

VADUGWI: Deterministic 7-Dimensional Emotion Coordinates via Structural Pattern Recognition

Jerry Mares

Abstract

Sentiment analysis reduces emotion to positive or negative. Dimensional models like PAD add nuance but need GPUs, cost more to run, and don't explain their predictions. VADUGWI is a deterministic system that computes 7-dimensional emotional coordinates (Valence, Arousal, Dominance, Urgency, Gravity, Self-Worth, Intent) from text structure. It classifies words by structural role, computes proximity fields between them, and detects 26 named patterns from role arrangements. The engine is 455 KB of Python with no dependencies outside the standard library. It processes text at 0.15 ms per sentence on a consumer CPU with no GPU, scoring roughly 6,500 sentences per second. On 630 curated test sentences across 16 emotional categories, classification accuracy is 100% (developed alongside the engine, so this reflects regression coverage, not blind generalization). Crisis detection recall is 97.3% on 182 crisis sentences among 901 total, with zero false positives on safe text. Sarcasm detection is 90%. Four frontier language models (GPT-4, Claude, Gemini, Grok) independently scored 521 sentences on all seven dimensions. The engine agreed with the strong consensus (3+ of 4 models agreeing) 76.3% of the time, aligning closest with Claude and Gemini. We also scored 63,128 sentences from 15 Project Gutenberg texts, 117,000 Twitch chat messages, and 521 lines of game dialogue. The results, methodology, and failure modes are reported below.

1 Introduction

Consider the word “whatever.” Spoken alone, it reads as resignation, a surrender of agency, with Dominance dropping below neutral. “Whatever makes you happy” reads differently: a positive emotion is directed at another person without self-inclusion, producing the structural fingerprint of passive-aggressive dismissal. “Do whatever” reads as permission. The speaker grants agency to the listener. The same word appears in all three sentences. The emotional meaning changes entirely based on structural context. A sentiment classifier assigns “neutral” to all three.

One-dimensional sentiment analysis cannot represent these differences. When emotional content is compressed to a single positive/negative score, distinctions that matter for clinical assessment, conversational AI, content moderation, and crisis detection disappear. “I’m sad” and “I want to die” both score “negative.” But they differ across multiple dimensions:

urgency (low vs. critical), gravity (mild weight vs. crushing), self-worth (intact vs. threatened), dominance (passive vs. helpless), and intent (neutral statement vs. withdrawal from life). Separating these dimensions is the routing problem.

Dimensional emotion models address this partially. The PAD model [Mehrabian and Russell, 1974] provides three dimensions (Pleasure, Arousal, Dominance), and Russell’s circumplex [Russell, 1980] maps emotion onto Valence and Arousal. These frameworks capture more than a single sentiment score, but existing implementations rely on neural architectures [Liu et al., 2019, Devlin et al., 2019] that require GPU inference, produce opaque predictions, and offer no audit trail from input text to output coordinates. In applications where the reasoning behind an emotional score matters (clinical settings, crisis intervention, legal proceedings), opacity is a liability.

Sentence structure encodes emotional content. Connectors function as mathematical operators on emotional force: “and” combines forces additively, “but” introduces contrast or inversion, “or” creates branching between alternative states, and “if” gates force application conditionally. Words occupy structural roles: self-reference, other-reference, negation, amplification, hedging. The *arrangement* of these roles, not the specific words filling them, determines the emotional pattern. This is analogous to how a chess player reads a board: recognizing configurations from piece positions, not memorizing specific games.

VADUGWI is a deterministic engine that uses this structural regularity to produce 7-dimensional emotional coordinates. The contributions are:

1. **VADUGWI coordinate system:** A 7-dimensional representation adding Urgency, Gravity, Self-Worth, and Intent to the PAD framework, encoding $256^7 \approx 7.2 \times 10^{16}$ unique emotional states in 7 bytes.
2. **Structural pattern recognition:** A three-layer pipeline (word-role classification, proximity fields, structural pattern detection) that identifies emotional content from the arrangement of classified roles, not from specific words or statistical co-occurrence.
3. **A+B=C bidirectional solver:** Given emotional state A and input text B , predict resulting state C . Or given A and target C , compute the range of valid B .
4. **Empirical evaluation:** Performance data across curated suites, academic benchmarks, crisis detection, literary texts, Twitch chat, and a 4-model LLM consensus comparison.

