreement and correlate that with hardness in classifying; and [RQ2] *Can the proposed taxonomy be used to reveal the main reasons for misclassification? Are there significant differences in the results between different domains?* In RQ2, drawing on the consensus achieved, we quantify and analyze the main reasons for the misclassification, highlighting the 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) between 50%-80% of the errors can be ascribed to the model for reasons further explained below; (iii) the evaluators found a significant amount of *incorrect labels in the dataset* –, i.e., there were incorrect labels in the gold datasets – 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 considered a major reason for incorrect prediction by the model, while (v) in product reviews, the main reason is ambivalence (40% of the cases) – we believe this is a particular characteristic of this domain.

The remaining *model* errors were considered instances of *"incorrect prediction despite available textual cues"* ascribable to a series of language phenomena, including amplified words, contrastive markers, comparative sentences, and world-knowledge regarding named entities. While for product review, the model errors are mostly associated with comparisons and contrastive cases, for movie scenarios, world knowledge, use of amplifiers and idiomatic/new expressions are issues in the model's incorrect predictions. Our results can be potentially leveraged for model enhancement focused on the application domain.

In sum, our main contributions include: (i) the development and evaluation of a *taxonomy* for categorizing the main causes for misclassification of the *hardest* documents; ii) a *fine-grained analysis* of the results of a comprehensive qualitative experiment applying the taxonomy to 3 different datasets in 2 different domains, with relevant implications for the improvement of the next generation of textual classifiers and representations; and (iii) a release of a *new dataset* of challenging documents manually annotated by humans.

2 Related Work

In Lee et al. (2017), five categories for misclassification of objects in images are explored (See Appendix A.3 for Evaluation Schema). Meek (2016) categorized prediction failures in textual documents by defining four error categories (see Appendix A.3 for schema), focusing on the lack of training information. Pandey et al. (2022) assesses the impact on labeling of (i) time allocated to evaluators; and (ii) the order of annotations in the labeling task. Unlike these works, we propose a taxonomy for ATC test errors and investigate a more comprehensive set of reasons, focusing on the hardest cases for classifiers using state-of-the-art,

[1]We ran 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.

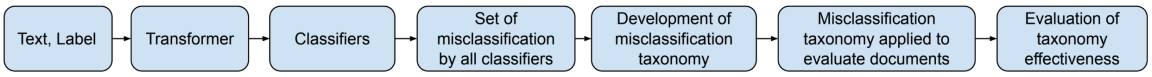


Figure 1: Methodological steps flowchart.

very separable (contextual) representations by means of qualitative human assessment.

Bras et al. (2020) remove bias in training to reduce misclassification. Pleiss et al. (2020) propose adapting the Area Under the Margin to identify training data that preclude generalization. Both focus on the training set to identify (challenging) documents that do not contribute to the learning process. Instead, we focus on misclassification at inference time (test set), aiming to identify common characteristics of misclassified documents.

Swayamdipta et al. (2020) present a tool to characterize and diagnose datasets regarding the behavior of the model on individual instances during training. Ethayarajh et al. (2022) seek to find challenging documents using V-usable information. Differently, we find challenging documents based on their incorrect classification by four classifiers using a very separable contextualized representation 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.

Martins et al. (2021) analyze a set of hard instances (evaluation schema in Appendix A.3) but, unlike ours, their study centers on evaluating the influence of challenging cases on the classifier's effectiveness when performing polarity detection using **one single movie review dataset**. Our study focuses on analyzing and quantifying the factors contributing to the misclassification of the hardest documents *using multiple datasets in two domains* with a more detailed taxonomy. We also have additional goals such as validating our taxonomy and contrasting the results in multiple different datasets and domains, running qualitative experiments engaging evaluators with different backgrounds.

Barnes et al. (2019) propose categories to understand model misclassifications. Unlike ours, their study: (i) did not focus on the hardest cases; (ii) did not detail how the data was evaluated; (iii) did not provide information on inter-rater agreement; and (iv) did not examine domain impact on results – all results for all datasets are analyzed in conjunction. We drew on their taxonomy, though, to develop the categories we used for focused (hierarchical) evaluation, as described in section 3.8.

3 Experimental Methodology

Our methodology, which comprises 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 was able to produce correct predictions (according to the labels assigned 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 will apply to evaluate documents in a second sample different from the first one ("Application of the misclassification taxonomy to evaluate documents" in Figure 1). Upon applying the taxonomy, we quantify the results and evaluate its efficiency. A detailed account of the steps follows.

