Kurdish Studies
March 2019
Volume: 07, No: 01, pp 121-134
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.67145/ks.v7i1.4134

A Statistical Profile of Linguistic Differences Between Objective and Subjective Sports Articles

Luis Eduardo Muñoz Guerrero[1]
[1]Universidad Tecnológica de Pereira. Email: lemunozg@utp.edu.co. ORCID ID : https://orcid.org/0000-0002-9414-6187

Abstract

Automated subjectivity detection in sports journalism has been studied since the early 2010s, yet the dataset most closely associated with this task — the Sports Articles for Objectivity Analysis dataset, released by Rizk and Awad in 2018 through the UCI Machine Learning Repository — remains largely unexplored beyond the work of its original creators. This paper presents a descriptive statistical analysis of the 59 syntactic and semantic features included in the dataset (N = 1,000 articles), comparing their distributions between articles labeled objective (n = 635) and subjective (n = 365). Non-parametric hypothesis testing (Mann-Whitney U) together with effect size estimation (rank-biserial correlation) identifies which linguistic markers differ significantly between the two classes, using Benjamini-Hochberg correction for multiple comparisons.

Of the 59 features tested, 53 (89.8%) show statistically significant differences. However, we identify an important confound: subjective articles are, on average, nearly twice as long as objective ones (1,006.6 vs. 519.2 words), which inflates the apparent effect size of virtually every raw frequency-count feature. After re-testing all features as length-normalized rates, 48 features (81.4%) remain significant, with possessive pronoun density, second-person pronoun usage, and imperative verb density emerging as the most robust discriminators once length is controlled for.

A logistic regression classifier trained on the statistically significant features, evaluated with nested 5-fold stratified cross-validation (feature selection re-run within each training fold to avoid data leakage), achieves 82.3% accuracy (F1 = 0.732) using raw features and 83.5% accuracy (F1 = 0.767) using length-normalized features, both well above the 63.5% majority-class baseline (a paired test across folds found this raw-vs-normalized gap statistically indistinguishable from noise, p = .19). A length-only control model (using article word count as the sole predictor) reaches 72.3% accuracy on its own, indicating that roughly 45% of the full model's improvement over baseline is attributable to article length alone rather than genuine per-word stylistic signal.

We further find severe multicollinearity among the raw features (66% with VIF ≥ 5, including several apparently redundant feature pairs), which is substantially reduced by length normalization and which cautions against interpreting individual logistic-regression coefficients as independent effects. These findings offer an accessible, non-technical reference point for researchers considering this dataset, and demonstrate that a resource with almost no citation history can still support a small, self-contained, and reproducible empirical contribution — including a genuine methodological caveat (the length confound) that prior classification-focused studies on this dataset did not report.

Keywords: subjectivity classification, sports journalism, text mining, linguistic features, descriptive statistics, natural language processing

1. Introduction

The proliferation of online sports journalism has made it increasingly difficult for readers to distinguish factual reporting from opinion-laden commentary. This distinction matters not only for media literacy but also for downstream applications such as sports betting analytics, automated news aggregation, and content moderation, where the reliability of a source depends in part on whether it presents information objectively or subjectively. Detecting subjectivity in text is a long-standing task in natural language processing, but sports writing presents a particular challenge: it frequently blends factual match reporting with evaluative language (e.g., praise, criticism, speculation about player performance), making the boundary between objective and subjective content less clear-cut than in other domains such as product reviews or political commentary.

In 2012, Rizk and Awad introduced one of the first datasets specifically constructed for this task, consisting of sports articles manually labeled as objective or subjective. This dataset was later extended and formally released in 2018 through the UCI Machine Learning Repository under the name Sports Articles for Objectivity Analysis (Rizk & Awad, 2018), comprising 1,000 articles and 59 pre-extracted syntactic and semantic features derived from part-of-speech tagging, punctuation counts, pronoun usage, and text complexity metrics. Despite being freely available for several years, the dataset has seen minimal use outside of the original research group, and no study to date has examined it from a purely descriptive statistical perspective, despite this being a natural and low-cost first step for understanding what the extracted features actually capture.

This paper addresses that gap. Rather than proposing a new classification algorithm, we ask a simpler question: which of the 59 linguistic features in this dataset actually differ between objective and subjective sports articles, and by how much? In the course of answering this question, we uncovered a methodological issue that, to our knowledge, has not been reported in

prior work on this dataset: subjective and objective articles differ substantially in raw length, which confounds any analysis based on raw frequency counts. We therefore extend the analysis with a length-normalized robustness check, and complement both analyses with a minimal predictive check — a logistic regression model restricted to the statistically significant features — to verify that the identified differences are not merely statistically significant but also practically informative.

The contribution of this paper is intentionally modest in scope. We do not claim state-of-the-art classification performance, nor do we introduce novel feature engineering. Instead, we provide a transparent, reproducible statistical baseline that can serve as a reference point for future work on this dataset, and we surface a length-confound caveat that should inform how future studies use these particular 59 features.

The remainder of this paper is organized as follows. Section 2 reviews related work on subjectivity classification in sports journalism and situates this study relative to prior use of the dataset. Section 3 describes the dataset. Section 4 details the statistical methodology. Section 5 presents the results. Section 6 discusses their implications and limitations. Section 7 concludes.

2. Related Work

Subjectivity classification is a well-established subtask of sentiment analysis, distinct from polarity detection in that it aims to determine whether a piece of text expresses a personal opinion or judgment, rather than whether that opinion is positive or negative (Pang, Lee, & Vaithyanathan, 2002). A more elaborate, span-level treatment of this distinction is offered by the MPQA framework, which defines subjectivity in terms of annotatable "private states" — opinions, emotions, and evaluations expressed toward a target — and provides one of the field's foundational annotated corpora for the task (Wiebe, Wilson, & Cardie, 2005). Early approaches relied on lexicon-based scoring methods, in which sentiment or subjectivity scores were assigned at the word level and aggregated at the document level (Heerschop et al., 2011). Later work shifted toward supervised machine learning approaches using bag-of-words, syntactic, and semantic features combined with classifiers such as Support Vector Machines and Naïve Bayes — an approach exemplified by Pang and Lee (2004), who trained an SVM-based subjectivity classifier on a corpus of subjective (movie review) and objective (plot summary) sentences using minimum-cut graph partitioning to incorporate cross-sentence context.

Turning from subjectivity classification in general to the sports domain specifically, sports journalism was identified early on as a domain where subjectivity detection carries practical value, particularly for applications related to sports betting and automated content curation (Rizk & Awad, 2012).

