

Sentiment Analysis: Relativity's Approach to Developing Responsible AI Solutions for e-Discovery and Investigation

*Brittany Roush, Ivan Alfaro, Roshanak Omrani, Nathan Reff,
Aron Ahmadia & Dilan Dubey*

Reprinted with permission.



Sentiment Analysis: Relativity's Approach to Developing Responsible AI Solutions for e-Discovery and Investigation

Authors: Brittany Roush, Ivan Alfaro, Roshanak Omrani, Nathan Reff, Aron Ahmadi, Dilan Dubey

[Executive Summary](#)

[Foreword](#)

[Sentiment Analysis in e-Discovery](#)

- [The Importance of Sentiment Analysis](#)
- [The Issue with Sentiment Analysis Models](#)

[Relativity's Approach to Sentiment Analysis](#)

- [Fit for Purpose Sentiment Models](#)
- [Built for a Legal Use Case](#)

[Model Performance and Assessment](#)

- [Relativity's Sentiment Analysis Model Performance](#)
- [Assessing Model Bias](#)
- [Mitigating Bias in Relativity's Sentiment Analysis Model](#)

[Conclusion](#)

Executive Summary

Relativity's sentiment analysis tool provides machine learning-powered labels that indicate the presence of positive or negative tone or emotions, such as desire and anger, within written documents. The tool also allows users to search for these behavioral signals to quickly pinpoint items of interest or scan emotionally charged sentences in documents in a case or investigation.

One of the biggest challenges to designing such a tool is the potential to create a performant but biased model. This can lead to not only inaccurate results, but also flawed decision-making that may harm protected groups. This white paper describes in technical detail how Relativity developed a high-quality sentiment analysis model that is industry-leading in its approach to mitigating bias.

Foreword

"What does it mean to achieve Relativity's vision to 'Power the Pursuit of Justice?'" That's a question our Product, Engineering, and Applied Science teams ask themselves every day. For every technological advancement we deliver, we must consider its impact on the people involved in legal matters. The work done in Relativity, one of the largest legal technology companies in the world, has shaped the legal, cultural, and political landscape over the last two decades. From empowering the Innocence Project, to investigating human trafficking or war crimes in Iraq, to the countless civil and criminal proceedings that pass through the platform, the work we do to support these matters has significant consequence.

This is especially true when it comes to AI and algorithms that evaluate communications and other people-generated content. These algorithms are historically rife with bias^{3,4} and with the advent of technologies such as OpenAI's GPT-4 and Google's Bard, the legal industry is facing ethical questions that will define it for decades to come. What technology is appropriate for use in these industries? What do the consumers have to consider when evaluating AI tools? Who is making AI and what are their goals with the algorithm?

Relativity plays an important role by acting responsibly in this new world of AI assistive technology. Given the critical work our customers do every day, the choices we make in development can have far-reaching

impacts on the human beings behind every litigation or investigation. This is not a role we take on lightly. After all, as a recent FTC report on the dangers of AI stated, “No matter how these harms are generated, technology and AI do not play a neutral role in their proliferation and impact.”⁵ Powering the Pursuit of Justice in an AI paradigm means reducing harm through ethically driven AI development, providing clear and understandable results, and being transparent about how the AI is built.

Powering the pursuit of justice is a calling all Relativians unequivocally embrace.

—Brittany Roush, Senior Product Manager, Relativity

Sentiment Analysis in e-Discovery

The Importance of Sentiment Analysis

Investigators may rely on communications that demonstrate emotions such as fear or anger in various matters involving harassment, fraud, or deception. For example, in employment-related litigation, identifying communication in which people made threats, expressed support, or relayed dissatisfaction with a particular person or organization could be crucial to the case. Understanding the tone and sentiment of key players can be essential to a successful outcome.

Identifying sentiment has long been a challenge for investigators and attorneys. Without a tool purposely built to identify sentiment in human communication, legal teams must rely on keywords and conceptual searches to find documents that include emotions. In English alone, there are over 3,000 recognized words to express emotion, and thousands more in slang. Searches to identify sentiment must be overinclusive by design, impacting the overall cost and risk for a matter. As a notable example, to identify negative sentiment in data sets, attorneys have long used the “profanity list,” (sometimes known as the “Carlin List”) a comprehensive list of vulgarities for keyword searching. Given the expansive number of terms on the profanity list and its various permutations, the results are usually overinclusive and must be reviewed and culled down. This inevitably leaves more opportunity for error and makes the process more costly.