3.1 Datasets

Our study draws on 3 datasets developed for binary sentiment classification. Although this task is considered less complex than, for instance, multi-label topic classification, our choice was strategically purposeful due to the very complex endeavor we made in our work to identify the potential reasons for misclassifications of the hardest cases - those that no classifier is able to predict correctly using state-of-the-art representations. Thus, even though current solutions for sentiment analysis are highly effective, with some solutions achieving F1≅90, one of our main goals is precisely to evaluate the current technologies' limits. With such high effectiveness, what are the reasons for the few errors still made by the very effective sentiment classifiers? Answers to this question, which our methodology helps clarify by pointing out and quantifying the main sources of misclassifications, are what we believe will provide necessary grounds for the improvements of the next generation of methods.

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. These datasets compile movie reviews comprising a text and a sentiment label (positive or negative). Table 3 in the Appendix presents some statistics of the datasets. As it can seen, class distribution into positive and negative instances is balanced in the three datasets.

3.2 Data Representation

We fine-tuned BERT, adapting this Transformer 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 (Cunha et al., 2021), 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. The 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 in terms of 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). The results are shown in Appendix A.8. As we can see, the effectiveness of these transformers is 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[3]. These differences, which mean just a few documents in practice, are potentially not relevant in a qualitative study as ours, which uses a sample of the documents that all classifiers predicted incorrectly. We believe the intuitions and insights gathered with the current methodology, representations, and models would not be substantially different if we used other Transformers[4].

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), as well as BERT model with the classification head as one of the classifiers. 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.8.

We decided 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 6. 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

[3]Indeed, some benchmarks such as GLUE do not make clear even if recent LLMs are better than RoBERTa, a remarkable sentiment classifier, see a discussion in the Appendix.

[4]We will evaluate different pre-trained representations in future studies to find out if the same error type, in similar proportions, occurs across different representations.

evaluation, as described next. Table 6 in the Appendix 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.

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. Drawing on the literature, we assumed that there could be incorrect labels in the gold datasets and hence decided to include human mislabelling as a potential reason for the mismatch between the model's prediction and our ground truth. Through this process, we agreed upon a set of potential reasons representing the bulk of the categories in our taxonomy of errors. We conducted 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 8 participants, 4 with expertise in Computer Science and 4 in Linguistics, all with prior experience in NLP annotation tasks. The participants comprised two professors holding a PhD in CS, one with a PhD in Linguistics, and five students pursuing their bachelor's or master's degrees, who performed the work voluntarily out of curiosity and with learning goals.

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 agreement within the same background groups and between the two groups with different backgrounds. Section 4 presents results considering all four evaluators – the impact of evaluators' background is analyzed in Appendix A.6.

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.

The form provided to evaluators presents 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"; and (ii) "Based on your answer to question 1, "why do you think the text was misclassified?", for which 1 out of 6 options could be chosen. Table 1 provides a description of the available options.

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 1. The first row shows an example of a text misclassified due to the model's incorrect prediction despite available textual cues. In this case, the model assigned a negative sentiment, though the text contains positive cues: "precious increments artfully.....". The second row shows an example of misclassification due to an incorrect label in the dataset. 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*, having both negative ("expensive") and positive cues ("won't oxidize" and "better than soap").

It should be noted that 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

Category	Description and Example
Sarcasm	**Description**: Text contains irony (words that express the opposite of what one means), humorous expressions, and figurative language (metaphors) **Example**: Final verdict: you've seen it all before.
Ambivalence	**Description**: Text contains both positive and negative opinions, neither being predominant over the other **Example**: Expensive but won't oxidize metal. Maybe better than soap
Lack of textual cues for label prediction	**Description**: Text is very brief or provides no cues for a human and a model to assign a predominant sentiment **Example**: Big biggg large shoes as expected and loose fitting
Incorrect prediction despite available textual cues	**Description**: Text provides textual cues but model fails to correctly assign the predominant sentiment **Example**: A film of precious increments artfully camouflaged as everyday activities
Incorrect label in the dataset	**Description**: Text has an incorrect gold label in original dataset **Example**: Rope completely broke off after a couple of months (positive in the gold standard)
None of the above	**Description**: None of the above categories can account for the misclassification