The original 2012 study by Rizk and Awad used a syntactic genetic algorithm trained on a small, manually collected corpus of 300 articles, relying on frequentist syntactic features such as punctuation, numerical values, and part-of-speech tag frequencies, and later released the expanded 1,000-article dataset used in the present study through the UCI Machine Learning Repository (Rizk & Awad, 2018). Despite this dataset's free availability since 2018, no further published research appears to make direct use of it — a gap that motivates the present study.

Unlike prior work on this dataset, which has consistently framed the task as a classification problem to be optimized, the present study takes a step back and asks a more fundamental descriptive question. Understanding which features differ significantly between classes — and by how much, and whether that difference survives controlling for a confound as basic as article length — is a prerequisite for meaningful feature engineering and model interpretation, yet this analysis has not previously been reported for this dataset. In this sense, the present paper is complementary to, rather than competitive with, the classification-oriented studies discussed above.

3. Dataset

This study uses the Sports Articles for Objectivity Analysis dataset (Rizk & Awad, 2018), obtained from the UCI Machine Learning Repository (https://doi.org/10.24432/C5801R). The dataset consists of 1,000 sports articles collected from various online sources, each labeled as either "objective" or "subjective." Labels were originally assigned through crowd-sourced annotation via Amazon Mechanical Turk, a non-expert annotation approach whose reliability for natural language tasks has been validated against expert gold-standard labels (Snow, O'Connor, Jurafsky, & Ng, 2008). Each article is represented not only by its raw text and source URL but also by 59 pre-extracted numerical features, including:

- Frequencies of part-of-speech tags (nouns, verbs, adjectives, adverbs, comparative and superlative forms, modal auxiliaries, determiners, etc.) obtained via the Stanford POS Tagger (Toutanova, Klein, Manning, & Singer, 2003), using Penn Treebank tag conventions (Marcus, Santorini, & Marcinkiewicz, 1993);
- Two semantic features (semanticobjscore, semanticsubjscore) reflecting the frequency of words with objective vs. subjective SentiWordNet scores (Baccianella, Esuli, & Sebastiani, 2010);
- Punctuation counts (exclamation marks, question marks, quotation pairs, commas, colons, semicolons, ellipses);
- Pronoun usage broken down by grammatical person (first, second, third);
- Verb-tense and mood features (past tense, imperative, present tense by person);
- Text complexity metrics (txtcomplexity) and sentence-class indicators for the first and last sentence of each article.

The dataset contains no missing values, which simplifies preprocessing considerably. The raw features.xls file is distributed as an Excel 2003 XML (SpreadsheetML) file rather than a binary .xls; we parsed it directly with Python's xml.etree.ElementTree module rather than a standard Excel-reading library, and verified correct column alignment by checking that the sum of all part-of-speech frequency columns closely tracked totalWordsCount for every article (mean ratio = 0.891, SD = 0.030, consistent across all 1,000 rows — the residual ~11% corresponds to tokens outside the specific POS categories tabulated, such as certain punctuation-adjacent tokens).

Turning to class distribution, of the 1,000 articles, 635 (63.5%) are labeled objective and 365 (36.5%) are labeled subjective — a moderate class imbalance (roughly 1.74:1) that we account for through stratified sampling in the predictive check (Section 4) and by reporting precision, recall, and F1-score alongside accuracy, since these measures are more informative than raw accuracy under class imbalance (Sokolova & Lapalme, 2009).

With respect to preprocessing, feature values were used as provided in the original release for the primary analysis, with no additional transformation applied prior to statistical testing, in order to preserve the interpretability of each variable in its original unit (raw frequency counts per article). As described in Section 4, we additionally repeated the entire analysis on length-normalized versions of the features (each divided by totalWordsCount) after discovering a substantial length difference between the two classes. For the confirmatory logistic regression models described there, all features were standardized (zero mean, unit variance) prior to model fitting, following standard practice for regularized linear models (Hastie, Tibshirani, & Friedman, 2009).

4. Methodology

For each of the 59 features, we tested whether its distribution differed significantly between the objective and subjective article groups. Given that most features are count-based and right-skewed rather than normally distributed, we used the Mann-Whitney U test (also known as the Wilcoxon rank-sum test), a non-parametric test that does not assume normality and is appropriate for comparing two independent groups on an ordinal or continuous variable (Mann & Whitney, 1947).

A significance threshold of $\alpha = .05$ was adopted. Given the large number of comparisons (59 tests), we additionally applied a Benjamini-Hochberg false discovery rate (FDR) correction to control for multiple comparisons (Benjamini & Hochberg, 1995), reporting FDR-adjusted p-values throughout.

Statistical significance alone does not indicate the practical magnitude of a difference, particularly with a sample size of 1,000, where even small differences can reach significance. To address this, we computed the rank-biserial correlation coefficient (r) for each feature as a measure of effect size, using the simple-difference formula (the proportion of favorable pairs minus the proportion of unfavorable pairs) (Kerby, 2014), following conventional thresholds adapted from Cohen's guidelines for effect size interpretation (negligible: $r < 0.10$; small: $0.10 \leq r < 0.30$; medium: $0.30 \leq r < 0.50$; large: $r \geq 0.50$) (Cohen, 1988).

During the initial descriptive pass, we observed that the two classes differ substantially in article length (see Section 5). Because nearly all 59 features are raw frequency counts, a longer article will mechanically tend to have higher counts of almost every feature regardless of any genuine stylistic difference — a form of confounding bias in text classification that has been documented more broadly, beyond this specific dataset (Landeiro & Culotta, 2016). To assess whether the observed differences reflect true linguistic style rather than simple document length, we repeated the entire testing procedure described above on length-normalized features, dividing each raw count feature by the article's totalWordsCount to obtain a per-word rate. Four features that are not frequency counts in the same sense (totalWordsCount itself, txtcomplexity, and the two categorical sentence-class indicators sentence1st and sentencelast) were excluded from normalization.

As a secondary analysis, we also trained logistic regression classifiers — one using the raw significant features, one using the length-normalized significant features — to assess whether the statistically identified feature sets retain predictive value when combined in a simple model. Model performance was assessed using nested 5-fold stratified cross-validation — stratification by class label is used throughout, following evidence that stratified cross-validation improves accuracy estimation over non-stratified k-fold or leave-one-out approaches (Kohavi, 1995): within each outer fold, feature selection (Mann-Whitney U + BH-FDR, as described above) was re-run using only that fold's training rows, and the resulting feature set was used to scale, fit, and evaluate the logistic regression on that fold's held-out test rows.

