Kurdish Studies
April 2020
Volume: 08, No: 02, pp 385-394
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.67145/ks.v8i2.4140

How Grammatical Register Distinguishes Objective Reporting From Subjective Commentary in Sports Journalism

Luis Eduardo Muñoz Guerrero[1]

[1]Universidad Tecnológica de Pereira, Pereira, Colombia. Email: lemunozg@utp.edu.co. ORCID: https://orcid.org/0000-0002-9414-6187

Abstract

Sports journalism blends factual narration with editorial commentary, making it a useful domain for studying linguistic markers of objectivity and subjectivity. This study presents a comparative statistical analysis of linguistic style between objective and subjective sports articles, using the Sports Articles for Objectivity Analysis dataset (Rizk and Awad, 2018; 635 objective, 365 subjective articles). Welch's t-tests with Bonferroni correction, complemented by Mann-Whitney U tests, were applied to 37 length-normalised grammatical categories. Twenty-three categories showed statistically significant differences, with effect sizes from small to large (Cohen's d:0.1--1.1). Subjective articles were significantly longer than objective ones.

Keywords: journalistic objectivity, subjectivity, linguistic analysis, sports journalism, natural language processing.

1. Introduction

The distinction between objective and subjective journalistic content is a classic problem in communication studies and, more recently, in natural language processing (NLP). An objective article tends to report verifiable facts neutrally, while a subjective one incorporates opinions, evaluations, or a more personal tone from the writer. Automatically detecting this bias has practical applications, from news curation to assessing the informational quality of a media outlet, although it is not a trivial task: the high-precision classifiers proposed by Riloff and Wiebe to identify subjective sentences, for example, achieved 91.5% precision but at the cost of a recall of only 31.9%, illustrating the usual trade-off between these two metrics in this task.

This trade-off is not a mere technical curiosity: it reflects a deeper asymmetry in how subjectivity manifests in text. Precision-oriented approaches tend to rely on a comparatively small set of strongly polarised lexical or syntactic cues—first-person pronouns, evaluative adjectives, hedging modals—that reliably signal a subjective stance whenever they occur, but which are far from exhaustive, since many subjective sentences are subjective only in context (through implicature, irony, or comparison) rather than through an isolable marker. Recall-oriented approaches, conversely, tend to cast a wider net and consequently capture more of the phenomenon at the cost of false positives. This asymmetry motivates a complementary strategy to the purely predictive one dominant in the field: rather than optimising a classifier's decision boundary, one can ask which individual linguistic features differ most systematically, and by how much, between the two registers, treating the labelled corpus as an object of descriptive and inferential analysis in its own right rather than as training data for a downstream model.

Sports journalism is particularly interesting for this type of analysis. Unlike political coverage, where subjectivity tends to be ideologically charged and entangled with partisanship, in sports it tends to be expressed through enthusiasm, opinions about a team's or player's performance, or speculation about future results: linguistically "purer" phenomena that are less contaminated by politically confounding variables. A sports report on a match result and a sports column on the same event can differ enormously in register while describing, in principle, the same underlying event, which makes the domain a comparatively clean testbed for isolating stylistic markers of objectivity from markers of ideological alignment, the latter being much harder to control for in political text.

This work has a deliberately narrow scope. It does not aim to build an automatic objectivity/subjectivity classifier—a task already addressed by prior work on this same dataset, see Section 2—but rather to answer a simpler, more direct question:

Research question: *Are there statistically significant and systematic differences in the use of grammatical categories between objective and subjective sports articles?*

As a secondary question, we explore whether article length is related to its objective or subjective character. Taken together, these two questions are intended as a descriptive foundation: establishing which features differ and how much is a necessary precursor to any subsequent predictive work that might use those same features as inputs, and it is also of independent interest to communication scholars who are not primarily concerned with classification performance but with characterising register itself.

2. Related Work

The dataset used in this study was originally introduced by Rizk and Awad as a resource for objectivity/subjectivity classification tasks in sports text. Being a relatively recent resource within the period covered by this review (through December 2018), no peer-reviewed works were identified that directly use it within that timeframe; non-academic tutorials do exist (for example, Medium blog posts) that train simple classifiers (SVM) on the same variables, reporting an accuracy of approximately 82–83%. This figure is consistent with what is reported in the classic literature on binary text classification: Pang, Lee, and Vaithyanathan, in their pioneering work on polarity classification of movie reviews, found that a manually constructed word lexicon achieved only ~70% accuracy, while trained automatic classifiers (Naive Bayes, maximum entropy, SVM) reached a range of 70% to the low 80s, consistently outperforming approaches based on predefined word lists. That early result is often cited as evidence that sentiment and subjectivity are not reducible to the presence or absence of a fixed set of "loaded" words, since a human-curated lexicon underperformed a statistical model trained on distributional patterns; the present study follows a related logic in that it treats grammatical rates, rather than individual lexical items, as the unit of comparison.