Table 1: First-level categories and their description with examples

Category	Description and Example
Amplifier	**Description**: Words such "really", "very", "super", "incredibly", "so", "pretty", "definitely", "too" tend to co-occur with instances of very negative or very positive sentiment and can be interpreted by the model as conveying a sentiment contrary to what they actually amplify. **Example**: Secretary is just too original to be ignored.
Comparative	**Description**: Comparisons ("more", "less", "higher", "lower", etc.) establish a relationship of inequality between two elements, requiring the model to interpret which of the two is being evaluated as positive or negative. **Example**: LaBute was more fun when his characters were torturing each other psychologically and talking about their genitals in public.
Contrastive	**Description**: Two distinct sentiments are expressed and explicitly signaled by conjunctions ("but", "yet", "on the other hand", "however", "yet", "still", "though", "despite this", "all the same"), one sentiment being dominant over the other. **Example**: Uneven, self-conscious but often hilarious spoof.
Idiom	**Description**: Meaning cannot be inferred from the meaning of each individual word in an expression. **Example**: A pleasurably jacked-up piece of action moviemaking.
Modality	**Description**: Modal expressions such as may, could, should, must, can, might, etc. imply that something is other than expected or desired. **Example**: Cattaneo should have followed the runaway success of his first film, the full monty, with something different.
Negation	**Description**: Polarity and negative markers (no, not, never, neither, etc) as well as negative words may be used in texts with positive sentiment. **Example**: Can't turn off the unit the fast charger work perfect.
Non-standard spelling	**Description**: Symbols such as #, words written together instead of apart, use of all caps, etc., may not be recognized as words by the model. **Example**: Much monkeyfun for all.
Reducer	**Description**: Reducers such as "kind of", "less", "lot less", "sort of", "so so", "about", "more or less", may shift classification towards a particular sentiment. **Example**: A subtle variation on i spit on your grave in which our purported heroine pathologically avenges a hatred for men.
World knowledge	**Description**: Facts, events, people, characters, etc., associated to positive and negative sentiment. **Example**: Granddad of Le Nouvelle Vague, Jean Luc Godard continues to baffle the faithful with his games of hide and seek.
New word / expression	**Description**: Newly-coined, mostly hyphenated words and expressions that may not be recognized by the model. **Example**: Even in this less-than-magic kingdom, reese rules.
None of the above	**Description**: None of the above categories can account for the misclassification

Table 2: Categories for fine-grained analysis of "Incorrect prediction despite available textual cues"

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.

3.8 Focused (Hierarchical) Categorization

The final step in our methodology comprises further evaluation of some of the "most complex errors", namely, those identified in the previous step as being *incorrect prediction despite available textual cues* could have led to assigning the correct sentiment. In this final analysis, we aim to identify reasons for those incorrect predictions. We opted for an increasingly focused evaluation process in order to manage the complexity of the annotation task, cognizant of the effort required by assessing documents with increasingly fine-grained categories. Hence, our methodology moved from a general, binary query (Question 1) to a more distilled, six-category query (Question 2), concluding with a ten-category query (Focused categorization).

In this last assessment round, all instances of *incorrect prediction despite available textual cues*

were evaluated using a fine-grained category set pertaining to linguistic phenomena reportedly not adequately captured by models. We designed a taxonomy based on ten particular linguistic phenomena potentially impacting a model's predictions. They cover words modifying the sentiment intensity (Amplifiers and Reducers); explicitly signaled comparisons which require identifying which of the two elements is decisive for a sentiment (Comparatives); explicitly contrasted arguments or aspects (Contrastive), with one of them being dominant; idiomatic expressions (Idiom); expressions of probability and obligation (Modality); negative polarity scope and negative words (Negation); symbols and characters rendering unrecognizable words (Non-standard spelling); newly-coined and idiosyncratic words unknown to the model (New word / Expression); and mentions to entities requiring world-knowledge to assign a correct sentiment (World-Knowledge). These categories are detailed described in Table 2 , along with the instructions provided to the evaluators, with a definition of each category and examples.

4 Results

Documents were assessed by 4 evaluators. Question 1 required selecting 1 out of 3 alternatives, whereas Question 2 had 6 alternatives. Focused categorization comprised 10 categories. Consensus was defined as one of the alternatives having the *majority* of votes – 4, 3, or 2 [5]. If there was no majority, a document was classified as "No consensus".

4.1 Taxonomy effectiveness

High consensus was achieved for the three levels of assessment: 86.7% for Question 1; 81.2% for Question 2 and 86.5% for focused assessment, allowing us to state that the taxonomy was effective for evaluation purposes[6]. We present a detailed effectiveness (consensus) analysis in the Appendix A.4.

4.2 Response Analysis

Given that a high consensus had been achieved, we proceeded to analyze the responses of the evaluators. Half of the misclassifications in the Amazon dataset were ascribed to the model (see Figure 5 in Appendix). 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" were very low in all datasets. Together with the option "No consensus", they achieved at most 18.3% in PangMovie (and 16.3% and 15% in Amazon and VaderMovie, respectively) of all analyzed documents in all datasets.

