"Do we like this, or do we *like* like this?" : Reflections on a Human-Centered Machine Learning Approach to Sentiment Analysis

Sourojit Ghosh[0000−0001−5143−6187], Murtaza Ali, Anna Batra, Cheng Guo, Mohit Jain, Joseph Kang, Julia Kharchenko, Varun Suravajhela, Vincent Zhou, and Cecilia Aragon[0000−0002−9502−0965]

University of Washington, Seattle

Abstract. Machine Learning is a powerful tool, but it also has a great potential to cause harm if not approached carefully. Designers must be reflexive and aware of their algorithms' impacts, and one such way of reflection is known as human-centered machine learning. In this paper, we approach a classical problem that has been approached through ML - sentiment analysis - through a Human-Centered Machine Learning lens. Through a case study of trying to differentiate between degrees of positive emotions in reviews of online fanfiction, we offer a set of recommendations for future designers of ML-driven sentiment analysis algorithms.

Keywords: human-centered natural language processing · natural language processing · sentiment analysis · qualitative coding.

1 Introduction

Participation in online communities is a common part of our digital lives, and brings about several benefits such as community building [20, 63] and informal learning [26, 36]. Emotional expression is one of the central features of interactions between members of online communities, and a large amount of effort from community designers goes into building in affordances for varied emotional expression. Consider, for instance, the range of emotions available at a single click on Facebook or LinkedIn. Emotional expression has been seen to be correlated with making connections in online communities [27, 38] and finding social support [29, 40], among several other benefits, making it an important topic of research in recent years. One vein in this field is concerned with identification of emotions expressed online, known as *sentiment analysis*. While sentiment analysis has commercial benefits, it has been criticized for its potential age bias [18], sexist nature [66], and several other problems (elaborated in Section 2.1).

While sentiment analysis can be done by human annotations, the more common approach is to use machine learning (ML) [10, 68]. While the ML approach has several benefits, such as being able to handle large quantities of data, there are also growing concerns about the accuracy of this approach, along with the

potential for introduction of bias, error, or contextual understanding. Such issues may involve reusing models trained on some specific dataset on unrelated datasets [47] or the models' failure to consider data in their contexts [6].

In this paper, we advocate for the adoption of a *human-centered machine learning* (HCML) [13] approach to sentiment analysis, as opposed to a standard ML approach. We present a case study of applying HCML to design an ML-classifier to detect and differentiate between degrees of positive emotions in a dataset of fanfiction reviews. We highlight and reflect upon the various stages of the HCML process, and evaluate the various successes and failures of our model. We conclude with recommendations for applying HCML to sentiment analysis.

2 Related Work

2.1 Sentiment Analysis using ML

Sentiment analysis is a computational technique used to determine the presence of emotions or feelings in pieces of text [53]. There are two broad types of sentiment analysis techniques: a lexical method and an ML approach.

In the lexical method, pieces of texts are typically assigned scores based on the cumulative presence of positive or negative words based on dictionaries of word-score pairs. Though simplistic in their approach, such approaches have been demonstrated to be highly accurate according to some metrics [5]. Such algorithms break down pieces of text through techniques such as stemming (removing prefixes and suffixes e.g. 'playing' is stemmed to 'play'), computing n-grams (breaking down text into phrases instead of individual words) and Parts-of-Speech Tagging (evaluating words/phrases in conjunction with associated parts of speech). Designers of lexical sentiment analysis algorithms can leverage existing word corpora such as WordNet, SenticNet or SentiFul [54, 49], though there are some known disadvantages of using them such as their failure to classify emotions into hyponymic relationships [58], low representation of words for nonbinary gender identities [32], and low accuracy outside of English [8].

The other popular approach is to use machine learning models. Such approaches typically begin with manually classifying a small chunk of data with emotions, training a model on this chunk of data (called training data), and then deploying the model on the remaining data. This approach is known as supervised learning, though variations of this exist with partially-labeled training data (semi-supervised learning) [52] or unlabeled training data (unsupervised learning) [34]. The algorithms vectorize the data similarly as their lexical counterparts (through techniques like stemming, POS tagging etc.) and then apply techniques such as Support Vector Machines [1], Naive Bayes [28], Maximum Entropy [37], Random Forest [31], K-nearest neighbors [17], or a combination of those [2, 35, 55]. The results of such classifications are calculated by observing the output metrics, such as Precision, Accuracy, Recall or f-score.

Both of these types of sentiment analysis methods are usually built to detect three types of emotions: Positive, Negative, or Neutral [48, 59]. Other sets

of emotions commonly used for sentiment analysis are Ekman's [22] six basic emotions: Anger, Disgust, Fear, Joy, Sadness and Surprise [41, 73].

However, such approaches to sentiment analysis has come under some criticism over the past few years, especially when they are used in conjunction with recommender systems that analyze users' reactions and use that data to determine what type of content to recommend. Users who are subjected to sentiment analysis processes dub them 'invasive and scary', associating them with loss of autonomy [4] as they are boiled down to single data points [39]. A similar criticism is of its potential for bias because of different types of bias baked into the training data, such as overrepresentation of content produced by men over women or gendernonbinary individuals [66] or annotators failing to recognize intentional stylistic choices inherent to specific Englishes [61]. Beyond technical challenges such as failure to recognize textual devices such as sarcasm or irony [45], ML-driven sentiment analysis algorithms should not simply be designed and deployed without careful introspection of their appropriateness, impacts and potential for harm. We believe that such an introspection is supported by a HCML approach, as we explain in the next section.

