

Interpretable Machine Learning for Narrative Framing Classification: A TF-IDF-Based Study of Frankenstein

Luthfiana Dwi Rakhmawati^{1*}, Rusdi Aulia Romadhon², Kiki Adnan Muzaki³, Anna Nurawalia⁴,
and Dewi Aprilia Ningrum⁵

¹English Literature, Faculty of Adab and Humanities, UIN Syarif Hidayatullah Jakarta, Indonesia

²Informatics Engineering, Faculty of Science and Technology, UIN Syarif Hidayatullah Jakarta, Indonesia

³Religious Studies Department, University of California, Riverside, USA

⁴English Language Department, Politeknik Bau-Bau, Indonesia

⁵Centre for Planners' Development, Education, and Training (Pusbindiklatren), Ministry of National Development Planning/Bappenas

ABSTRACT

Open Access

Type: Research Paper

Article:

Received: February 12, 2026

Revised: May 16, 2026

Accepted: September 13, 2026

Issued: October 30, 2026

© Rakhmawati et al. (2026).



This is an open-access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

*Correspondence Address:

luthfiana.dwi@uinjkt.ac.id

This study evaluates interpretable supervised machine learning for three-class narrative framing classification of literary text, using Mary Shelley's Frankenstein (1818) as a case study; narrative framing is the evaluative position a passage assigns to the Creature: Threat, Victim, or Ambivalent. The methodology comprised five stages: (1) extraction of passages narrated by Victor Frankenstein, the Creature, and Robert Walton from the 1818 Project Gutenberg text; (2) annotation of 195 passages, with a blinded second annotator re-labeling 68 passages (Cohen's $\kappa = 0.6926$); (3) sublinear TF-IDF unigram and bigram representation; (4) classification with Multinomial Naive Bayes, Linear SVM, and Logistic Regression against Majority and VADER baselines under stratified 5-fold cross-validation; and (5) a designation-masking ablation and coefficient analysis. The results show that Logistic Regression obtained the highest mean scores (66.67% accuracy; 47.75% Macro-F1), although its margin over Linear SVM (46.18%) and VADER (44.85%) lies within fold-level variability. Pooled F1 reached 69.0% for Threat and 74.3% for Victim, whereas none of the 29 Ambivalent passages was correctly identified. Masking explicit designations reduced Macro-F1 only to 45.20%, indicating partial robustness to this procedure rather than discourse-level modeling. It is concluded that interpretable classical classifiers can separate Threat from Victim framing using transparent lexical evidence but cannot yet detect ambiguous framing; given the single-novel, small, and imbalanced dataset, the findings are a preliminary proof of concept. Future work should enlarge and balance the corpus, apply group-based and repeated cross-validation, and compare context-aware models, including large language models, on the Ambivalent class.

Keywords: narrative framing; text classification; TF-IDF; interpretable machine learning; computational literary studies; Frankenstein

1. INTRODUCTION

Mary Shelley's *Frankenstein; or, The Modern Prometheus*, first published in 1818, is widely recognized as a foundational work of world literature [1]. For more than two centuries, literary scholars have actively debated the moral status and narrative representation of Victor Frankenstein's nameless Creature [2]. Consistently, Victor labels his creation through vilifying terms such as monster, daemon, and fiend, whereas the Creature presents himself as an initially benevolent being driven to despair by relentless social rejection [1], [2]. Analyzing this character representation is complicated by the novel's nested tripartite narrative structure: the outermost epistolary frame of Robert Walton encapsulates Victor Frankenstein's confessional narrative, which in turn encloses the Creature's autobiographical account [3]. In conventional literary criticism, characterization has been evaluated through close textual reading of selected passages [2]. Yet, insightful as it is, manual close reading presents methodological limitations: it is susceptible to confirmation bias and is difficult to extend into a systematic and reproducible measurement of how framing varies across a whole text [4], [5], [37].

From the standpoint of literary theory, this study defines narrative framing as the evaluative position a passage assigns to the Creature through the configuration of its narrative discourse. Following Genette [33], three narratological categories make this definition operational. First, voice answers the question 'who speaks?', which Genette distinguishes from mood and its question 'who sees?' [33, p. 186]; this distinction separates Victor's narration from the Creature's speech that Victor reports. Second, focalization, Genette's more abstract replacement for 'point of view', specifies through whose knowledge or perception the Creature is presented [33, p. 189]. Third, narrative levels describe how any event a

narrative recounts lies at a diegetic level immediately higher than the narrating act that produces it [33, p. 228]: Walton's epistolary narrating is extradiegetic, Victor narrates at the intradiegetic level so that his story is metadiegetic, and the Creature's autobiography, told within Victor's story, forms a third degree that Genette terms a meta-metanarrative [33, p. 228]. Complementing this structural framework, Brooks [34] argues that the central issues of *Frankenstein* are joined in the opposition of sight and speech: the Creature is the most eloquent figure in the novel and, as a verbal creation, "the very opposite of the monstrous", whereas his visible body is the source of his monstrosity [34, pp. 199–202]. Brooks's reading explains how a single figure can be framed as both victim and threat, and it provides the interpretive lens for the close reading of cross-framed passages in Section 3.6.

In this study, the narratological framework informs annotation and interpretation rather than feature design. Voice, focalization, and narrative level guide how passages are assigned to narrators and how annotators judge whose evaluation of the Creature a passage conveys, whereas the classifiers see only TF-IDF word and bigram frequencies, which do not encode voice or focalization directly. The models can therefore approximate theory-guided labels only through their lexical correlates, such as first-person pronouns in the Creature's self-narration or evaluative vocabulary in Victor's report, and Section 3.8 examines where this approximation fails. The three classes follow the critical debate outlined above, which reads the Creature as a monster, as a victim, or as a figure in whom both readings coexist [2], [34]: Threat and Victim capture the two poles, and Ambivalent captures passages in which the tension between them is itself the evaluative content, as in Brooks's opposition between the persuasive tongue and the hideous body [34]. The labels are treated as

mutually exclusive at the passage level, and passages with mixed cues were resolved by the priority rules in Section 2.4 (Table 1).

Recent advances in Natural Language Processing (NLP) and Computational Literary Studies (CLS) provide empirical methods for modeling narrative phenomena [3], [4], and several lines of work come close to the problem addressed here. The most common computational entry point to literary characters has been sentiment and emotion analysis, which surveys of the field show to be widely used for tracking emotional development and character relations in fiction [10], [11]. Beyond sentiment, one line is character-centric narrative understanding. Brahman et al. [6] introduced LiSCU, a dataset that pairs literary summaries with descriptions of their characters, and evaluated pre-trained language models on character identification and description generation. Yang et al. [7] extended this direction to characters' inner lives: their LitCharToM benchmark, drawn from 20 classic books, shows that temporal knowledge graphs of characters' beliefs, emotions, intentions, and desires improve the theory-of-mind reasoning of large language models. Both studies model what a character is or thinks, but neither asks how a narrator evaluates that character. Narrative perspective itself is the target of Brei et al. [8], who classified types of unreliable narrators in TUNA, a corpus of 817 narratives from four domains, using zero- and few-shot prompting, fine-tuning, and curriculum learning; their object, however, is the reliability of the narrator rather than the stance the narrator takes toward another character. Toro Isaza et al. [9] analyzed the gendered event chains of characters in children's fairy tales, but their analysis is descriptive and does not classify how characters are framed. Weimer et al. [5] came closer to interpretive categories by annotating and automatically detecting narratorial comment in literary texts, yet their classifiers reached F-scores of only 28–

59%, which shows how difficult it is to operationalize literary concepts even with a dedicated annotation scheme. The framing scheme nearest to the present task comes from outside literature: Frermann et al. [13] modeled the narrative framing of entities as heroes, villains, or victims in English news articles, using supervised, semi-supervised, and retrieval-based multi-label prediction.