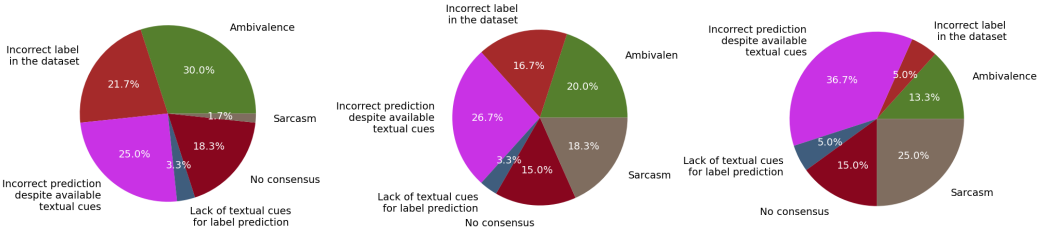
Though lower than errors ascribed to the model, the percentage of errors ascribed to the "Human" category is significant, mainly in the Amazon dataset (33%) (See Appendix A.5). This means that in 33% of the misclassifications, 3 or 4 evaluators (majority of the cases) considered that the model classified the document correctly and there was an error in the gold dataset. Though lower in the movie domain, human mislabeling is not negligible – 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 and prone to errors (Zhu et al., 2023).

Figure 2 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).

We see a different result in the movie domain, 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

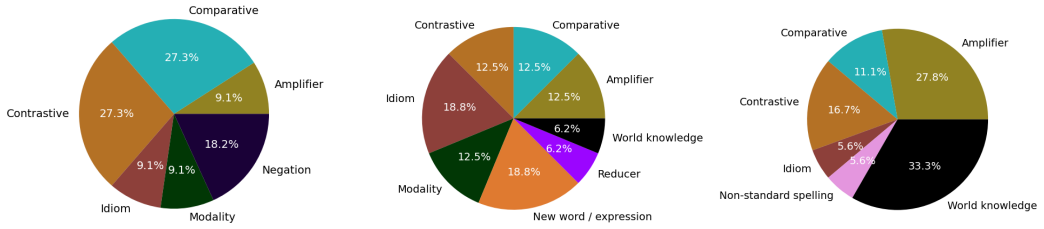
[5] In case of two votes, provided that the remaining two alternatives have one vote each.

[6] An effective taxonomy has high consensus among evaluators upon the defined categories and 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



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

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



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

Figure 3: Percentages for answers when breaking down the category "Incorrect prediction despite available textual cues".

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.

4.2.1 Focused (Hierarchical) Analysis

A major reason for errors in both movie datasets (36.7% in Vader and 26.7% in PangMovie) and the second most frequent for products (25% of the cases) for Question 2 was "Incorrect Prediction Despite Available Textual Cues" (Figure 2). Indeed, if we look at the reasons why evaluators selected Model failure in Question 1 (Figure 10 in the Appendix), almost half of the errors are ascribed to this category for the three datasets. Evaluators considered textual cues were available to predict the correct sentiment, but for reasons other than "Ambivalence" or "Sarcasm", the model failed to do it.

The final step in our methodology was devoted precisely to understanding the reasons for those errors. In a new round of assessment, we evaluated 52 documents that had been assigned this category in the first round: 14 in Amazon, 16 in Pang Movie and 22 in VaderMovie. Like the first round, we also obtained a high overall percentage of agreement– 86.5% – which can be considered quite high considering that (i) there are more categories to assign (10 in total) and (ii) these are some of the hardest cases to evaluate.

Figure 3 shows the results of this final focused (hierarchical) analysis. As we can see, in Amazon, 54.6% of the model errors are due to explicit comparisons and contrastive cases where one aspect is dominant over the other. This is expected as these are product reviews. Negation (e.g., "Can't turn off the unit the fast charger work perfect." in Table 2) is also a major reason for errors. The remainder of the errors are roughly evenly spread over the categories related to idiomatic expressions, modality (e.g., "Cattaneo should have followed the runaway success of his first film with something different." in Table 2) and errors due to amplifiers.