2.2 Human Centered Machine Learning

Human-centered machine learning (HCML) is defined as 'a set of practices for building, evaluating, deploying, and critiquing ML systems that balances technical innovation with an equivalent focus on human and social concerns.' [13]. It is a rising sub-field that seeks to examine the impacts of ML on individuals and communities, and how designers of such technologies can evaluate their own practices [9]. HCML is a timely area of research, as more and more ML algorithms are abandoning the previously popular publicly available corpora of text data such as WordNet [46] in favor of human-generated data that users might not even be aware is getting used for such purposes [24]. At its core lies the directive of being responsible with the 'silver bullet' that ML is widely considered to be [13] in 'a deliberate, careful, and inclusive way that will set a standard for the future of algorithmic accountability' [51].

Past researchers of HCML have put forward practices of being more human-centered in applications of ML. Such practices include considering whether ML is the right solution for a given question [13, 47], recognizing that ML algorithms contain and reflect the opinions of their annotators [13, 15], and that they might cause harm and perpetuate negative biases [65, 70]. At their core, such practices are rooted in human-centered design and design justice principles of knowing who a design is for and how designs might impact them [14, 16]. In this paper, we adopt such practices as we design a ML model to differentiate between degrees of positive emotions in online fanfiction reviews. We present this as a case study of lessons learned in a HCML approach to sentiment analysis.

3 Case Study: Detecting Degrees of Positive Sentiment in Fanfiction Reviews

Fanfiction is "writing in which fans use media narratives and pop cultural icons as inspiration for creating their own texts." [7] Online fanfiction communities are some of the most active text-based communities, supporting over 2.5 million users daily [43] in 2020, at the last time of counting, with numbers most likely having grown since then. A majority of users on popular online fanfiction communities like Fanfiction.net and Archive of Our Own (AO3) are female or gendernonbinary young adults [43, 44] and their active participation has been seen to be impactful in the formation of informal mentorship networks [12], helped them explore gender identities [21] and strengthen language skills [7].

On online fanfiction communities, users can upload stories or add chapters to existing stories, and other users can respond to them by leaving reviews. These reviews are one of the most common methods of communication between users, especially since platforms like Fanfiction.net lack robust direct messaging affordances. Prior research into reviews [12, 23] has found the tone of reviews to be overwhelmingly positive, with as few as 2% reviews being negative. However, with the abundance of positive reviews, it becomes important to determine degrees of positivity between them because, while reviews such as 'I like this' and 'I love this so much, this has changed my life!!' are both positive, there are very different degrees of positivity expressed in them. In this case study, we aimed to build an ML classifier to differentiate between different degrees of positive sentiment in fanfiction reviews. We work with a dataset of fanfiction reviews which contains over 176 million reviews scraped from Fanfiction.net [72].

4 Methods

4.1 Considering Appropriateness of ML

Prior to the design of our classifier, we began with considering whether ML was an appropriate approach towards the question we were trying to answer, and how our approach could potentially be harmful. Such a consideration was especially important given that that fanfiction reviews often known contain sensitive content [67], and mostly come from young adults [42].

We recognized that classifying degrees of positive emotions in a dataset of over 176 million pieces of text would not be humanly possible. An ML approach would make our task achievable, but we had to come up with some considerations of mitigating the potential harms we could cause. We decided that we would not report any reviews verbatim, such that they could either be searched in the dataset or on Fanfiction.net. We also decided that for any reviews we consider reporting, even if we are obfuscating them for anonymization, we would first check if the user who published that review still had an active account and if not, then we would not report their review in this paper. These considerations aim to protect the users' choices, since they likely did not author fanfiction reviews

knowing that those reviews might one day be analyzed for research purposes [24]. We also determined that our work has a low potential for harm, since we are not producing actionable results with respect to online fanfiction communities, or using our findings to suggest interventions or changes to how such communities currently operate. Based on these considerations, we decided that our HCML approach would be appropriate for the question we aim to answer.

4.2 Positionality

We begin our case study with a consideration of our positionalities. ML models have positionality [11] because of the positions and biases, both conscious and unconscious, of their creators [19, 60] and we believe that any representation of ML work is incomplete without including mention of its creators' positionalities.

All the authors of this paper have experience reading or writing fanfiction in online fanfiction communities. Some did not have this experience at the start of the project, but acquired it by immersing themselves into reading, reviewing and, in some cases, writing their own works of fanfiction over the course of this research. The first author has over six years of experience with reading and writing in online fanfiction communities, and two years' experience with studying them. Some of the fandoms to which we are most connected to are Harry Potter, Naruto, Lord of the Rings, the Marvel Comics, Batman, and Twilight, and our connections with these fandoms likely influenced how we analyzed reviews on stories related to them. All of the authors are fluent in English (though not all are native speakers or have English as their first language) but share no other language between them, a fact that prompted us to analyze only English reviews in our study. We can account for some other subconscious biases that likely impacted our manual annotation (discussed in next Section), such as annotators' preferences towards particular fandom characters, but acknowledge that there likely were several other individual or collective subconscious biases that impacted our data annotation.

4.3 Data Collection and Annotation

As mentioned before, in this study we used a dataset of over 176 million fanfiction reviews [72]. We began with a random sample of 15,000 reviews from that dataset, and removed all the non-English reviews from it to arrive at a reduced dataset of 11,292 reviews, hereafter referred to as Dataset 0. We then made two copies of this dataset, hereafter referred to as Datasets I and II.