To compound the problem, keywords can have multiple meanings depending on their context. The emotion of a particular word or phrase heavily depends on the surrounding words and sentences. For example, “I hate you” could be used earnestly or jokingly between friends. Context is further nuanced by the speaker’s age, culture, nationality, gender, social status, etc. This can make it challenging for keyword searches to identify the tone of written communication across different socio-economic and cultural contexts.

In these situations, a sentiment analysis tool can detect emotional signals to shed light on important evidence. That evidence can lead investigators to other information, such as the involvement of additional parties. Sentiment analysis can help attorneys build case strategy by providing insights into the attitudes, biases, and opinions surrounding issues or people.

Sentiment analysis also reduces cost. Corporations can expect to field 50-100 cases a year⁶ related to investigations, while service providers, law firms, and government entities, whose organizational purposes are to conduct investigations and manage cases, can expect many more. Sentiment analysis helps save money by decreasing the time required to identify high quality evidence.

The Issue with Sentiment Analysis Models

Sentiment analysis has its origin in marketing. Marketers that understand emotional responses to advertisements can better target their campaigns. In these scenarios, sentiment was generally captured through well-defined inputs from users, such as survey results, or through unstructured inputs like comments. Even if the model returned biased results, the harm caused by targeted advertisement was generally low risk.¹

Conversely, using a sentiment analysis model with this marketing foundation poses a high risk in Relativity's ecosystem. In testing historical models for legal use cases, we found that without defined, specific language interpretation, key material could be excluded from searches, or humans could misinterpret evidence due to automatic detection results. In one notable example regarding a sexual harassment investigation, an investigator discovered that all communications sent by one woman were scored positively, even when that person was clearly complaining about the behavior of a male colleague. The model scored this person's language differently, based on statistical representations of words relating to her gender. Reviewing only the documents with a negative sentiment would exclude key evidence and ultimately lead to unfair outcomes.

Relativity's Approach to Sentiment Analysis

Relativity's approach to building any new tool starts with customers and their use cases. We seek to understand the problems customers are trying to solve and the risks they are trying to mitigate. Ultimately, we work to deliver responsible products that reduce workloads for customers.

Relativity followed the below approach to deliver Sentiment Analysis:

1. Documented the process of understanding customer use cases, the intended purpose of the model, the model build mechanisms, and what steps were taken to validate the model, to provide transparency and develop customer understanding of the product.
2. Defined the use cases for Sentiment Analysis through extensive customer discovery efforts.
3. Experimented with existing models and ultimately designed the model to be fit for the purposes of the intended use cases.
4. Investigated the model's training data for bias that would skew the results of the algorithm.
5. Evaluated the output of the model for bias and adjusted it as required.
6. Worked directly with customers to test the model output and verify they were obtaining desired results that provided value to their workflows.

Traditionally, sentiment analysis is defined as: "... common text categorization task, sentiment analysis, the extraction of sentiment, the positive or negative orientation that a writer expresses toward some object. A review of a movie, book, or product on the web expresses the author's sentiment toward the product, while an editorial or political text expresses sentiment toward a candidate or political action. Extracting consumer or public sentiment is thus relevant for fields from marketing to politics." [1]

"..., also called opinion mining, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics. The field represents a large problem space. Many related names and slightly different tasks – for example, sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, and review mining – are now all under the umbrella of sentiment analysis." [2]

¹ There are notable exceptions, like marketing around adult content, which is specifically called out in Sofiya Noble's book *Algorithms of Oppression*.

7. Engaged in a continuous improvement cycle through client feedback and AI experimentation.

In the process of developing sentiment analysis, Relativity spent over a year in product discovery and development. This included numerous customer interviews and granting a select group early access to versions of the product to understand how best to deliver value.

During the evaluation of initial sentiment analysis models, Relativity identified results that exhibited bias regarding certain protected groups. Due to these findings, Relativity decided to develop a new model. The second version was provided to the same group of customers for testing. It delivered valuable results while eliminating the bias issues from the previous version. The sections under “Model Performance” go over these assessments and results in detail.