This nested procedure avoids a data-leakage issue we identified during model checking: selecting significant features once on the full dataset before cross-validating would let each test fold indirectly influence which features were chosen to predict it, since every row participates in the univariate significance tests — the same bias that motivates nested cross-validation for feature/model selection more generally (Varma & Simon, 2006). We report both the corrected (nested) estimate and, for comparison, the naive (non-nested) estimate that has this leakage — see Section 5.

Three further diagnostic checks accompany the main models. To assess whether the raw-feature and length-normalized models differ reliably in accuracy, we compared their per-fold nested-CV accuracies (evaluated on identical outer folds) with a paired t-test, cross-checked with a Wilcoxon signed-rank test (Wilcoxon, 1945) reported for transparency only (see Section 5 for its statistical-power caveat). As a check on the interpretability of the individual model coefficients, we also computed pairwise Pearson correlations and Variance Inflation Factors (VIF) among the significant raw features, discussed further below. Finally, for interpretability, we report results from a single stratified 80/20 train/test split (random seed fixed for reproducibility), including the resulting confusion matrix and feature coefficients; this single-split result is reported as an illustrative example only, since a single split can be unrepresentatively favorable or unfavorable relative to the cross-validated estimate. In all cases, performance is compared against a majority-class baseline. This step is not intended as a contribution to classification performance in itself, but as a check on whether the features identified through descriptive testing retain predictive value when combined in a simple model.

Finally, all analyses were conducted in Python 3, using pandas for data handling, scipy.stats for hypothesis testing, statsmodels for multiple-comparison correction, and scikit-learn for the logistic regression models and evaluation metrics.

5. Results

Before testing individual features, we compared overall article length between the two classes. Objective articles contain a mean of 519.2 words (SD = 417.7; median = 389), while subjective articles contain a mean of 1,006.6 words (SD = 544.5;

median = 930) — subjective articles are, on average, nearly twice as long as objective ones. This difference is itself statistically significant and has a large effect size (rank-biserial r = 0.604; see Table 1).
Because 55 of the 59 available features are raw frequency counts (not rates), this length disparity is a critical methodological consideration: a feature could appear to differ strongly between classes simply because subjective articles contain more words overall, independent of any genuine difference in linguistic style — precisely the kind of confounding variable that has been shown to inflate apparent classifier performance in text classification tasks more generally (Landeiro & Culotta, 2016). We treat this as a substantive finding in its own right and structure the remainder of the results around it.

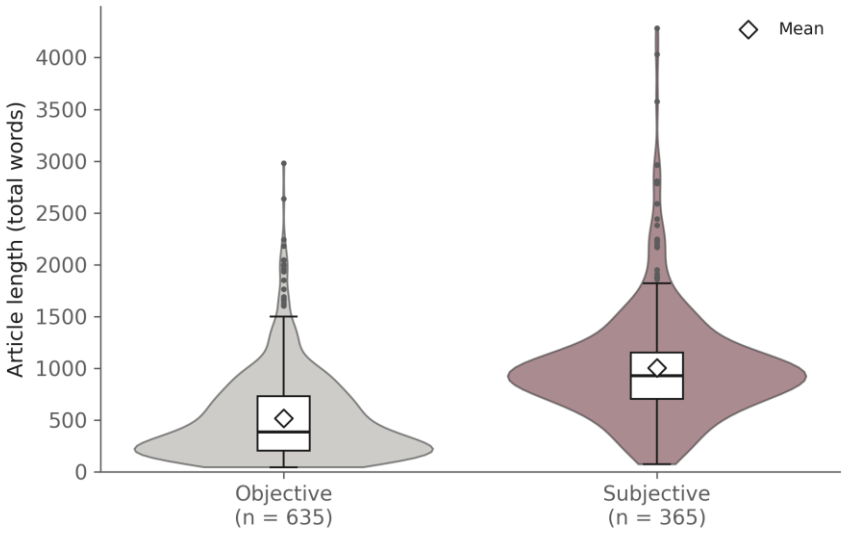

Figure 1. Distribution of total article length (words) by class. Violin plots show the full density shape; inset boxplots show the median and interquartile range; diamonds mark the class means. The two distributions overlap substantially in the 500–1,000-word range, but subjective articles skew markedly longer and show a heavier right tail.

Turning to the raw frequency-count analysis, using the raw (un-normalized) features, 53 of the 59 features tested (89.8%) show statistically significant differences between objective and subjective articles after FDR correction. Table 1 shows the ten features with the largest effect sizes.

Table 1. Top 10 features by effect size (rank-biserial correlation, raw counts), Mann-Whitney U test, FDR-adjusted p-values (N = 1,000; n_objective = 635; n_subjective = 365).

Rank	Feature	Description	Median (Obj.)	Median (Subj.)	U statistic	p (adjusted)	Effect size (r)
1	LS	List-item markers	2	11	36,375.5	8.1×10^{-72}	0.686
2	VBP	Present-tense verbs, non-3rd-person	6	26	36,903.5	9.2×10^{-71}	0.682
3	present3rd	Present-tense verbs, 3rd person	6	26	37,105.0	1.2×10^{-70}	0.680
4	PRP$	Possessive pronouns	12	50	37,114.0	1.2×10^{-70}	0.680
5	semanticsubjscore	SentiWordNet subjective-word frequency	6	20	41,053.5	5.5×10^{-64}	0.646
6	VBN	Past-participle verbs	4	18	42,419.5	7.8×10^{-62}	0.634

Rank	Feature	Description	Median (Obj.)	Median (Subj.)	U statistic	p (adjusted)	Effect size (r)
7	present1st2nd	Present-tense verbs, 1st/2nd person	4	18	42,661.0	1.7×10^{-61}	0.632
8	baseform	Infinitive (base-form) verbs	10	30	42,881.5	4.3×10^{-61}	0.630
9	POS	Genitive markers	9	39	43,494.5	4.0×10^{-60}	0.625
10	imperative	Imperative verbs	1	7	44,441.5	1.1×10^{-59}	0.617

Note: all ten features are higher in subjective articles, consistent with the length confound identified above — see below for the length-normalized re-analysis.