To determine a slate of positive emotions to analyze our datasets with, we drew inspiration from a slate of review types identified by Evans et al. [23] in their work on reviews in online fanfiction communities. They identified two mutually-exclusive review types expressing different degrees of positivity (Shallow Positive and Targeted Positive) and a third non-mutually exclusive review type to label reviews excitedly asking for updates (Update Encouragement). From Ghosh et al.'s [27]'s slate of emotional expression in online fanfiction communities, we adopt three emotions to mirror the topics: Like, Joy/Happiness

and Anticipation/Hope. Like and Joy/Happiness are mutually exclusive positive emotions, with Like being the milder of the two, and Anticipation/Hope is non-mutually exclusive to Like or Joy/Happiness. Due to this partially mutually and non-mutually exclusive nature of the codes, we measure agreement using the Generalized Cohen's kappa [25].

Dataset I was manually annotated entirely by the first author, and Dataset II was annotated by the remaining members of the team. During the process of annotation of Dataset II, the annotators demonstrated over 95% agreement for Anticipation/Hope, but only 52% and 63% agreements respectively for Like and Joy/Happiness. Annotations in Dataset II were consolidated by two-thirds agreement within the team. We discuss comparisons between annotations of Datasets I and II in Sections 5.1 and 6.2.

Annotation processes involved checking whether a review contained Like or Joy/Happiness, and whether it contained Anticipation/Hope or not. Thus, each review had 6 possible annotations: (Like), (Joy/Happiness), (Anticipation/Hope), (Like, Anticipation/Hope), (Joy/Happiness, Anticipation/Hope), and (None). None was used for all negative reviews (e.g. 'I hate this'), as well as those stating facts (e.g. 'Harry Potter was the son of Lily and James Potter').

During the annotation processes, a debate arose of how annotators could distinguish between Like and Joy/Happiness, since such distinctions could be incredibly subjective. Anticipation/Hope was easier to recognize, with the presence of phrases such as 'I hope...', 'I'm excited for the next update' or 'Please update soon'. Similarly, annotators agreed upon which reviews did not contain any of the three positive emotions.

After annotating the first 1000 reviews, annotators identified seven common aspects of reviews expressing Like or Joy/Happiness, mentioned in Table 1. For each of these, we believed that their presence in a review indicated a higher positive emotion than an absence, because the reviewer made a conscious decision to use such aspects. We annotated another 1000 reviews and, having observed patterns of reviews being labeled as Like or Joy/Happiness, determined that any positive review that contained 4 or more of the aspects mentioned in Table 1 could be strong candidates for Joy/Happiness. However, annotators still used their own interpretations of review contents to determine whether a given review with less than 4 of these aspects could still be considered Joy/Happiness, or whether a review with 4 or more of these could still be annotated with Like. We did so to retain the human element in annotation, rather than designing a rule that could be mechanically applied.

We completed our annotations of Datasets I and II using these rules, and compared differences at the end.

4.4 Model Design and Training

Our ML model is a supervised Naive Bayes classifier. Naive Bayes algorithms are some of the most popular for text classification [57, 30], and for our use case, its assumption of independence of features across data points holds true. We also considered Linear Support Vector Classification (SVC) approaches, but Naive

Table 1. Aspects of Positive Reviews, used for Annotation.

Aspect	**Sample Usage**
Intentional Capitalization	I LOVE this
Emoticons	I <3 this
Exclamation Points	I love this!
Repetition	I love love love this
Intentional Misspelling	I looooove this
Actions	*dies* I love this
Keyboard Smashing	askhwwifnwwervbu I love this

Bayes performs better over datasets containing short pieces of text over those containing longer pieces of text [69], making it more appropriate for our use. We experimented with the Bernoulli Naive Bayes model and the Multinomial Naive Bayes models, but abandoned those because of their poor performance [56] to deal with unbalanced datasets such as ours where we expect far more reviews to be classified with at least one positive emotion than not. We thus chose a Complement Naive Bayes model, because of its known effectiveness in accurately classifying unbalanced text data [56].

We preprocessed our data by filtering out all non-English reviews, and tokenized them with TF-IDF Vectorizers [62]. Given the gulf in size between our training data (11,292) and test data (over 176 million), we performed Laplace smoothing by setting the alpha hyperparameter to 0.07, a value experimentally determined. Based on our annotation, we identified and built a list of stop words appropriate for our dataset.

We trained two different models, hereafter referred to as Prototype Models I and II, for testing purposes. Prototype Model I was trained on a random sample of 2000 reviews from Dataset I, and used to classify the remaining reviews in Dataset I. Similarly, Prototype Model II was trained on a random sample of 2000 reviews from Dataset II, and used to classify the remaining reviews in Dataset II. In these classifications, Prototype Model II returned an overall f-score of 0.67, far superior to Prototype Model I's score of 0.23. Prototype Model II's better performance was also confirmed by manual examination of the reviews classified by both Prototype Models I and II. We thus adopted Dataset II as our 'ground truth' dataset, both because it led to more accurate classifications and because it represented the opinions of a larger majority within the research team. We discuss this further in Section 6.2.

After identifying the ground truth dataset, we continued fine-tuning Prototype Model II. This process consisted of modifying both the algorithm code and re-auditing Dataset II to be more consistent in the annotations. Through this process, we identified a few issues that we could fix, such as a bug in emoticon recognition, and some misclicked annotations in Dataset II. A few iterations through these issues resulted in an increased overall f-score of 0.76. We tested it further by obtaining and classifying a random sample of 1000 reviews from the master dataset, and manual examination of these classified reviews showed over 90% true positives. At this stage, we were satisfied with the model.

