Beyond Polarity: A Python-Based Corpus Framework for Examining Foreign Media Coverage of Africa

Xingchi Ma
Xi'an Fanyi University, Xi'an, Shaanxi, China

Abstract: This paper develops a Python-based corpus framework for examining sentiment and representation in foreign media coverage of Africa. Rather than equating negative vocabulary with bias, the framework distinguishes event-related negativity, editorial evaluation, topic selection, source attribution, and narrative agency. A multi-outlet corpus is organized by headline, standfirst, article body, quotation, country, topic, and publication. VADER and TextBlob provide transparent baseline scores, while human annotation addresses irony, negation, attribution, and domain-specific language. Topic and outlet comparisons are controlled for article volume and issue mix, and concordance reading links numerical patterns to their textual contexts. The study proposes a layered route from lexical polarity to claims about framing: automated results identify clusters and anomalies; contextual coding explains who evaluates whom; comparative analysis tests whether patterns are specific to Africa or common to international reporting. The framework offers a reproducible foundation for studying foreign media narratives without treating Africa or foreign media as homogeneous categories.

Keywords: Africa; Foreign Media; Sentiment Analysis; Corpus Linguistics; Media Framing; Python

1. Introduction

Foreign news is one of the main channels through which distant regions become meaningful to international audiences. Reports do not merely transmit facts. They select events, name actors, distribute responsibility, and establish which developments deserve sustained attention. This process is especially consequential in coverage of Africa, a continent of fifty-four states that is still frequently compressed into a single news category. Stories about conflict, debt, public health, migration, elections, economic reform, culture, sport, and technological innovation may therefore contribute to broader assumptions about African societies even when the article concerns only one country.

Critiques of foreign media have long identified recurring images of crisis, dependency, corruption, disorder, and external rescue. These concerns are important, but the supporting evidence has not always matched the breadth of the claims. Scott's review shows that research on Africa's media image has often concentrated on a limited set of outlets, countries, genres, and dramatic events [1]. Nothias likewise argues that criticism should be empirically differentiated and open to findings that complicate the familiar account of uniformly negative representation [2]. A method that begins by collecting only crisis stories or by classifying every adverse word as prejudice risks reproducing the conclusion it intends to test.

Python-based text analysis creates an opportunity to examine larger and more transparent collections of reporting. It can identify sentiment patterns across thousands of sentences, compare headlines with article bodies, and locate cases that deserve close reading. Yet a sentiment score does not reveal by itself whether a report is fair, stereotypical, or hostile. Negative language may describe a genuinely harmful event; positive language may celebrate investment while presenting African institutions as passive. Evaluation may also belong to a quoted politician rather than to the journalist. Computational scale is valuable only when it remains connected to context, attribution, and news selection.

This paper develops a corpus-assisted framework organized around three questions. First, how does evaluative tone vary across outlets, topics, countries, and textual layers? Second, whose voice carries the evaluation, and how are African actors represented as decision-makers, interpreters, beneficiaries, or victims? Third, when automated scores and

human judgments diverge, which linguistic or editorial features explain the difference? The central argument is that sentiment analysis should function as an evidential map. It can reveal concentrations, contrasts, and anomalies, while framing analysis and close reading determine what those patterns mean.

2. Literature Review

Classic news-value research explains why international coverage is uneven. Conflict, elite actors, unexpectedness, magnitude, and cultural proximity affect whether an event becomes news [3]. Later studies emphasize that news values are not only selection criteria but are also produced through language, images, headlines, and narrative emphasis [4]. Coverage may therefore develop a crisis orientation even when individual articles contain little explicit judgment. The choice to report war repeatedly and local innovation rarely is analytically different from using negative language inside a report, although both choices may shape public perception.

Research on Africa's representation documents persistent narratives of danger and deficiency, but it also identifies change, diversity, and contradiction. The collection edited by Bunce, Franks, and Paterson describes an evolving media environment in which older stereotypes coexist with reporting on business, culture, technology, and African agency [5]. Scott cautions against treating a small group of highly visible examples as proof of an unchanging continental image [1]. Nothias finds that some established criticisms receive limited systematic support, while linguistic conflation and biased representations of political leadership remain important concerns [2]. These findings make comparison and country specificity essential.