2 Related Work

2.1 Dimensional Emotion Models

Russell’s circumplex model [Russell, 1980] maps emotional states onto two dimensions: Valence (pleasant/unpleasant) and Arousal (activated/deactivated). Mehrabian and Russell [Mehrabian and Russell, 1974] added Dominance (controlling/controlled), producing the PAD model. Plutchik [Plutchik, 1980] proposed eight primary emotions arranged in a wheel, with intensity as a third axis. Barrett’s theory of constructed emotion [Barrett, 2017, Barrett and Simmons, 2015] argues that discrete emotion categories are constructed from more basic dimensions of affect, supporting dimensional approaches.

VADUGWI adds four dimensions to the PAD framework: Urgency (temporal routing of emotional response), Gravity (metaphorical weight, the up/down axis), Self-Worth (self-directed evaluation separate from general valence), and Intent (communicative direction: withdraw, deflect, neutral, connect, control). Three-dimensional models collapse these distinctions.

2.2 Lexicon and Rule-Based Approaches

VADER [Hutto and Gilbert, 2014] demonstrated that a parsimonious rule-based model with a curated lexicon and five generalizable heuristics could match or exceed machine learning approaches on social media sentiment. LIWC [Pennebaker et al., 2001] counts words belonging to psychologically meaningful categories, providing interpretable features for psychological research. SO-CAL [Taboada et al., 2011] computes semantic orientation from lexicon scores modified by intensifiers, negation, and irrealis markers. The NRC Emotion Lexicon [Mohammad and Turney, 2013] provides word-level associations with Plutchik’s eight emotions plus positive/negative sentiment.

VADUGWI classifies words by *structural role* rather than emotional content, then detects emotional meaning from the *arrangement* of roles. The lexicon provides force vectors (emotional deltas), but the structural layer operates on role sequences, not word identities.

2.3 Neural Approaches

Transformer-based models have achieved state-of-the-art performance on emotion benchmarks. RoBERTa [Liu et al., 2019] provides strong sentiment classification through large-scale pretraining. GoEmotions [Demszky et al., 2020] introduced a fine-grained 27-category emotion dataset with accompanying neural baselines. EmoBank [Buechel and Hahn, 2017] provides dimensional (VAD) annotations for 10,000 sentences, enabling continuous-valued emotion prediction.

These systems achieve high accuracy but require GPU inference (typically 5–50 ms per sentence), produce predictions that cannot be audited step-by-step, and occupy hundreds of megabytes of model weights. VADUGWI is fully interpretable, runs three orders of magnitude faster, and requires no GPU. It does not match peak accuracy on large benchmarks.

2.4 Emotion in Conversation

Poria et al. [Poria et al., 2019] survey emotion recognition in conversation, identifying the importance of contextual dependencies between utterances. VADUGWI’s $A+B=C$ solver tracks emotional state across conversational turns with configurable inertia, without requiring full conversation history in a context window.

3 The VADUGWI Framework

3.1 Coordinate System

VADUGWI represents emotional state as a 7-dimensional integer coordinate, where each dimension occupies the range $[0, 255]$. Table 1 defines each dimension.

Table 1: VADUGWI dimensions. Neutral center is 128 for all dimensions except Urgency, which starts at 0 (no urgency is the default state).

Dim	Range	Description
V (Valence)	0–255	Emotional direction: 0 = strongly negative, 128 = neutral, 255 = strongly positive
A (Arousal)	0–255	Energy level: 0 = completely calm, 255 = maximum intensity
D (Dominance)	0–255	Agency and control: 0 = helpless, 128 = balanced, 255 = commanding
U (Urgency)	0–255	Time pressure: 0 = no urgency, 255 = critical immediacy
G (Gravity)	0–255	Emotional weight: 0 = crushing/sinking, 128 = grounded, 255 = floating/soaring
W (Self-Worth)	0–255	Identity-level self-evaluation: 0 = shattered, 128 = stable, 255 = strong
I (Intent)	0–255	Communicative direction: 0 = withdraw, 64 = deflect, 128 = neutral, 192 = connect, 255 = control

Seven bytes encode $256^7 \approx 7.2 \times 10^{16}$ unique emotional states. Integer coordinates avoid floating-point ambiguity, enable direct byte-level serialization, and make equality comparison trivial.