The case is more complex in the Movie domain, where the errors evidence a different pattern. In the Vader dataset, lack of world knowledge (e.g. a movie name, a director/actor, a real-world event (e.g., "Granddad of Le Nouvelle Vague, Jean Luc Godard continues to baffle the faithful with his games of hide and seek." in Table 2) accounts for $\frac{1}{3}$ of the errors, followed by amplifiers, which are popular among movie reviewers. In the PangMovie dataset, we see a more complex, almost even distribution of errors among all categories with a high impact (37.6% of the cases) of idiomatic expressions (e.g. "A pleasurably jacked-up piece of action moviemaking." in Table 2) and newly coined words/expressions, also popular among movie reviewers, which may occur in a single or just a few documents and do not have enough support in the training data for the model to learn properly.

The few errors that remain unexplained may be due to distinct reasons, such as lack of training data and borderline cases. Although it is possible to perform this analysis in open models such as BERT, which is not the case for closed-source solutions such as GPT, it is 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 (Appendix A.11) presents the TSNE visualization of the misclassified BERT-based document vectors – many of them lie on class borders.

In this work, we investigated the reasons for misclassification, highlighting the issues found and enabling the implementation of strategies to address these problems. For example, if an instance is found to be wrongly classified due to sarcasm, this implies that before the actual classification, the sentiment classifier should be given information that this message is possibly sarcastic (using, for instance, a sarcasm/irony classifier) so that the sentiment classifier can use this information in the decision process. If a document is found to be ambivalent, segments with polarity clash should be located and assigned a separate label for each polarity, or the full sentence be assigned the polarity of the stronger sentiment. If a sentence has two polarities and there is an overt contrasting connector, polarity inversion may be performed, as it is done in Vader's shell. If an instance is incorrectly classified due to idiomatic expressions, lack of world knowledge, or the occurrence of newly coined words, the solution involves enhancing the model with further training instances that provide the missing knowledge, including idioms and new expressions. Similar strategies can be developed regarding the other categories.

As a **final remark**, we would like to emphasize the *complexity* of the performed analysis. Our misclassification assessment prioritizes fine-grained analysis of a 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. The focused (hierarchical) categorization required yet, a new round of evaluation considering 10 linguistic categories. 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. This very complex task constrains sample size and number of participants, a not uncommon issue in qualitative experiments (Sharp et al., 2019) and justifies our current choice of a single task - - sentiment analysis.

5 Conclusion

We addressed the hard task of unveiling the reasons why models misclassified the hardest documents, those which no classifier using very separable contextual representations could correctly classify. For this, we devised an error taxonomy and ran qualitative experiments requesting 8 evaluators with distinct backgrounds to use the taxonomy to qualify the errors using 3 datasets in 2 domains – prior work has been limited to a *single domain or dataset*. The high consensus among the evaluators emerged as an interesting finding. 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 a high proportion of wrong labels in the gold dataset and a noteworthy number of incorrect model predictions due to various linguistic phenomena, including comparisons, contrastive constructions, negation and instances requiring world knowledge. No single category emerged as dominant.

Future work includes explaining the few remaining unexplained cases; applying our methodology/taxonomy to other domains (e.g., topic classification); and using acquired knowledge to improve models. Additionally, we intend to investigate current models, such as Large Language Models (LLMs), in classification tasks, assessing the potential of these models to address the issue of misclassification presented in this paper.

Acknowledgements

This work was partially supported by CNPq, CAPES, FAPEMIG, AWS, UNIMED, NVIDIA, CIIA-Saúde, and FAPESP.

References

Jun Bai, Xiaofeng Zhang, Chen Li, Hanhua Hong, Xi Xu, Chenghua Lin, and Wenge Rong. 2023. How to determine the most powerful pre-trained language model without brute force fine-tuning? an empirical survey. In *Findings of the EMNLP 2023*.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. ACL.

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*. PMLR.

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

Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023a. An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *Proceedings of the 46th International ACM SIGIR'23*.

Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, et al. 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481.

Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2023b. A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM CSUR*.

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*, 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*.

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*, pages 4038–4051.

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*.

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*.

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)*.

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 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.

While our current work covers only one classification task, in our study, we devise a general-purpose taxonomy for text classification designed to be useful in more than one scenario. Our first question aims to answer whether the source of the misclassification is human or the model—a question that applies to any ATC task where we have a label and a model's prediction. Our second question inquires about the reason for the misclassification - Incorrect Prediction Despite Available Textual Cues; or incorrect label in the dataset - lack of textual cues for label prediction, ambivalence, and sarcasm. Likewise, the first two categories are not restricted to the sentiment analysis task but apply to other ATC tasks. At the first level of the proposed taxonomy, two categories (ambivalence and sarcasm) can be said to be task-related, but the taxonomy needs them for analytical purposes; otherwise, it would be too general. Nonetheless, if used for evaluation in other tasks, these two more task-oriented categories may be adapted, the core of the taxonomy remaining as it is.