Sentiment analysis estimates evaluative orientation from lexical and grammatical cues. Lexicon-based tools are attractive because their rules can be inspected and their outputs reproduced, but performance depends heavily on domain and context [6]. VADER combines a sentiment dictionary with rules for intensification, capitalization, punctuation, contrast, and negation [7]. TextBlob provides polarity and subjectivity measures through an accessible Python interface. Research in political communication suggests that automated tone measures are most defensible when they are validated against human judgments rather than treated as neutral facts [8].

Framing analysis addresses a related but broader problem. Entman defines framing through selection and salience: texts define problems, diagnose causes, evaluate actors, and suggest remedies [9]. A frame may be restrictive even when its vocabulary is neutral. Likewise, a critical report may preserve African agency by foregrounding local disagreement and locally proposed solutions. Corpus analysis and close reading therefore answer different questions. The former tests frequency and distribution; the latter explains rhetorical function. Their combination offers a stronger basis for evaluating foreign media coverage than either a collection of striking quotations or a single continental sentiment score.

3. Research Design

3.1. Corpus Scope and Sampling

For this study, foreign media refers operationally to news organizations headquartered outside Africa that produce English-language international reporting for audiences beyond a single African state. The category is methodological rather than ideological. It should not imply that British, American, European, Asian, or transnational outlets share one editorial position. A balanced corpus should include several publications with different ownership structures, audiences, political traditions, and reporting formats. News reports, analysis pieces, editorials, and features should either be sampled separately or identified by genre so that unlike texts are not compared without qualification[10].

Corpus construction begins with a documented search strategy covering African country names, regional organizations, major cities, and continent-wide terms. An article enters the main corpus only when an African country, African institution, or Africa-wide development is central to the report. Passing references are excluded. Podcasts, videos, newsletters, wire-service duplicates, and sponsored content form separate datasets because their production and editorial conditions differ. The sampling period should be long enough to include routine reporting as well as major events; otherwise, one election, conflict, or epidemic may dominate the entire picture[11].

Each record includes a stable identifier, outlet, date, URL, headline, standfirst, byline where available, article section, genre, country references, principal topic, and lawfully

accessible body text. Web and print versions are linked and counted once. Excluded items remain in an audit file with a stated reason. This record is not administrative detail: transparent inclusion decisions prevent researchers from quietly favoring the most obviously negative headlines[12]. Where copyright prevents redistribution of full articles, a replication package can still provide metadata, code, dictionaries, annotation guidelines, and derived scores[13].

3.2. Python-Based Processing Pipeline

The workflow preserves raw text before any cleaning. Python scripts remove navigation labels, subscription notices, image credits, and unrelated recommendations while retaining punctuation, capitalization, quotation marks, and paragraph boundaries. Stop-word deletion is avoided during sentiment scoring because negators and contrast markers affect polarity. The cleaned version is checked against the source, and each transformation is logged. Articles are then segmented into headline, standfirst, body paragraph, sentence, direct quotation, and non-quotation layers. This structure allows the analysis to ask where evaluation appears rather than treating the article as one undifferentiated block.

Python also supports systematic metadata enrichment. Regular expressions and named-entity recognition can identify country names, organizations, public officials, and geographic labels, but automated entities require manual review because names may be ambiguous and regional terms may conceal country differences. Topic categories can be assigned through a supervised coding scheme covering politics, conflict, economy, public health, environment, migration, culture, science, education, and sport. The code should preserve the original category and any later correction so that the development of the dataset remains visible.

