Do fans' and media opinions influence NBA player contracts relative to performance?

s1159164
Radboud University, Nijmegen, Netherlands

1 INTRODUCTION

NBA contracts are negotiated in a setting where teams balance on-court production with entertainment value, brand fit, and public interest. Front offices have access to detailed statistics and models, but they also operate in an environment shaped by fans, journalists, and online narratives. This raises a central question for this project: are contract outcomes linked only to past performance, or also to how players are talked about in public before they sign?

To study this, the project combines structured basketball data with unstructured text. On the structured side, I use publicly available season statistics and advanced metrics from Basketball-Reference.com, together with contract information (years and total value) for players who sign during one free-agency period. This data is used to model a baseline relationship between past performance and expected contract size.

On the text side, I collect fan discussions from basketball focused subreddits (e.g., r/NBA) and media coverage from major sports outlets in a pre-signing window for each player. Named entity recognition and simple filtering are used to isolate clauses or sentences that refer to a single player, and sentiment analysis is then applied to obtain player level fan and media sentiment scores.

The goal is to quantify how pre-free-agency sentiment relates to contracts *relative* to performance. Concretely, I ask whether players with more positive fan or media sentiment tend to sign contracts that look larger than what their statistical track record would predict, and explore how such patterns differ between fan and media perspectives.

2 RELATED WORK

2.1 Sentiment and entity-level opinion mining

Sentiment analysis is commonly framed as a supervised text-classification problem, where models learn to distinguish positive and negative opinions from labeled examples [6]. For short, informal social-media messages, lexicon-based approaches such as VADER are widely used, as they are fast, robust to noisy language, and tuned to online communication [2]. In this project, such a rule-based model serves mainly as a transparent baseline for scoring fan and media comments about players.

To link opinions to individual players, sentiment needs to be computed at the entity level. Modern neural named-entity recognition (NER) models make it possible to identify person names reliably even in noisy, multi-entity contexts [4]. Applying NER to basketball discussions allows sentences and clauses to be split and assigned to specific players, so that sentiment can be aggregated per player over the pre-signing window.

TxMM'25-'26, Radboud University, Nijmegen, Netherlands

2.2 Sentiment and perception of NBA players

A small but growing literature uses social media sentiment to study perceptions of NBA players. One line of work analyzes Twitter messages mentioning individual players and applies sentiment analysis to relate online opinions to performance indicators and derived metrics [5]. Another line analyzes Reddit discussions about players and teams, showing that fan conversations on basketball focused subreddits provide rich, fine grained signals about narratives and reputation over time [8]. These studies demonstrate that fan sentiment can be measured at the player level and that it captures aspects of perception not visible in box score statistics, but they focus on performance evaluation rather than on contract size or overpayment.

2.3 Popularity, status, and economic outcomes in sports

Beyond on court performance, research in management and sports economics shows that reputation, status, and visibility can make players more money. Studies of organizational careers report that higher status individuals receive higher compensation even when controlling for measurable performance [1]. Sports analytics work similarly documents the use of social media metrics to quantify "off-court" or "media" value for NBA players, for example by relating follower counts and engagement to income or sponsorship outcomes [3, 7]. Comparable effects have been reported in football, where club revenues and brand value are linked to star player popularity [9]. Together, these findings suggest that public visibility and fan interest can affect financial decisions in professional sports, motivating an investigation of whether pre-signing window sentiment relates to contracts that deviate from what past performance would predict.

3 APPROACH

3.1 Study design and variables

The fundamental unit of analysis for this study is the individual player. For every player who signed a new contract during the 2025 offseason, we construct three distinct categories of data points: an outcome variable representing the size of the contract, a baseline control variable representing performance on the court, and a set of predictor variables derived from text that capture how the player was discussed in the public sphere prior to signing. To ensure that information from all sources is aligned temporally, we utilize a fixed global start date of May 1, 2025. The end date for data collection is specific to each player, corresponding to the actual date they signed their new contract. Consequently, all text collected from both fan forums and media outlets is strictly restricted to this window of time leading up to the signing.

This specific design is intended to target sentiment as a preliminary signal of public perception, rather than capturing reactions that occur after the contract details have been made public. The primary outcome variable we model for each player is the average annual value (AAV) of their deal, provided that the duration of the contract and its total dollar value are publicly available. In cases where the average annual value cannot be precisely computed due to missing data fields, we retain the player in the sample by using a deterministic proxy, such as the salary for the immediately upcoming season. Making this choice allows us to avoid discarding players from the dataset due to incomplete contract information while still preserving a consistent monetary outcome variable for modeling purposes.