Our spreadsheet validation only allowed annotators to choose a single category to answer each question. A column for annotators to freely state their Remarks was available in case the categories should present any annotation problem. No remarks were placed by annotators, which suggests no overlapping was felt by them. While theoretically some of the categories could be felt to overlap, our results did not support this.

A.2 Datasets Statistics

Table 3 presents statistics of the datasets in terms of the number of documents and average document length, and the class distribution into positive and negative instances is balanced in the three datasets.

Dataset	Documents	Avg words	Positive	Negative
Amazon	168000	33	84000	84000
PangMovie	10662	19	5331	5331
VaderMovie	10568	19	5242	5326

Table 3: Datasets Statistics

A.3 Summary of Evaluation Schemas Reported in Related Work

Table 4 shows a summary of the evaluation schemas reported in related work. Compared to them, our schema is much more robust and comprehensive.

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 an 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 4: Summary of Evaluation Schemas Reported in Related Work.

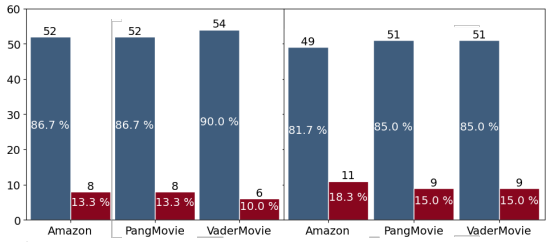


Figure 4: Consensus and No Consensus on Q1 (left) and Q2 (right).

A.4 Taxonomy Effectiveness (Consensus) Analysis

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.

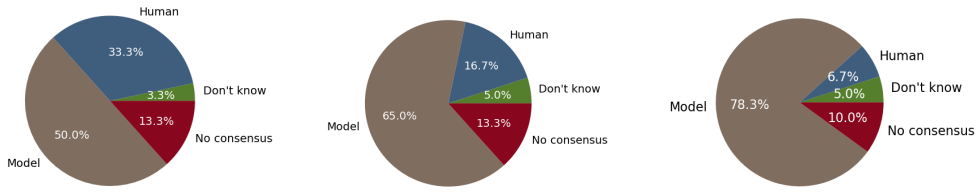
Figure 4 shows the consensus percentages obtained for Questions 1 and 2 in the three evaluated datasets. For Question 1, out of 60 documents, 54 attained high inter-evaluator agreement in VaderMovie, and 52 in Amazon and PangMovie. In other words, in at least 86.7% of the cases (52/60), consensus was 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 Section A.4.1 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 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 Section A.4.1, we show examples (in Table 5) of documents that posed difficulties for evaluators.

A.4.1 Consensus Distribution

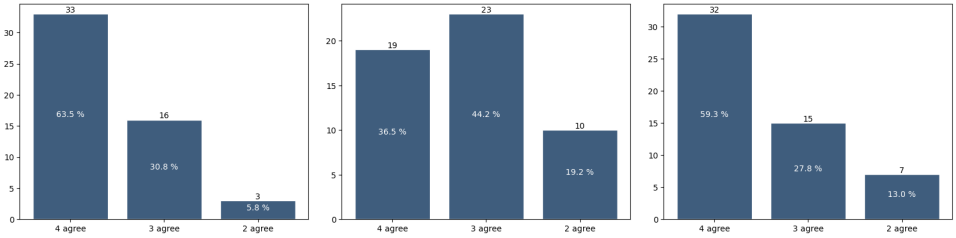
This subsection presents the evaluator consensus distribution for Questions 1 and 2, which is analyzed in Section 4.1. Regarding Question 1, as can be seen in Figure 6a, out of the 52 documents that achieved evaluator consensus in the Amazon dataset, 33 reached full agreement among all four evaluators, 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, 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 7. 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. Similarly to Figures 6 and 7, Figure 8 shows the distribution of consensus among evaluators for hierarchical categorization.