We then applied the trained model over full dataset of over 176 million reviews. The algorithm was executed on the Mox supercomputer at the University of Washington, Seattle. Once it returned an output dataset of classified reviews, we randomly selected a sample of 10,000 reviews to manually examine.

5 Results

5.1 Comparing Datasets I and II

Because the performance of supervised ML models are direct results of the data they are trained on, we began our analysis by taking a look at this training data. We compared annotations in Dataset I (annotated entirely by the first author) and Dataset II (annotated by the rest of the researchers) to examine similarities and differences between the two. These results are depicted in Table 2. All numbers are out of 11,292 reviews.

Table 2. Comparison of Annotations of Datasets I (annotated by first author) and II (annotated by other members of research team. All numbers out of 11,292 reviews.

Emotions	**Reviews Annotated with Emotion in Dataset I**	**Reviews Annotated with Emotion in Dataset II**
Like	4,537	3,966
Joy/Happiness	3,268	2,912
Anticipation/Hope	851	803
Like, Anticipation/Hope	1,167	1,662
Joy/Happiness, Anticipation/Hope	923	1,631
None	346	318

We also present a set of examples of reviews where annotators of Datasets I and II differed in their applications of Like and Joy/Happiness, along with their rationale for their annotation. These are depicted in Table 3.

Through our observations in Table 2, we find that the first author interpreted more reviews as expressing a single emotion (e.g. Like or Joy/Happiness), while the rest of the researchers identified more reviews with a combination of emotions (e.g. Like, Anticipation/Hope or Joy/Happiness, Anticipation/Hope). We also observe that annotators of both Datasets have high agreement for reviews that either show no positive emotion (i.e. annotated with None) or express Anticipation/Hope, but differ greatly in their interpretations of Like and Joy/Happiness.

Table 3. Examples of Disagreement between Annotations of Like and Joy/Happiness in Datasets I and II.

Review Text	**Annotation in Dataset I**	**Annotation in Dataset II**
It's such a great story for your first time. Thank you for sharing your stories with us!	Like	Joy/Happiness
LOVE this story. Very very well written!	Joy/Happiness	Like
Absolutely fell in love with this story. Please with a cherry on top continue!	Joy/Happiness, Anticipation/Hope	Like, Anticipation/Hope
Hurray! Congrats for becoming a published author! I am waiting to buy your book! :D.	Like, Anticipation/Hope	Joy/Happiness, Anticipation/Hope

5.2 Model Performance

We report the performance of our model in terms of its f-score, given the unbalanced nature of our dataset. Our model returned an overall f-score of 0.694, with individual f-scores of 0.829 for Anticipation/Hope, 0.601 for Joy/Happiness and 0.622 for Like. Based on our manual examination of 10,000 randomly selected reviews, we identified 77.2% reviews where we agreed with the model's classification. The breakdown for this random set is shown in Table 4, and some sample disagreements of annotation are shown in Table 5.

Table 4. Results from manual examination of classified reviews.

Emotions	**Number of Annotations**	**Number of Manually-determined True Positives**
Like	4,691	3,462 (74%)
Joy/Happiness	3,378	2,744 (81%)
Anticipation/Hope	699	628 (90%)
Like, Anticipation/Hope	302	117 (39%)
Joy/Happiness, Anticipation/Hope	886	730 (82%)
None	44	42 (95%)

We thus observe that our model is moderately successful at identifying the emotions Like, Joy/Happiness, Anticipation/Hope or some combination of those in online fanfiction reviews. We observe that it does better for Anticipation/Hope

Table 5. Examples of Disagreement between ML Classifier and Human Annotation

Review Text	Model Classification	Human Annotation
It ends here? But I also love the happy ending. Do a sequel!	Anticipation/Hope	Like, Anticipation/Hope
LOVE THIS! I read it all in one day, it's so amazing!!	Like	Joy/Happiness
this is great	Joy/Happiness	Like
I'm so angry you stopped here	None	Anticipation/Hope
oh my gosh this is SO SO good! :)) I love this, I hope you update soon! *eager waiting*	Like, Anticipation/Hope	Joy/Happiness, Anticipation/Hope

than for Like or Joy/Happiness, both in terms of performance metrics and upon manual observation. The high performance for Anticipation/Hope can be attributed to the high agreement between annotators of Dataset II, which led to high-quality of training data annotated with Anticipation/Hope. This implies the model's success in being able to identify a positive emotion when it does not need to distinguish between two mutually-exclusive ones.

The fact that human annotators (both within Dataset II and across Datasets I and II) could not consistently identify Like and Joy/Happiness explains the model's relatively lower confidence in classifying reviews with either of those emotions. We cannot be sure about the model's confidence in classifying reviews with a combination of Anticipation/Hope and either Like or Joy/Happiness because of the lack of a numerical metric representing the entire dataset, but from the disagreements between human annotators and the possibly moderately-low quality of training data supplied, we believe that such classifications are going to be moderately accurate across the dataset.

6 Reflecting on our Design Process

6.1 Pre-Design Stage: Consideration of Appropriateness

One of the fundamental principles in human-centered design is to consider the potential harms that can come from designing a piece of technology, and whether it should be designed at all. Chancellor [13] encapsulates this within HCML practices as a directive to 'ensure ML is the right solution and approach to take'. We believe that it is important to have this conversation *before* beginning to approach the solution, since it can reveal potential harms of using ML and

therefore inform the approach of mitigating such harms. This practice is rooted in design justice [16] principles of prioritizing the impact of design over the intent of the designers.