Beyond this particular dataset, linguistic bias detection has been studied in other textual domains. Recasens, Danescu-Niculescu-Mizil, and Jurafsky analysed real edits made by Wikipedia editors to remove bias from articles, distinguishing between framing bias (the choice of a word that presupposes a contested viewpoint) and epistemological bias (the choice of a word that subtly undermines or bolsters a proposition's credibility); in their study, external human annotators correctly identified the word causing the bias in only 37% of cases, underscoring how subtle this type of linguistic marker can be even for human readers, and reinforcing the value of a systematic statistical analysis such as the one proposed here over mere reader intuition. A related approach, although lexicon-based rather than relying on statistical tests over grammatical categories, is that of Wilson, Wiebe, and Hoffmann, who proposed first determining whether an expression is neutral or polar and only then disambiguating its orientation, obtaining results significantly better than baseline; this distinction between neutrality and polarity is analogous to the problem faced here by the dataset's raw semantic scores (semanticobjscore, semanticsubjscore), which turned out to be confounded with article length (see Section 5).

A further, older strand of work is concerned less with subjectivity as a text-classification target and more with objectivity as a professional practice. Sociologists of journalism have long argued that objectivity is not a property a text simply "has" but a set of ritualised procedures that reporters follow in order to be able to claim it (see Section 6 and the discussion of Tuchman's work). This perspective is complementary rather than competing: it suggests that some of the grammatical patterns identified through purely statistical means in the present study—reliance on modal hedging, preference for direct quotation over first-person assertion—may correspond to concrete newsroom conventions rather than to an emergent, unintentional stylistic drift.

In all of the above cases the goal was predictive: to build a model or lexicon that classifies new text. None focuses on a descriptive/inferential statistical analysis of which grammatical categories distinguish the two groups, nor do they report effect sizes or robustness tests on individual variables, which is precisely the approach adopted here. At the time of writing, the UCI repository itself reports 0 formal citations for the dataset, suggesting that it remains an academically underexploited resource, and that a purely descriptive characterisation of its linguistic structure—independent of any classifier built on top of it—has, to the best of our knowledge, not previously been published.

2. Data Source

Field	Value
Name	Sports Articles for Objectivity Analysis
Authors	Yara Rizk & Mariette Awad (American University of Beirut)
Repository	UCI Machine Learning Repository
Donation date	8 April 2018
DOI	10.24432/C5801R
License	Creative Commons Attribution 4.0 International (CC BY 4.0)

Field	Value
Link	https://archive.ics.uci.edu/dataset/450/sports+articles+for+objectivity+analysis
Instances	1000 sports articles
Labelling	Objective (n = 635) / Subjective (n = 365), via Amazon Mechanical Turk
Variables	59 (includes raw text, URL, and 56 numeric variables)
Missing values	None (confirmed by UCI and by this study)

Suggested citation:
Rizk, Y. & Awad, M. (2018). Sports articles for objectivity analysis [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5801R

Before the analysis, data integrity was verified: the 1000 records were checked on the following aspects:

- Duplicates by TextID: 0.
- Duplicates by URL: 2 (two articles share a source URL; possibly a republication or scraping error in the original dataset).
- Full-row duplicates (all numeric variables identical): 5.
- Null values: 0, consistent with what UCI reports.

These seven potentially problematic rows represent 0.7% of the sample. They were not removed from the main analysis, in order to maintain comparability with future studies using the full dataset, but are documented here as a relevant data-quality observation for any subsequent work.

This dataset was selected over more widely used subjectivity corpora (for example, generic movie-review or newswire subjectivity datasets) for two reasons. First, its domain specificity — sports journalism — allows the research question to be posed without the ideological confound that affects political text, as discussed in Section 1. Second, its near-total absence of prior citation, noted above, meant that a systematic descriptive characterisation of its linguistic structure had genuine value independent of any predictive application, rather than merely replicating an already well-documented analysis on a well-studied corpus.