After dividing each raw feature by totalWordsCount and repeating the testing procedure, 48 of the 59 features (81.4%) remain statistically significant. Five features that were significant in the raw analysis — VBD (past-tense verbs), INs (subordinating prepositions/conjunctions), NNPS (plural proper nouns), VBG (gerunds), and colon — lose significance once length is controlled for, indicating that their apparent raw-count differences were driven primarily by article length rather than a genuine per-word stylistic difference.
Critically, only 4 of the 10 top raw-count features (LS, PRP$, imperative, and semanticsubjscore) remain in the top 10 once features are expressed as rates, confirming that much of the raw-count ranking in Table 1 reflected length rather than style. This shrinkage is not confined to the top 10: across all 59 features, mean |r| drops from 0.41 (raw) to 0.25 (length-normalized) — a systematic reduction, not an artifact of a handful of extreme cases (Figure 2). Table 2 shows the length-normalized top 10.

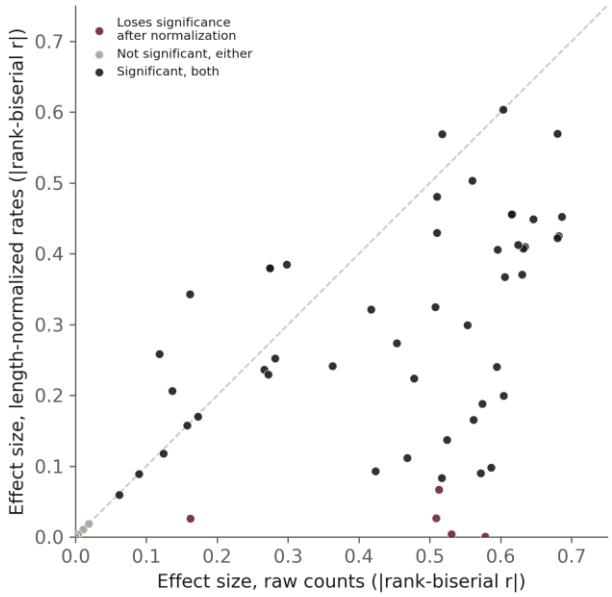

Figure 2. Effect size (|rank-biserial r|) for each of the 59 features, raw counts versus length-normalized rates. The dashed line marks equality (no change after normalization). The large majority of features fall below the line, visualizing the systematic effect-size shrinkage summarized above; burgundy points denote the five features that lose statistical significance entirely once length is controlled for. Full per-feature values are reported in Table S1 (Appendix).

Table 2. Top 10 features by effect size (rank-biserial correlation, length-normalized rates), FDR-adjusted p-values.

Rank	Feature	Description	Rate (Obj., median)	Rate (Subj., median)	p (adjusted)	Effect size (r)
1	totalWordsCount	Article length (raw, for reference)	389	930	3.3×10^{-55}	0.604
2	PRP$	Possessive pronouns (per word)	0.0331	0.0493	1.5×10^{-49}	0.569
3	MD	Modal auxiliaries (per word)	0.2468	0.1971	1.5×10^{-49}	−0.569
4	pronouns2nd	Second-person pronouns (per word)	0.0000	0.0033	2.6×10^{-44}	0.503
5	questionmarks	Question marks (per word)	0.0000	0.0012	2.5×10^{-49}	0.481
6	imperative	Imperative verbs (per word)	0.0033	0.0070	1.0×10^{-32}	0.456
7	UH	Interjections (per word)	0.0033	0.0070	1.1×10^{-32}	0.456
8	LS	List-item markers (per word)	0.0058	0.0115	5.2×10^{-32}	0.452
9	semanticsubjscore	SentiWordNet subjective score (per word)	0.0147	0.0215	1.5×10^{-31}	0.449
10	DT	Determiners (per word)	0.0000	0.0014	3.9×10^{-34}	0.429

The length-normalized results (Table 2) point to a more interpretable and arguably more genuine picture of stylistic difference. Subjective articles use possessive pronouns, second-person pronouns, question marks, imperative verbs, and interjections at a significantly higher rate — even after accounting for their greater length — consistent with the theoretical expectation that opinionated writing relies more heavily on direct address, rhetorical engagement, and personal framing (Pang et al., 2002) — a pattern that closely mirrors Biber's (1988) "involved vs. informational production" dimension of register variation, in which first/second-person pronouns and direct engagement mark one pole and nominal, informational style marks the other. Notably, modal auxiliary verbs (MD) show the only reversed effect among the top 10: objective articles use modals at a higher relative rate (0.247 vs. 0.197 per word, $r = -0.569$) than subjective ones, despite subjective articles containing more modals in absolute terms. This suggests that hedging language conveyed through modal verbs (e.g., "could," "may," "should") — the linguistic device by which writers calibrate their commitment to the certainty of a claim (Hyland, 1998) — is, proportionally, more characteristic of factual sports reporting than of opinion pieces in this dataset — a nuance that raw-count analysis alone would not reveal.

Turning to the confirmatory logistic regression check outlined in Section 4, two models were trained: one restricted to statistically significant raw-count features, and one restricted to statistically significant length-normalized features. Under nested 5-fold cross-validation — our primary and leakage-free evaluation protocol, in which feature selection is re-run within each training fold rather than once on the full dataset — the raw-feature model achieved a mean accuracy of 82.3% (SD = 1.4 percentage points across folds; F1 = 0.732, precision = 0.816, recall = 0.666), and the length-normalized model achieved a mean accuracy of 83.5% (SD = 0.8 percentage points; F1 = 0.767, precision = 0.790, recall = 0.748). The number of features selected varied only slightly across the five outer folds (53–54 for raw features; 44–49 for normalized features), indicating that the Mann-Whitney/FDR feature selection is fairly stable with respect to which rows are held out.

Is the 1.2-point gap between the raw-feature model (82.3%) and the length-normalized model (83.5%) a real difference, or fold-to-fold noise? Because both models were evaluated on the same five outer folds, we tested this directly with a paired t-test on the per-fold accuracies (raw: 84.5%, 80.5%, 83.0%, 82.5%, 81.0%; normalized: 84.5%, 84.5%, 83.0%, 83.0%, 82.5%), which gave $t(4) = 1.60$, $p = .19$ — not significant at the conventional $\alpha = .05$ threshold. A Wilcoxon signed-rank test on the

same pairs agreed (p = .25), though with only five paired folds this non-parametric test has essentially no power to detect anything (the smallest two-sided p-value attainable with n = 5 is .0625, so a non-significant result here is uninformative on its own and is reported only for transparency). Taken together, the raw and length-normalized models should be treated as statistically indistinguishable in predictive accuracy given the current evidence — the normalized model's numerically higher accuracy and better-balanced precision/recall (F1 = 0.767 vs. 0.732) are suggestive but not confirmed by this test, and a larger number of folds or repeated cross-validation would be needed to detect a difference of this size reliably, if one exists.