In our case, we considered ML to be an appropriate approach because it would be able to handle our large dataset better than manual annotation could and, more importantly, we did not foresee much potential for harm because we are not deploying our algorithm to a real-world setting. Were we to do so, and if our algorithm would interpret fanfiction reviewers' words and determine what content to recommend them next, then we would have had to make very different considerations before we built the algorithm.

However, we do not believe that refusing to pursue an ML approach to sentiment analysis is a viable answer, though such notions of design refusal are present both in Chancellor's work [13] and in design justice principles [16]. ML-driven sentiment analysis processes are fairly well-deployed in several systems used by millions around the world, and refusal to design future systems would not alleviate the issues caused by the existing ones. Instead, we call for present and future designers of such systems to be more human-centered in their approach, reflecting upon the power and potential for harm within their design, and strive for algorithmic fairness [3]. These reflections can occur by asking questions like who participated in the design process and who did not, and who benefits from or is harmed by the design [64].

6.2 Training Stage: Working with Subjective Interpretations of Emotions

That ML models reflect and replicate the biases and opinions embedded within its training data has been repeatedly demonstrated over the past few years. Chancellor [13] calls upon HCML designers to 'acknowledge that ML problem statements take positions' and recommends designers publish position statements along with their work or documentation of the influence of individual perspectives on the algorithm. We take this one step further by directly demonstrating the differences between two models trained on the same dataset but annotated differently.

We prepared Datasets I and II as the same copies of Dataset 0, but with one fundamental difference: they were annotated by different people. Dataset I was annotated by the first author and therefore contained only their interpretation of the three positive emotions and differences between them. Dataset II was annotated by the rest of the research team and consolidated by two-thirds majority, meaning that it reflected a strong majority opinion of the research team. A manual examination of the two (Table 2) shows big differences in Like and Joy/Happiness, and this is confirmed by the gulf in scores between Prototype Models I and II trained on Datasets I and II respectively. Examination of specific reviews where annotators differed (shown in Table 3) give specific insights into why such differences occurred, particularly between Like and Joy/Happiness. For instance, the review in the first row of Table 3 is considered Like by the first author in Dataset I, but the fact that the reviewer acknowledges the greatness

of the story given it's the author's writing debut was considered by annotators of Dataset II to warrant a label of Joy/Happiness.

By designing and observing the differences between Datasets I and II, we gain an insight into a core issue with sentiment analysis – if interpretations of human emotions are so subjective and if two groups of human annotators can annotate the same dataset so differently, how can sentiment analysis algorithms claim to have reliable understandings of human emotion? The differences across Datasets I and II remind us to continually be aware that our ML model replicates *our* interpretation of the differences between Like and Joy/Happiness. Such interpretations might not align with those of others and so, a different group of designers working with the same problem statement and dataset might have produced completely different results. In our advocacy for a HCML approach to sentiment analysis, we encourage future designers of ML-driven sentiment analysis algorithms (elaborated in Section 7) to consider the inherent ambiguity and human subjectivity of sentiment analysis [15] and thus adopt a sociotechnical approach to sentiment analysis which centers and seeks input from the direct users of the algorithms [4].

This exercise also allows us to imagine how Prototype Models I and II would have behaved very differently were they both applied to the entire dataset of 176 million reviews, based on the differences in their respective training datasets. It demonstrates how subjectivities that are embedded within the training data of ML models are replicated in the application of the models, such that their results are also subjective. We believe that in working with ML, it is important to understand just how closely design is tied to designers, especially when systems designed by the few affect the lives of many. This is especially important when making claims about the accuracy or effectiveness of sentiment analysis algorithms, because a model can only accurately identify emotions that the annotators of its training data would agree with, and to claim any accuracy beyond that is unfair (discussed further in Section 7.2).

6.3 Design Stage: Prototyping and Iteration

Another aspect of human-centered design is prototyping and iteration, which we consider important to HCML processes. After we determined the annotated Dataset II to be our ground truth dataset, we performed a few rounds of prototyping and iteration before running it on the full dataset. These were relatively low-cost processes, such as spot-checking the accurate functioning of the emoticon recognition or performing automated searches through the annotated datasets to verify no annotations were mistyped. Though low-cost, these processes greatly benefitted our design, with the final version of Prototype Model II demonstrating better performance than its original one.

We would like to see more importance given to prototyping and iteration within the HCML process. This is especially relevant to ML-driven sentiment analysis algorithms which are deployed in conjunction with recommender systems. In such cases, iteration can involve large user testing or signing up potential users to evaluate and comment on small sections of the system.

6.4 Evaluation Stage: Measuring Performance through Manual Examination and Anticipating Failure

Finally, we focus on the evaluation of our algorithm, moving beyond quantitative metrics of performance and focusing on manual examination of classified reviews. For sentiment analysis problems, we hold manual verification to be especially important, preferably by involving people who might be eventual users. Since we did not have any future users of our algorithm, we performed the manual verification ourselves.

Our moderately successful classifier performance is underwhelming, especially in light of most ML algorithms reporting high successes [33, 71], because our model is not able to reliably distinguish between degrees of positive emotions. However, the result is not surprising because we can attribute it to the researchers' inability to, between themselves, could not determine a rigid rule for what constituted Like and what elevated it to Joy/Happiness. The closest we came was to determine a set of common aspects of reviews we believed expressed Joy/Happiness (Table 1), but even then we did not feel that it would be accurate to say that for a review to be considered Joy/Happiness, it had to have more than half of those. Such a rule would have taken away from the fact that different people express degrees of positivity in different ways, and their typed-out text might not always encapsulate the exact strengths of their emotions. It would not have been realistic to expect a model trained on data that did not have a clear pattern to be able to infer such nuances, and thus we went in expecting moderate results.