Turning to the variables used, the dataset includes 37 grammatical frequency variables based on the Penn Treebank Project tags, extracted using the Stanford POS Tagger (for example: NN = singular common nouns, MD = modal verbs, PRP$ = possessive pronouns, etc.), as well as punctuation variables (Quotes, questionmarks, exclamationmarks, fullstops, commas, semicolon, colon, ellipsis), grammatical-person variables (pronouns1st, pronouns2nd, pronouns3rd), and two raw semantic scores (semanticobjscore, semanticsubjscore) based on SentiWordNet, the lexical resource that assigns each WordNet synset a numeric score of objectivity, positivity, and negativity.

4. Methodology

As a preprocessing step, and since article length varies considerably (from 47 to 4283 words), the absolute frequencies of each grammatical category are not comparable across articles. For this reason, each count variable X was transformed into a normalised rate:

$$X_{\text{rate}} = \left(\frac{X}{\text{totalWordsCount}}\right) \times 100$$

interpreted as "occurrences of X per 100 words of the article."

For the group comparison itself, the mean of each of the 37 normalised grammatical categories was compared between the objective group (n=635) and the subjective group (n=365) using:

1.Welch's t-test, which does not assume equal variances between groups — appropriate given the imbalance in sample sizes between groups. The statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2.Effect size (Cohen's d), calculated as the difference in means divided by the pooled standard deviation. It supports interpretation follows the conventions proposed by Cohen: 0.2 (small), 0.5 (medium), and 0.8 (large).

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}}$$

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

3.Mann-Whitney U test, as a non-parametric backup, since a normality check (Shapiro-Wilk test) on the variables with the largest effect rejected the normality assumption in most cases ($p<0.001$ in 4 of 5 variables reviewed).

A correction for multiple comparisons was also required: since 37 simultaneous tests were performed, the probability of obtaining false positives by chance increases considerably. Bonferroni correction was applied (adjusted $\alpha=0.05/37\approx0.00135$) to both the t-test and Mann-Whitney results. This correction is conservative by design — it controls the family-wise error rate rather than the false discovery rate — which means that, if anything, the analysis is biased towards under-reporting rather than over-reporting significant categories; a category that survives Bonferroni correction under this scheme is unlikely to be a false positive.

As a secondary analysis, total length (totalWordsCount) was also compared between both groups using an additional Welch's t-test, and the Pearson correlation between length and the dataset's raw semantic scores (semanticobjscore, semanticsubjscore) was calculated, in order to verify whether these scores are confounded with text length.

A robustness check was performed as well: since manual inspection of individual articles revealed implausible values in some grammatical categories (see Section 7), the group comparison was repeated using only the punctuation variables (Quotes, questionmarks, exclamationmarks, fullstops, commas, semicolon, colon, ellipsis). Their extraction is mechanical (character counting) and therefore much less susceptible to tagging errors than the grammatical categories, which depend on a probabilistic tagger. This allows evaluating whether the general pattern of differences holds when restricted to high-reliability variables.

As for the tools used, all analysis was carried out in Python 3.12, using pandas for data manipulation and scipy.stats for the statistical tests. The complete, reproducible script is included as supplementary material (analisis_ttest.py).

5. Results

As a general descriptive statistic, the following table summarises article length by group:

	Objective (n=635)	Subjective (n=365)
Mean length (words)	519.2 (SD = 417.7)	1006.6 (SD = 544.5)
Minimum length	47	76
Maximum length	2981	4283
Median	389	930

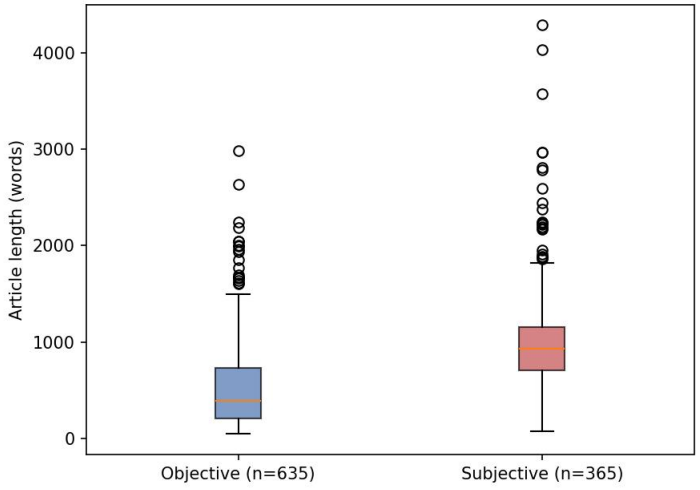

Figure 1. Distribution of article length (word count) by group. Circles represent outlier values.