To establish a baseline model for the contract size that one would expect based solely on production on the court, we utilize performance data from the previous season. We summarize this performance using Win Shares (WS) as a single comprehensive predictor. Win Shares is an aggregated statistical measure that attempts to estimate the total number of wins a player contributes to their team over the course of a season by translating productive statistics from the offensive and defensive box scores into marginal wins. We selected Win Shares as the sole measure of performance because it provides a unified estimate of total value on the court for a single season. Furthermore, it is additive across different players and allows for direct comparisons across different teams, which aligns well with how an overall contribution is reflected in the actual success of a team. As a cumulative metric, Win Shares also implicitly incorporates both the effectiveness of a player when they are on the court and the total amount of playing time they receive, both of which are jointly relevant factors when determining contract valuation.

3.2 Text collection: Reddit and media

Discussions among fans are collected from chat forums located on the Reddit platform. Specifically, we target subreddits whose primary focus is the NBA, such as r/nba, r/nbadiscussion, and r/NBAtalk. This data is retrieved using the Reddit API via PRAW. For each individual player in our target set, the search query combines all known aliases for that player using an OR operator. Any aliases consisting of multiple words are enclosed in quotes. We retrieve all posts that fall within the defined window leading up to the signing. To reduce potential ambiguity and filter out threads that are off-topic, we prioritize and filter candidate posts based on the number of distinct individuals mentioned within them. We utilize spaCy to identify `PERSON` entities and discard posts that exceed a specified threshold for the maximum number of people mentioned. We further establish boundaries for the sample size for each player by implementing caps on the total number of posts collected and the number of comments selected per post.

The process for selecting comments gives priority to those that explicitly mention the target player by name. If there is an insufficient number of such comments, the remainder of the quota is filled with "neutral" comments. These are defined as comments that mention neither the tracked players nor any other entities identified as `PERSON`, which helps to reduce information leakage from conversations involving multiple players.

Media coverage regarding the players is discovered through the use of the GDELT Doc 2.1 API operating in ArtList mode. Queries are executed on a per-player basis and are restricted to the same pre-signing time window used for Reddit. To ensure we capture reporting that is relevant to contracts and player valuation, the queries sent to GDELT combine player aliases with the term "NBA" alongside short sets of keywords reflecting specific themes. These themes include language related to the market or rumors, language concerning valuation (such as "overpay," "underpay," or "worth"), framings related to opinions or analysis (such as "grades" or "preview"), and general terms related to free agency.

The URLs retrieved through this process are deduplicated both within individual query runs and across different runs. For every unique URL discovered, we fetch the content of the page using a user agent designed to mimic a standard web browser, including fallbacks to AMP versions when they are available. We then extract the main text of the article using the `trafilatura` library, with a robust fallback mechanism to raw HTML text extraction in cases where the primary extraction fails. Finally, we apply sentence segmentation using spaCy.

A significant limitation regarding the collection of media text must be noted because it affects the reliability of the dataset. While the Reddit data was sourced directly from subreddits dedicated specifically to basketball, ensuring a high degree of topic relevance, the media data was retrieved from a broad range of websites identified by GDELT. We did not manually verify every source website to guarantee it was exclusively a sports or NBA related outlet. Consequently, there is a non-negligible risk that some retrieved articles mentioning a common name might refer to an individual other than the targeted NBA player, despite our keyword filters. For example, a search for a common name might pull up news about a businessman or a local politician instead of the athlete. This introduces a layer of noise into the media dataset that is likely less prevalent in the more focused Reddit environment, where the context of the subreddit makes it more certain that the discussion is about basketball.

3.3 Entity resolution: linking sentences to players

A central technical challenge encountered in this project is the reliable mapping of free-form mentions in noisy text to canonical player identities. We implement a practical entity resolution layer that combines several approaches: (1) a carefully curated dictionary of aliases for the targeted players, (2) the use of spaCy NER to measure noise from non-target entities, (3) a set of deterministic rules for selecting a primary player in instances where multiple tracked players match the same text, and (4) a lightweight heuristic based on pronouns to recover sentences containing only pronouns, which typically occur immediately following sentences with direct mentions of the players. We construct the dictionary of player aliases by combining canonical full names taken from the signing list with additional surface forms. These include unique last names (included only when they are unambiguous within the complete set of players) and nicknames scraped from nickname lists on Wikipedia, which are then filtered down to match the signing population.

This set of aliases is applied using the `PhraseMatcher` component of spaCy to mark mentions of tracked players in a consistent manner across both Reddit posts and news articles. Because a large number of posts and sentences mention multiple players alongside other public figures, we record the presence of other `PERSON` entities for each unit of text. During the stage of collecting data from Reddit, this information is used as a measure for filtering and sorting. At the sentence level, it is recorded as metadata to quantify ambiguity and to support later sensitivity analyses, such as restricting the analysis to sentences containing lower levels of noise.