In Indonesian applied informatics, Term Frequency-Inverse Document Frequency (TF-IDF) combined with classical supervised machine learning algorithms (Naive Bayes, Support Vector Machines, and Logistic Regression) has been widely applied to text classification across diverse national corpora and is computationally efficient [16], [17], [18], [19], [20], following well-established practice in supervised text categorization [29], [35]. Shakti et al. [16] applied TF-IDF with Naive Bayes to administrative document classification. Saraswati et al. [17] and Adinata et al. [18] applied SVM and Logistic Regression to public review classification. Fitriani and Hidayat [19] as well as Fauzi and Kharisma [20] applied TF-IDF and feature selection to opinion and narrative texts. However, these studies address contemporary functional texts (official letters, application reviews, and public opinion) and do not model evaluative framing across nested narrative perspectives.

Taken together, these studies leave three gaps that motivate the present work. First, research on how literary characters are evaluated still tends to use sentiment polarity as a proxy for framing [10], [11]. General-purpose lexicons such as VADER were developed for social media text [24] and represent evaluation on a single positive–negative scale, whereas in nineteenth-century fiction the passages that frame the Creature as an oppressed victim are saturated with negative affective tokens (misery, tears, sorrow, anguish); a polarity lexicon therefore has

no principled way to separate sympathetic suffering from hostility [10], [11]. Second, none of the studies reviewed above classifies the evaluative framing of a single character across the nested narrators of one canonical novel. Character-understanding work [6], [7] targets descriptions and mental states, narrator-oriented work [8] targets reliability, the fairy-tale study [9] remains descriptive, and the hero–villain–victim scheme of Frermann et al. [13], although conceptually the closest, was designed for multi-label news framing, in which no embedded narrators mediate the evaluation. Third, the most recent character and narrator studies rely increasingly on large language models (LLMs) [7], [8], whose predictions are difficult to attribute to specific textual evidence [12], so it remains unclear what transparent lexical models can and cannot capture in a small, interpretively annotated literary corpus. The novelty of this study lies at the intersection of these gaps: it transfers the hero–villain–victim logic of narrative framing [13] from news discourse to a literary text, recasts it as a three-class Threat–Victim–Ambivalent task for one character told through three nested narrators, and evaluates it with interpretable classical classifiers against a sentiment-lexicon baseline. This study does not claim that LLMs are unsuitable for the task; it deliberately restricts its scope to transparent classical models whose decisions can be audited against lexical features [12], [14], and it leaves an LLM comparison to future work (Section 4).

To address these gaps, this study makes three contributions: (1) it formulates a three-class narrative framing task (Threat, Victim, Ambivalent) for one character across the nested narrators of a canonical novel, supported by a documented annotation codebook and inter-annotator agreement; (2) it evaluates classical supervised classifiers against majority and sentiment-lexicon baselines and uses a designation-masking ablation to test how far

performance depends on explicit epithets; and (3) it uses logistic regression coefficients and error analysis to examine which lexical cues the models rely on and where they fail. The work is positioned as an exploratory, single-novel proof of concept rather than a general framework. Accordingly, it addresses three research questions (RQs):

RQ1: To what extent can TF-IDF-based classical classifiers reproduce human Threat, Victim, and Ambivalent framing labels, relative to majority and sentiment-lexicon baselines?

RQ2: How far does classification performance depend on explicit designations of the Creature (e.g., monster, fiend, daemon), as tested by masking them?

RQ3: Which lexical features are associated with each framing class, and how are the human-assigned labels distributed across the narrators Victor Frankenstein, the Creature, and Robert Walton?

2. METHODS

2.1 Research Framework and Methodological Flow

As illustrated in Figure 1, the research methodology operates across seven sequential stages: (1) Corpus Acquisition and Cleaning; (2) Rule-Based Passage Extraction; (3) Manual Annotation, Pilot Calibration, and Gold Standard Construction; (4) TF-IDF Feature Engineering and Lexical Leakage Masking; (5) Supervised Model Training and Baseline Modeling; (6) 5-Fold Stratified Cross-Validation; and (7) Lexical Interpretability and Cross-Narrator Analysis.

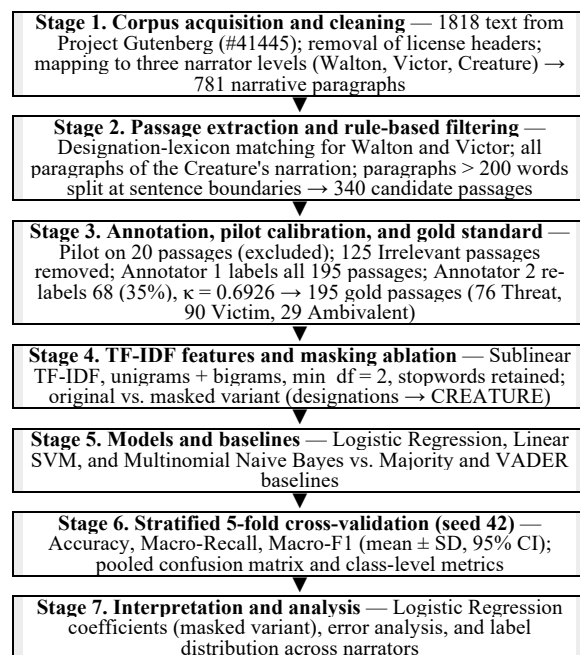


Figure 1. Methodological workflow of the study, with the number of passages retained at each stage

2.2 Corpus Acquisition and Narrative Segmentation

The analysis is based on the first (1818) edition of *Frankenstein* as made available by Project Gutenberg (EBook #41445) [1]. This edition was selected as it maintains the original three-volume structure prior to Shelley's extensive revisions of 1831. Since the two editions differ in both chapter numbering and philosophical emphasis, they are treated here as distinct texts. Gutenberg headers, license metadata, and transcriber notes were removed. The cleaned corpus contains 72,118 words across 781 narrative paragraphs.

Each paragraph was mapped to its respective narrator level: Robert Walton (Letters I–IV and closing journal), Victor Frankenstein (Vol. I Chapters I–VII, Vol. II Chapters I–II & IX, Vol. III Chapters I–VII), and the Creature (Vol. II Chapters III–VIII). In Genette's terms [33], these tiers correspond to the extradiegetic (Walton), metadiegetic (Victor), and meta-

metadiegetic (Creature) levels of the novel's nested narrative.

2.3 Passage Extraction and Rule-Based Filtering

The unit of analysis was defined at the paragraph-level passage. In 19th-century Gothic fiction, paragraphs frequently extend to 500–800 words and contain multi-emotional shifts (topic drift). Consequently, paragraphs exceeding 200 words were split into consecutive segments at the sentence boundary closest to each 200-word mark, so that no sentence was divided. Each resulting segment was annotated independently rather than inheriting the label of its source paragraph, segments whose framing still changed internally were labeled with the priority rules in Table 1, and every segment retained the identifier of its source paragraph. Because framing can still shift within a segment, this unitization remains a source of error (Section 3.8).