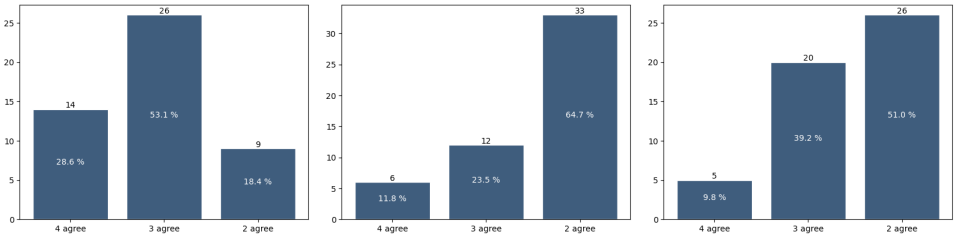
(a) Amazon (b) Pang Movie (c) Vader Movie

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



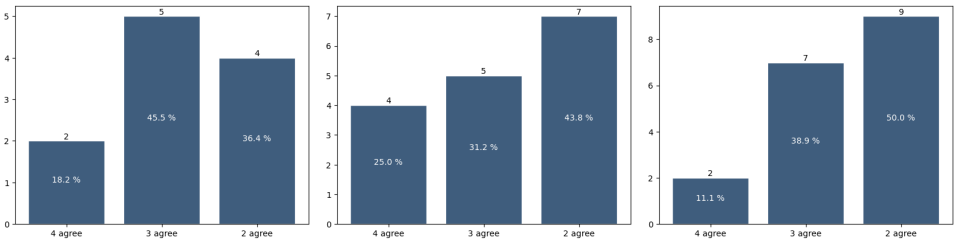
(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 6: Consensus for Question 1.



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

Figure 7: Consensus for Question 2.



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

Figure 8: Consensus for hierarchical categories.

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 5: Texts illustrating the "No consensus" category

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 5.

The first row in Table 5 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 agreement.

A.5 Inter-evaluator agreement for Question 2 in cases of "Human Mislabeling"

In Figure 9, similar to Figure 10, 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 errors in the gold standard.

Additionally, we can observe a high prevalence of the category Incorrect label in the dataset, 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 Incorrect label in the dataset.

A.6 Differences in Evaluation carried out by Computer Scientists and Linguists

We carried out an additional analysis focusing on the evaluators' backgrounds. Since two evaluators rated each document 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.7 Set of misclassifications by all classifiers

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

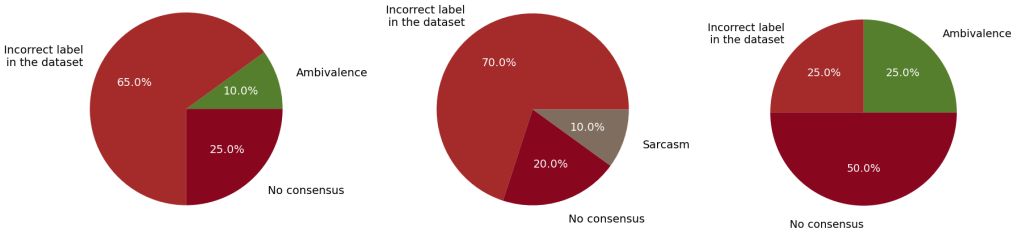
Table 6: Set of misclassifications by all classifiers.

A.8 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.

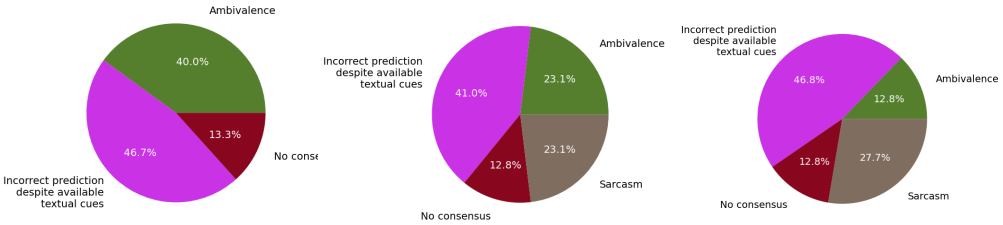
A.9 Comparison Among Transformers

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



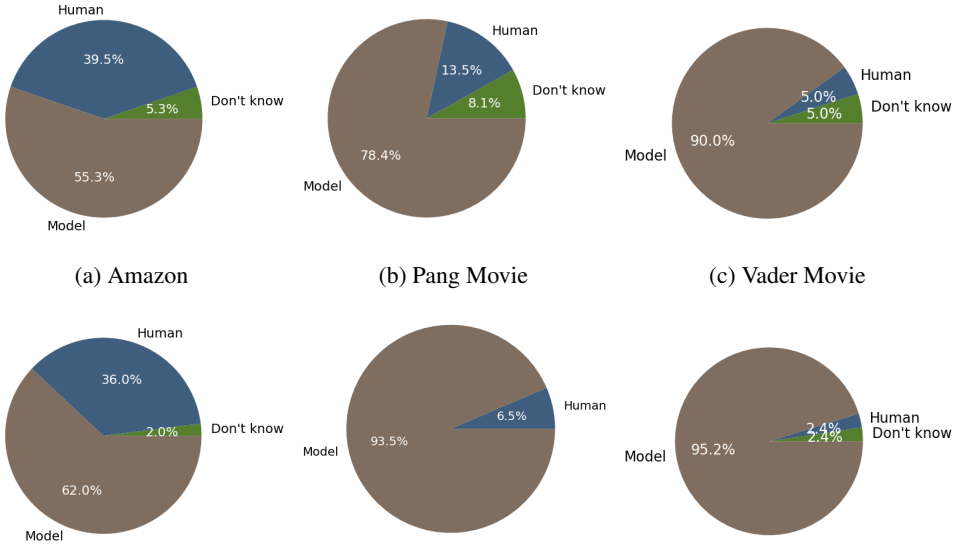
(a) Amazon (b) Pang Movie (c) Vader Movie