When a single sentence matches multiple players being tracked, we assign a single `primary_player` using a set of deterministic tie-breaking rules. The preference regarding these rules is as follows: the longest span of the alias (preferring matches of the full name), followed by the highest number of mentions, then the earliest position of a mention within the sentence, and finally alphabetical order to guarantee that the results are reproducible. This rule helps us avoid the issue of counting the same sentence for multiple different players while ensuring that the assignment process remains interpretable.

Many sentences that carry significant sentiment about players omit explicit names entirely, for example, "he was washed last year." Therefore, we apply a deliberate methodological compromise: if a sentence contains male pronouns but does not contain an explicit mention of a tracked player, it can be assigned to a context player defined by the surrounding text. For Reddit, this context is defined as the last explicitly mentioned tracked player within the same block of text (title, body, or comment); otherwise, it defaults to the player that was originally used to retrieve the post. For media text, the context is defined as the last explicitly mentioned tracked player within the article, with an optional fallback to the player associated with the discovered URL. It is important to note that this is not full coreference resolution. It may lead to wrongful assignments in contexts involving multiple people, so we treat it merely as a heuristic that improves coverage and discuss it as a limitation.

3.4 Sentiment scoring and feature construction

We compute sentiment scores for individual sentences and aggregate them into player level predictors within the pre-signing window. Reddit and media are treated as distinct sources because they differ significantly in language patterns and coverage. As shown in Figure 1 (Appendix), the volume of sentences is strongly skewed where a few players dominate the corpus while most have sparse coverage. Consequently, we handle sentence volume primarily as a reliability metric through robustness restrictions rather than using it as a substantive predictor.

Each sentence is scored using two complementary sentiment models. First, we use VADER to produce a continuous compound score ranging from -1 to +1, which serves as a transparent baseline well suited for informal language. Second, we apply the transformer based DeBERTa ABSA model in a target aware setup by pairing each sentence with the player name to ensure the sentiment score reflects opinions directed specifically toward the player.

These models behave quite differently in practice as DeBERTa assigns a larger fraction of sentences as neutral and tends to score Reddit text more negatively than VADER. The relationship between the two scores seems weak rather than tightly aligned as illustrated in Figure 3 (Appendix), indicating that sentiment measurement itself is a source of uncertainty. This observation motivates us to treat the choice of sentiment scorer as a robustness check in the downstream analysis rather than as a single fixed decision.

A central challenge is correctly attributing sentiment in sentences that mention multiple individuals. We track metadata describing how each sentence is assigned to a player to manage the ambiguity of whether an opinion targets the specific player or someone else in the context. Figure 2 (Appendix) displays the distribution of attribution methods where the majority relies on pronoun backfilling or direct single player matches. This motivates a robustness check where we restrict analysis to explicitly matched sentences containing only a single player to reduce noise from multi-person contexts.

We construct features at the player level by aggregating these sentence scores over the pre-signing window separately for each source and model. The primary predictor used in the contract analysis is the mean sentiment per player. We treat additional summaries such as non-neutrality merely as secondary diagnostics because the exploratory analysis indicates that model disagreement and uneven coverage already introduce substantial uncertainty at the measurement stage.

3.5 Contract modeling and definition of "overpay"

To study contracts relative to production on the court, we begin by establishing a baseline salary model that depends solely on performance statistics. Salary is measured as the average annual value (AAV) and is modeled on a logarithmic scale. This is done to reduce skew in the data and to stabilize comparisons across different salary tiers. Performance is summarized by Win Shares (WS) from the last season, which provides a single-season estimate of total contribution that incorporates both effectiveness on the court and playing time. Rather than using ratio-based measures such as pay per win share, which can produce extreme values when WS is small or close to zero, we define "overpay" as the residual derived from the baseline regression equation $\log(\text{AAV}) \sim \text{WS}$.

A positive residual in this context indicates that a player is paid more than would be expected given their WS, while a negative residual indicates they are underpaid relative to that expectation. This definition based on residuals yields a continuous and interpretable outcome variable that avoids instability driven by effects from low denominators. Sentiment is evaluated in relation to this baseline in two distinct ways. First, we compare a model based only on performance to an extended model that adds sentiment at the player level. This is estimated separately for Reddit and media sources, and separately for VADER and DeBERTa scoring systems. Second, we perform descriptive comparisons based on groups by labeling players as relatively overpaid or underpaid based on the ranks of their residuals and then comparing their aggregated sentiment scores. Throughout this process, the choice of sentiment model and the precision of attribution are treated as dimensions for robustness checking rather than as fixed design decisions, reflecting the uncertainty revealed by the exploratory analysis.