Both distributions are right-skewed, as indicated by the gap between the mean and the median in each group and confirmed visually in Figure 1; a small number of unusually long articles pull the mean above the median in both cases, which is why the group comparison below additionally reports Cohen's d alongside the t-statistic, since d is less sensitive than a raw mean-difference to this kind of skew when the two groups share a broadly similar distributional shape.

Turning to the main result, differences in grammatical categories: of the 37 grammatical categories evaluated, 23 showed statistically significant differences after Bonferroni correction using Welch's t-test. The Mann-Whitney U test, more conservative given the lack of normality, identified 25 significant categories, with only 2 discrepancies relative to the t-test (RBR and WP, not significant in the t-test but significant in Mann-Whitney): a 94.6% agreement between both methods, supporting the robustness of the main finding regardless of the statistical test chosen.

Table 1. Grammatical categories with the largest effect size (ordered by |Cohen's d|).

Category	Description	Obj. mean	Subj. mean	Cohen's d	p (Bonferroni)
PRP$	Possessive pronouns	3.34	5.00	−1.12	4.3×10^{-55}
MD	Modal verbs	25.29	20.30	1.01	1.1×10^{-52}
UH	Interjections	0.40	0.73	−0.82	5.9×10^{-29}
POS	Genitive marker	2.68	4.02	−0.75	5.5×10^{-26}
VBP	Present tense verb (3rd pers. plural)	1.95	2.88	−0.72	1.6×10^{-25}
WP$	Possessive interrogative pronoun	0.33	0.58	−0.72	2.6×10^{-23}
LS	List marker	0.74	1.21	−0.72	8.9×10^{-27}
VBN	Past participle verb	1.28	1.96	−0.70	1.1×10^{-22}
VB	Base form verb	4.96	3.61	0.67	8.1×10^{-24}
baseform	Infinitive verb (preceded by "to")	2.54	3.11	−0.64	9.98×10^{-22}

(Full table of all 37 comparisons, including means, standard deviations, t-statistic, and both p-values, available in ttest_results.csv, mannwhitney_results.csv, and in the accompanying tables_objectivity_analysis.xlsx workbook.)

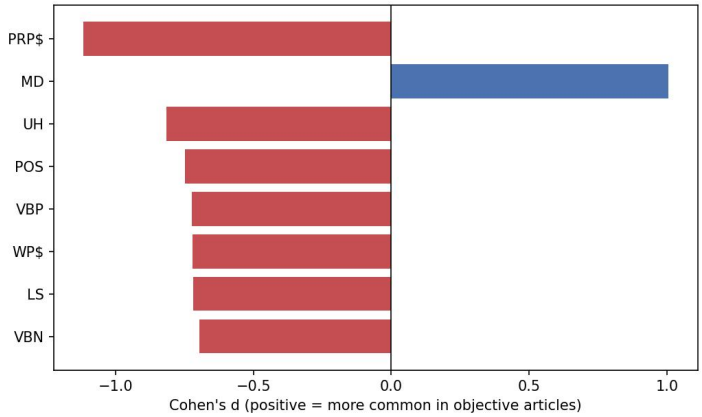

Figure 2. Grammatical features with the largest effect size (Cohen's d); positive indicates greater presence in objective articles.