Fit-for-Purpose Sentiment Models

In machine learning, “fit for purpose” refers to the ability of a model to meet the specific requirements and objectives of a particular use case,^{7,8} meaning:

1. The model was designed to meet the specific needs of the task it was intended for;
2. The model was trained with a dataset that is representative of the use case it was meant to solve;
3. The level of accuracy and bias of the model does not create harmful or adverse results for the use case it was meant to solve.

Research showed that most sentiment analysis tools available in the market were trained on datasets not relevant to the legal industry, including the IMDB Movie Review dataset, the Yelp Review dataset, and the Amazon Review dataset. BERT, a commonly used AI model, is partially trained based on the BooksCorpus and English Wikipedia dataset.⁹

To overcome these limitations and further refine Relativity’s sentiment analysis tool, we trained our model using datasets that exclusively represent communication among individuals. This included two publicly available labeled datasets created from Twitter and Reddit content: TweetEval¹⁰ and GoEmotion.¹¹ The TweetEval dataset contains approximately 60,000 tweets in English that are labeled into one of the three sentiment categories of negative, neutral, and positive. The GoEmotion dataset is a human-annotated dataset of 58,000 Reddit comments which have been curated through collaborative efforts of researchers at Stanford Linguistics, Google Research, and Amazon Alexa. GoEmotions contains 27 emotion categories and has been used to train both anger and desire models. These datasets are more representative of the communications upon which legal teams would apply sentiment analysis, such as emails, chats, or text messages.

Built for a Legal Use Case

Relativity’s Sentiment Analysis tool is designed to meet the needs of legal teams and identify representative documents that should be promoted for further review. It conducts sentiment analysis at the sentence level, rather than the document level, which offers several benefits, including:

Accuracy: Individual sentences can express different sentiments, even if the overall sentiment of the document is positive or negative.

Context: Legal teams can more easily identify the specific objects, events, or people discussed when sentiment analysis results are provided at the sentence level.

Granularity: Documents may contain mostly positive sentiments, but if there are a few strongly negative sentences, the sentiment results could be skewed inappropriately to reflect the “stronger” sentiment, leading to a misinterpretation in analysis. Sentence-level sentiment ensures that an accurate representation of sentiment is captured.

To further aid in finding emotionally driven communications, Relativity's sentiment analysis tool provides different fields and dimensions upon which to analyze information. This includes the proportion of emotional sentences, the distribution of the presence of different emotions, and a summary of the most relevant sentences found in a document.

Model Performance and Assessment

During the development of this sentiment analysis product, Relativity evaluated different models including a sentiment analysis model provided by a global cloud vendor and another academic large language model (LLM). Given the level of bias exhibited by these models towards protected groups, Relativity decided to develop its own model for sentiment analysis that addressed this issue while still meeting the needs of a legal use case through purpose-built design principles. Relativity iterated on the modelling approach through various architectures and conducted multiple experiments focused on model performance, bias assessment, and bias reduction. The goal was to ensure that the sentiment analysis model released by Relativity was within reasonable expectations for classification performance and contained minimal bias.

Relativity's Sentiment Analysis Model Performance

For the purpose of this paper, model performance refers to the ability of a model to produce correct results. We will focus on four common metrics:

Precision: The percentage of the sentences correctly classified by the model as representative of a specific emotion (e.g., anger) across the entire population of documents. Higher precision means the corpus of documents are more likely to contain relevant hits.

Recall: The percentage of the total relevant sentences correctly identified by the model across the subpopulation of documents where sentiment has been identified. Higher recall means there are less cases of false negatives and customers are less likely to miss relevant content.

F1-score: Combines precision and recall into a single number (the harmonic mean of both precision and recall). A higher F1-score translates to both precision and recall metrics performing well and penalizes situations where these two metrics are either greatly unequal or low in value.

Depth of Recall: A measure of how many documents one must review to find the number of relevant documents one is looking for. This is similar to the evaluation of search engines. The fewer results that one is required to look through to find what one is looking for, the better the engine. Informally, one can think of this as a measure of review effort.

Precision and recall are appropriate metrics in situations where users will randomly review documents out of a collection. However, we found that customers would sort documents based on sentiment analysis results and then begin to review those documents with the highest sentiment scores. Where documents are reviewed based on a ranking order, depth of recall is an appropriate metric on which to evaluate the model.

