Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU34146

Selection Bias in Sentiment Data of Social Media in LLM Training Datasets

Shenghao Liu

Business School, Jiangnan University, Wuxi, China
riucheahyo@gmail.com

Abstract. Big language models are possibly becoming positively biased–not on purpose. The systematic literature review methodology was adopted to analyse the survival of negative sentiment content through the aggressive filtering pipelines that convert raw web crawls into training corpora, like C4 and RefinedWeb. A systematic depletion of negative sentiment was reported, through a qualitative synthesis of eleven empirical studies, following a standard filtering procedure, whereas positive and neutral content are relatively preserved. This skew is not consistent--it changes steeply depending on the architecture of the datasets, changing depending on whether using general-purpose and domain-specific classifiers, and increasing when using a more demanding toxicity detection threshold. It was found that the depletion was attributable to four conflating negative emotion with harmful language, length-based heuristics, short negative posts, deduplication with low-repetition emotional posts, and lexical diversity filtering influencing lexical diversity survival. These results give empirical signals to transparency audits and can give practical suggestions on how bias can be reduced in commercial sentiment analysis deployments.

Keywords: social media, sentiment analysis, selection bias, LLM training data

1. Introduction

Most large language models (LLMs) training corpora are web-scale, filtered by multi-stage filtering pipelines, commonly on Common Crawl. Pre-training datasets of mainstream, such as C4, RefinedWeb, and The Pile, are deduplicated, language identified, and quality scored then toxicity detected and entered into model training [1-3]. Though these heuristics are created to enhance data quality, they can also unintentionally cause changes in the distribution of sentiment-bearing content, and systematic selection bias is introduced into the downstream applications.

The content on social media is a significant part of these corpora due to its affective richness and high turnover. This means any skew added in the process of filtering has a direct effect on the accuracy of sentiment analysis, brand tracking, and surveillance of the public opinion that relies on proper polarity identification [4-6]. Recent audits show that quality and toxicity filters are overly discriminative against low-resource, informally written, or emotionally charged text -categories which intersect with negative social media expression. However, the current literature is still disjointed: the literature is mostly made up of single-dataset studies, measurement instruments are

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU34146

diverse, and the causal relationship between particular filtering rules and sentiment depletion is poorly studied.

This paper will use systematic literature review research design to generalise findings on selection bias in social media sentiment data in LLM training datasets. The specific emotion factor of negative sentiment is analyzed to follow the evolution of the negative sentiment representation across the types of data sets, types of measurement methods and platform architecture. The analysis is done in three phases. To begin with, the measurement instruments used by other researchers to measure sentiment are described. Second, patterns of bias direction and magnitude are synthesised that are reported. Third, the most important moderating factors, including data structure, domain of classifiers, and features of the platform, are determined, and recommendations that can be implemented at the level of a business are obtained.

The methodology used was a systematic literature review that was guided by PRISMA 2020. Records published between January 2022 and March 2025 were searched in Google Scholar, arXiv, ACL Anthology, Semantic Scholar and Web of Science. The search query was: C4 dataset (or RefinedWeb or Common Crawl) and (LLM training data or pre-training corpus) and (sentiment or emotion) and (bias or selection bias) and (social media or Twitter). Detailed inclusion and exclusion criteria were developed prior to screening. Based on 187 records including 187 records following the deduplication process, 11 empirical studies were identified, which met the predefined eligibility criteria, and were incorporated to undergo qualitative synthesis. A standardized coding scheme was used to extract the data on the type of datasets, sentiments measurement techniques, direction of bias, and factors that affect sentiments.

2. Literature review

An empirical review methodology using systematic literature review was used to combine empirical evidence on sentiment selection bias in training corpora of LLM. The review structure is based on three analytical themes: composition of LLM training datasets, filtering mechanisms, and their disproportional influence on negative sentiment, and the methodologies used to measure sentiment that drives the patterns of observed bias. Eleven studies are synthesised, including dataset audits, toxicity filtering experiments and cross platform sentiment analyses.

2.1. In the training corpora in LLM, the dataset is composed of two parts: the input and the output

A study concerning the composition of the training dataset on LLM training indicates that there are a lot of platform biases that come before sentiment-specific analysis. Dodge et al. analyzed the source distribution of C4 by document sources, and found that personal blogs and news websites are over-represented, with the social media content (Twitter and Reddit posts) being under-represented compared to its prevalence on the web [1]. Their effort presents critical background information about the compositional imbalances but is descriptive in nature, and does not look at how filtering rules unequally influence sentiment-bearing content. Bender et al. proposed the term documentation debt, and stated that, due to opaque filtering decisions, researchers cannot trace bias to its origin [7]. This observation contextualizes transparency as a condition to comprehend bias in datasets, but neither paper measures the directional impact of curation on negative sentiment retention.