Such algorithmic failure was anticipated given the subjectivity of our annotations and instead of considering it a defeat, we use it as an opportunity to improve our understanding of how HCML should work. In line with Chen et al. [15], we advocate for using manual examination as a means to better understand how and where annotators disagree. For sentiment analysis problems, such manual examination can reveal scenarios where two annotators disagree with the machine classification of sentiment, which can be then pursued to examine how these two annotators' opinions are embedded within the ground truth dataset.

Above all, we imagine a HCML approach to set realistic and human-centered expectations, instead of imagining ML as a magical force with unlimited potential. Designers of ML algorithms must remember at every step of the way that models only amplify their own opinions, and that it can be very easy to think of their results as 'highly accurate' if the results confirm the designers' opinions. Therefore, designers must involve direct users of the algorithms throughout the design process, especially during the training and evaluation stages, to see if the opinions of designers align with those of users. For sentiment analysis, where the opinions are as subjective as the interpretation of human emotions, this might lead to results that score low on quantitative metrics and show large disagreements between annotators and machine classifications. We advocate for such results to be considered valid, instead of designers manipulating the training data to chase higher scores.

7 Recommendations for Applying HCML to Sentiment Analysis

7.1 Consider Alternative Approaches

Given the aforementioned criticisms against both ML in general and specifically sentiment analysis via ML, we recommend that before taking such an approach, designers carefully consider whether other alternative approaches could be appropriate. Some potential alternatives could be manual analysis of affect in text, or directly asking users to elaborate on their own emotions when they author some text. Such considerations are especially important when sentiment analysis is used to predict user behavior, such as in recommender systems.

7.2 Manually Annotate Your Own Data

We believe that it is important for researchers to train their ML models on data that resembles the data to which the model would be eventually applied, instead of relying on existing datasets of words assosciated with emotions. The reasons for this are twofold. Firstly, manual annotation of data which closely resembles the target data will give designers some insights inherently unique to their data. For our case study, one example of such an insight was the emergence of a set of stopwords, prompting us to use them instead of existing libraries. Secondly, such manual annotation inserts the researchers' own positions into the algorithm such that the views of the model represent the views of the designers instead of those of the creators of other datasets. This creates a sense of algorithmic accountability with emotion recognition [4], which is important in cases of breakdown and failure. This should also serve to help designers represent their work in the most honest way. Rather than saying 'my classifier accurately identifies positive emotions', a more accurate representation would be 'my classifier accurately identifies emotions I think are positive'. We believe that such representations of ML work, centering around the designers or data handlers, must become more common, so that consumers of the work can be more informed.

7.3 Manually Examine Classifier Performance and Involve End-Users

While most ML classifiers report quantitative metrics of evaluation such as f-scores, we recommend that designers of ML-driven sentiment analysis algorithms evaluate their work through manual examination of classified reviews. This can begin during the training stage and reveal insights into potential errors in the process, as it did for us when it showed us that some annotators mislabeled a few reviews. Identifying and rectifying these labels led to a more accurate algorithm, something we would not have achieved were it not for manual examination.

However, manual examination must be considered of highest importance during the classification stage, just prior to deployment. At that stage, the model is likely operating on real people's data, and any issues that exist at this stage

will later effect real users if not rectified. This is also a good exercise in identifying potentially harmful algorithmic classifications, such as perpetuating racist stereotypes by labeling speech from Black women as 'angry' [50].

Thus, this is an opportune moment to subject the algorithm to user testing, one of the most important pieces of the human-centered design process. Designers at this stage can recruit a panel of potential users and test whether they agree with the machine classifications. Low agreement between the user and the classifier can be identified early, before the algorithm is deployed and assigns incorrect labels to user emotions.

7.4 Accept and Expect Underwhelming Results

Finally, we ask that developers of ML-driven sentiment analysis algorithms prepare for and accept underwhelming performances from their algorithms. Even though today's world is driven by demands of high speed and high accuracy, designers must realize that interpretations of sentiments are inherently subjective and trying to predict just how much positivity or negativity is embedded within a piece of text might not be accurate for a large number of people. Realizing the limits of one's model might result in more conservative application and deployment, and careful consideration of its ability to predict human behavior.

8 Conclusion

In this paper, we performed a case study of a Human-Centered Machine Learning approach to a sentiment analysis problem over a very large text corpus. We attempted to differentiate between different degrees of positive emotions in over 176 million reviews of online fanfiction, defining two mutually-exclusive positive emotions of different degrees (Like and Joy/Happiness), and a third emotion not mutually-exclusive to those (Anticipation/Hope). We began with a consideration of whether ML would be appropriate for this problem and, once we resolved it, continued with trying to identify the various positions we were bringing to the data. We demonstrated the impact of differently encoded data on a model's performance, through the creation of two datasets annotated by different researchers. Our model demonstrates underwhelming results, which points to the inherently subjective nature of differentiating between degrees of positive emotions. We concluded with some recommendations for future designers of ML-driven sentiment analysis algorithms including going through stages of the human-centered design process by manually annotating their own data and examining the classified data through user testing, as we hope that they would consider a HCML approach.

As ML becomes more and more popularly used as a solution to problems in most fields, we believe that it is important for designers to consider the very real possibilities of causing harm. A Human-Centered Machine Learning practice might alleviate some of those possibilities, and do right by the communities that are subject to these algorithms.