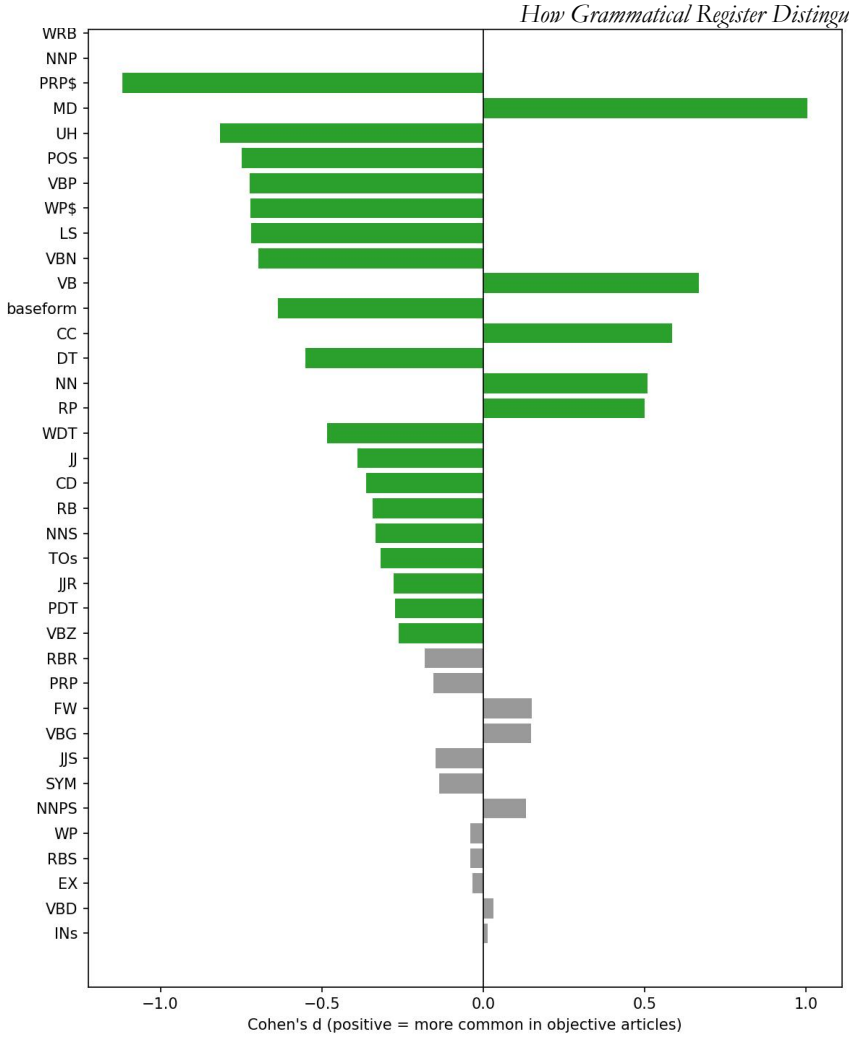

Figure 3. Effect size (Cohen's d) of the 37 grammatical categories evaluated. Categories in green are statistically significant after Bonferroni correction.

Multicollinearity among the main variables was checked next: the correlation matrix among the eight features with the largest effect size (Table 1) was examined to assess whether they capture redundant signals.

Table 3. Pearson correlation matrix among the normalised rates of the eight features with the largest effect size.

	PRP$	**MD**	**UH**	**POS**	**VBP**	**WP$**	**LS**	**VBN**
PRP$	1.00	-0.64	0.39	0.56	0.26	0.40	0.27	0.40
MD	-0.64	1.00	-0.39	-0.57	-0.20	-0.40	-0.27	-0.39
UH	0.39	-0.39	1.00	0.38	0.23	0.29	0.26	0.30
POS	0.56	-0.57	0.38	1.00	0.39	0.49	0.28	0.45
VBP	0.26	-0.20	0.23	0.39	1.00	0.22	0.27	0.32
WP$	0.40	-0.40	0.29	0.49	0.22	1.00	0.18	0.28
LS	0.27	-0.27	0.26	0.28	0.27	0.18	1.00	0.29
VBN	0.40	-0.39	0.30	0.45	0.32	0.28	0.29	1.00

The most notable correlations are:

●PRP$ and MD: moderate-to-high negative correlation (r=−0.64).
●PRP$ and POS: moderate positive correlation (r=0.56).
●The remaining pairs show low-to-moderate correlations (|r| between 0.18 and 0.49).

None reaches a level suggesting full redundancy (r>0.9), so the variables contribute partially independent information about linguistic style, although they share some common signal, probably related to a latent "formality/personalisation" factor of the text. The full matrix is included in correlacion_top8.csv.
A further robustness analysis focused on punctuation variables. Repeating the analysis using only punctuation variables—whose extraction is mechanical and was manually verified against the raw text—showed the same general pattern of significant differences:

Table 2. Differences in punctuation variables (rate per 100 words).

Variable	Obj. mean	Subj. mean	Cohen's d	p-value
Quotes	1.06	0.35	0.67	4.0×10^{-28}
questionmarks	0.04	0.21	−0.92	2.6×10^{-28}
fullstops	5.25	4.90	0.25	6.1×10^{-5}
commas	3.66	4.05	−0.26	3.1×10^{-5}
colon	0.31	0.21	0.20	3.3×10^{-4}
exclamationmarks	0.02	0.05	−0.19	1.6×10^{-2}
semicolon	0.02	0.05	−0.15	1.1×10^{-2}
ellipsis	0.0004	0.0018	−0.09	0.32 (n.s.)