The Pile is statistically described by Gao et al., who reported a change in text length, domain diversity and linguistic complexity across sources [2]. Although their audit confirms the non-random selection patterns in curated corpora, it fails to examine the impact of filtering heuristics on

DOI: 10.54254/2755-2721/2026.GU34146

emptying certain emotion categories. Combined, these papers prove that there are platform biases during pre-training corpora and that filtering transparency hides their causes. Nevertheless, none of them discuss negative sentiment depletion particularly, and there is a glaring gap in the knowledge of how emotive content makes it through data curation.

2.2. Mechanisms of filtering and negative sentiment depletion

Recent data curation practice audits show that quality and toxicity filters can significantly change content distributions, and content with negative sentiment is disproportionately impacted. Longpre et al. quantitatively assessed the data quality and toxicity filtering impact on language model training and showed that unkind heuristics eliminate significant chunks of emotive or informally written text [4]. Their experiments indicate that training sets with much stricter toxicity thresholds yield training sets with statistically different sentiment profiles to unfiltered baselines, and that content moderation pipelines are not emotion-neutral. This analysis was generalized to multilingual web-crawled data by Kreutzer et al., who concluded that the quality heuristic removes large amounts of low resource and non-standard language-types, which often overlap with the affective expression of social media [5]. Their audit report that filtering policies to sift high-resource English content selectively underrepresent emotively marked talk in various language groups.

Garg et al. conducted a survey on the patterns of bias in automated toxic speech detectors, reporting the increased rate by which negative emotional language is tagged as harmful by moderation systems [6]. They show that the common lexical overlaps between sadness, anger, and frustration and the toxic categories, like harassment and hate speech, are often mistaken due to their similarity. Taken together, these three studies determine that filtering pipelines produce systematic representational biases in the negative sentiment, but they fail to determine the rates of depletion in a specifically polarity-dimension-focused manner, a gap which can be filled by synthesizing specifically.

2.3. Sentiment measurement methodologies

The sentiment classifier used has a significant impact on the amount of bias observed, especially in the negative polarity classification. In a survey of modern sentiment analysis techniques, Wankhade et al. reported the continued existence of both lexicon-based and neural architecture-based approaches to social media research [8]. Hutto and Gilbert presented VADER, a rule-based sentiment lexicon which became the de facto standard in early research [9]. Their scanty adjustability to platform-specific expressions, including irony, slang, and posts with emojis, present classification risk to negative sentiment, which tends to use unconventional lexical markers to express the intensity of affect.

A single benchmark called TweetEval by Barbieri et al. was established that involves sentiment polarity tasks in addition to tweet classification [10]. Their analysis shows that transformer models that are fine-tuned on Twitter data (Twitter-RoBERTa) achieve significantly better performance on social media sentiment detection compared to general-domain lexicons, but the performance depends on training corpus and filtering history. This performance difference is important to detect bias: a tool unscaled to social media language can systematically errors in labelling negatively charged posts as neutral or toxic, which mixes real effects of filtering with measurement artifacts. Therefore, the differences in sentiment classifiers between the studies reviewed can be considered a significant source of inter-study heterogeneity in depletion patterns reported.

DOI: 10.54254/2755-2721/2026.GU34146

2.4. Research gaps and contributions

Current studies have three limitations that drive the current review. First, in most studies, the data of a single dataset or platform are analyzed, which does not allow conducting a systematic comparison of the effects of negative sentiment filters across datasets. Second, the heterogeneity reported by disparate sentiment measurement tools is probably overstated, but no research has explicitly compared the impact of these tools on the bias detection of the LLM corpora. Third, prior research is mostly descriptive and little has been explored on the commercial implications or how negative polarity loss can be mitigated.

The gaps are filled by conducting a systematic review that synthesises eleven empirical studies under four dimensions of analysis, i.e., type of datasets, sentiment measurement methodology, bias direction and factors that influence it. Three contributions are obtained in the synthesis. Negative sentiment depletion is shown first to be a common and heterogeneous phenomenon in filtered corpora, with the degree of depletion depending on filtering architecture. Second, the rules of dataset construction, domain adaptation of classifiers, and platform text-length features are discovered as the important moderators of the degree of depletion, which offers the empirical basis of mitigation measures. Third, it provides real-life advice on how to pick and optimize sentiment analysis pipelines in business scenarios like brand tracking and social opinion monitoring.

3. Discussion

3.1. Direction and magnitude of negative sentiment bias

In eleven of the eleven studies synthesised there is a consistent directional depletion in negative sentiment following standard filtering and relatively stable polarities of positive and neutral. According to Longpre et al., aggressive toxicity filtering minimizes retention of emotionally colored text at significant levels with negative posts having higher removal rates than neutral and positive posts [4]. Kreutzer et al. confirm this trend in a multilingual environment, reporting that quality heuristics eliminate negative emotional expressions more frequently in low resource and informally written varieties of language [5]. This directional pattern, which seems to be negative depletion and relative preservation of positive content and neutral content, is often consistent, indicating that filtering pipelines encode a programmatic positivity bias, which is not programmed but arises as a result of the interaction of heuristic rules and affective language properties.