3.3. Sentiment Scoring and Human Validation

VADER returns negative, neutral, positive, and compound scores, while TextBlob produces polarity and subjectivity estimates. Continuous scores are stored before any categories are imposed. Results are calculated separately for headlines, standfirsts, full bodies, quoted sentences, and journalistic narration. The two tools are not merged automatically. Agreement may indicate lexically explicit evaluation; disagreement can reveal negation, irony, long syntax, metaphor, or domain-specific vocabulary. A headline-body difference is retained because promotional or compressed headlines may sound more judgmental than the report they introduce.

Automated classification is validated on a stratified sample covering outlets, regions, topics, score ranges, and cases in which the two tools disagree. At least two trained coders independently assess polarity, intensity, target, and attribution. They identify whether evaluation is directed toward an event, a government, an institution, a social group, or Africa as a generalized category. They also record whether the stance belongs to the journalist, a quoted source, or both. Reliability can be measured with Krippendorff's alpha, and held-out cases can be evaluated through macro-F1, confusion matrices, and rank correlation.

3.4. Framing, Voice, and Agency Coding

Sentiment becomes more meaningful when it is linked to contextual variables. Each article is coded for dominant topic, country specificity, causal attribution, proposed response, source composition, and representation of agency. Source categories may include African government, opposition, civil society, scholar, business, citizen, regional organization, foreign government, and international organization. Agency coding asks whether African actors appear as decision-makers and interpreters or mainly as recipients, locations, and objects of external action. This distinction prevents positive language about aid or investment from being mistaken automatically for empowering representation.

Geographic coding also tests homogenization. Researchers should distinguish reports that use "Africa" accurately for continent-wide institutions or trends from those that generalize evidence drawn from a few countries. Country counts must precede country comparisons, and countries represented by only one or two articles should not receive stable image labels. Concordance lines around terms such as "Africa," "African," "crisis," "reform," and "opportunity" can reveal recurring associations, but interpretation must examine the sentence, paragraph, speaker, and topic rather than relying on isolated keywords.

3.5. Comparison, Robustness, and

Reproducibility

The analysis begins with distributions rather than one overall average. Median polarity, dispersion, and the proportion of positive, neutral, and negative units are reported by outlet, topic, country, genre, and textual layer. Article counts accompany every comparison. Because sentiment values are often skewed, non-parametric tests or regression models with robust standard errors are preferable to untested assumptions of normality. Topic, article length, month, genre, and outlet can be introduced as controls when examining geographic differences. Matched comparisons with non-African reporting on similar topics help determine whether a pattern is Africa-specific or part of a publication's wider editorial style.

Robustness checks vary thresholds, units of analysis, and treatment of quotations. Results based on sentence means are compared with article-level scores and alternative length adjustments. The analysis is repeated after removing direct quotations and after separating topics with predictably adverse vocabulary, such as armed conflict or epidemic disease. Bootstrap confidence intervals communicate uncertainty without implying that software output is error-free. Search queries, inclusion decisions, cleaning rules, package versions, random seeds, coding manuals, and corrections are preserved in a versioned audit trail. Failed classifications are reported alongside successful examples so that reproducibility includes the limits of the method.

4. Analytical Framework

4.1. Event Negativity and Representational Negativity

The most important interpretive distinction is between the negative character of an event and the negative construction of a place. Reports on violence, displacement, hunger, disease, repression, or disaster necessarily contain adverse vocabulary. A classifier may score such an article correctly while offering no evidence of prejudice. The relevant question is whether the language is proportionate to the event, whether comparable events elsewhere receive similar treatment, and whether the report extends criticism beyond identifiable actors. Describing a failed policy is not equivalent to portraying a country as inherently incapable.

Selection nevertheless matters. An outlet may write accurate individual reports and still create a narrow image if it repeatedly selects crisis topics while overlooking ordinary politics, culture, research, local enterprise, or social change. Topic distributions must therefore be read beside sentiment distributions. Sparse coverage can appear neutral simply because little is published, whereas sustained reporting from conflict zones may appear more negative because the outlet is present. Silence is not balance, and publication volume is not bias; both require interpretation in relation to topic and geographic range.