Objective articles quote more often (more quotation marks: Quotes), consistent with using statements from players and coaches as factual support, while subjective articles use considerably more question marks, consistent with a style that poses rhetorical or speculative questions ("can the team come back?"). This pattern, obtained with high-reliability variables, corroborates the general direction of the main findings obtained with the POS tagger's grammatical categories.
As a secondary finding, article length itself differed sharply between groups: subjective articles were significantly longer than objective ones (t=−14.78, p<0.0001; Cohen's d≈1.0, a large effect). Furthermore, the dataset's raw semantic scores correlate strongly with article length (semanticobjscore: r=0.96; semanticsubjscore: r=0.89; both p<0.001), which confirms that these scores are absolute counts not normalised by length and reinforces the methodological decision to work with normalised rates instead of using these variables directly.

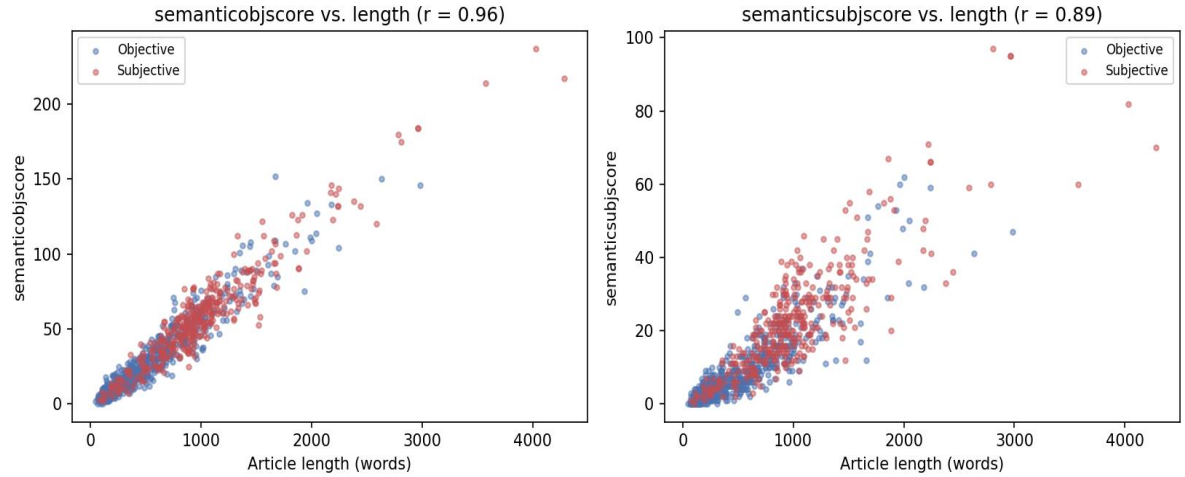

Figure 4. Original dataset's raw semantic scores as a function of article length, by group. The near-linear relationship confirms that both scores are, in practice, a proxy for text length.

6. Discussion

The findings are, in general, consistent with linguistic intuition about the objective versus subjective register:

●**Subjective articles** use significantly more possessive pronouns (PRP$) and interjections (UH), features associated with a personal and expressive style, and more question marks, consistent with posing rhetorical questions or speculating about future results: a common editorial device in sports commentary. This link between pronoun frequency and personal style is not coincidental: Pennebaker and King showed, based on thousands of writing samples, that the frequency of word categories such as pronouns constitutes a consistent and psychologically meaningful marker of linguistic style, and Newman, Pennebaker, Berry, and Richards found that the frequency of pronominal self-references can distinguish communicative styles with an accuracy of 61–67%, providing empirical support for using pronouns1st/2nd/3rd and PRP$ as valid markers of authorial stance.

●**Objective articles** use more modal verbs (MD) and more direct quotes (Quotes), consistent with a narrative that relies on statements from sources (players, coaches) and on constructions that express possibility or necessity in measured terms ("could," "should") rather than categorical statements of opinion. This pattern aligns with what the sociology of journalism has described for decades: Tuchman identified precisely the practice of quoting others instead of offering one's own opinion as one of the ritual procedures through which journalists claim objectivity for their work, alongside the presentation of conflicting possibilities and the structured use of quotation marks — a "strategic ritual" that protects the reporter from accusations of bias by attributing evaluative content to a named source rather than asserting it directly. Read against that framework, the elevated modal-verb rate in objective articles is not merely a stylistic tic but a linguistic trace of that same ritual: hedged, source-attributed claims ("the coach said the team could improve") substitute for direct evaluative assertions.

●**That subjective articles are, on average, nearly twice as long** suggests that subjectivity in this corpus is not limited to the use of certain function words, but is accompanied by greater argumentative development: those who offer an opinion elaborate more than those who simply report a result. This is broadly consistent with the intuitive expectation that argumentation — building a case for why a team's performance was disappointing, or why a player deserves criticism or praise — requires more textual space than the comparatively compressed act of reporting a final score or a sequence of events.