For comparison, we also computed the naive (non-nested) cross-validated accuracy, in which features are selected once on the full 1,000-row dataset before cross-validating — the leakage-prone protocol we initially ran, of the general kind cautioned against when feature or model selection is performed outside the cross-validation loop (Varma & Simon, 2006). This naive estimate was 82.5% for raw features and 83.4% for normalized features — within 0.2 percentage points of the leakage-corrected nested estimates in either direction. In this particular dataset, then, the data leakage we identified turned out not to meaningfully inflate the reported accuracy, most likely because the feature-selection step is stable across folds (as shown above) and because, at n = 1,000, each individual test fold (n ≈ 200) has limited influence on which features pass a stringent FDR-corrected significance threshold. We report the nested estimate as the primary figure throughout this paper on methodological grounds, but note that both protocols converge on essentially the same conclusion here.

Finally, we report a single stratified 80/20 train/test split (n_test = 200) for interpretability: the raw-feature model reached accuracy = 85.5%, F1 = 0.794 (confusion matrix: 115 true negatives, 12 false positives, 17 false negatives, 56 true positives). We flag this single-split figure explicitly as illustrative rather than headline evidence of model performance: with other random seeds, single-split accuracy for the raw-feature model ranged from about 82% to 86.5%, which is precisely why cross-validated (and, more rigorously, nested cross-validated) estimates are reported as the primary figures in this paper rather than a single train/test partition. The three largest logistic-regression coefficients in the single-split raw-feature model belonged to PRP$ (possessive pronouns, coefficient = 1.11), CD (numerals/cardinals, coefficient = 0.87), and INs (subordinating prepositions, coefficient = 0.82), broadly corroborating the descriptive results reported earlier in this section.

The results above establish that article length is a substantial confound in the raw-feature analysis, but this does not by itself tell us how much of the classifier's predictive power is attributable to length alone versus genuine per-word stylistic signal. To answer this directly, we trained a third logistic regression model using totalWordsCount as its sole feature — no other linguistic information — evaluated with the same 5-fold cross-validation protocol used elsewhere in this paper.

This single-variable, length-only model achieved a mean accuracy of 72.3% (SD = 1.5 percentage points; F1 = 0.543, precision = 0.681, recall = 0.455) — 8.8 percentage points above the 63.5% majority-class baseline, using nothing but a word count. Framed against the two full models reported above, article length alone accounts for 46.8% of the raw-feature model's improvement over baseline (82.3% − 63.5% = 18.8 points; length alone contributes 8.8 of those points) and 44.0% of the length-normalized model's improvement (83.5% − 63.5% = 20.0 points).

This result puts a concrete number on the confound identified earlier: nearly half of what a simple classifier "learns" to distinguish objective from subjective sports articles in this dataset is explained by how long the article is, not by any specific pattern of word choice, pronoun use, or syntax. The remaining slightly-more-than-half of the classifier's discriminative power reflects genuine per-word stylistic differences — consistent with the length-normalized features described earlier remaining statistically significant even after the length confound is removed from the feature values themselves (as opposed to removed from the classifier's inputs entirely, as done here). Notably, the fact that the length-normalized model (83.5%) still outperforms the length-only model (72.3%) by 11.2 points confirms that per-word stylistic features carry real, non-redundant signal beyond simply knowing how long an article is — length is a genuine confound worth reporting, but it is not the whole story.

Table 3 consolidates accuracy, F1, precision, and recall across all six evaluation protocols discussed above, for direct comparison.

Table 3. Classifier performance across all evaluation protocols (5-fold cross-validation unless noted). Nested CV is the primary, leakage-free protocol (Section 4); naive CV is the non-nested comparison protocol described above.

Model	Accuracy	SD	F1	Precision	Recall
Majority-class baseline	63.5%	—	—	—	—
Length-only (totalWordsCount)	72.3%	1.5	0.543	0.681	0.455
Raw features — nested CV	82.3%	1.4	0.732	0.816	0.666
Raw features — naive CV	82.5%	1.5	0.735	0.819	0.668
Length-normalized — nested CV	83.5%	0.8	0.767	0.790	0.748
Length-normalized — naive CV	83.4%	1.2	0.766	0.787	0.748

Turning finally to multicollinearity among the raw features, the logistic regression coefficients reported above (PRP$, CD, INs) invite a natural but risky reading: that these three features are individually "the most important" predictors of subjectivity. To check whether that reading is warranted, we examined pairwise correlations and Variance Inflation Factors (VIF) among the 53 significant raw features used to fit that model. For a given feature i, VIF is defined as

$$VIF_i = \frac{1}{1 - R_i^2}$$

where Ri2 is the coefficient of determination obtained from regressing feature i on all other features in the model.
A VIF of 1 indicates no correlation with the other predictors, while larger values indicate that an increasing share of feature i's variance is already explained by the other features — a VIF of 2,900, for instance, corresponds to Ri2 ≈ 0.9997, meaning that feature is almost perfectly predictable from the rest of the feature set.
The result is a severe degree of multicollinearity. Of the 53 raw features, 35 (66%) have a VIF at or above the conventional concern threshold of 5, and 29 (55%) exceed the stricter threshold of 10 (Kutner, Nachtsheim, Neter, & Wasserman, 2004); several features are so close to perfectly collinear with others that their VIF is effectively unbounded.
It is worth noting, however, that these commonly-cited VIF thresholds (5, 10) are themselves rules of thumb rather than fixed statistical cutoffs, and high VIF values do not automatically invalidate a regression model — they indicate that individual coefficients are less precisely estimated and more sensitive to which correlated predictors are included, which is exactly the caution we apply to coefficient interpretation below rather than treating high VIF as disqualifying in itself (O'Brien, 2007). totalWordsCount alone has a VIF above 2,900, unsurprising given that it mechanically drives nearly every other raw count, as discussed above.
But the collinearity is not limited to features versus length: 116 feature pairs show $|r| \geq 0.80$, and several pairs are essentially duplicates of one another — VB and past correlate at r = 1.00 (identical values in 99.9% of articles), as do VBN/present1st2nd and VBP/present3rd (r ≈ 0.9999 each), and imperative/UH (r = 0.998). These near-identical pairs almost certainly reflect overlapping or redundant feature definitions in the original dataset's feature-extraction pipeline (e.g., two tag categories that count largely the same tokens) rather than a coincidental stylistic relationship, and future users of this dataset should be aware that several of its 59 nominally distinct features are not, in practice, independent measurements.