Economic and development reporting illustrates the limits of average polarity. A story may describe growth as promising but fragile, reform as ambitious but politically difficult, or investment as beneficial but externally controlled. Opposing evaluations can cancel each other numerically even though the later clause carries greater argumentative weight. Contrast markers, sentence order, and causal explanation therefore matter. Concordance reading can distinguish conditional optimism from genuine neutrality and can show whether risk is attributed to specific policies or naturalized as a permanent feature of African societies.

4.2. Headlines, Article Bodies, and Attributed Speech

Headlines deserve independent analysis because they compress complex arguments and compete for attention. They may use metaphor, irony, wordplay, or categorical judgment that general sentiment dictionaries misread. A strongly negative headline followed by a qualified body may reflect editorial promotion rather than a stable article-level stance. Headline-body gaps should therefore be compared across outlets and topics. If similar gaps occur in reporting on Europe, Asia, or the Americas, they may indicate a general headline style rather than an Africa-specific tendency.

Attribution creates a second difficulty. A minister may call a policy successful, an opposition figure may describe it as disastrous, and the journalist may evaluate evidence without adopting either claim. When quotation boundaries are removed, these positions collapse into one score. Separating quoted and unquoted sentences is a necessary first step, but source order, paraphrase, and concluding placement also distribute authority. Human coding should ask not only whether African voices appear, but

which voices are selected, whether their claims are contextualized, and whether they are allowed to interpret events rather than merely react to foreign initiatives.

4.3. Outlet, Topic, and Geographic Differences

Foreign media should never be treated as a single speaker. Publications differ in audience, ownership, political orientation, regional expertise, genre, and access to correspondents. An outlet specializing in finance will select different African stories from a public broadcaster or foreign-policy magazine. Raw sentiment differences may therefore reflect issue mix rather than evaluation. Outlet comparisons become credible only after controlling for topic and genre and after checking whether the same country or event is being compared. Multi-level models can separate article-level features from outlet-level tendencies when the corpus is large enough.

Geographic imbalance is equally consequential. A small group of heavily covered countries can dominate an apparent continental pattern because of elections, conflict, market size, diplomatic importance, or reporting infrastructure. Regional summaries should show which countries contribute to each result. The term "Africa" should be examined as a linguistic choice: sometimes it names an appropriate continental institution or shared policy issue; elsewhere it may erase national differences. Evidence of homogenization becomes stronger when continental labels recur alongside limited country specificity, narrow sourcing, and generalized causal explanations.

4.4. From Computational Pattern to Critical Claim

A defensible critical argument should move through four levels of evidence. The lexical level identifies the words and constructions that produce a score. The textual level determines who speaks, what is evaluated, and whether the evaluation is qualified. The corpus level establishes how frequent and widely distributed the pattern is. The comparative level asks whether similar patterns appear in non-African coverage under comparable conditions. Skipping any level invites overstatement: a striking sentence may be unrepresentative, while a statistical difference may disappear after topic or outlet is controlled.

Methodological caution does not require avoiding criticism. If crisis topics are disproportionately selected, African sources are less visible, continental labels replace country names, and negative headline-body gaps persist after matched comparisons, the combined evidence would support a strong conclusion about representational imbalance. Conversely, diverse topics, substantial country specificity, varied African voices, and mixed evaluative patterns would require a differentiated account. The value of the framework is that either result can be reported. It turns broad concerns about media power into claims that can be traced, compared, challenged, and refined.

5. Limitations and Future Research

The framework cannot eliminate the limitations of automated language analysis. VADER and TextBlob rely heavily on lexical cues and do not fully resolve sarcasm, cultural references, complex negation, or long-range argument. Human annotation corrects many errors but introduces judgment, particularly when coders decide whether criticism of a government constructs a wider national or continental image. English-language reporting is also only one part of foreign media discourse. Translation, local editions, images, video, and platform-specific distribution may alter how evaluative meaning is produced.