Figure 9: Results for Question 2 in cases where Response to Question 1 was “Human Failure".



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

Figure 10: Response analysis for Question 2 in cases where “Model Failure” was selected for Question 1.



(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).

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.

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.10 Comparison between Transformers and LLM's

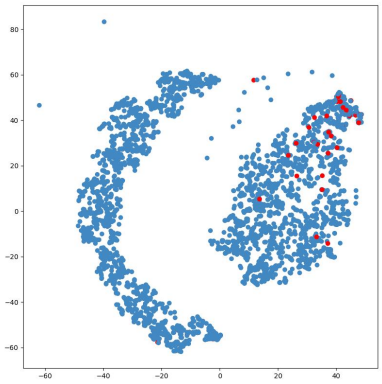
Applying our methodology to other stronger LLMs would be interesting and we will do it in the near future. However, we would like to call the reader's attention to the fact in GLUE's benchmark, for the sentiment analysis task, SST-2, a dataset similar to the ones used in our work, has an accuracy of 97.9 (Vega v1), whereas RoBERTa obtains 96.7 (Facebook AI). Without a statistical method for comparison, these results are not enough to claim that Vega V1 is clearly superior to RoBERTa. In other words, it is not always true that LLMs are better than 1st or 2nd generation Transformers for all tasks.

Several studies show that RoBERTa is a very strong model for sentiment analysis (Cunha et al., 2023b; Bai et al., 2023). Indeed, recent benchmarks (Cunha et al., 2023a) have shown that the differences among the latest versions of these Transformers (including RoBERTa, BERT, DistilBERT, BART, AlBERT, and XLNet) in some of the datasets we use in our experiments are very small. More specifically, in (Cunha et al., 2023b), RoBERTa achieved the highest effectiveness on 12 out of 22 datasets compared to other Transformer-based alternatives. On the remaining datasets, RoBERTa's performance was statistically equivalent to the best method, with marginal differences ranging from 0.10% to 1.09% (on average, 0.82%).

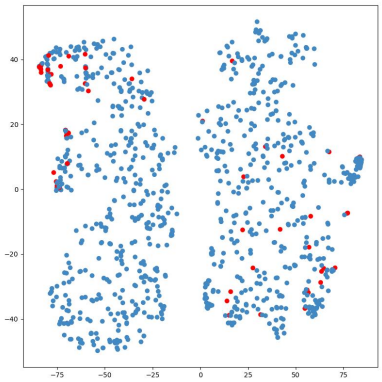
Furthermore, our proposed endeavor of analyzing the hardest misclassification cases (those that no classifier can correctly assign using very separable contextual embeddings (de Andrade et al., 2023)) is a challenging one. So we decided to start with strong methods for the (sentiment analysis) task, which is better documented, allowing us to understand certain premises, over which we also have better control regarding training and fine-tuning. Moreover, these analyses can be done at a much reduced cost than used Large Language Models.

A.11 TSNE Visualization of the Errors in the Dataset

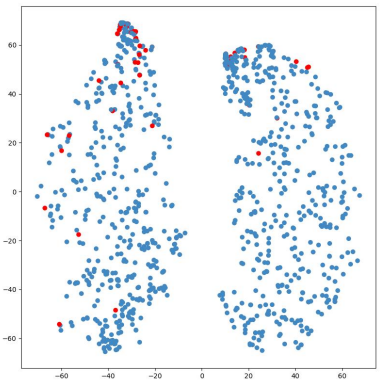
Figure 12 presents the TSNE visualization of the documents in the analyzed datasets using the BERT-based vectors. We marked in red the misclassified documents. We can see that many misclassified documents lie on the class borders, but there are other cases demanding further investigation.



(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.