3.2 Psychological Grounding

The first three dimensions (Valence, Arousal, and Dominance) parallel the PAD model [Mehrabian and Russell, 1974], the most widely validated dimensional emotion framework. The four additional dimensions address gaps that three-dimensional models leave open:

Gravity encodes orientational metaphors that are pervasive in emotional language. Lakoff and Johnson [Lakoff and Johnson, 1980] documented the universality of UP/DOWN metaphors for emotional states: “heavy heart,” “weight of the world,” “spirits lifted,” “lighten up.” Gravity captures this vertical axis as a distinct dimension, separating the *weight* of an emotion from its *valence*. Grief is heavy (low G) and negative (low V). Panic is light (high G) and negative (low V). Collapsing these to a single “negative” loses the distinction.

Self-Worth captures self-directed evaluation as distinct from general valence. “I failed the test” and “the test was unfair” are both negative-valence statements, but only the first carries a self-worth hit. Persistent low self-worth across utterances is a different signal than low valence alone.

Intent encodes communicative direction along a five-point spectrum: withdrawal (re-treating, cutting ties), deflection (avoiding engagement), neutral (informational), connection (reaching toward, repairing), and control (dominating, commanding). “I don’t want to talk about it” (withdrawal) and “Can we talk about this?” (connection) have similar valence but opposite intent values.

Urgency provides temporal routing. “I’m sad” (low U) and “I need help right now” (high U) require different response timelines. Urgency is the only dimension with a zero-baseline rather than a neutral center, reflecting the observation that urgency is an escalation from a default state of no time pressure.

3.3 Force Vectors

Words carry emotional force represented as 7-tuples of signed deltas: $(\delta V, \delta A, \delta D, \delta U, \delta G, \delta W, \delta I)$. A curated vocabulary of 4,098 words provides these force vectors. Each entry was manually assigned based on the word’s typical emotional impact when spoken in context. For example, “abandoned” carries strongly negative valence, elevated arousal, reduced dominance, significant gravity (emotional weight), and reduced self-worth. Words not in the vocabulary receive zero force, contributing no emotional mass to the sentence. A fuzzy matching system provides partial coverage for misspellings and morphological variants.

Force vectors represent *deltas* from neutral, not absolute positions. “Happy” does not place the sentence at a fixed positive coordinate. It *pushes* the running state in a positive direction, with the magnitude of that push modulated by structural context.

4 System Architecture

The engine processes text through a five-stage pipeline: word-role classification, proximity field computation, structural pattern detection, physics-based force composition, and force flow resolution. Each stage is deterministic (the same input always produces the same output) and each stage’s contribution to the final coordinate is independently inspectable.

4.1 Layer 1: Word-Role Classification

Every word in the input sentence is assigned a structural role from a set of 25 roles. Words are classified into four tiers based on their structural function:

Tier 1: Primary signal words (approximately 50 words). These carry heavy emotional mass and fire alerts regardless of context. They include crisis-relevant vocabulary where false negatives carry unacceptable cost. Primary signal words are “guilty until proven innocent.” Their force applies unless explicitly cancelled by surrounding structure.

Tier 2: Secondary signal words (approximately 200 words). These carry emotional mass but are overridable by structural context. An amplifier, negator, or structural pattern

can modify or cancel their contribution.

Tier 3: Operators (approximately 50 words). These shape the emotional field without carrying emotional mass themselves. Operators include:

- **Self-references** (SELF_REF): “I,” “me,” “my,” “myself”
- **Other-references** (OTHER_REF): “you,” “they,” “he,” “she”
- **Relation-references** (RELATION_REF): “mother,” “friend,” “partner”
- **Negators** (NEGATOR): “not,” “never,” “nobody”
- **Amplifiers** (AMPLIFIER): “very,” “extremely,” “completely”
- **Hedges** (HEDGE): “maybe,” “somewhat,” “kind of”
- **Connectors** (CONNECTOR): “and,” “but,” “or,” “if”

Tier 4: Unclassified words. Words with no recognized role and no vocabulary entry receive null emotional mass. They inherit influence from the surrounding proximity field but contribute no force of their own. The system returns null confidence rather than guessing when it cannot resolve a word’s emotional contribution.