4 RESULTS AND ANALYSIS

Exploratory analysis indicates that the two text sources differ qualitatively. Media contributes far more sentences than Reddit and shows more frequent co-mentions, increasing attribution ambiguity. Both sources also exhibit strong long-tail coverage, where a small set of players accounts for most sentences (Figure 1, Appendix). As a result, player-level sentiment is more stable for high-coverage players but noisy for low-coverage ones. The analysis therefore treats sentence volume as a reliability constraint and evaluates Reddit and media separately rather than pooling them.

At the sentence level, VADER and DeBERTa yield meaningfully different score distributions, with only weak to moderate correspondence. Figure 3 (Appendix) shows a diffuse relationship, including many near-neutral DeBERTa scores where VADER assigns moderate polarity. This motivates the use of two separate sentiment scorers as parallel measurements: if a sentiment effect is real and sizable, it should be visible under both scoring families, while effects that appear only under a single model are more plausibly driven by idiosyncrasies in measurement.

The media attribution pipeline is largely driven by pronoun-based backfilling, with a smaller share assigned via explicit single-player matches (Figure 2, Appendix). In contract modeling, the baseline performance specification explains a meaningful share of salary variation, consistent with Win Shares capturing an important component of contract valuation.

However, adding sentiment does not yield consistent incremental explanatory power once Win Shares is included. Residual diagnostics make this most transparent: Figure 5 (Appendix) plots baseline residuals (interpreted as relative overpay) against mean player sentiment. In media, the fitted relationship is essentially flat with wide dispersion. In Reddit, the fitted trend is modestly positive but highly uncertain, and it does not persist across sentiment methods, consistent with the disagreement in Figure 3.

A descriptive comparison of players categorized as 'overpaid' or 'underpaid' based on their salary residuals provides further insight, as visualized in Figure 4 (Appendix). In the media dataset, the average sentiment scores for these two groups are nearly identical, suggesting that the formal media does not necessarily talk more positively about players who get big contracts. However, the Reddit data presents a slightly counterintuitive and weird pattern: players who were underpaid relative to their performance statistics actually had slightly lower, or more negative, average sentiment scores compared to those who were overpaid. This is the opposite of what one might expect if positive public sentiment strongly drove salaries upward, as the overpaid group should theoretically have the highest sentiment. The fact that the differences are small and inconsistently directional across sources reinforces the conclusion from the regression analysis that sentiment is not a reliable indicator of whether a player will receive a contract that exceeds performance expectations.

Overall, the results support a cautious negative finding: on-court performance strongly predicts contract size, while pre-signing sentiment from media and Reddit adds little stable information beyond that baseline. This is plausible given long-tail coverage (Figure 1), heuristic-heavy media attribution (Figure 2), and substantial measurement differences between sentiment models (Figure 3), all of which reduce the likelihood of detecting a subtle effect within a single offseason sample.

5 DISCUSSION AND OUTLOOK

This study examined whether public sentiment expressed before contract signings helps explain why some players are paid above or below what their performance on the court would predict. The results suggest that performance, as measured by Win Shares, is a strong determinant of salary, whereas features derived from sentiment add little consistent explanatory value once performance is accounted for. The absence of a strong sentiment effect is informative in itself. It indicates either that public narratives largely reflect information already embedded in performance metrics, or that any independent sentiment signal is too weak or too noisy to be detected reliably in this specific setting.

The exploratory analysis highlights several plausible sources of this weakening. Coverage is highly uneven across the player population, with sentiment estimates for many players based on relatively few sentences. Attribution is significantly more difficult in media text due to frequent co-mentions of multiple people and a heavy reliance on resolution based on pronouns. Furthermore, as noted in the Approach section, the uncertainty regarding the source domains of media articles adds another layer of potential noise to player identification. Because we searched a large list of websites that were never explicitly checked to be NBA-related, we may have matched names with people who are not actually the players in question. While Reddit data is guaranteed to be relevant because it comes from specific subreddits, the media data remains more chaotic.

Sentiment models also disagree substantially on the same sentences, particularly through DeBERTa's tendency toward neutrality and its more negative scoring of text from Reddit. In addition, upstream noise in entity matching—such as collisions between names or references to entities that are not players—can further dilute aggregates of sentiment at the player level. Despite all our attempts to tame the unstructured text data and organize it into a predictive model, the data was simply too chaotic to provide a clear answer. We essentially came out empty-handed, as any potential predicting power of a player's salary based on the sentiment around him was not statistically significant in either of our samples.