Precision, Recall, and F-1 Scores

We conducted an experiment on EmoBank,¹² a representative dataset for sentiment analysis. EmoBank contains 10,000 short messages, comprised of 2,500 with negative sentiment, 5,000 with neutral sentiment, and 2,500 with positive sentiment. Using the EmoBank dataset, we evaluated the performance of our current model in production against both a vendor model and an academically inspired large language model (LLM). See Figure 1 for the precision, recall and F1-scores of each of these models.

		Precision			Recall			F1-Score		
Classifier		Relativity	LLM	Vendor	Relativity	LLM	Vendor	Relativity	LLM	Vendor
Category	Negative	0.5	0.75	0.59	0.65	0.5	0.63	0.55	0.6	0.61
	Positive	0.55	0.7	0.62	0.45	0.5	0.39	0.5	0.6	0.48

Figure 1: Performance metrics on the EmoBank dataset.

The LLM performed well under the precision metric but exhibited low recall. The Relativity model performed the best at recall in the negative category but was outperformed by both LLM and vendor models in the combined F1-score.

Depth of Recall

For this experiment we also captured depth of recall performances for all three models (Figure 2). The x-axis is the number of relevant documents one would like to find. The y-axis represents the number of documents one would need to review using the various models in order to find their x-value target. An ideal model is as perfect as possible, meaning that if there are 200 negative sentiment documents available, then it will only require 200 documents of the overall population to be reviewed. This is reflected in Figure 2.

A high-performant model would be as close as possible to this ideal performance marker (i.e., lower is better in Figure 2). The random model is also illustrated here. It will randomly assign sentiment using the same class balance as the EmoBank test set. Specifically, there is a 25% chance the sentiment is negative, 50% chance the sentiment is neutral, and 25% chance the sentiment is positive. Lastly, there is also a line for the expected random performance, called Random (Theoretical). We expect the random model to follow this line, but there are chances the random model will deviate slightly from this from time to time. These random model indicators are added in to give measurable insight into how much closer these models are to the ideal model rather than a random model.

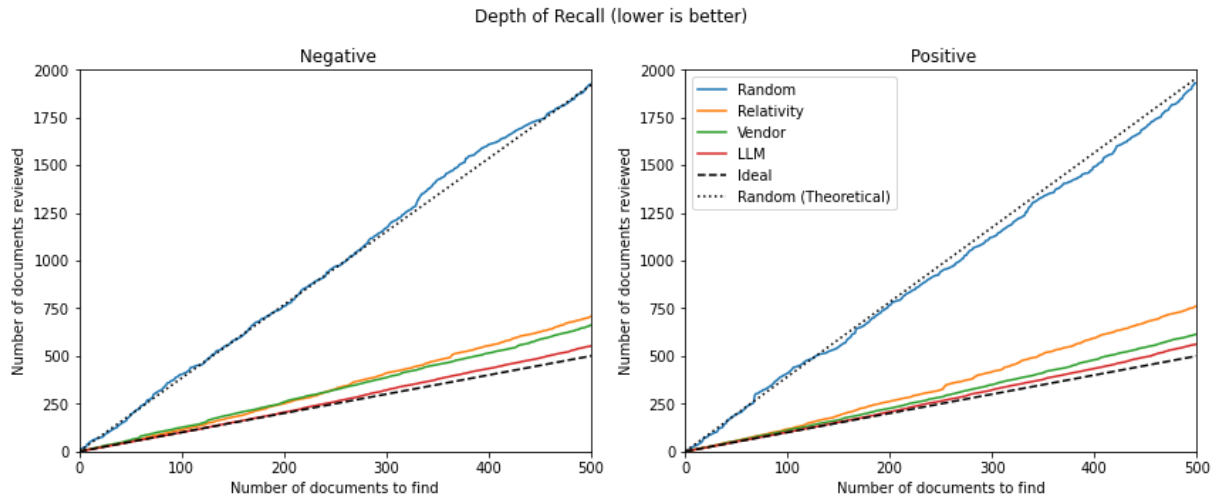


Figure 2: Rank Performance metrics for review effort (depth-of-recall) on the EmoBank dataset.