In Walton's and Victor's narrations, passages were extracted if they contained at least one Creature designation keyword from a domain lexicon (*designations.txt*, comprising *creature*, *monster*, *daemon*, *fiend*, *wretch*, *being*, *devil*, *ogre*, *spectre*, *insect*, and *creation*; the complete list is provided with the released materials, Section 2.7) matched case-insensitively with plurals. Conversely, third-person pronouns (*he*, *him*, *his*, *she*, *her*, *it*) were deliberately excluded as automated triggers due to severe coreference ambiguity: '*she*' exclusively references female characters (Elizabeth, Justine, Safie), '*it*' predominantly references natural forces or scientific concepts, and '*he*' far more frequently refers to other male characters (such as Henry Clerval: 'He was a boy of singular talent...' or M. De Lacey: 'He was old and blind...'). Including third-person pronouns would pull over 90% of the entire novel as noise.

In the Creature's narration, all paragraphs were extracted because the Creature refers to

himself in the first person (I, me, my). A designation lexicon would miss most of this self-referential narration. The two strategies are therefore deliberately asymmetric, but the resulting samples are not directly comparable: Victor's and Walton's passages are, by construction, those in which the Creature is explicitly designated, whereas the Creature's sample covers all of his narration, including passages about the De Lacey family. The label distributions reported in Table 2 are therefore conditional on this extraction procedure and are not estimates for the novel as a whole; a sensitivity analysis with a uniform extraction strategy is left to future work. In total, 340 candidate passages resulted from the extraction process.

2.4 Annotation Scheme, Pilot Calibration, and Inter-Annotator Agreement

Manual annotation categorized candidate passages into three distinct narrative framing classes:

1. Threat: Passages that depict the Creature as an aggressive monster, murderer, demonic entity, or an agent of terror and revenge.
2. Victim: Passages that present the Creature as a recipient of intense physical or emotional distress, unjust social rejection, or deserving sympathy.
3. Ambivalent: Passages that reveal moral duality, holding sympathy and horror in tension, framing the Creature's goodness as conditional, or linking his suffering causally to his malevolence (e.g., the Creature's admission: 'I was benevolent and good; misery made me a fiend. Make me happy, and I shall again be virtuous' [1]).

A fourth category, Irrelevant (Exclude), was applied during the filtering stage to remove passages containing no characterization of the

Creature, such as the backstory of Safie's family in Constantinople or purely geological descriptions of the Chamounix valley.

Table 1. Annotation codebook: inclusion criteria, boundary rules, and examples

Class	Assign when (positive example)	Do not assign when (boundary rule)
Threat	The passage presents the Creature primarily as dangerous: violence, murder, menace, revenge, or demonic agency. Example: "Nothing in human shape could have destroyed that fair child. He was the murderer!"	Danger is mentioned only as a consequence of suffering on which the passage dwells (→ Ambivalent), or hostile epithets come from other characters while the passage sympathizes with him (→ Victim).
Victim	The passage presents the Creature primarily as suffering, rejected, excluded, or deserving sympathy. Example: "I was a poor, helpless, miserable wretch; ... I sat down and wept."	The passage culminates in a vow or act of violence that dominates its evaluation (→ Threat), or sympathy and horror receive comparable weight (→ Ambivalent).
Ambivalent	Sympathy and horror are held in tension, his goodness is made conditional, or his suffering is causally linked to his malevolence. Example: "I was benevolent and good; misery made me a fiend."	The passage merely contains both positive and negative words while one evaluation clearly dominates (→ Threat or Victim).
General rules	One label per passage, judged on the evaluation of the Creature conveyed by the passage as a whole, whoever narrates it. When cues conflict, the evaluation that dominates the passage decides; Ambivalent is used only when neither evaluation dominates or a causal link is explicit.	Passages that do not characterize the Creature (e.g., Safie's family history, descriptions of the Chamounix valley) are labeled Irrelevant and excluded.

Before full annotation began, a calibration pilot was carried out on 20 randomly selected candidate passages by Annotator 1 (English literature lecturer) and Annotator 2 (linguistics lecturer, serving as annotation validator). The

pilot served to align the interpretation of the guidelines, to resolve borderline cases, and to fix the boundary rules summarized in Table 1. These 20 passages were then excluded from all subsequent annotation, agreement, and evaluation steps.

The dataset was constructed as follows (Figure 1): 340 candidate passages were extracted; 20 were reserved for the pilot, leaving 320; 125 of these were labeled Irrelevant and removed; and the final gold standard comprised 195 passages (76 Threat, 90 Victim, and 29 Ambivalent).

Blinding was limited in scope: annotators labeled the passage text with the narrator and extraction-trigger columns hidden. This prevented the metadata from guiding the labels, but it did not prevent annotators from inferring the narrator from the text itself (e.g., first-person self-reference in the Creature's narration or references to Victor's family). The labels should therefore be regarded as narrator-aware interpretive judgments rather than narrator-independent ones.

Annotator 1 labeled all 195 passages. Annotator 2 independently labeled a random 35% subset (68 passages) drawn from the 195 gold-standard passages after the pilot passages had been removed, and agreement was computed on these 68 passages only. The reliability sample contained 21 Threat, 36 Victim, and 11 Ambivalent passages (based on Annotator 1's initial labeling; Annotator 2 identified 20 Threat, 32 Victim, and 16 Ambivalent); the two annotators disagreed on 13 passages (19.1%), most frequently between Victim and Ambivalent (8 instances: 6 where Annotator 1 marked Victim and Annotator 2 marked Ambivalent, and 2 vice versa). Disagreements were resolved through discussion, and the adjudicated label was retained. Inter-annotator reliability was

calculated using Cohen's kappa (κ) in Equation (1) [21]:

$$\kappa = (p_o - p_e) / (1 - p_e) \quad (1)$$

Here, p_o represents the observed proportion of agreement, while p_e represents the agreement expected by chance. The annotation yielded a Cohen's kappa $\kappa = 0.6926$, which falls within the “substantial agreement” range ($\kappa \geq 0.61$) under the Landis and Koch benchmark [22]. Following Artstein and Poesio [15], this level of agreement is adequate but not high for an interpretive task, and it bounds the performance that a classifier can be expected to reach; disagreements in literary annotation can also be informative about how a concept is operationalized [31].

2.5 TF-IDF Feature Representation and Lexical Masking Ablation

Each passage was transformed into a numerical feature vector using Term Frequency–Inverse Document Frequency (TF-IDF) weighting [23], [32], [35] with unigrams and bigrams ($ngram_range = (1, 2)$), minimum document frequency $min_df = 2$, and sublinear term frequency scaling, implemented with scikit-learn's `TfidfVectorizer` ($sublinear_tf = True$, $smooth_idf = True$, $norm = 'l2'$, $lowercase = True$) [25]. The feature weight $w(t, d)$, the smoothed $idf(t)$, and the L2 normalization of each passage vector are defined in Equations (2)–(4), which correspond to this implementation; $w(t, d) = 0$ when $tf(t, d) = 0$:

$$w(t, d) = (1 + \ln tf(t, d)) \times idf(t) \quad (2)$$

$$idf(t) = \ln((1 + n) / (1 + df(t))) + 1 \quad (3)$$

$$v(d) = w(d) / \|w(d)\|_2 \quad (4)$$

In contrast to the extraction stage, where pronouns were excluded as retrieval filters, stopwords were intentionally preserved during TF-IDF vectorization. Personal pronouns (e.g., I, my, he, his) encode important stylistic signals that separate first-person introspective speech

(the Creature) from third-person observational narration (Victor).