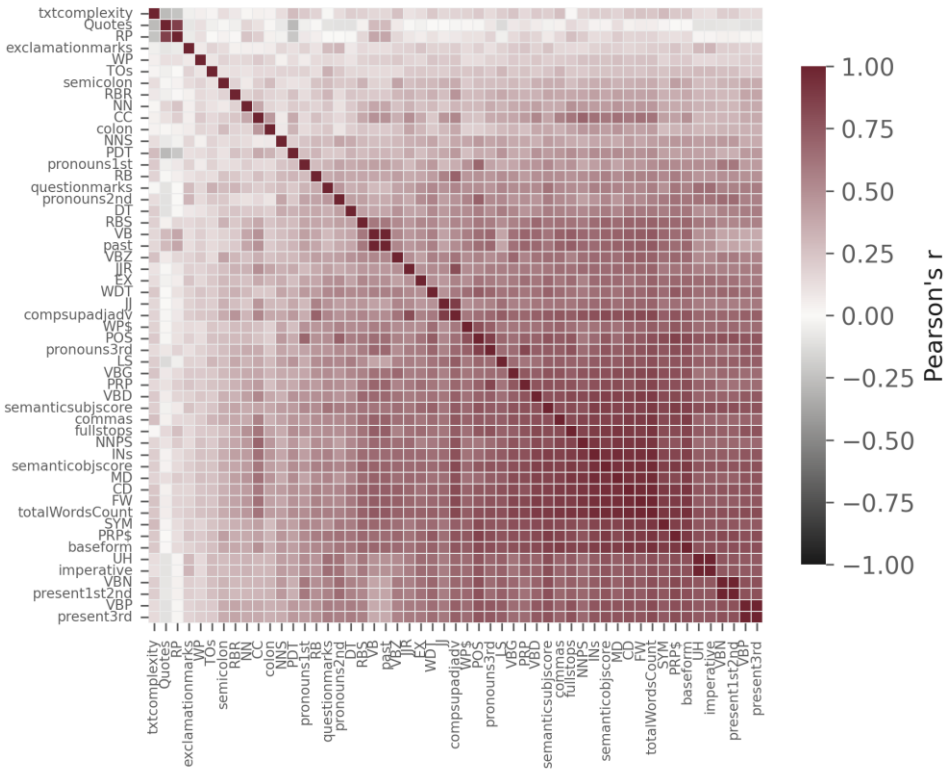

Figure 3. Pairwise Pearson correlation matrix among the 53 significant raw features, with rows and columns reordered by hierarchical clustering on correlation distance so that groups of mutually correlated features appear as visible blocks. The near-perfectly correlated pairs discussed in the text (VB/past, VBN/present1st2nd, VBP/present3rd, imperative/UH) appear as the small, near-black 2×2 blocks along the diagonal at the bottom right; the broad reddish band covering most of the matrix reflects the pervasive length-driven collinearity described above.

Length normalization (Section 4) substantially reduces, but does not eliminate, this problem: among the 48 significant length-normalized features, only 16 (33%) have VIF ≥ 5 and 14 (29%) have VIF ≥ 10 — roughly half the rate seen in the raw features — consistent with much (though not all) of the raw-feature collinearity being a downstream consequence of the shared length confound rather than a property of the linguistic signal itself.
The near-duplicate pairs identified above (VB/past, etc.) remain essentially perfectly correlated even after normalization, since dividing two nearly-identical columns by the same denominator preserves their near-identical relationship.

The practical implication is that the model-level accuracy figures reported above are unaffected by this collinearity — a logistic regression's overall predictive accuracy is robust to correlated predictors — but the individual coefficients for any one feature should not be over-interpreted as isolated, independent effects. A given coefficient's sign and magnitude can shift considerably depending on which correlated features happen to be included, so the coefficients in logreg_coefficients.csv are best read as describing a jointly predictive set of features rather than as clean, independent effect-size estimates for each individual linguistic marker. This is consistent with, and reinforces, the descriptive (Mann-Whitney/rank-biserial) results reported earlier in this section as the more reliable source for claims about which individual features differ between classes, since those univariate tests do not share this multicollinearity problem.

6. Discussion

The central empirical finding of this study is two-layered. First, at the raw-count level, the overwhelming majority of the 59 available features (89.8%) differ significantly between objective and subjective sports articles — but this result is substantially inflated by the fact that subjective articles are, on average, nearly twice as long as objective ones. Second, once this length confound is controlled for through per-word normalization, a smaller but still substantial majority of features (81.4%) remain significant, and the specific features that emerge as most discriminative shift meaningfully: possessive and second-person pronoun usage, question marks, imperative verbs, and interjections — markers of direct, rhetorically engaged, opinion-driven writing — outperform features that had appeared dominant only because of their raw frequency in longer texts.
This length-confound finding is, to our knowledge, not previously discussed for this dataset, since prior work using it worked directly with the raw features as provided. This does not invalidate the classification results obtainable from raw features — a supervised classifier can legitimately exploit article length as a predictive signal, and length itself may be a genuine, stable correlate of subjectivity in sports writing rather than a nuisance variable to be removed — consistent with the long-standing observation that sentence/text length can itself be a genuine and stable stylistic characteristic, rather than mere noise, in studies of authorship and prose style (Yule, 1939). However, it does mean that claims about which specific linguistic markers distinguish subjective from objective sports writing should be made with reference to length-normalized rates rather than raw counts, since the latter conflate style with sheer document length.
The length-only control model reported in Section 5 makes this concrete: a classifier given nothing but word count reaches 72.3% accuracy — nearly half of the full model's total improvement over the majority-class baseline.
Length is therefore not a minor nuisance variable to footnote — it is doing a substantial share of the classification work — yet it is also demonstrably not the whole story, since length-normalized stylistic features add a further 11.2 points of accuracy beyond what length alone provides.
From a practical standpoint, these findings offer three contributions. First, they provide an interpretable reference point for researchers who may wish to use this dataset in future work, clarifying which of the 59 available features carry genuine per-word discriminative signal as opposed to signal that is an artifact of length.
Second, the modal-verb reversal reported in Section 5 — objective articles use modal auxiliaries at a higher relative rate than subjective ones, despite the latter's greater absolute modal count — illustrates concretely why length normalization matters and would have been invisible under a raw-count-only analysis.
Third, the fact that a simple logistic regression model restricted to a reduced, statistically justified feature set achieves 82.3–83.5% accuracy suggests that much of the classification signal in this task can be captured through relatively simple, interpretable linear combinations of surface-level linguistic features.
Several limitations qualify these findings. First, the dataset itself is relatively small (1,000 articles) and drawn from a limited set of online sources circa the mid-2010s, which may limit the generalizability of findings to contemporary sports journalism, particularly given stylistic shifts in online writing over the past decade.
Second, the subjectivity labels were obtained through crowd-sourced Amazon Mechanical Turk annotation in the original data collection process, and no inter-annotator agreement statistics are available to us to assess label reliability — a real concern given that even validated non-expert annotation pipelines can show variable agreement across tasks (Snow, O'Connor, Jurafsky, & Ng, 2008).
Third, the multiple comparisons correction applied here (Benjamini-Hochberg FDR) is conservative but does not eliminate the possibility of residual false positives given the exploratory nature of testing all 59 features. Fourth, the length-normalization approach used here (simple division by word count) is a standard but not the only possible way to control for length; alternative approaches (e.g., regression-based residualization, or matching articles on length before comparison) might yield somewhat different rankings.
Fifth, as detailed in Section 5, the significant raw features exhibit severe multicollinearity (66% with VIF ≥ 5), including several near-duplicate feature pairs that appear to reflect redundant definitions in the dataset's original feature-extraction pipeline rather than genuine independent signal; this does not affect the model's overall accuracy but means the individual logistic-regression coefficients reported there should not be read as isolated effect-size estimates for single features.