Specific depth of recall results from this experiment demonstrated that for small investigations, the Relativity model is on par with the vendor and LLM models. If a reviewer is looking for 100 documents with negative sentiment, they would need to review 114 documents with Relativity's model, 126 with the vendor model, and 100 using the LLM. If a reviewer is looking for 200 documents with negative sentiment, they would need to review 250 documents with Relativity's model, 259 with the vendor model, and 205 using the LLM.

For larger reviews, the differences between depth of recall begins to become more prominent. For example, if a reviewer is looking for 500 documents, it would require a review of 706 documents in Relativity, 662 with the vendor mode, and 552 using the LLM. If Relativity had assessed models solely on performance, the Vendor Model and LLM model may have been better choices for larger reviews, but as mentioned above, bias is a key consideration for model assessment.

Assessing Model Bias

Bias in machine learning refers to a systematic error in a model's predictions or decisions. When a model is used in situations where bias can cause harm, we refer to it as a high-risk model. For our customers, most use cases are high-risk as AI bias in litigation and investigations can lead to several negative, unfair outcomes. If biased against protected classes, such as minorities or women, the model may not return evidence that could exonerate an individual, or it may only return a partial scope of evidence, leading legal teams to draw inappropriate conclusions.

While no one model will be completely exempt of bias, the bias should be minimized as much as possible. Using a biased sentiment analysis model in litigation also increases legal risk. If a model used in analysis is found to be biased, this could be used as evidence to challenge the validity of conclusions and the defensibility of processes and procedures. The analysis could be excluded from court proceedings, or the case could be dismissed.

In a sentiment analysis model with minimal bias, the model's confidence that a sentence has a negative connotation should not change across specific nationalities, religions, genders, or other protected classes. Relativity used Amazon's Generalized Fairness Metric Framework¹³ to assess bias in our current model production and compare the results against both a vendor model and an academically inspired large language model (LLM). These are the same models against which we compared performance.

This framework provides templates to evaluate whether the model's predictions would change if certain features that are associated with sensitive attributes were different, while keeping other features constant. For example, the sentence *I like {person}*, should be predicted as a positive sentences independently of the name of the person used in the template. The attributes evaluated included: age, nationality, disability status, gender, race, religion, and sexual orientation.

If we consider the templates for each category (protected attributes), we can look at the difference between the probabilities produced by a model to get a sense of how much variability there is within each category. For each template, we will consider the difference between the maximum and minimum probabilities produced and look at their distribution. In Figure 3, we see box plots for each protected attribute and the distribution of these probability differences. A less biased model will have these box plots as close to the bottom as possible, ideally all flat to 0.0 to indicate no variability in a model's prediction, and therefore less likely to be biased under our assumptions. The figure only considers negative scores, but the results are very similar for positive scores.

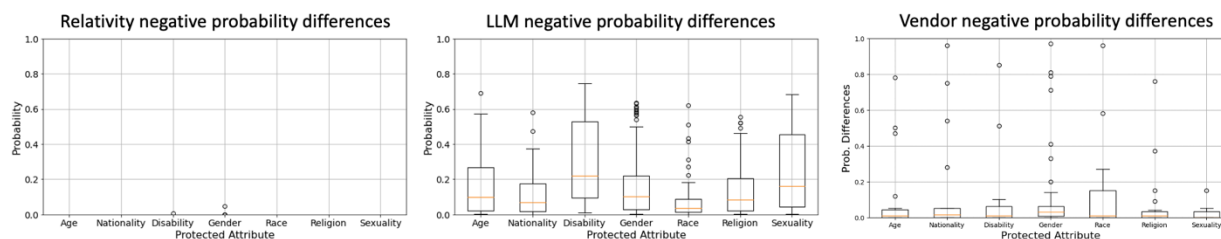


Figure 3: Distribution of negative probability differences predicted for all three models tested over various template substitutions across protected attributes

Results indicate that both the large language model (LLM) and vendor models exhibit higher levels of bias towards seven protected groups: age, nationality, disability status, gender, race, religion, and sexual

orientation. For example, when these two models evaluated a sentence like “*I don’t trust [religious group] people*” the results of the models had significantly different scores across religious groups. In other cases, the mention of specific nationalities made these models incorrectly classify sentences as positive or negative. Comparatively, Relativity’s sentiment analysis model exhibits minimum bias towards these protected groups in this experiment.