These differences are statistically highly significant, which is expected given the sample size, but their effect sizes range from small (ellipsis, semicolon) to large (PRP$, MD, UH). The practical relevance of each feature should be interpreted with this magnitude in mind, not only its statistical significance. A further point worth noting is that none of the individual features, however large its effect size, would be sufficient on its own to reliably separate the two classes at the level of an individual article; the value of the present analysis lies in identifying a coherent bundle of markers that move together in the expected direction, rather than in proposing any single feature as a stand-alone diagnostic of subjectivity.

It is also worth situating this finding relative to the classification literature reviewed in Section 2. Studies that build predictive classifiers on this kind of feature typically report the relative importance of individual variables only indirectly, through model coefficients or feature-importance scores that are contingent on the particular algorithm used and on interactions with other features in the model. The present analysis instead reports, for each of the 37 categories independently, a directly interpretable quantity — the standardised mean difference between groups — together with an explicit correction for the fact that 37 comparisons were run simultaneously. This makes the ranking in Table 1 comparatively robust to the choice of downstream model: whichever classifier a future study builds on this corpus, PRP$, MD, UH, and the other high-ranking categories identified here are likely to remain among its most informative inputs, since their separation between groups does not depend on any particular modelling assumption beyond those of the two statistical tests applied.

7. Limitations

1.Quality of automatic grammatical tagging. When manually inspecting individual articles against their grammatical counts (for example, Text0001.txt, a 95-word article with no modal verbs in the actual text but with MD = 29 recorded), clearly implausible values were observed in several categories: counts of nouns and adjectives close to zero in texts where they clearly appear, and elevated counts in rare categories such as NNPS or MD. It was verified, cell by cell against the original XML file, that this does not correspond to a processing error in this study, but rather to a pre-existing artefact, probably from the automatic POS tagging pipeline (Stanford POS Tagger) used by the original authors in 2018. It is worth noting that even the reported performance of that same tagger under ideal conditions is 97.24% word-level accuracy on the Penn Treebank corpus; some margin of error is therefore expected, but the magnitude of the cases detected here (for example, MD = 29 in a text with no modal verbs) clearly exceeds that normal margin and points to a specific artefact rather than typical statistical noise. For this reason, the robustness analysis reported in Section 5 was restricted to manually verified punctuation variables, which showed the same general pattern: this mitigates, but does not eliminate, the limitation.

2.Label origin. The objective/subjective labels were assigned via crowdsourcing (Amazon Mechanical Turk), a method subject to the annotator's own subjectivity and, as far as could be verified, without any publicly reported inter-annotator agreement metric for this specific dataset. As a general reference on the reliability of this method, Snow, O'Connor, Jurafsky, and Ng evaluated non-expert Mechanical Turk annotations against expert reference labels across five different NLP tasks, finding high agreement in all five; this provides indirect—not dataset-specific—support for the viability of crowdsourcing for linguistic annotation tasks such as the one used here.

3.Minor duplicates. 2 duplicate URLs and 5 rows with identical numeric values were identified (0.7% of the sample). They were not excluded from the analysis, but could slightly inflate or distort the results if they correspond to genuinely repeated articles.

4.Thematic and language scope. The corpus is composed exclusively of English-language sports articles collected through 2018. The findings should not be generalised without further validation to other journalistic domains, languages, or

text produced after that date, including AI-generated or AI-assisted text, a phenomenon that postdates the construction of the dataset.

5.Correlational/descriptive nature. This study identifies statistical associations, not causal relationships. It cannot be concluded that using a possessive pronoun "causes" subjectivity; it simply co-occurs with it in this corpus.

8. Future Work

- Repeat the punctuation-based robustness analysis reported in Section 5, extending it to a subsample of articles re-tagged with a modern POS tagger (e.g., spaCy or a transformer-based model) to more precisely quantify the impact of the tagging artefact identified in Section 7. A direct comparison between the original Stanford-tagger counts and counts obtained from a contemporary tagger on the same raw text would allow the residual uncertainty flagged in that section to be bounded empirically rather than only qualitatively.
- Explore whether the differences found hold when segmenting by specific sport (football, basketball, etc.), since the dataset does not explicitly include this variable but it could be inferred from the URLs or the text. Different sports carry different journalistic conventions (for instance, individual combat sports may invite more evaluative commentary on a competitor's character than team sports), so this segmentation could reveal whether the present findings generalise across sub-genres of sports writing or are partly driven by the composition of the corpus.
- Extend the descriptive analysis towards a simple predictive model (e.g., logistic regression using the punctuation variables, which proved reliable) as a natural next step following this exploratory work, using the present study's effect-size ranking to motivate feature selection rather than treating all 37 categories as equally informative inputs.
- Replicate the methodology on a more recent corpus to evaluate whether the linguistic patterns identified remain stable over time, particularly given the emergence, since 2018, of large-scale AI-assisted sports writing, which may exhibit a different or more homogeneous stylistic signature than the human-authored articles analysed here.
- Investigate inter-annotator agreement directly, by having a small subsample of the corpus independently re-labelled by multiple annotators, to obtain a dataset-specific estimate of labelling reliability that is currently unavailable (see Limitation 2).