References

1. Ahmad, M., Aftab, S., Ali, I.: Sentiment analysis of tweets using svm. Int. J. Comput. Appl **177**(5), 25–29 (2017)
2. Al Amrani, Y., Lazaar, M., El Kadiri, K.E.: Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science **127**, 511–520 (2018)
3. Altman, M., Wood, A., Vayena, E.: A harm-reduction framework for algorithmic fairness. IEEE Security & Privacy **16**(3), 34–45 (2018)
4. Andalibi, N., Buss, J.: The human in emotion recognition on social media: Attitudes, outcomes, risks. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–16 (2020)
5. Annett, M., Kondrak, G.: A comparison of sentiment analysis techniques: Polarizing movie blogs. In: Conference of the Canadian Society for Computational Studies of Intelligence. pp. 25–35. Springer (2008)
6. Aragon, C., Guha, S., Kogan, M., Muller, M., Neff, G.: Human-Centered Data Science: An Introduction. MIT Press (2022)
7. Black, R.W.: Language, culture, and identity in online fanfiction. E-learning and Digital Media **3**(2), 170–184 (2006)
8. Bond, F., Foster, R.: Linking and extending an open multilingual wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1352–1362 (2013)
9. Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society **15**(5), 662–679 (2012)
10. Burnap, P., Rana, O.F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L.: Detecting tension in online communities with computational twitter analysis. Technological Forecasting and Social Change **95**, 96–108 (2015)
11. Cambo, S.A., Gergle, D.: Model positionality and computational reflexivity: Promoting reflexivity in data science. In: CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2022)
12. Campbell, J., Aragon, C., Davis, K., Evans, S., Evans, A., Randall, D.: Thousands of positive reviews: Distributed mentoring in online fan communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. pp. 691–704 (2016)
13. Chancellor, S.: Towards practices for human-centered machine learning. arXiv preprint arXiv:2203.00432 (2022)
14. Chancellor, S., Baumer, E.P., De Choudhury, M.: Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–32 (2019)
15. Chen, N.C., Drouhard, M., Kocielnik, R., Suh, J., Aragon, C.R.: Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(2), 1–20 (2018)
16. Costanza-Chock, S.: Design justice: Community-led practices to build the worlds we need. The MIT Press (2020)
17. Daeli, N.O.F., Adiwijaya, A.: Sentiment analysis on movie reviews using information gain and k-nearest neighbor. Journal of Data Science and Its Applications **3**(1), 1–7 (2020)

18. Díaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D.: Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 chi conference on human factors in computing systems. pp. 1–14 (2018)
19. Draude, C., Klumbyte, G., Lücking, P., Treusch, P.: Situated algorithms: a sociotechnical systemic approach to bias. Online Information Review (2019)
20. Dym, B., Brubaker, J.R., Fiesler, C., Semaan, B.: "coming out okay" community narratives for lgbtq identity recovery work. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–28 (2019)
21. Dym, B., Brubaker, J.R., Fiesler, C., Semaan, B.: "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–28 (2019). https://doi.org/10.1145/3359256, https://dl.acm.org/doi/10.1145/3359256
22. Ekman, P.: All emotions are basic. The nature of emotion: Fundamental questions pp. 15–19 (1994)
23. Evans, S., Davis, K., Evans, A., Campbell, J.A., Randall, D.P., Yin, K., Aragon, C.: More than peer production: Fanfiction communities as sites of distributed mentoring. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. pp. 259–272 (2017)
24. Fiesler, C., Proferes, N.: "participant" perceptions of twitter research ethics. Social Media+ Society **4**(1), 2056305118763366 (2018)
25. Figueroa, A., Ghosh, S., Aragon, C.: Generalized cohen's kappa: A novel inter-rater reliability metric for non-mutually exclusive categories. In: Proceedings of the Human Interface and the Management of Information Thematic Area in the context of the 25th International Conference on Human-Computer Interaction (HCI International). Springer (2023)
26. Ghosh, S., Figueroa, A.: Establishing tiktok as a platform for informal learning: Evidence from mixed-methods analysis of creators and viewers. Proceedings of the 56th Hawaii International Conference on System Sciences pp. 2431–2440 (2023)
27. Ghosh, S., Froelich, N., Aragon, C.: "i love you, my dear friend": Analyzing the role of emotions in the building of friendships in online fanfiction communities. In: Proceedings of the 15th International Conference on Social Computing and Social Media in the context of the 25th International Conference on Human-Computer Interaction (HCI International). Springer (2023)
28. Goel, A., Gautam, J., Kumar, S.: Real time sentiment analysis of tweets using naive bayes. In: 2016 2nd International Conference on Next Generation Computing Technologies (NGCT). pp. 257–261. IEEE (2016)
29. Gui, X., Chen, Y., Kou, Y., Pine, K., Chen, Y.: Investigating support seeking from peers for pregnancy in online health communities. Proceedings of the ACM on Human-Computer Interaction **1**(CSCW), 1–19 (2017)
30. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd international conference on World Wide Web. pp. 729–736 (2013)
31. Hegde, Y., Padma, S.: Sentiment analysis using random forest ensemble for mobile product reviews in kannada. In: 2017 IEEE 7th international advance computing conference (IACC). pp. 777–782. IEEE (2017)
32. Hicks, A., Rutherford, M., Fellbaum, C., Bian, J.: An analysis of wordnet's coverage of gender identity using twitter and the national transgender discrimination survey. In: Proceedings of the 8th Global WordNet Conference (GWC). pp. 123–130 (2016)
33. Hong, L., Doumith, A.S., Davison, B.D.: Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 557–566 (2013)

34. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd international conference on World Wide Web. pp. 607–618 (2013)
35. Huq, M.R., Ahmad, A., Rahman, A.: Sentiment analysis on twitter data using knn and svm. International Journal of Advanced Computer Science and Applications **8**(6) (2017)
36. Kassens-Noor, E.: Twitter as a teaching practice to enhance active and informal learning in higher education: The case of sustainable tweets. Active Learning in Higher Education **13**(1), 9–21 (2012)
37. Kaya, M., Fidan, G., Toroslu, I.H.: Sentiment analysis of turkish political news. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 174–180. IEEE (2012)
38. Kivran-Swaine, F., Brody, S., Diakopoulos, N., Naaman, M.: Of joy and gender: emotional expression in online social networks. In: The ACM Conference on Computer Supported Cooperative Work Companion. pp. 139–142 (2012)
39. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al.: Computational social science. Science **323**(5915), 721–723 (2009)
40. Levonian, Z., Dow, M., Erikson, D., Ghosh, S., Miller Hillberg, H., Narayanan, S., Terveen, L., Yarosh, S.: Patterns of patient and caregiver mutual support connections in an online health community. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW3), 1–46 (2021)
41. López-Chau, A., Valle-Cruz, D., Sandoval-Almazán, R.: Sentiment analysis of twitter data through machine learning techniques. In: Software engineering in the era of cloud computing, pp. 185–209. Springer (2020)
42. Lulu: The slow dance of the infinite stars (2013)
43. Lulu: Archive of our own: 2020 statistics (Nov 2020)
44. Lulu: Archive of our own: Overall gender and sexuality of ao3 users (Nov 2020)
45. Maynard, D.G., Greenwood, M.A.: Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In: Lrec 2014 proceedings. ELRA (2014)
46. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
47. Mohammad, S.M.: Ethics sheet for automatic emotion recognition and sentiment analysis. Computational Linguistics **48**(2), 239–278 (2022)
48. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd international conference on Knowledge capture. pp. 70–77 (2003)
49. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Sentiful: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing **2**(1), 22–36 (2011)
50. Noble, S.U.: Algorithms of oppression. In: Algorithms of Oppression. New York University Press (2018)
51. O'neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway books (2016)
52. Ortigosa-Hernández, J., Rodríguez, J.D., Alzate, L., Lucania, M., Inza, I., Lozano, J.A.: Approaching sentiment analysis by using semi-supervised learning of multidimensional classifiers. Neurocomputing **92**, 98–115 (2012)
53. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval **2**(1–2), 1–135 (2008)

54. Poria, S., Gelbukh, A., Cambria, E., Yang, P., Hussain, A., Durrani, T.: Merging senticnet and wordnet-affect emotion lists for sentiment analysis. In: 2012 IEEE 11th international conference on signal processing. vol. 2, pp. 1251–1255. IEEE (2012)
55. Rana, S., Singh, A.: Comparative analysis of sentiment orientation using svm and naive bayes techniques. In: 2016 2nd International Conference on Next Generation Computing Technologies (NGCT). pp. 106–111. IEEE (2016)
56. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 616–623 (2003)
57. Roback, A., Hemphill, L.: "i'd have to vote against you" issue campaigning via twitter. In: Proceedings of the 2013 conference on Computer supported cooperative work companion. pp. 259–262 (2013)
58. Rudnicka, E., Bond, F., Grabowski, Ł., Piasecki, M., Piotrowski, T.: Lexical perspective on wordnet to wordnet mapping. In: Proceedings of the 9th Global Wordnet Conference. pp. 209–218 (2018)
59. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: International semantic web conference. pp. 508–524. Springer (2012)
60. Scheuerman, M.K., Wade, K., Lustig, C., Brubaker, J.R.: How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. Proceedings of the ACM on Human-computer Interaction **4**(CSCW1), 1–35 (2020)
61. Shen, J.H., Fratamico, L., Rahwan, I., Rush, A.M.: Darling or babygirl? investigating stylistic bias in sentiment analysis. Proc. of FATML (2018)
62. Singh, A.K., Shashi, M.: Vectorization of text documents for identifying unifiable news articles. International Journal of Advanced Computer Science and Applications **10**(7) (2019)
63. Stanoevska-Slabeva, K., Schmid, B.F.: A typology of online communities and community supporting platforms. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences. pp. 10–pp. IEEE (2001)
64. Sterling, S., Marton, H.: Design justice: An exhibit of emerging design practices. vol. 2. The Allied Media Conference (2016)
65. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Equity and access in algorithms, mechanisms, and optimization, pp. 1–9 (2021)
66. Thelwall, M.: Gender bias in sentiment analysis. Online Information Review (2018)
67. Tosenberger, C.: " oh my god, the fanfiction!": Dumbledore's outing and the online harry potter fandom. Children's Literature Association Quarterly **33**(2), 200–206 (2008)
68. Venigalla, A.S.M., Chimalakonda, S., Vagavolu, D.: Mood of india during covid-19-an interactive web portal based on emotion analysis of twitter data. In: Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing. pp. 65–68 (2020)
69. Wang, S.I., Manning, C.D.: Baselines and bigrams: Simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 90–94 (2012)
70. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al.: Do no harm: a roadmap for responsible machine learning for health care. Nature medicine **25**(9), 1337–1340 (2019)

71. Yang, X., Steck, H., Liu, Y.: Circle-based recommendation in online social networks. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1267–1275 (2012)
72. Yin, K., Aragon, C., Evans, S., Davis, K.: Where no one has gone before: A metadataset of the world's largest fanfiction repository. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 6106–6110 (2017)
73. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. Empirical Methods in Natural Language Processing (2017)