Inspired by this framework for bias assessment of negative and positive sentiment, we further generalized the process to assess bias in Relativity’s anger and desire models. This generalization required creating new templates. After creating these templates and completing the assessments, we obtained similar results. Our in-house models exhibited minimum bias without a reduction in performance metrics.

Given the impact that using a biased model to make decision in the legal industry could have to outcomes of litigation and investigation cases, Relativity decided to release its in-house model which exhibited the lowest levels of bias and had performance only marginally impacted in larger matters.

Mitigating Bias in Relativity’s Sentiment Analysis Models

Achieving the bias results described above in the Relativity models required substantial cross-functional collaboration and model iteration. The first iteration of the model leveraged an academically inspired large language model (LLM) architecture. This initial choice was due to the potential high performance achievable with such a model. However, as described in the previous section, there were concerns on how to mitigate bias under this LLM architecture. This model has a pretrained component with over 30,000 vocabulary terms. Although tedious, it is feasible to manually identify and flag terms that describe sensitive attributes as special tokens to prevent the model from encoding them. However, this model has a specific tokenizer for parsing sentences and can handle words that are out-of-vocabulary by breaking them down into vocabulary terms when possible. Therefore, out-of-vocabulary terms associated with protected groups could be encoded by this tokenizer despite a revised vocabulary, and the model is still prone to propagating the social bias that it may have learned in pre-training and fine-tuning phases. Consequently, we chose to build our current sentiment analysis model using a more classical architecture.

The model Relativity currently has in production uses a logistic regression architecture that is trained with an underlying fixed debiased vocabulary. This approach avoids the pitfalls of the LLM architecture in terms of bias mitigation. However, there is a trade-off in terms of performance metrics (as documented above) since this classical approach is more limited in capturing nuanced semantic information, and in its context encoding abilities. For example, a LLM is much better equipped to capture complex meaning of words that may have multiple definitions by using surrounding words for context.

We found that a simple way to mitigate bias was to remove all terms that allowed the model to learn bias from training materials with stereotyping and denigrating content. A collaborative initiative with [Relativity’s Community Resource Groups](#) (CRGs) gathered a diverse group of volunteers to provide insight on words that could inflict bias. The task focused on identifying terms that should be treated as neutral because they are associated with sensitive attributes that are qualified for protection under United States federal anti-discrimination law, i.e.:

1. Gender
2. Sexual Orientation
3. Religion
4. Nationality
5. Race
6. Age
7. Disability

Example of Approach

Training example: "Jane is struggling in math." | **Sentiment:** Negative

The model needs to learn to associate negative sentiment with "struggle in math", irrespective of the subject of the sentence. However, if the vocabulary contains "Jane", "struggling", and "math", the model learns that coincidence of these three tokens yields negative sentiment. As such, it will probably assign less negativity at inference to the unseen sample "John is struggling in math." compared to "Jane is struggling in math." because one of the elements in the negative trio, i.e., "Jane", "struggling", "math", is missing from sentence "John is struggling in math.".

To avoid such biased behavior, "Jane" is excluded from the vocabulary; the model learns that coincidence of "struggling" and "math" yields negative sentiment and assigns the same score to "Jane is struggling in math." and "John is struggling in math."

Note: if a "curse" word or offensive term is often used with respect to a protected group, it is kept in the vocabulary because the model should be able to catch the negative sentiment and toxicity in phrases where such words are used. For example, the offensive term "bitch" is often used with respect to a particular gender. "Bitch" is retained in the model to catch the negative sentiment in the phrase "His colleague is a bitch."

On the other hand, the term "feminist" although neutral, may appear in either negative or positive context, which can cause the model to associate polarity with a supposedly neutral term. Therefore, "feminist" is excluded from vocabulary, so that the polarity associated with it in the training sample does not affect the output of model at inference and a phrase like "His colleague is a feminist." is treated as neutral.

To better clarify, with a vocabulary that includes "bitch" but excludes "feminist", the model learns from the negative training sample "His colleague is a feminist bitch." that the negative sentiment is solely due to the presence of "bitch" in the sentence.