9. Conclusion

This study, of deliberately narrow scope, found robust statistical evidence—confirmed by two distinct tests and by an independent robustness analysis on high-reliability variables—that systematic differences exist in linguistic style between objective and subjective sports articles, both in the use of specific grammatical categories (possessive pronouns, modal verbs, interjections, punctuation marks) and in text length. These findings are consistent with the general literature on linguistic markers of subjectivity and constitute a simple but solid empirical basis for subsequent descriptive or predictive studies in this same domain.

Beyond its immediate empirical contribution, this analysis illustrates a broader methodological point: a dataset originally built for a predictive task can still yield useful descriptive knowledge when subjected to inferential statistics rather than machine-learning evaluation alone. The 37 grammatical rates examined here were, in this dataset, engineered as classifier inputs; treating them instead as the outcome variables of a group comparison surfaced both a substantive finding (the coherent bundle of markers separating objective from subjective sports writing) and a methodological caution (the tagging artefact discussed in Section 7) that a purely predictive evaluation, focused only on classification accuracy, would have had no particular reason to expose. For editors, style-guide authors, and journalism educators, the practical takeaway is comparatively modest but concrete: objectivity in sports writing, at least in this corpus, is not achieved through the avoidance of any single word class but through a cluster of interacting choices — fewer possessive pronouns and interjections, more modal hedging, more direct quotation, and, on average, a shorter and more compressed text.

References

1. Cohen, Jacob. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
2. Dunn, Olive Jean. "Multiple Comparisons among Means." Journal of the American Statistical Association 56, no. 293 (1961): 52–64.
3. Esuli, Andrea, and Fabrizio Sebastiani. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining." In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), 417–422. Genoa: European Language Resources Association, 2006.
4. Mann, Henry B., and Donald R. Whitney. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." Annals of Mathematical Statistics 18, no. 1 (1947): 50–60.
5. Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. "Lying Words: Predicting Deception from Linguistic Styles." Personality and Social Psychology Bulletin 29, no. 5 (2003): 665–675.
6. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, 79–86. Stroudsburg, PA: Association for Computational Linguistics, 2002.
7. Pennebaker, James W., and Laura A. King. "Linguistic Styles: Language Use as an Individual Difference." Journal of Personality and Social Psychology 77, no. 6 (1999): 1296–1312.
8. Recasens, Marta, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. "Linguistic Models for Analyzing and Detecting Biased Language." In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1, Long Papers, 1650–1659. Sofia: Association for Computational Linguistics, 2013.

9. Riloff, Ellen, and Janyce Wiebe. "Learning Extraction Patterns for Subjective Expressions." In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 105–112. Stroudsburg, PA: Association for Computational Linguistics, 2003.
10. Rizk, Yara, and Mariette Awad. "Sports Articles for Objectivity Analysis" (dataset). UCI Machine Learning Repository, 2018. https://doi.org/10.24432/C5801R.
11. Shapiro, Samuel S., and Martin B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." Biometrika 52, no. 3–4 (1965): 591–611.
12. Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and Fast—But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 254–263. Stroudsburg, PA: Association for Computational Linguistics, 2008.
13. Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 252–259. Stroudsburg, PA: Association for Computational Linguistics, 2003.
14. Tuchman, Gaye. "Objectivity as Strategic Ritual: An Examination of Newsmen's Notions of Objectivity." American Journal of Sociology 77, no. 4 (1972): 660–679.
15. Welch, Bernard L. "The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved." Biometrika 34, no. 1–2 (1947): 28–35.
16. Wiebe, Janyce, Theresa Wilson, and Claire Cardie. "Annotating Expressions of Opinions and Emotions in Language." Language Resources and Evaluation 39, no. 2–3 (2005): 165–210.
17. Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis." In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 347–354. Stroudsburg, PA: Association for Computational Linguistics, 2005.