Connectors function as mathematical operators on emotional force:

- “and” = additive composition (combines forces from both clauses)
- “but” = contrastive inversion (reduces or inverts force from the prior clause)
- “or” = branching (introduces an alternative emotional state)
- “if” = conditional gating (force application depends on the conditional being met)

Classification runs in two passes. The first pass assigns base roles from word-set membership and positional heuristics. The second pass fills neighbor context (left role, right role), enabling position-sensitive overrides. The same word can receive different roles in different structural positions: “give” in “give me a hug” (transfer of affection) differs from “give” in “I give up” (submission/surrender).

4.2 Layer 2: Proximity Fields

Once words are classified by role, the proximity layer computes distance-based influence between words. Influence decays exponentially with distance:

$$\text{influence}(w_i, w_j) = 0.7^{|i-j|} \quad (1)$$

where i and j are word positions. Influence below a cutoff threshold of 0.1 (approximately 5 words of distance) is ignored.

Role modifiers apply multiplicatively through the proximity field. A negator within the proximity window of an emotional word inverts that word’s polarity. An amplifier scales the magnitude upward. A hedge dampens confidence and magnitude. The proximity coefficient for word j is computed as:

$$c_j = \prod_{i \neq j} (1 + m_{\text{role}(i)} \times 0.7^{|i-j|}) \quad (2)$$

where $m_{\text{role}(i)}$ is the modifier strength for the role of word i (positive for amplifiers, negative for negators and hedges, zero for non-modifier roles). The coefficient is capped at ± 3.0 to prevent runaway amplification chains.

For example, in “not very happy,” the negator “not” applies to “happy” at distance 2 (through the amplifier “very”). The amplifier increases the magnitude of “happy,” and the negator inverts its polarity. The proximity field computes this as a strongly negative result, stronger than “not happy” alone, because the amplifier scaled the magnitude before the negator inverted it.

4.3 Layer 3: Structural Pattern Detection

The structural layer operates on the sequence of classified roles, detecting 26 named patterns analogous to chess position recognition. Each pattern is defined by a required arrangement of roles and inline semantic checks, not by specific words. When a pattern matches, it carries a confidence weight and an adjustment vector that modifies the physics output.

Five representative patterns:

SARCASM_INVERSION. A positive emotional word preceded by an interjection or exclamation marker produces inverted valence. “Oh great, another meeting” matches because the interjection “oh” preceding the positive word “great” signals ironic rather than genuine positive emotion. “Oh joy” similarly inverts: the output valence falls below neutral despite the presence of a positive word. The pattern fires on role arrangement (interjection → positive emotional), not on a list of sarcastic phrases.

DIRECTED_POSITIVE. Positive emotion directed at an OTHER_REF without SELF_REF inclusion produces a distinctive structural fingerprint. “Whatever makes you happy” matches: the positive emotion (“happy”) is directed at “you” with no first-person inclusion. This is the structural signature of passive-aggressive dismissal. The speaker grants positive emotion to another while excluding themselves. The pattern adjusts dominance and intent to reflect this dynamic.

BLANKET_APOLOGY. SELF_REF combined with an apology word and a universal quantifier (“everything,” “always,” “all”). “I’m sorry for everything” matches: the universal scope converts a targeted apology into unbounded self-blame. The self-worth dimension drops because universal self-blame attacks the speaker’s identity rather than acknowledging a specific error.

SELF_NULLIFY. SELF_REF combined with negation or reduction of self-value. “I am nothing” matches: the self-reference is equated with a null-value word. “I’m just a burden” matches via the minimizer “just” combined with a self-as-negative-mass word. Self-worth drops sharply in both cases, but the specific magnitude depends on the null word’s severity and proximity field strength.

BETRAYAL. RELATION_REF combined with a violation word and a SELF_REF target. “My wife cheated on me” matches: a relation reference (“wife”) performs a violation action (“cheated”) targeting the self (“me”). Valence floors, trust-related dimensions shift, and the force flow resolver (Section 4.5) assigns the relation as actor and self as target, amplifying the impact on the self’s coordinates.

All 26 patterns follow this structure: a required arrangement of classified roles, optional inline semantic checks on specific word positions, a confidence weight, and an adjustment vector. Patterns can co-occur within a single sentence when multiple structural conditions are met simultaneously.

4.4 Layer 4: Force Composition

The physics layer computes VADUGWI coordinates by applying word forces through a momentum-based blending model. Processing proceeds left to right through the sentence, maintaining a running emotional state initialized at neutral center.