Future research should compare languages, media systems, time periods, and regions using compatible sampling and coding rules. Transformer-based classifiers may improve contextual performance, but they should be evaluated against human judgments and interpretable lexical baselines. Visual framing and audience reception also require separate study. Textual patterns matter because they may organize public knowledge, yet effects on readers cannot be inferred from content alone. The present framework is best understood as a transparent foundation for later empirical work rather than a universal measure of media bias.

6. Conclusion

Python can extend the scale and transparency of research on foreign media coverage of Africa, but computation is most persuasive when it supports rather than replaces interpretation. A rigorous design separates headlines, article bodies, and quotations; records topic, country, source, and agency; validates automated scores

with human coding; and compares like with like across outlets and regions. This approach avoids two symmetrical errors: assuming in advance that foreign reporting is uniformly negative, and treating apparently neutral averages as proof that representation is balanced.

The framework moves analysis beyond polarity by linking numerical patterns to language, speakers, topics, geography, and comparison sets. It preserves critical inquiry while making evidence reproducible and claims precise. Foreign media and Africa remain diverse categories. Respecting that diversity shows where reporting narrows understanding or challenges established assumptions.

Acknowledgments

This paper is supported by the 2025 University-level Scientific Research Project of Xi'an Fanyi University titled Python Technology-Driven Sentiment Analysis of Foreign Media Coverage of Africa (Project No.: 2025B07), with the author serving as the principal investigator. It also acknowledges the 2026 General Research Project of the Shaanxi Provincial Sports Bureau titled Enhancing the Social Influence and Cultural Communication Capacity of Shaanxi Competitive Sports within and beyond the Province: A Study of Localized Communication Pathways for Shaanxi Tennis from an International Perspective (Project No.: 20261005), and the 2024 Special Project for Research Institutions of Xi'an Fanyi University titled Corpus-Based Pathways for Building a China-Africa Community with a Shared Future in the New Era from the Perspective of Area and Country Studies (Project No.: 2024Z08), both of which are headed by the author as principal investigator.

References

[1] Scott M. The myth of representations of Africa: A comprehensive scoping review of the literature[J]. Journalism Studies, 2017, 18(2):191-210.

[2] Nothias T. How Western journalists actually write about Africa: Re-assessing the myth of representations of Africa [J]. Journalism Studies, 2018, 19 (8):1138-1159.

[3] Galtung J, RUGE M H. The structure of foreign news[J]. Journal of Peace Research, 1965, 2(1):64-90.

[4] Harcup T, O'Neill D. What is news? News values revisited (again)[J]. Journalism Studies, 2017, 18 (12):1470-1488.

[5] Bunce M, Franks S, Paterson C, eds. Africa's Media Image in the 21st Century [M]. London: Routledge, 2017.

[6] Taboada M, Brooke J, Tofiloski M, voll K, Stede M. Lexicon-based methods for sentiment analysis[J]. Computational Linguistics, 2011, 37(2):267-307.

[7] Hutto C J, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2014:216-225.

[8] Young L, Soroka S. Affective news: The automated coding of sentiment in political texts[J]. Political Communication, 2012, 29(2):205-231.

[9] Entman R M. Framing: Toward clarification of a fractured paradigm[J]. Journal of Communication, 1993, 43(4):51-58.

[10] Yang J, Cui H, Zhang Y, et al.From "Angels" to "Role Models of the Times": A Discourse-Historical Analysis of Ningxia's Media Coverage of Medical Aid to Benin (2006–2024)[J].Scientific and Social Research,2026,8(3):150-157.

[11] Hong A, Hu D. Analyzing Foreign Media Coverage of China During the 2022 Beijing Winter Olympics Opening and Closing Ceremonies: A Corpus-Assisted Critical Discourse Analysis[J].Journalism and Media,2025,6(3):145-145.

[12] KumarP. Analysing foreign media coverage of India's Chandrayan-3 mission[J].Asian Politics & Policy,2024,16(2):294-297.

[13] Li Y ,Ruan X. Exploring the Role of Media Framing in the Self-construction of China's National Image: An Analysis of News Coverage About Foreign Language Education in China Daily[J].The Educational Review, USA,2025,9(1):103-111..