Future work could address these limitations in several ways. Improved entity resolution and handling of coreference would reduce noise related to attribution, especially in media sources. Expanding the analysis across multiple offseasons would increase statistical power and reduce sensitivity to idiosyncratic market conditions of a single year. Hierarchical models that explicitly incorporate uncertainty in sentiment estimates at the player level as a function of sentence volume could further stabilize inference. Finally, comparing sentiment to alternative signals off the court, such as measures of visibility or attention, may help disentangle whether it is sentiment itself, rather than public attention more broadly, that relates to contract outcomes.

REFERENCES

[1] Gokhan Ertug and Fabrizio Castellucci. 2013. Getting What You Need: How Reputation and Status Affect Team Performance, Hiring, and

Salaries in the NBA. *Academy of Management Journal* 56, 2 (2013), 407–431. doi:10.5465/amj.2010.1084

[2] C. J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, Vol. 8. AAAI Press, 216–225. doi:10.1609/icwsm.v8i1.14550

[3] Maksymilian Kloc, Mateusz Tomanek, and Wojciech Cieśliński. 2020. Social Media and the Value of Contracts Based on the Example of the NBA. *Journal of Physical Education and Sport* 20, 5 (2020), 3063–3069. doi:10.7752/jpes.2020.s5416

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 260–270.

[5] Qiwen Li, Jiarui Zhang, Jiayu Guo, Jiaqing Li, and Chenhao Kang. 2021. Evaluating Performance of NBA Players with Sentiment Analysis on Twitter Messages. In *Proceedings of the 2021 2nd European Symposium on Software Engineering (ESSE 2021)*. Association for Computing Machinery, New York, NY, USA, 150–155. doi:10.1145/3501774.3501796

[6] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 79–86.

[7] Zack Stice. 2020. Athlete Branding: Does Social Media Presence Impact NBA Players' On-Court Salaries? Samford University Center for Sports Analytics blog. https://www.samford.edu/sports-analytics/fans/2020/Athlete-Branding-Does-Social-Media-Presence-Impact-NBA-Players-On-Court-Salaries Published April 6, 2020. Accessed November 23, 2025.

[8] Rohan Tummala. 2023. Analyzing NBA Player Sentiment through Reddit. Samford University Center for Sports Analytics blog. https://www.samford.edu/sports-analytics/fans/2023/Analyzing-NBA-Player-Sentiment-through-Reddit Published August 4, 2023. Accessed November 23, 2025.

[9] Karl Valentini. 2020. *Transfer Pricing: An Analysis of the Impact of Player Brand Value on Transfer Fees in European Football*. Master's thesis. The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. https://repository.upenn.edu/joseph_wharton_scholars/94 Joseph Wharton Scholars undergraduate thesis.

APPENDIX

A WORK REPORT

I initiated the project by defining the research scope and identifying feasible data sources for measuring pre-signing public discourse. I compiled a dataset of 2025 NBA offseason signings with contract outcomes and prior-season performance and validated that relevant fan text could be collected via the Reddit API (PRAW). To improve coverage, I aggregated player aliases from Wikipedia. I then implemented a heuristic Reddit crawler that favored posts with fewer entity mentions to reduce noise from having too many entities in the sentences while still retaining multi-player discussions. In parallel, I used GDELT to query for relevant media coverage and crawled the resulting sources. I developed a custom processing pipeline to segment text into sentences and attribute each sentence to players using consistent heuristics across both fan and media corpora. After exploratory analysis, I computed sentiment using VADER and a DeBERTa-based model and aggregated the scores at player level for regression analysis, which ultimately did not yield statistically significant evidence of a relationship between either fan or media sentiment and NBA player salaries.