To examine whether the models capture broader framing discourse rather than simply memorizing pejorative labels (lexical leakage), two feature variants were established:

1. Original Variant: Standard text without modification.
2. Masked Variant: All terms in the designation lexicon of Section 2.3 (including monster, fiend, daemon, devil, wretch, and insect, with their plural forms) were replaced with the single placeholder token 'CREATURE' by a Python script before vectorization. Because the vectorizer lowercases text, the placeholder merges with the word creature, so that in the masked variant the token creature stands for any designation. The placeholder can therefore itself become a predictive feature (Section 3.5), words outside the lexicon may still signal the labels, and only this one masking strategy was tested. The vectorization step was embedded within scikit-learn Pipelines [25] to prevent test-data leakage during cross-validation.

2.6 Supervised Machine Learning Models, Baseline Formulations, and 5-Fold Cross-Validation

Three supervised machine learning algorithms were assessed: Multinomial Naive Bayes (MultinomialNB), Linear Support Vector Machine (LinearSVC), and Logistic Regression (LR) [25], [29]; their configuration is given in Section 2.7. Two baselines were set up for comparison:

1. Majority Baseline: A Dummy Classifier that labels every passage with the most frequent class (Victim), serving as a minimal reference point for performance.

2. VADER Lexicon Baseline: the rule-based VADER sentiment analyzer [24], applied without any training.

Compound scores were calculated for each sentence and then averaged across the passage. The compound score normalizes valence to the range $[-1.0, +1.0]$ using $x = \text{sum}(v) / \sqrt{(\text{sum}(v))^2 + 15}$ [24]. Standard mapping thresholds were applied: compound ≤ -0.05 mapped to Threat; compound ≥ 0.05 mapped to Victim; and $-0.05 < \text{compound} < 0.05$ mapped to Ambivalent. These ± 0.05 cut-offs are the conventional thresholds distributed with the VADER implementation [24]. Mapping negative polarity to Threat, positive polarity to Victim, and the neutral band to Ambivalent operationalizes the polarity-as-framing assumption questioned in Section 1. The VADER baseline is therefore a diagnostic test of that assumption rather than a competitive framing classifier; a framing-specific rule-based baseline could not be defined without reusing the annotation criteria themselves.

All configurations were evaluated using stratified 5-fold cross-validation with a fixed random seed (42), using the same folds for every configuration. Each test fold contains 39 passages, with about 15–16 Threat, 18 Victim, and 5–6 Ambivalent passages, so a single Ambivalent prediction changes fold-level Ambivalent recall by roughly 17–20 percentage points. The folds were stratified by class but not grouped by source paragraph or episode; segments split from the same long paragraph, or adjacent passages from the same episode, may therefore fall into different folds and share distinctive vocabulary, which can make the estimates optimistic (Section 3.9). For each configuration, Accuracy, Macro-Recall, and Macro-F1 (Equations 5–6) are reported as the mean $\pm$ standard deviation over the five test folds [25], [30], and 95% confidence intervals for the fold means are computed as $\text{mean} \pm t(0.975, 4) \cdot s / \sqrt{5}$. In addition, the out-of-fold predictions of

the best configuration were pooled into one confusion matrix (Figure 2), in which each passage appears exactly once, and class-level precision, recall, and F1 were derived from it (Table 4):

$$F1_k = 2 \times (P_k \times R_k) / (P_k + R_k) \quad (5)$$

$$Macro-F1 = (1/K) \times \sum F1_k \quad (6)$$

2.7 Experimental Configuration and Reproducibility

Experiments were implemented in Python 3.12 with scikit-learn 1.7.2 [25] and the vaderSentiment package 3.3.2 [24]. No hyperparameter tuning was performed: all classifiers used scikit-learn default settings except where stated, so no data were used for model selection, and the cross-validation estimates are not affected by tuning. Logistic Regression used L2 regularization with $C = 1.0$, the lbfgs solver, and `class_weight = 'balanced'`; LinearSVC used L2 regularization with $C = 1.0$ and `class_weight = 'balanced'`; MultinomialNB used additive smoothing $\alpha = 1.0$ (it does not support class weights); and the majority baseline used `DummyClassifier(strategy = 'most_frequent')`. The random seed (42) was used for fold assignment and, where applicable, for the classifiers.

To support reproducibility, the annotation codebook, passage identifiers with source-paragraph and chapter indices, gold labels, the designation and masking lexicon, the extraction and masking scripts, and the model configuration are publicly available at <https://github.com/rusdi313/frankenstein-framing-nlp>. Because the 1818 text is in the public domain (Project Gutenberg EBook #41445), the processed passages can be shared in full.

3. RESULTS AND DISCUSSION

3.1 Gold Standard Dataset Characteristics and Cross-Narrator Distribution

Of the 340 candidate passages initially extracted, 20 were reserved for pilot calibration and 125 were labeled Irrelevant, leaving a final gold standard of 195 passages ($340 - 20 - 125 = 195$). Table 2 summarizes the distribution of the human-assigned labels across narrator tiers. These proportions describe the annotated sample produced by the extraction procedure in Section 2.3, not the distribution of framing across the whole novel. Walton contributes only five passages, so his row is reported for completeness and is not interpreted further.

Table 2. Distribution of gold-standard labels across narrator tiers (conditional on the extraction procedure)

Narrator Level	Total Passages (n)	Threat (n / %)	Victim (n / %)	Ambivalent (n / %)
Victor Frankenstein	109	60 (55.0%)	29 (26.6%)	20 (18.3%)
The Creature	81	14 (17.3%)	59 (72.8%)	8 (9.9%)
Robert Walton	5	2 (40.0%)	2 (40.0%)	1 (20.0%)
Total Corpus	195	76 (39.0%)	90 (46.2%)	29 (14.9%)

3.2 Classification Performance and Baseline Comparisons

Table 3 presents the stratified 5-fold cross-validation results for all model configurations and baselines. Logistic Regression on the original variant obtained the highest mean scores (Macro-F1 $47.75\% \pm 3.64\%$; accuracy $66.67\% \pm 5.13\%$), followed by Linear SVM ($46.18\% \pm 4.95\%$; $64.10\% \pm 7.25\%$). The 95% confidence intervals of the fold means overlap substantially (Macro-F1: Logistic Regression 43.2–52.3%, Linear SVM 40.0–52.3%, VADER 35.5–54.2%), so the ranking among these three configurations should not be read as a meaningful difference. With only five folds, a paired comparison such as the

corrected resampled t-test [36] has very low power; Logistic Regression is therefore reported as the configuration with the highest mean score rather than as a superior model. In accuracy, Logistic Regression exceeded VADER by 12.3 percentage points, but in Macro-F1 the gap was only 2.9 points. All supervised models clearly exceeded the majority baseline (21.05% Macro-F1); however, this baseline is deliberately minimal, and exceeding it does not establish that the approach is competitive with framing-specific methods.