Finally, the confirmatory logistic regression analysis is intended only as a sanity check on the descriptive findings, not as a competitive classification benchmark; more sophisticated modeling approaches would likely yield different — potentially higher — performance.
Several additional caveats concern the evaluation protocol specifically.
As discussed in Section 5, single train/test splits produced accuracy figures that varied by several percentage points depending on the random seed, which is why nested cross-validated estimates are treated as the primary evidence of model performance throughout this paper — nested cross-validation should be used by default whenever feature selection and model evaluation are performed on the same data, even though in this particular dataset the naive and nested protocols happened to converge.
Relatedly, the 1.2-point edge of the length-normalized model over the raw-feature model (83.5% vs. 82.3%) should not be over-interpreted: as shown earlier, this difference was not statistically significant, so this paper's claim is that per-word stylistic features retain genuine signal beyond length alone, not that length-normalization improves classification accuracy over raw counts.
Looking ahead, future research could extend this analysis in several directions: (1) applying model-agnostic interpretability techniques such as SHAP to a more complex, non-linear classifier trained on this dataset, to validate whether the length-normalized features identified here through descriptive statistics also emerge as most influential in a non-linear model; (2) re-collecting a comparable dataset from contemporary sports journalism sources to assess whether these linguistic markers of subjectivity — and the length confound itself — have remained stable over time; (3) exploring alternative length-control strategies, such as residualizing each feature on article length via regression rather than simple division; and (4) extending the analysis to other sports-adjacent text domains, such as social media commentary or podcast transcripts, where subjectivity may manifest differently and length distributions may differ.

7. Conclusion

This paper presented a descriptive statistical analysis of the Sports Articles for Objectivity Analysis dataset, a small, publicly available, and largely uncited resource for subjectivity classification in sports journalism.
By systematically testing all 59 available linguistic features using non-parametric hypothesis testing and effect size estimation, this study found that 89.8% of raw features differ significantly between objective and subjective articles — but also uncovered a substantial length confound (subjective articles average nearly twice the word count of objective ones) that had not been previously reported for this dataset.
After controlling for this confound through length normalization, 81.4% of features remain significant, with pronoun usage, question marks, imperative verbs, and a notable reversal in modal-verb rate emerging as the most robust and interpretable markers of subjectivity.
A simple logistic regression model restricted to the statistically significant features achieves 82.3–83.5% accuracy under nested 5-fold cross-validation — and a length-only control model shows that roughly 45% of this classifier's advantage over the baseline comes from article length alone, a reminder that document length accounts for a large share of what looks, at first glance, like a purely stylistic signal.
While the analysis is intentionally modest in technical scope, it fills a small but genuine gap in the literature, surfaces a methodological caveat relevant to any future use of these features, and provides a transparent, reproducible statistical baseline for researchers considering this dataset for future subjectivity detection studies.

8. Data Availability

Dataset. The underlying dataset — Sports Articles for Objectivity Analysis (Rizk & Awad, 2018) — is publicly available from the UCI Machine Learning Repository at https://doi.org/10.24432/C5801R, and is distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits redistribution and adaptation with attribution.
The original download consists of features.xls (an Excel 2003 XML/SpreadsheetML file containing the 59 pre-extracted linguistic features, article labels, and source URLs for all 1,000 articles) and a Raw data/ folder of 1,000 individual article text files; only features.xls was used in this study.

References

1. Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 2200–2204). European Language Resources Association.
2. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.
3. Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press.
4. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.
6. Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., & de Jong, F. (2011). Polarity analysis of texts using discourse structure. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (pp. 1061–1070).
7. Hyland, K. (1998). Hedging in Scientific Research Articles. John Benjamins.
8. Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. Comprehensive Psychology, 3, Article 11.

9. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) (Vol. 2, pp. 1137–1143).
10. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied Linear Regression Models (4th ed.). McGraw-Hill/Irwin.
11. Landeiro, V., & Culotta, A. (2016). Robust text classification in the presence of confounding bias. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.9997
12. Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18(1), 50–60.
13. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330.
14. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6
15. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (pp. 271–278).
16. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 79–86).
17. Rizk, Y., & Awad, M. (2012). Syntactic genetic algorithm for a subjectivity analysis of sports articles. In Proceedings of the 11th IEEE International Conference on Cybernetic Intelligent Systems, Limerick, Ireland.
18. Rizk, Y., & Awad, M. (2018). Sports Articles for Objectivity Analysis [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5801R (CC BY 4.0)
19. Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). 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 (EMNLP) (pp. 254–263).
20. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.
21. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). 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 (HLT-NAACL) (pp. 252–259).
22. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91
23. Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2–3), 165–210. https://doi.org/10.1007/s10579-005-7880-9
24. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83.
25. Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. Biometrika, 30(3–4), 363–390.