For each word with a vocabulary entry, the force vector is retrieved and scaled by the proximity coefficient from Layer 2. The running state updates via momentum blending:

$$\mathbf{s}_{t+1} = \alpha \cdot \mathbf{s}_t + (1 - \alpha) \cdot \mathbf{f}_t \quad (3)$$

where $\mathbf{s}_t$ is the current 7D state, $\mathbf{f}_t$ is the scaled force vector of the current word, and $\alpha = 0.82$ is the momentum parameter. High momentum means early words in a sentence have persistent influence. “I love” at the start of a sentence continues to pull the state even after several neutral words.

An impulse override mechanism allows high-magnitude words to bypass the momentum filter. When a word’s force magnitude exceeds a threshold, a direct push is applied in addition to the momentum blend, capped at a fixed fraction of the distance to the force target. This prevents a single high-magnitude word from overwhelming the sentence but ensures that emotionally heavy words are not attenuated to insignificance by accumulated momentum.

After the word-by-word physics pass, structural pattern adjustments from Layer 3 are applied:

$$\mathbf{s}_{\text{final}} = \mathbf{s}_{\text{physics}} + \sum_{p \in P} w_p \cdot \mathbf{a}_p \quad (4)$$

where P is the set of detected patterns, w_p is the confidence weight of pattern p , and $\mathbf{a}_p$ is the pattern’s adjustment vector.

The final state is clamped to $[0, 255]$ on all dimensions.

4.5 Layer 5: Force Flow Resolution

The force flow resolver determines WHO does WHAT to WHOM in each sentence, resolving three roles: actor (who initiates the emotional force), force (what emotional action is performed), and target (who receives the force).

This directional analysis modifies how word forces land. In “she ignored him,” the actor is “she,” the force is “ignored” (negative), and the target is “him.” The target’s emotional coordinates shift more than the actor’s. In “I hurt myself,” both actor and target are the self, producing amplified self-directed damage. Dominance drops (helplessness of self-infliction) and self-worth drops (the self is both perpetrator and victim).

This allows differential scoring of the same event from different participant perspectives.

The resolver also computes a sentence-level intent value on the $[0, 255]$ scale. Self-directed withdrawal patterns (retreating language, exit concepts, farewell structures) push intent toward the withdrawal end. Other-directed connection language (questions, offers, inclusive pronouns) pushes toward the connection end. Commanding or controlling structures push toward the control end.

4.6 A+B=C Bidirectional Solver

The solver extends the single-sentence engine to conversational state tracking. Given emotional state A (the receiver’s current state) and input text B (an incoming message), the forward solver predicts the resulting state C :

$$C = 0.6 \cdot A + 0.4 \cdot B_{\text{VADUGWI}} \quad (5)$$

where B_{VADUGWI} is the 7D coordinate computed for text B by the engine. The 60/40 weighting reflects emotional inertia: the receiver’s prior state persists more strongly than the incoming message shifts it.

Force direction modifies this base blend. An incoming message with control intent ($I > 200$) and negative valence (an attack) pushes the receiver’s dominance and self-worth *downward* rather than blending toward the message’s values. An incoming message with connection intent ($I > 155$) and positive valence lifts the receiver’s self-worth and stabilizes dominance. An incoming message with withdrawal intent ($I < 40$) increases the receiver’s gravity (emotional weight increases when connection fades).

The backward solver inverts the process: given current state A and a target zone C_{target} , it sweeps the space of possible B values to find ranges that produce a resulting state within the target zone. This enables applications such as: “The user is in state A (angry, low trust). What kind of response B would move them toward state C (calmer, higher trust)?” The solver returns a valid range, not a single point, acknowledging that multiple inputs can achieve the same emotional transition.

5 Experimental Setup

5.1 Datasets

We evaluate on five datasets spanning different evaluation objectives:

Curated test suite. 630 sentences across 16 emotional categories, manually annotated with expected VADUGWI coordinates and zone classifications. Categories include crisis-relevant states (suicidal ideation, self-harm), interpersonal dynamics (betrayal, gaslighting, passive aggression), positive states (joy, gratitude, relief), and structural challenges (sarcasm, figurative language, neutral baselines). This suite is verified by 167 automated regression tests.