The CRG volunteers were given the full vocabulary list from the model and asked to manually code words as neutral, but volunteers were enabled to give their opinion on what should be considered positive or negative. For example, a volunteer from the Relativity Pride CRG, which focuses on LGBTQIA+ support, may score "queer," "lesbian," and "transgendered" as neutral but "fairy" as negative, as it is a common slur used against LGBTQIA+ people. Volunteers from the BRel CRG, which focuses on Black employee support, may score "Black" as neutral but "articulate" as negative since being described as "articulate" is a common microaggression experienced by Black employees in a workplace.

Note: Having more than one perspective in this process is also important, as marginalized groups of people are not monoliths. While "fairy" may be a slur to one LGBTQIA+ employee, it may be how another employee self-identifies. Multiple inputs ensure a higher degree of accuracy in the model.

The input provided by the volunteers was used to create a more limited model. By considering the lived experiences of diverse, protected classes of people, Relativity's sentiment analysis model is better able to differentiate truly neutral sentiment from negative and positive sentiments, resulting in more fair outcomes for protected classes.

Diverse Teams Generate Better Outcomes

In her book *The Algorithms of Oppression*¹⁴, Dr. Safiya Nobel states that the creators of algorithms are just as important as the training data. Historically, AI has been built by cis-male-dominated teams with limited diversity in race, gender, age, and socio-economic upbringing. As a result, training sets may not be scrutinized, and model language may not be investigated and corrected as diversity becomes an

afterthought. This leads to the unintentional introduction of bias into both the models and the outcomes associated with their use. Dr. Timnit Gebru, AI expert and 2022 Relativity Fest keynote speaker, has written and spoken extensively about the dangers of homogenous teams building AI: “The people creating the technology are a big part of the system. If many are actively excluded from its creation, this technology will benefit a few while harming a great many.”¹⁵

Who makes models is something Relativity takes into consideration to combat bias problems. The team that created Relativity’s sentiment analysis solution has at least one representative from every protected class, often with intersecting facets of marginalization. Having multiple experiences and perspectives enabled Relativity to have difficult conversations on how race, gender, sexual orientation, socio-economic status, disability, religion, and many other factors could influence our AI, and in turn, empowered the team to make the right technology choices to maximize fairness in our models.

Conclusion

As AI systems become more pervasive and impactful in the legal industry, it is critical to consider the ethical implications of their use and to ensure that they are designed and developed responsibly. One of the most significant challenges in doing so is mitigating bias in AI systems. This challenge is particularly relevant for sentiment analysis tools, where biased models can lead to inaccurate and unfair results, especially in criminal cases, civil cases, and investigations that involve issues such as discrimination or harassment or where the person or people under suspicion are in a protected class. When bias is not mitigated it can introduce new and significant legal risks to customers and require more work to identify relevant materials, reducing the value of such a tool. It is imperative for legal technology companies to develop models with as much bias mitigation as possible to ensure fair outcomes.

Results from our depth of recall indicate that Relativity’s sentiment analysis model is not only on par with the performance of both a vendor model and an academically inspired large language model, but also does not exhibit the level of bias found in those models. The results depicted in this paper suggest that Relativity’s sentiment analysis model is the first in-market that is appropriate for use in sensitive and high-risk legal matters where AI fairness is critical.

References

1. Keselj, V. (2009). *Speech and Language Processing* Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+ 988 pp; hardbound, ISBN 978-0-13-187321-6.
2. Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.
3. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.
<https://science.sciencemag.org/content/356/6334/183.full>
4. Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
5. Federal Trade Commission. (2020). *Combatting Online Harms Through Innovation: Federal Trade Commission Report to Congress*.
https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf
6. Business Wire. (2019, October 24). New Survey Reveals Corporate Investigations Expected to Rise, Increasing Strain on Internal Resources.
<https://www.businesswire.com/news/home/20191024005674/en/New-Survey-Reveals-Corporate-Investigations-Expected-to-Rise-Increasing-Strain-on-Internal-Resources>
7. Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA: MIT Press.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
10. Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.
11. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
12. Buechel, S., & Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.
13. Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9, 1249-1267.
14. Noble, S. U. (2018). Algorithms of oppression. In *Algorithms of oppression*. New York University Press.

15. The New York Times. (2021, March 15). Who Is Making Sure the A.I. Machines Aren't Racist?
<https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html>