Table 3. Classification performance (mean $\pm$ standard deviation over five stratified folds)

Classification Model	Feature Variant	Accuracy (%)	Macro-Recall (%)	Macro-F1 (%)
Logistic Regression	Original (Unmasked)	66.67 $\pm$ 5.13	51.47 $\pm$ 3.95	47.75 $\pm$ 3.64
Linear SVM	Original (Unmasked)	64.10 $\pm$ 7.25	49.69 $\pm$ 5.60	46.18 $\pm$ 4.95
Linear SVM	Masked (Ablation)	63.59 $\pm$ 5.23	49.25 $\pm$ 4.02	45.67 $\pm$ 3.54
Logistic Regression	Masked (Ablation)	63.59 $\pm$ 4.10	48.91 $\pm$ 3.02	45.20 $\pm$ 2.96
Multinomial Naive Bayes	Original (Unmasked)	61.03 $\pm$ 4.10	46.44 $\pm$ 3.23	42.54 $\pm$ 3.05
Multinomial Naive Bayes	Masked (Ablation)	58.46 $\pm$ 4.41	44.25 $\pm$ 3.61	39.88 $\pm$ 3.83
VADER Lexicon (Baseline)	Sentiment Lexicon	54.36 $\pm$ 7.50	45.80 $\pm$ 7.37	44.85 $\pm$ 7.55
Majority Class (Baseline)	Majority Baseline	46.15 $\pm$ 0.00	33.33 $\pm$ 0.00	21.05 $\pm$ 0.00

3.3 Aggregate Confusion Matrix and Class-Level Analysis

Figure 2 shows the confusion matrix of Logistic Regression (original variant), obtained by pooling the out-of-fold predictions of the five test folds so that each of the 195 passages is counted once. Its accuracy (130/195 = 66.67%) and Macro-F1 (47.78%) closely match the fold means in Table 3; the small Macro-F1 difference arises because pooled and fold-averaged metrics are computed differently. Table 4 reports the class-level metrics derived from this matrix.

Victim passages were recognized with high recall (81/90 = 90.0%) but moderate precision (63.3%), because 27 Threat and 20 Ambivalent passages were also predicted as Victim; Threat passages were recognized with 64.5% recall and 74.2% precision. The Ambivalent class was not recognized at all: all 29 Ambivalent passages were predicted as Victim (20) or Threat (9), and the model predicted Ambivalent only once, for a Victim passage, so its precision, recall, and F1 are all zero. The Macro-F1 of about 48% is therefore the average of two moderately recognized classes and one failed class rather than uniformly moderate performance. This is a serious limitation because the Ambivalent category is central to the claim that the Creature can be framed in conflicting ways; Section 3.8 analyzes its likely causes.

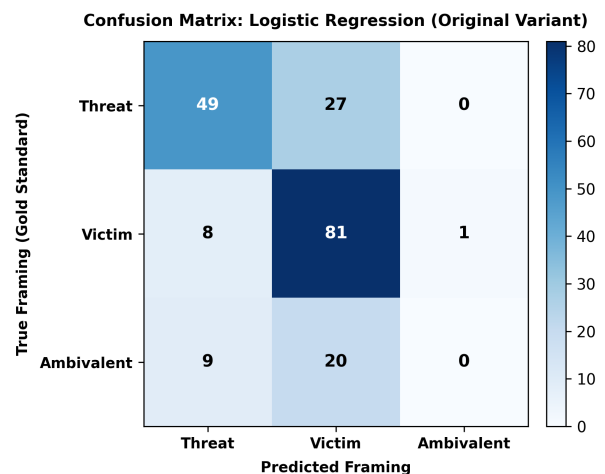


Figure 2. Confusion matrix of Logistic Regression (original variant), pooled over the out-of-fold predictions of the five test folds

Table 4. Class-level metrics of Logistic Regression (original variant) derived from the pooled confusion matrix in Figure 2

Class	Precision (%)	Recall (%)	F1 (%)	Support (n)
Threat	74.24	64.47	69.01	76
Victim	63.28	90.00	74.31	90
Ambivalent	0.00	0.00	0.00	29
Macro average	45.84	51.49	47.78	195

3.4 Impact of Lexical Masking (Ablation of Lexical Leakage)

The masking ablation in Table 3 shows that replacing all designation-lexicon terms (e.g., monster, fiend, daemon, wretch) with 'CREATURE' reduced Macro-F1 by 2.55 percentage points for Logistic Regression (from 47.75% to 45.20%) and by 0.51 points for Linear SVM (from 46.18% to 45.67%); both changes are smaller than the fold-level standard deviations. This indicates that the classifiers do not depend only on the masked epithets. It does not show that they capture discourse-level structure: other lexical correlates of the labels (e.g., domestic vocabulary in the De Lacey episodes or first-person pronouns in the Creature's narration), narrator-specific vocabulary, and the selection effects described in Section 2.3 could equally explain the result. In addition, the placeholder merges with the word creature and itself becomes the strongest Threat feature (Table 5), so the ablation tests a single masking strategy; alternatives such as deleting designations or masking other label-indicative words were not tested.

3.5 Lexical Markers of Narrative Framing (Regression Feature Weights)

Table 5 lists the highest-ranked n-gram features for each class, ordered by their positive logistic regression coefficients in the masked variant. The coefficients indicate associations within the fitted model, not causal effects, and the same term may play different roles in different passages; they are therefore read below alongside passages quoted in Section 3.6.

Table 5. *Top-ranked lexical features by Logistic Regression for each framing class (Masked Variant)*

Framing Class	Top Lexical Features (Regression Coefficients)
Threat	creature (0.55), my creature (0.55), the creature (0.52), to me (0.42), do (0.41), you (0.33), existence (0.33), do not (0.31), enemy (0.30), destroy (0.28)

Framing Class	Top Lexical Features (Regression Coefficients)
Victim	gentle (0.51), the old (0.43), old (0.41), old man (0.39), from the (0.36), amiable (0.34), virtue (0.33), man (0.32), affection (0.30), cottage (0.28)
Ambivalent	misery (0.44), from (0.41), alone (0.41), behold (0.40), still (0.40), of my (0.37), love (0.37), be the (0.37), pity (0.35), could (0.33)

The coefficients suggest the following tendencies:

- Threat vocabulary: the token 'creature' carries the highest weight (0.55). This reflects the masking procedure: all designations (monster, fiend, daemon) were mapped to 'creature', which is frequent in Victor's hostile denunciations. The same token, however, also occurs in Victor's sympathetic reflection on 'the duties of a creator towards his creature' (quoted in Section 3.6 as an instance of Victor's Victim framing), so a positive Threat coefficient does not fix the word's meaning. The phrases 'to me' (0.42), 'you' (0.33), and 'do not' (0.31), together with 'destroy' (0.28) and 'enemy' (0.30), plausibly reflect direct address and personal confrontation between the Creature and Victor.
- Victim vocabulary: this class is associated with terms of moral virtue and domestic life (gentle, amiable, virtue, old man, cottage, affection), which point to the Creature's observations of the De Lacey family, where he presents himself as an admirer of human kindness who is nonetheless excluded.
- Ambivalent vocabulary: this class is associated with existential and affective terms (misery, alone, behold, love, pity), as in 'misery made me a fiend'. Several of these terms also appear in Victim passages, such as 'from which I alone am irrevocably excluded' (Section 3.6), which helps

explain why the classifier could not separate Ambivalent from Victim passages (Section 3.8).