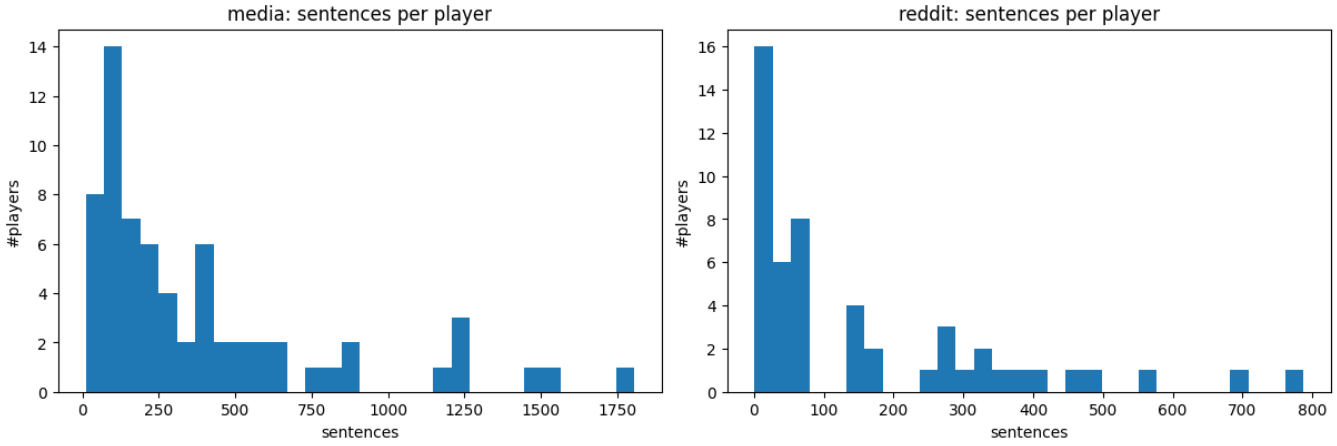

Figure 1: Sentence coverage by player for media (left) and Reddit (right). Both sources show long-tail coverage: a few players dominate the corpus while most have limited mentions. This imbalance motivates source-separated analysis and minimum-sentence thresholds.

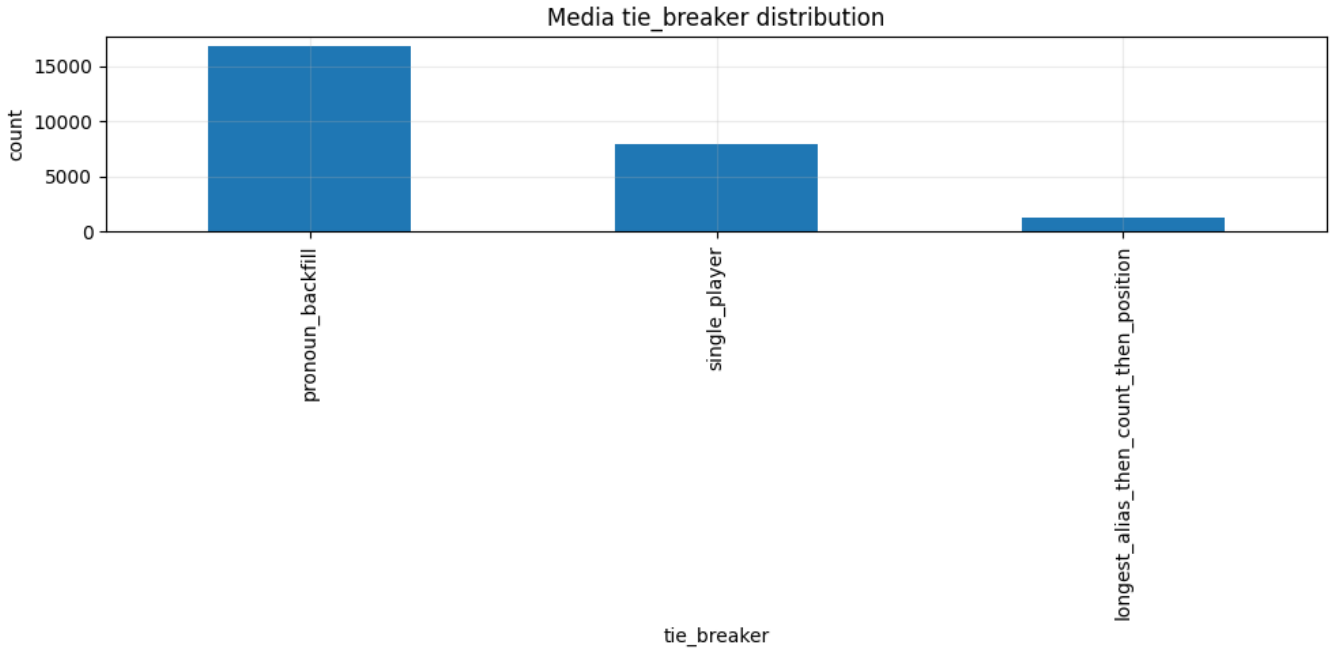

Figure 2: Media sentence attribution mechanisms. Pronoun-based backfilling dominates, followed by single-player matches; only a small subset of multi-player sentences rely on deterministic tie-breaking. This distribution underlines the importance of attribution precision checks.