Appendix

Table S1. Complete Mann-Whitney U test results for all 59 features (raw and length-normalized). This table reports, for every one of the 59 linguistic features tested in this study, the rank-biserial effect size (r) and Benjamini-Hochberg FDR-adjusted p-value from the Mann-Whitney U test comparing objective vs. subjective articles, computed separately on raw frequency counts and on length-normalized (per-word) rates (Sections 4–5). Rows are sorted by the magnitude of the raw-feature effect size, descending; the top 10 by this ordering are also summarized in-text in Section 5 and in top10_features.csv. A checkmark (✓) indicates FDR-adjusted $p < .05$ (statistically significant). This table corresponds to the machine-readable file table_s1_mannwhitney_full_59_features.csv, which in turn is derived directly from mannwhitney_results_full.csv and mannwhitney_results_normalized.csv (Section 8).

#	Feature	r (raw)	p_FDR (raw)	Sig. (raw)	r (norm.)	p_FDR (norm.)	Sig. (norm.)
1	LS	0.686	8.09e-72	✓	0.452	5.20e-32	✓
2	VBP	0.682	9.15e-71	✓	0.425	2.13e-28	✓
3	present3rd	0.680	1.24e-70	✓	0.422	4.17e-28	✓
4	PRP$	0.680	1.24e-70	✓	0.569	1.54e-49	✓
5	semanticsubjscore	0.646	5.55e-64	✓	0.449	1.49e-31	✓
6	VBN	0.634	7.85e-62	✓	0.410	1.33e-26	✓
7	present1st2nd	0.632	1.69e-61	✓	0.407	2.57e-26	✓
8	baseform	0.630	4.34e-61	✓	0.370	4.78e-22	✓
9	POS	0.625	3.99e-60	✓	0.412	7.52e-27	✓
10	imperative	0.617	1.12e-59	✓	0.456	1.01e-32	✓
11	UH	0.615	1.89e-59	✓	0.456	1.06e-32	✓
12	compsupadjadv	0.606	3.01e-57	✓	0.367	8.46e-22	✓
13	CD	0.604	1.94e-56	✓	0.199	2.32e-07	✓
14	totalWordsCount	0.604	2.37e-56	✓	0.604	3.32e-55	✓
15	WP$	0.596	2.70e-56	✓	0.406	1.83e-26	✓
16	semanticobjscore	0.594	8.82e-55	✓	0.241	4.08e-10	✓
17	FW	0.587	2.02e-53	✓	-0.098	0.013	✓
18	INs	0.578	6.45e-52	✓	-0.000	1.000	
19	pronouns3rd	0.574	2.95e-51	✓	0.188	1.05e-06	✓
20	SYM	0.572	6.63e-51	✓	0.090	0.022	✓

#	Feature	r (raw)	p_FDR (raw)	Sig. (raw)	r (norm.)	p_FDR (norm.)	Sig. (norm.)
21	commas	0.561	3.98e-49	✓	0.166	1.90e-05	✓
22	pronouns2nd	0.560	2.86e-55	✓	0.503	2.62e-44	✓
23	WDT	0.553	1.30e-48	✓	0.299	4.99e-15	✓
24	VBG	0.530	4.45e-44	✓	-0.004	0.961	
25	fullstops	0.524	4.88e-43	✓	-0.137	4.17e-04	✓
26	MD	0.517	5.25e-42	✓	-0.569	1.54e-49	✓
27	PRP	0.517	5.25e-42	✓	0.083	0.035	✓
28	NNPS	0.513	2.55e-41	✓	-0.067	0.091	
29	DT	0.510	4.52e-49	✓	0.429	3.90e-34	✓
30	questionmarks	0.510	2.70e-56	✓	0.481	2.53e-49	✓
31	VBD	0.509	7.90e-41	✓	-0.027	0.512	
32	JJ	0.508	1.14e-42	✓	0.325	6.75e-18	✓
33	VBZ	0.478	1.99e-37	✓	0.224	3.79e-09	✓
34	EX	0.468	7.32e-35	✓	0.112	0.004	✓
35	JJR	0.453	1.12e-34	✓	0.274	2.85e-13	✓
36	RBS	0.423	3.15e-29	✓	0.093	0.018	✓
37	RB	0.417	3.29e-33	✓	0.321	8.31e-20	✓
38	pronouns1st	0.363	1.20e-22	✓	0.242	1.11e-10	✓
39	Quotes	-0.298	3.05e-16	✓	-0.385	1.16e-25	✓
40	NNS	0.282	7.39e-22	✓	0.252	2.10e-17	✓
41	VB	0.275	5.82e-13	✓	-0.379	4.97e-23	✓
42	past	0.274	6.39e-13	✓	-0.379	4.92e-23	✓
43	PDT	0.272	2.00e-17	✓	0.229	1.20e-12	✓
44	RBR	0.266	1.31e-19	✓	0.237	2.05e-15	✓

#	Feature	r (raw)	p_FDR (raw)	Sig. (raw)	r (norm.)	p_FDR (norm.)	Sig. (norm.)
45	semicolon	0.173	1.53e-14	✓	0.170	6.82e-14	✓
46	colon	0.162	7.06e-06	✓	0.026	0.509	
47	CC	0.162	2.48e-05	✓	-0.343	3.98e-19	✓
48	txtcomplexity	0.157	3.81e-05	✓	0.157	4.65e-05	✓
49	RP	-0.136	1.00e-04	✓	-0.206	5.19e-09	✓
50	exclamationmarks	0.124	8.39e-10	✓	0.118	6.70e-09	✓
51	NN	0.118	0.002	✓	-0.258	1.58e-11	✓
52	TOs	0.090	5.01e-10	✓	0.089	8.42e-10	✓
53	WP	0.062	8.34e-04	✓	0.059	0.002	✓
54	sentence1st	-0.019	0.286		-0.019	0.308	
55	JJS	0.011	0.058		0.011	0.064	
56	sentencelast	0.004	0.459		0.004	0.493	
57	ellipsis	0.003	0.202		0.003	0.217	
58	NNP	0.000	1.000		0.000	1.000	
59	WRB	0.000	1.000		0.000	1.000	

Note. Four features (totalWordsCount, txtcomplexity, sentence1st, sentencelast) were excluded from length-normalization (Section 4) and therefore retain identical raw and normalized values/effect sizes in this table. Rank-biserial r ranges from −1 to 1, with positive values indicating higher values among subjective articles; magnitude interpretation follows the thresholds in Section 4 (negligible: $r < 0.10$; small: $0.10 \leq r < 0.30$; medium: $0.30 \leq r < 0.50$; large: $r \geq 0.50$).