3.6 Textual Evidence Across Narrative Levels

Within the annotated sample (Table 2), Victor's passages are most often labeled Threat (55.0%), whereas the Creature's are most often labeled Victim (72.8%). These are descriptive findings about human-assigned labels on selectively extracted passages. They are distinct from the classification results in Sections 3.2–3.3, which measure how well the models reproduce those labels, and they are not independent confirmation that the models capture the novel's framing. Representative passages illustrate the contrast:

1. Victor Frankenstein's Threat Framing (Vol. I Chapters IV & VI): "A mummy again endued with animation could not be so hideous as that wretch. I had gazed on him while unfinished; he was ugly then; but when those muscles and joints were rendered capable of motion, it became a thing such as even Dante could not have conceived." [1]. "Nothing in human shape could have destroyed that fair child. He was the murderer! I could not doubt it." [1].
2. The Creature's Victim Framing (Vol. II Chapters II, III & VII): "I was a poor, helpless, miserable wretch; I knew, and could distinguish, nothing; but, feeling pain invade me on all sides, I sat down and wept." [1] "I am full of fears; for if I fail there, I am an outcast in the world for ever." [1] "Every where I see bliss, from which I alone am irrevocably excluded." [1]
3. Ambivalent Framing (Vol. II Chapter II): "I was benevolent and good; misery made me a fiend. Make me happy, and I shall again be virtuous." [1]

4. Cross-Framed Passages: Table 2 also contains counter-intuitive cases: 29 of Victor's passages are labeled Victim (26.6%) and 14 of the Creature's passages are labeled Threat (17.3%). Read through Genette's distinction between voice and focalization [33], Victor's Victim framing can be understood as moments in which the speaking voice remains Victor's while the perspective passes temporarily to the Creature's claims on him. Following the Creature's plea on the *Mer de Glace*, Victor reflects: "For the first time, also, I felt what the duties of a creator towards his creature were, and that I ought to render him happy before I complained of his wickedness" (Vol. II Chapter II) [1]. Brooks interprets this moment as the point at which the Creature's speech first casts Victor in a parental role [34, p. 202]. A similar tension recurs in Victor's reaction to the Creature's request for a female companion: "His words had a strange effect upon me. I compassionated him, and sometimes felt a wish to console him; but when I looked upon him, when I saw the filthy mass that moved and talked, my heart sickened, and my feelings were altered to those of horror and hatred" (Vol. II Chapter IX) [1]. Sympathy is produced by listening and is withdrawn by looking, which is the opposition of "the hideous body and the persuasive tongue" that Brooks identifies [34, p. 201].

By contrast, the Creature's own Threat framing emerges at the point where his narrative turns from rejection to revenge. After his expulsion from the De Lacey cottage, he describes himself through a Miltonic self-image: "I, like the arch fiend, bore a hell within me; and, finding myself unsympathized with, wished to tear up the trees, spread havoc and destruction around me" (Vol. II Chapter VIII) [1], and he

declares "everlasting war against the species" [1]. When William, seeing his form, calls him "monster! ugly wretch!", the Creature kills the boy and exclaims, "I, too, can create desolation" [1]. In these passages the Creature narrates himself as a threat after he has been seen and named as a monster, so that his self-narration absorbs the designations imposed by others' gaze. Brooks's observation that a monster is "that which calls into question all our cultural codes, including language itself" [34, p. 219] helps explain why the Creature's self-narration cannot remain consistently victim-framed: the same eloquence that pleads for sympathy also articulates vengeance. These cross-framed passages are also connected to the embedded-dialogue errors discussed in Section 3.8, where the framing of a passage depends on whose voice and whose perception dominate it.

To connect these close readings with the model, the out-of-fold predictions of Logistic Regression for the gold-standard passages analyzed above are as follows:

- Victor Threat passages: Victor's expressions of dread and enmity in PAS_0035 ('frightful fiend') and PAS_0039 ('my enemy had indeed fled') are correctly classified as Threat (gold Threat, predicted Threat). Conversely, PAS_0030 ('breathless horror and disgust'), while labeled Threat, is misclassified as Victim (gold Threat, predicted Victim) because its opening reflection contains affective terms ('infinite pains', 'beauty') that confuse bag-of-words weighting. (The illustrative quote "Nothing in human shape..." from Vol. I Chapter VI corresponds to an excluded candidate passage, PAS_0053).
- Creature Victim passages: The Creature's awakening confession in PAS_0092 ('poor, helpless, miserable wretch') is correctly classified as Victim (gold Victim, predicted Victim). In contrast, his plea to De Lacey in PAS_0186 ('unfortunate and deserted creature; I am full of fears') is misclassified as Threat (gold Victim,

predicted Threat), where words of social rupture dominate.

- Ambivalent passages: As reflected in the zero Ambivalent recall, the model collapses all ambivalent passages into one of the two polar classes: PAS_0084 (containing both 'misery made me a fiend' and 'Every where I see bliss...') is misclassified as Threat (gold Ambivalent, predicted Threat), whereas Victor's acknowledgment of creator duties in PAS_0089 and the Creature's Miltonic lament in PAS_0197 ('like the arch fiend, bore a hell within me') are both misclassified as Victim (gold Ambivalent, predicted Victim).
- Creature Threat passages: The Creature's violent confrontation with William in PAS_0224 ('monster! ugly wretch!'), his oath of vengeance in PAS_0227 ('sworn eternal revenge'), and his declaration of ruin in PAS_0229 ('I, too, can create desolation') are all correctly classified as Threat (gold Threat, predicted Threat). (In Victor's narrative, his reflection "I compassionated him..." represents a brief 86-word contextual passage not included in the 195 gold corpus). Correctly classified passages demonstrate where salient lexical cues align directly with close reading, whereas misclassified passages highlight where narrative framing depends on voice, focalization, or inter-clausal causal relations that a bag-of-words representation cannot capture (Section 3.8).

3.7 Discussion and Theoretical Implications

Within the limits set by a single novel and a small dataset, the results have the following methodological and theoretical implications for computational literary studies and applied NLP.

First, all supervised models exceeded the majority baseline, and Logistic Regression obtained the highest mean Macro-F1 (47.75%). This suggests that Threat and Victim framing are

partly recoverable from lexical features [26], [29]; it does not show that the three-class task is solved, given the failure on Ambivalent passages. VADER's accuracy (54.36%) was clearly lower than that of the supervised models, but its Macro-F1 (44.85%) was within the variability of the best model. The result is consistent with the argument that sentiment polarity and narrative framing are not equivalent [10], [28], since victimization in the novel is expressed through intensely negative vocabulary. It does not establish that all lexicon-based or sentiment-based approaches are unsuitable for literary analysis, nor that the present models outperform framing-specific approaches such as [13].

Second, the small change under masking indicates partial robustness to the specific masking procedure used, not that sublinear n-gram TF-IDF captures higher-level discourse structure (Section 3.4).

Third, linear models expose their feature weights directly (Table 5), so that the lexical associations behind each decision can be inspected without post hoc explanation methods of the kind needed for black-box models [27]. Interpretability in this sense means transparency of the fitted model; it does not guarantee that the features are semantically valid indicators of framing [12], [14]. Annotation-based work in literary computing likewise suggests that disagreements and model errors can be informative about how a literary concept has been operationalized [5], [31].

Fourth, the narratological analysis in Section 3.6 suggests that framing changes with voice, focalization, and narrative level [33], and that the Creature's persuasive language and monstrous body pull in opposite directions [34]. These are interpretive findings from close reading; the classifier does not yet capture this tension, as its failure on the Ambivalent class shows. The quantitative labels are therefore best

read alongside close reading rather than as a substitute for it, and the literary and computational findings should be treated as distinct forms of evidence.