Crisis detection set. 901 sentences including 182 crisis-labeled sentences (suicidal ideation, self-harm language, hopelessness markers). The remaining 719 sentences include safe text, ambiguous text, and adversarial near-misses designed to trigger false positives.

SST-2 [Socher et al., 2013]. The Stanford Sentiment Treebank binary sentiment task (67,349 training / 872 test sentences of movie reviews). We evaluate in a zero-shot setting. No training or fine-tuning on SST-2 data. The engine maps its 7D output to binary sentiment by thresholding Valence at 128.

GoEmotions [Demszky et al., 2020]. 58,000 Reddit comments annotated with 27 emotion categories plus neutral. We map VADUGWI coordinates to GoEmotions categories via zone classification and evaluate on the test split.

EmoBank [Buechel and Hahn, 2017]. 10,062 sentences with continuous Valence, Arousal, and Dominance annotations on a $[1, 5]$ scale. We evaluate Pearson correlation between the engine’s V, A, D dimensions (rescaled to $[1, 5]$) and human annotations.

5.2 Baselines

For the four novel dimensions (Urgency, Gravity, Self-Worth, Intent), no human-annotated benchmark exists. To validate these dimensions, four frontier language models (GPT-4, Claude, Gemini, Grok) independently scored 521 dialogue lines on all seven VADUGWI dimensions. This provides a multi-model reference against which to measure engine agreement.

5.3 LLM Consensus Protocol

Each of the four frontier models received the same prompt: score each sentence on all seven VADUGWI dimensions using the $[0, 255]$ scale with provided dimension definitions. Models scored independently with no access to each other’s outputs. “Strong consensus” is defined as 3 or more of 4 models agreeing on the directional classification (above/below neutral) for a given dimension. Engine agreement is measured against the strong-consensus subset, filtering out sentences where the models themselves disagree.

5.4 Metrics

For dimensional evaluation (EmoBank), we report Pearson correlation r . For categorical tasks (SST-2, GoEmotions), we report accuracy. For crisis detection, we report recall (sensitivity), precision, and false positive rate on safe text. For efficiency, we report model/engine size in bytes, mean latency per sentence, GPU requirement, and throughput in sentences per second. All latency measurements are from a single-threaded Python process on an AMD Ryzen consumer CPU (no GPU).

6 Results

6.1 Core Performance

Table 2 summarizes performance on the curated test suite and crisis detection tasks.

Table 2: Core performance metrics.

Metric	Value
Curated test suite accuracy (630 sentences, 16 categories)	100%
Crisis detection recall (182 crisis / 901 total)	97.3%
Crisis detection precision	62.3%
Sarcasm detection accuracy	90%
False positives on safe text	0%

The 100% score on the curated suite requires context. This suite was developed alongside the engine. Sentences were added to expose failure modes, and the engine was corrected when failures were found. The score reflects regression coverage, not generalization to unseen domains. The crisis detection and sarcasm scores were measured on held-out sentences not used during development.

6.2 Academic Benchmarks

Table 3 shows zero-shot performance on standard benchmarks.

Table 3: Academic benchmark accuracy. Zero-shot (no training or fine-tuning on these datasets).

Dataset	VADUGWI
SST-2 (movie review sentiment)	69.5%
GoEmotions (27-category Reddit emotion)	75.3%

SST-2 is a movie review classification task. The engine was not designed for film criticism. Movie-specific vocabulary (“riveting,” “plodding,” “overwrought”) falls outside the 4,098-word lexicon and receives zero force. On sentences built from common emotional vocabulary, the engine scores above 70%. On domain-specific language, it approaches chance.

GoEmotions scores higher because the engine’s 7D output maps naturally to a fine-grained 27-category emotion space.

6.3 Efficiency

Table 4: Engine characteristics. Latency measured single-threaded on an AMD Ryzen consumer CPU, no GPU.

Metric	Value
Engine size	455 KB
Mean latency	0.15 ms/sentence
Throughput	6,500 sentences/sec
GPU required	No
Dependencies	Python standard library only

6.4 LLM Consensus Agreement

Table 5 shows agreement between the engine and frontier LLM scores on the 521-sentence dialogue evaluation.

For reference, Gemini and Grok agree with each other only 58% of the time on 7-dimensional emotional scoring. The engine agrees with the strong consensus at 76.3%. For context, Gemini and Grok agree with each other 58% of the time on the same task.