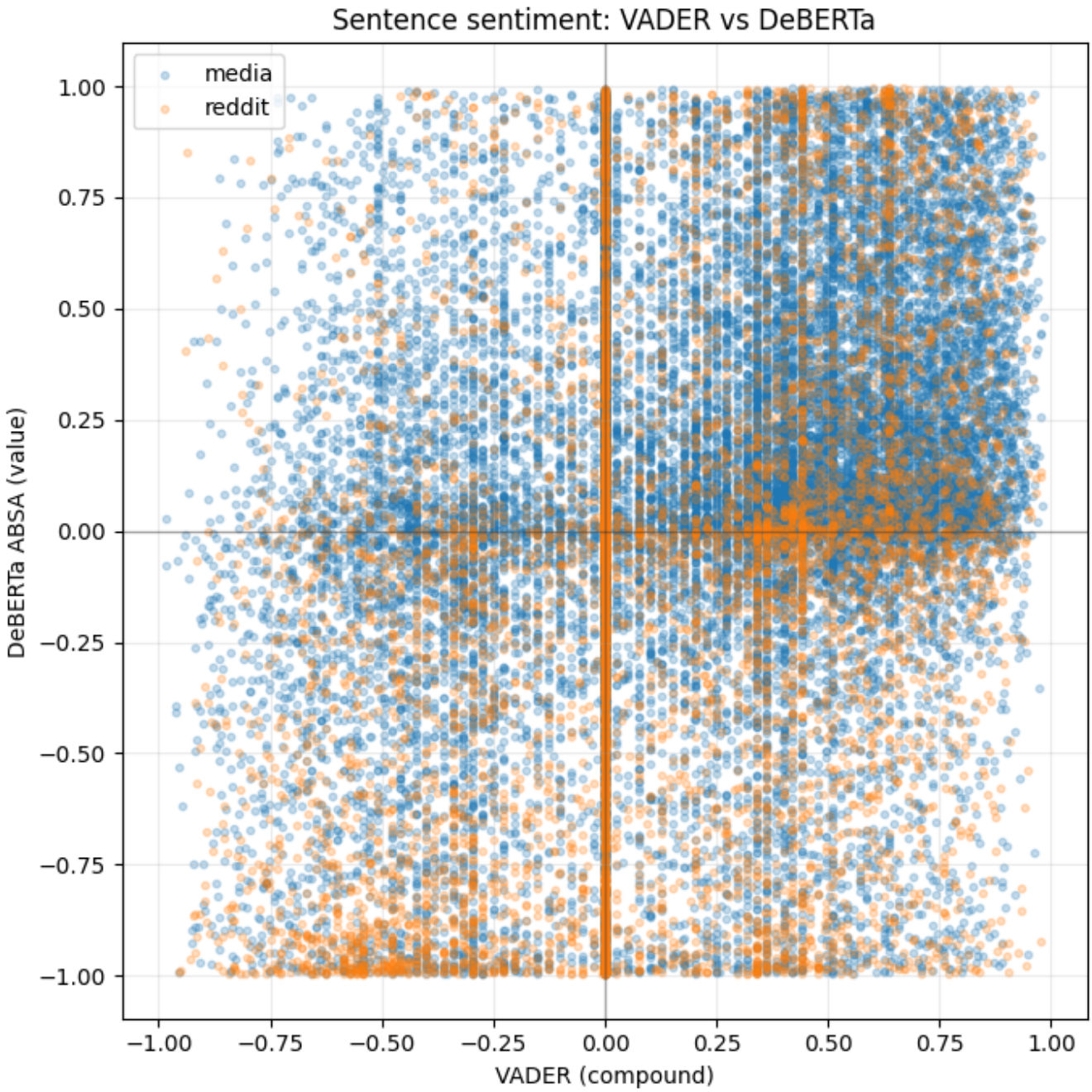

Figure 3: Sentence-level sentiment comparison between VADER (x-axis) and DeBERTa (y-axis), colored by source. The diffuse scatter and concentration near zero for DeBERTa highlight weak agreement and the model's neutrality bias.