The loss in negative sentiment is however significantly different between studies. The depletion rates reported are moderate to severe based on the strictness of the filtering tool, measurement tool and platform source [4-6]. This heterogeneity suggests that there is no one depletion figure that can appropriately account to the conditional nature of filtering effects. Rather, the degree of bias is tempered by a set of architectural and methodological conditions that are cut up below.

3.2. Effect of type of data set on the level of depletion

Comparisons across datasets indicate significant differences in depletion of negative sentiment severity, which is attributable to differences in philosophy of filtering. It is documented by Dodge et al. that C4 is used to do aggressive toxicity identification and text quality screening when constructing and this seems to yield greater negative sentiment removal than other architectures [1]. In comparison, Penedo et al. define RefinedWeb as using refined heuristics that were optimised to maintain content variety, which can potentially buffer negative expression of emotions by removing

DOI: 10.54254/2755-2721/2026.GU34146

them [3]. The Pile is in a medium state: the multi-source composition consists of Reddit posts, tweets, and web text, but no sub-specific sentiment analysis has singled out the negative polarity depletion rate in its curation pipeline [2].

There are practical implications to this architectural contrast. Strictness in the C4 toxicity-first method, although it works well in weeding out truly harmful content, seems to have a disadvantage of greater false-positives of negative emotional expression [6]. The diversity-conserving heuristics at RefinedWeb, on the other hand, store more affectively marked content, trading off the possible inclusion of borderline toxic content. The lack of sentiment-specific reporting in the documentation of The Pile does not allow to compare it directly, which highlights the debt of documentation that Bender et al. find to be intrinsic to the practices of dataset construction [7]. When practitioners choose pre-training corpora, such differences between architectures suggest that downstream sentiment classifiers will have different polarity distributions based on the underlying dataset they are trained on.

3.3. Bias detection with measurement methodology

The sentiment detecting measurement tools used have significant effects on the depletion magnitudes experienced, especially when of negative polarity. The literature using general-purpose lexicons like VADER shows a stronger negative sentiment loss in comparison with that using domain-specific models like Twitter-RoBERTa [9, 10]. This descriptive trend is consistent with TweetEval results: domain-adapted models are more effective at capturing social media syntax, emojis, and emotional subtlety, and confounding the effects of genuine filtering with measurement artifacts.

It is theorized that the domain-specific models have a higher ability to separate toxic language and legitimately negative sentiment since during their training, they are exposed to the colloquial expressions and figurative devices that lexicons have trouble parsing. The selection of tools is however somewhat confounded with dataset type in the sample: VADER-based audits tend to consider older or more aggressively filtered corpora, whereas RoBERTa-based studies consider recently curated and with less heuristic datasets. However, measurement methodology seems to be a factor in inter-study heterogeneity and needs to be standardised in subsequent studies of negative sentiment retention.

3.4. Differences among platforms and moderating factors

Platform architecture turns out to be a vital moderator of negative sentiment depletion, and Twitter and Reddit have a contrasting pattern of survival. Thapa contrasted the performance of sentiment classification on Twitter and on Reddit in the context of cybersecurity and discovered that the longer discussion threads and more comprehensive contextual embedding in Reddit limit negative polarity detection errors compared to the short posts on Twitter [11]. Generalizing this observation to dataset curation, the length and richness of the discussion on Reddit generally cushions it against aggressive heuristics of filtering, which are sensitive to superficial characteristics like the number of tokens or the n-gram toxicity scores [5, 6]. Conversely, the short-text nature of Twitter makes concentration of negative emotional expression in less tokens, which is more likely to be mistakenly identified as toxicity by its filters.

Four pipeline-stage mechanisms, which disproportionately eliminate negative sentiment, are identified in the literature. In the context of toxicity detection, negative emotional language is often mixed with the definition of toxicity since these two classes have lexical units that represent anger,

DOI: 10.54254/2755-2721/2026.GU34146

frustration, or distress [6]. Length-based heuristics also overrepresentively sample out short texts, and negative tweets are also shorter than neutral or positive ones due to affective intensity often being correlated with shortness of spontaneous expression [5]. Deduplication also decreases the relative percentage of negative content since there is a lower rate of repetition of emotionally colored posts than informational or promotional posts. Lastly, language filtering eliminates large quantities of non-English content [5]; since negative sentiment in high-resource languages exhibits a higher degree of lexical diversity, filtering can change its chances of survival when undergoing curation. These mechanisms act in a sequence, building up depletion pressure on negative polarities which do not exist with neutral or positive counterparts.