3.8 Error Analysis

Logistic Regression misclassified 65 of the 195 passages (Figure 2). As shown in the pooled confusion matrix (Figure 2), these errors consist quantitatively of: 29 Ambivalent passages misclassified (20 as Victim, 9 as Threat); 27 Threat passages misclassified as Victim; and 9 Victim passages misclassified (8 as Threat, 1 as Ambivalent). A qualitative linguistic review of these 65 error cases identifies four primary textual mechanisms:

1. Intra-Paragraph Emotional Shifts:

Passages that begin with empathetic reflection or vulnerability and conclude with violent oaths combine vocabulary from opposing poles. Because n-gram TF-IDF weights words independently of clause ordering, it cannot capture how an adversative connective (such as 'but', 'yet', or 'however') pivots the overall evaluative stance of the passage. For instance, in PAS_0030, Victor opens with aesthetic dedication and care ('infinite pains and care I had endeavoured to form') before pivoting sharply to revulsion ('breathless horror and disgust filled my heart'), causing a gold Threat passage to be mispredicted as Victim. Similarly, in PAS_0084, the Creature's Montanvert plea opens with vulnerability ('I am thy creature... I ought to be thy Adam') and conditional virtue ('I was benevolent and good; misery made me a fiend. Make me happy, and I shall again be virtuous'), causing a gold Ambivalent passage to be misclassified as Threat.

2. Embedded Dialogue (Voice vs. Focalization): In passages where Victor quotes the Creature's threats verbatim or the Creature recounts human hostility, the bag-of-words representation cannot separate the grammatical narrator (Genette's category of voice [33]) from the quoted character's evaluative perspective (focalization). For instance, in PAS_0240, Victor quotes the Creature pleading ('How is this? I thought I had moved your compassion...'), resulting in an Ambivalent passage being predicted as Threat due to hostile surrounding dialogue.

3. Ambivalent Boundary Collapse: The 29 Ambivalent passages account for 29 of the 65 errors (20 predicted as Victim, 9 as Threat). Ambivalent framing is fundamentally expressed through inter-clausal relations—such as the causal bond in 'misery made me a fiend' (PAS_0084), Victor's tension between duty and hatred in PAS_0089, or the Creature's mixture of misery and destructive rage in PAS_0197. Lacking syntactical awareness, the model simply defaults to whichever lexical pole carries greater cumulative TF-IDF weight. Furthermore, inter-annotator agreement was lowest precisely on this boundary, as all 13 disagreements in the reliability sample involved Ambivalent (5 between Threat and Ambivalent, 8 between Victim and Ambivalent; none between Threat and Victim). Balanced class weighting did not overcome these factors, since the model predicted Ambivalent only once.

4. Threat Predicted as Victim: This is the largest single error type between opposing poles (27 passages), plausibly because Threat passages narrated by the Creature (14 in Table 2) share first-person, affect-laden vocabulary ('misery', 'tears', 'horror')

with his Victim passages (e.g., PAS_0030, where Victor's disgust at the Creature is accompanied by words of care and beauty, or PAS_0201, where the Creature's declaration of war follows an account of bodily fever and suffering). Representative misclassified passages for each category are documented with complete text, gold annotations, and model predictions in the supplementary evaluation dataset.

3.9 Limitations

The findings are subject to the following limitations, which determine how far they can be generalized.

- (1) Corpus and sample size: the dataset contains 195 passages from a single novel, which is adequate for an exploratory case study but not for claims about narrative framing in general or in nineteenth-century Gothic fiction, and no formal sample-size analysis was performed.
- (2) Narrator representation and selection: Walton contributes only five passages, and the extraction procedure differs between narrators (Section 2.3), so narrator-level proportions are conditional on that procedure.
- (3) Annotation: framing labels involve interpretive judgment; agreement was substantial but not high ($\kappa = 0.6926$), was measured on 68 passages only, and the narrator could often be inferred from the text despite blinding.
- (4) Validation: stratified 5-fold cross-validation with 5–6 Ambivalent passages per test fold yields unstable class-level estimates; passages from the same paragraph or episode were not grouped, so the estimates may be optimistic; and

neither repeated nor group-based cross-validation was performed.

- (5) Ambivalent framing: the model did not identify any Ambivalent passage, so the three-class task is only partially solved.
- (6) Masking and baselines: only one masking strategy was tested, and the baselines include neither a framing-specific classifier nor an LLM.

Despite these limitations, two conclusions remain credible: lexical features separate Threat from Victim framing far better than a majority baseline, and explicit epithets are not the only predictive cue. Claims about Walton's framing, about the distribution of framing across the whole novel, and about the model's capture of discourse structure should be regarded as tentative. Extending the findings to the wider body of nineteenth-century Gothic fiction will require validation on a multi-author corpus.

CONCLUSION

This study operationalized the critical debate over the representation of the Creature in Mary Shelley's *Frankenstein* (1818) as an interpretable three-class text classification task and evaluated classical classifiers on 195 annotated passages. Sublinear TF-IDF features with Logistic Regression and Linear SVM distinguished Threat and Victim framing considerably better than a majority baseline (pooled F1 of 69.0% for Threat and 74.3% for Victim with Logistic Regression), but they did not identify any Ambivalent passage, and their Macro-F1 scores (47.75% and 46.18%) were within fold-level variability of each other and of the VADER baseline. Performance changed little when explicit epithets were masked, which indicates that these epithets are not the only predictive cue, although it does not show that the models capture discourse-level structure.

Within the annotated sample, Victor's passages were most often labeled Threat (55.0%) and the Creature's most often Victim (72.8%), which is consistent with reading the monster-victim dichotomy as a discursive construction that varies across narrative levels rather than as a fixed trait of the Creature. Because passages were selectively extracted and Walton contributes only five passages, these proportions are conditional on the sampling procedure, and no conclusion is drawn about Walton's position. The comparison with VADER supports the narrower point that sentiment polarity is not equivalent to narrative framing. Overall, the results are a preliminary proof of concept based on a single novel, a small and imbalanced dataset, uneven narrator representation, and interpretive annotation, and they do not establish a general framework for digital narratology.

Future work should: (1) expand the corpus within *Frankenstein*, including a larger and uniformly extracted Walton sample, and extend it to other nineteenth-century Gothic and science fiction novels (e.g., by Bram Stoker and H. G. Wells); (2) strengthen validation through repeated and group-based cross-validation, keeping passages from the same paragraph or episode in the same fold, together with paired statistical comparison of models; (3) refine the annotation scheme, for example by adding annotators, reporting agreement per class, and testing a multi-label or graded alternative to the single Ambivalent class; (4) test whether framing cues can be separated from narrator-specific vocabulary, for instance by training and testing within a single narrator or by masking pronouns and narrator-specific terms; (5) compare the models with stronger baselines, including a zero- or few-shot LLM; and (6) use dependency parsing or quotation detection to separate embedded dialogue from narratorial report.

DECLARATIONS

Author Contributions: The first author conceived and designed the study, developed the annotation codebook, conducted the passage annotation and narratological analysis, interpreted the results, wrote the original draft, and coordinated the revision of the manuscript. The second author processed the data, including passage extraction, text preprocessing, TF-IDF feature extraction, and the implementation and evaluation of the machine learning models, and prepared the tables and figures. The third author contributed to the theoretical framing and the interpretation of the findings and reviewed and edited the manuscript. The fourth author served as the second independent annotator, validated the annotation scheme, and reviewed and edited the manuscript. The fifth author contributed to the literature review and the methodological design, verified the consistency of the reported results, and reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Funding: This research received no external funding. All costs associated with the study were borne by the authors.