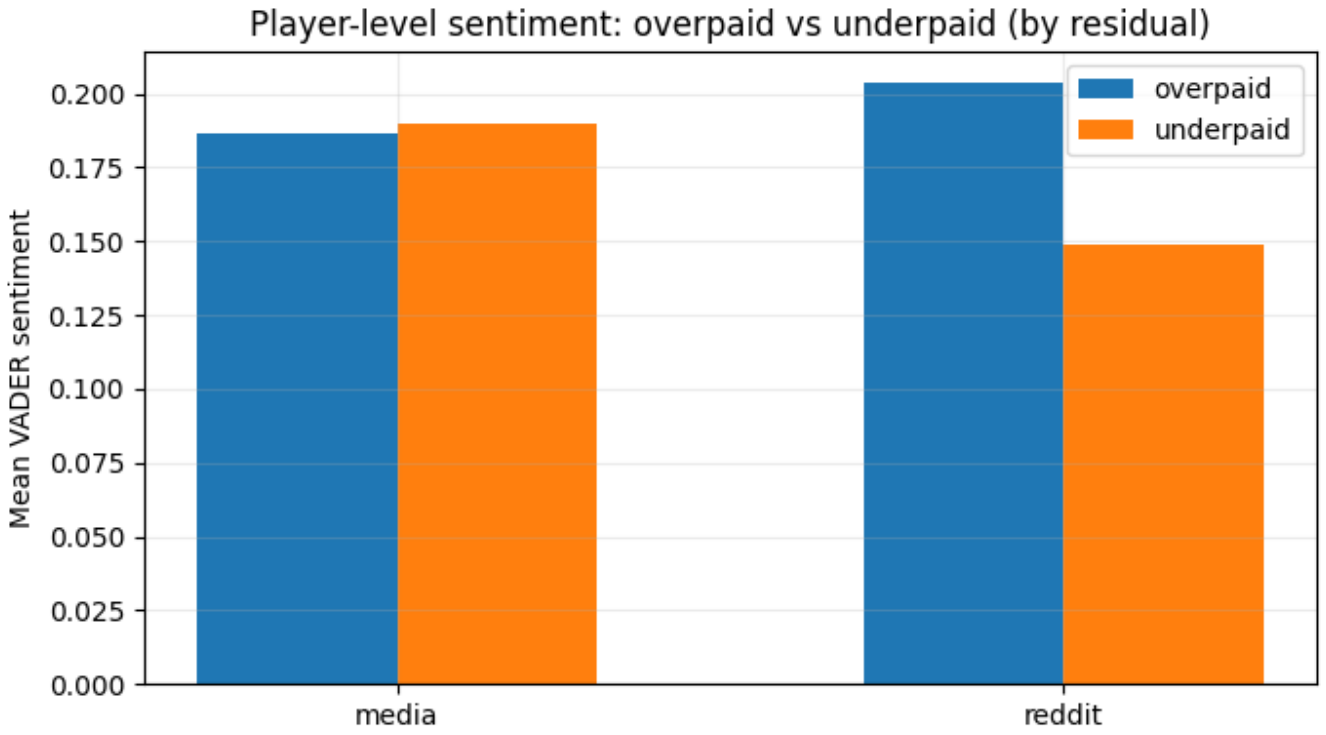

Figure 4: Mean VADER sentiment for overpaid and underpaid player groups (based on residuals from $\log(\text{AAV}) \sim \text{WS}$). Group means differ only slightly and inconsistently across sources, mirroring the regression result that sentiment is not a strong discriminator of contract deviation.

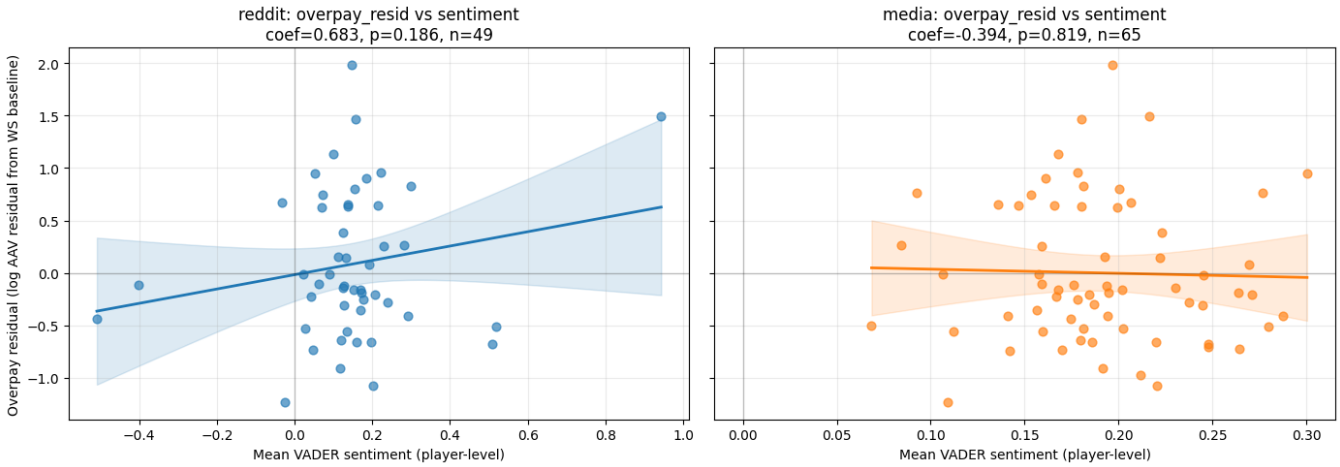

Figure 5: Scatterplots of overpay residuals versus mean VADER sentiment, shown separately for Reddit (left) and media (right). The fitted trends illustrate the negligible relationship between sentiment and residual salary once performance (WS) is controlled.