3.5. Business implications

The results have direct implications on brand monitoring and consumer analytics. The synthesis shows that negative sentence sentiment is underrepresented systematically in filtered training corpora. When the model sensitivity to negative polarity is dependent on training exposure, which is a conservative heuristic, downstream classifiers can have high rates of false negatives with consumer complaints [4, 6]. A brand that gets a lot of negative mentions might therefore become blind to the posts that are critical and this will cripple crisis early-warning systems and postponing the detection of reputational threat.

Where used in consumer insights, depleted corpora-trained classifiers can tend to push towards the neutral distribution, flattening the polarity distribution. Longpre et al. demonstrate that models that have been trained on aggressively filtered data have lower downstream task accuracy than their counterparts that have been trained on balanced corpora, indicating that market analytics trained on conventional pre-training data are vulnerable to systematically distorting demand cues [4]. Undercounting negative product feedback silently can cause organisations to spend resources in inappropriate ways, to hold on to take corrective measures, or to overestimate customer satisfaction.

Practitioners are followed by three actionable recommendations. Audit foundation models In the first step, datasheet or model card-based assessment of training-corpus sentiment distributions are used to evaluate the training-corpus sentiment distributions prior to deployment [7]. Second, to deploy organisation-specific models, augment pre-trained models with in-domain negative examples with targeted fine-tuning, especially when the application domain is a complaint handling or risk monitoring domain. Third, where training-data selection is possible, use corpora that has been described to have reduced negative sentiment depletion; the various-content-preserving heuristics of RefinedWeb seem to better keep affective content intact than the aggressive-toxicity filtering used by C4 [1, 3]. Such measures do not remove bias brought about by filtering, but they minimize its spread in business decision making.

4. Conclusion

This research uses the methodology of the systematic literature review to synthesise the evidence on the selection bias in the sentiment data of social media in the context of LLM training corpora. The combination of eleven empirical studies demonstrates uniformity in the depletion of sentiment negativity following typical filtering, as well as the level of severity depending on the architectures of the datasets, measurement procedures and platforms. In particular, C4 aggressive toxicity-first pipeline seems to cause more negative polarity loss than diversity-optimised heuristics; general-purpose lexicons like VADER magnify observed depletion rates compared to domain-specific

DOI: 10.54254/2755-2721/2026.GU34146

transformers like Twitter-RoBERTa; and short text format on Twitter focuses negative emotional expression in ways that hallucinate more false removal by superficial filters.

These results will offer empirical standards of auditing negative sentiment bias in pre-training corpora and the choice of downstream models to use in commercial applications. There are two constraints that are recognized. The existing evidence base is Anglophone and Twitter-focused and limits the extrapolation to multilingual and image-dominant platforms. Moreover, mechanisms using causation between particular filtering regulations and the elimination of negative content must be empirically proved directly instead of the observational designs that prevail in the literature. Further efforts are required in terms of extending coverage to the Instagram and TikTok platforms, disaggregating negative sentiment into finer-grained groups, such as anger, sadness, fear, and testing whether filtering has a disproportionate impact on a specific subgroup of negative sentiment, and creating negative-sentiment-aware curation pipelines that would maintain emotional variety without compromising on safety levels.

References

[1] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Gardner, A case study on the colossal clean crawled corpus. Proc. Conf. Empir. Methods Nat. Lang. Process., 1286–1305 (2021).
[2] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800GB dataset of diverse text for language modeling. arXiv 2009.00031 (2020).
[3] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, The RefinedWeb dataset for Falcon LLM: Fine-tuning web data on curated corpora and web data alone. arXiv 2306.01116 (2023).
[4] S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, N. Robinson, D. Mimno, D. Ippolito, A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity. arXiv 2305.13169 (2023).
[5] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Es, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Kasai, S. Chiruzzo, et al., Quality at a glance: An audit of web-crawled multilingual datasets. Trans. Assoc. Comput. Linguist. 10, 50–72 (2022).
[6] T. Garg, S. Masud, T. Chakraborty, Handling bias in toxic speech detection: A survey. ACM Comput. Surv. 55(13s), 1–32 (2023).
[7] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? Proc. ACM Conf. Fairness Account. Transpar., 610–623 (2021).
[8] M. Wankhade, A. C. S. Rao, C. Kulkarni, A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. 55, 5731–5780 (2022).
[9] C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proc. Int. AAAI Conf. Web Soc. Media 8, 216–225 (2014).
[10] F. Barbieri, J. Camacho-Collados, L. Neves, L. E. Anke, TweetEval: Unified benchmark and comparative evaluation of tweet classification. arXiv 2010.12421 (2020).
[11] Sentiment analysis of cyber security content on Twitter and Reddit. arXiv 2204.12267 (2022).