Conflict of Interest: The authors declare no conflict of interest.

Generative Artificial Intelligence Use Statement: Generative artificial intelligence tools were used only in a limited, supporting capacity during the preparation of this manuscript, such as language refinement in parts of the text. They were not used to generate data, annotate passages, conduct the analyses, or interpret the results. The authors reviewed and verified all content and take full responsibility for the integrity and accuracy of this article.

Ethics Approval and Informed Consent: Not applicable. This study analyzed a public-domain literary text and did not involve human participants, personal data, or animals.

Data Availability: The source text of *Frankenstein* (1818) is in the public domain and

is freely available from Project Gutenberg (EBook #41445). The annotation codebook, passage identifiers, gold labels, lexicons, scripts, and model configurations supporting the findings of this study are publicly available at <https://github.com/rusdi313/frankenstein-framing-nlp> and are also available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENT

The authors would like to express gratitude to the Faculty of Adab and Humanities and the Faculty of Science and Technology, Universitas Islam Negeri Syarif Hidayatullah Jakarta, for supporting this interdisciplinary research collaboration between English Literature, Linguistics, and Informatics. The English literature lecturer conducted the passage annotation and narratological analysis, the linguistics lecturer validated the annotation as the second independent annotator, and the Informatics student developed the TF-IDF feature-extraction pipeline and the machine learning models.

REFERENCES

- [1] M. W. Shelley, *Frankenstein; or, The Modern Prometheus*. London, U.K.: Lackington, Hughes, Harding, Mavor, & Jones, 1818.
- [2] A. K. Mellor, *Mary Shelley: Her Life, Her Fiction, Her Monsters*. New York, NY, USA: Routledge, 1988.
- [3] A. Piper, R. J. So, and D. Bamman, "Narrative theory for computational narrative understanding," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic, 2021, pp. 298–311, doi: 10.18653/v1/2021.emnlp-main.26.
- [4] H. O. Hatzel, H. Stierner, C. Biemann, and E. Gius, "Machine learning in

- computational literary studies," *Information Technology*, vol. 65, no. 4–5, pp. 200–217, 2023.
- [5] A. M. Weimer, F. Barth, T. Dönicke, L. Gödeke, H. Varachkina, A. Holler, C. Sporleder, and B. Gittel, "The (in-)consistency of literary concepts. Operationalising, annotating and detecting literary comment," *Journal of Computational Literary Studies*, vol. 1, no. 1, 2022, doi: 10.48694/jcls.90.
- [6] F. Brahman, M. Huang, O. Tafford, C. Zhao, M. Sachan, and S. Chaturvedi, "'Let your characters tell their story': A dataset for character-centric narrative understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, 2021, pp. 1734–1752.
- [7] B. Yang, H. Xu, J. Du, Z. Li, Y. He, and C. Lin, "EvolvTrip: Enhancing literary character understanding with temporal theory-of-mind graphs," *arXiv preprint arXiv:2506.13641*, 2025.
- [8] A. Brei, K. Henry, A. Sharma, S. Srivastava, and S. Chaturvedi, "Classifying unreliable narrators with large language models," in *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria, 2025, pp. 20766–20791, doi: 10.18653/v1/2025.acl-long.1013.
- [9] P. Toro Isaza, G. Xu, T. Oloko, Y. Hou, N. Peng, and D. Wang, "Are fairy tales fair? Analyzing gender bias in temporal narrative event chains of children's fairy tales," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2023, pp. 6509–6531, doi: 10.18653/v1/2023.acl-long.359.
- [10] S. Rebora, "Sentiment analysis in literary studies. A critical survey," *Digital Humanities Quarterly*, vol. 17, no. 2, 2023.
- [11] E. Kim and R. Klinger, "A survey on sentiment and emotion analysis for computational literary studies," *Zeitschrift für digitale Geisteswissenschaften*, vol. 4, 2019, doi: 10.17175/2019_008_v2.
- [12] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.
- [13] L. Frermann, J. Li, S. Khanehzar, and G. Mikolajczak, "Conflicts, villains, resolutions: Towards models of narrative media framing," in *Proc. 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada, 2023, pp. 8712–8732.
- [14] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book/>
- [15] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008, doi: 10.1162/coli.07-034-R2.
- [16] F. W. S. Shakti, E. Purwanto, and A. I. Pradana, "Klasifikasi surat dinas berbasis text mining menggunakan TF-IDF dan algoritma Naive Bayes Classifier," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, pp. 1152–1158, 2023.
- [17] D. Y. Saraswati, M. R. Handayani, K. Umam, and M. I. Mustofa, "Sentiment classification of MyPertamina reviews using TF-IDF and support vector machine

- algorithm," Jurnal Teknik Informatika (JUTIF), vol. 5, no. 1, pp. 177–184, 2024.
- [18] R. B. Adinata, Supriyono, and D. L. Fithri, "Sentiment classification of MyTelkomsel reviews using SVM and logistic regression," Jurnal Infotel, vol. 15, no. 4, pp. 312–320, 2023.
- [19] A. S. Fitriani and N. Hidayat, "Klasifikasi teks opini publik terhadap kebijakan menggunakan TF-IDF dan multinomial Naive Bayes," Jurnal SISFO (Sistem Informasi), vol. 12, no. 2, pp. 85–94, 2023.
- [20] M. A. Fauzi and A. P. Kharisma, "Penerapan feature selection dan SVM pada analisis teks naratif Bahasa Indonesia," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK), vol. 10, no. 4, pp. 789–796, 2023.
- [21] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.
- [22] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977.
- [23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.
- [24] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. 8th International AAAI Conf. Weblogs and Social Media (ICWSM), Ann Arbor, MI, USA, 2014, pp. 216–225.
- [25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [26] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd ed. draft, 2024. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [27] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD International Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 1135–1144.
- [28] S. M. Mohammad, "Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text," in Emotion Measurement, 2nd ed., H. L. Meiselman, Ed. Elsevier, 2021, pp. 323–379, doi: 10.1016/B978-0-12-821124-3.00011-9.
- [29] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002, doi: 10.1145/505282.505283.
- [30] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
- [31] E. Gius and J. Jacke, "The hermeneutic profit of annotation: On preventing and fostering disagreement in literary analysis," International Journal of Humanities and Arts Computing, vol. 11, no. 2, pp. 233–254, 2017, doi: 10.3366/ijhac.2017.0194.
- [32] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of

Documentation, vol. 60, no. 5, pp. 493–506, 2004.

- [33] G. Genette, *Narrative Discourse: An Essay in Method*, J. E. Lewin, Trans. Ithaca, NY, USA: Cornell Univ. Press, 1980.
- [34] P. Brooks, "What is a monster? (According to Frankenstein)," in *Body Work: Objects of Desire in Modern Narrative*. Cambridge, MA, USA: Harvard Univ. Press, 1993, pp. 199–220.
- [35] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [36] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.
- [37] T. Underwood, *Distant Horizons: Digital Evidence and Literary Change*. Chicago, IL, USA: Univ. of Chicago Press, 2019.