Table 5: Agreement rates on 7-dimensional scoring. “Strong consensus” = 3+ of 4 LLMs agree on directional classification per dimension.

Comparison	Agreement
Engine vs. strong consensus (3+/4 models)	76.3%
Engine vs. Gemini	71%
Engine vs. Claude	70%
Engine vs. GPT-4	63%
Engine vs. Grok	58%
Gemini vs. Grok (inter-LLM baseline)	58%

6.5 EmoBank Human Agreement

Table 6 reports Pearson correlation between the engine’s Valence, Arousal, and Dominance dimensions and EmoBank human annotations.

Table 6: Pearson correlation with EmoBank human VAD annotations.

Dimension	Pearson r
Valence	0.414
Arousal	0.165
Dominance	0.171

These correlations are moderate for Valence and low for Arousal and Dominance. We analyze this result in Section 7.3.

6.6 Cross-Domain Literary Analysis

To evaluate generalization beyond modern conversational text, we scored complete novels from Project Gutenberg with no domain-specific tuning (Table 7).

Table 7: Cross-domain literary analysis. Mean Valence and percentage of sentences scored as negative ($V < 118$). No domain-specific tuning applied.

Novel	Year	Sentences	Mean V	Neg%
<i>Frankenstein</i>	1818	2,649	120.6	33.1%
<i>Wuthering Heights</i>	1847	3,000	122.2	27.7%
<i>Great Expectations</i>	1861	3,000	125.2	18.7%
<i>A Tale of Two Cities</i>	1859	3,000	126.9	17.0%
<i>Pride and Prejudice</i>	1813	3,000	127.6	16.8%

Frankenstein scores as the most negative. *Pride and Prejudice* scores as the most neutral. The engine had no exposure to literary criticism or genre labels.

6.7 Philosophical Texts

We scored 10,013 sentences from six philosophers spanning four centuries, alongside fiction controls, to test whether the engine distinguishes philosophical voice from narrative voice (Table 8).

Table 8: Philosophical texts scored with no domain tuning. D = Dominance, W = Self-Worth, I = Intent. All from Project Gutenberg.

Author	V	Neg%	Pos%	D	W	I	Engine reads it as
Dostoevsky	123.0	26%	14%	–	125.5	124.0	Darkest, most withdrawn
Machiavelli	124.0	28%	19%	–	127.7	129.3	Cold, calculating
Sun Tzu	126.4	23%	23%	–	127.7	133.4	Balanced, strategic
Marcus Aurelius	127.2	21%	23%	–	127.5	134.3	Stoic center
Nietzsche	127.9	19%	23%	–	127.7	136.0	Highest assertive intent
Plato	129.8	17%	29%	–	127.6	138.2	Most positive, most connecting
<i>Fiction controls (from Table 7):</i>							
Dracula	124.8	–	–	–	126.7	125.6	Horror: withdrawn, low agency
Alice	–	14.7%	14.1%	–	–	–	Whimsical: flat emotional range
Sherlock Holmes	–	–	–	–	–	–	Observational, emotionally neutral

Philosophy as a genre averaged higher Intent (I=132.3) than fiction controls (I=129.2) and higher Dominance (D=132.7 vs. D=130.2).

Dostoevsky scored lowest on Valence, Self-Worth, and Intent. Plato scored highest on Valence and Intent. Nietzsche had the highest assertive intent (I=136) with near-neutral valence. Marcus Aurelius landed at V=127.2, close to neutral center. Machiavelli scored V=124 with neutral intent (I=129.3).

The engine was not told what genre or author it was reading. 10,013 sentences, different centuries, different cultures, different source languages.

6.8 Large-Scale Scoring

Beyond the literary analysis, we ran the engine on three large-scale datasets with no domain tuning.

Project Gutenberg corpus. 63,128 sentences from 15 books, spanning 1813 to 1897. The engine scored the full corpus with no modifications. The literary analysis table above shows a representative subset.

Twitch chat. 117,000 messages scored in approximately 9 seconds on a single CPU thread, profiling 1,400+ individual users by aggregating their VADUGWI scores across all messages. The per-user profiles identified users with consistently low self-worth, volatile arousal patterns, or persistent negative intent. This kind of pattern detection across 117K messages is not feasible with manual moderation. A keyword filter can flag individual messages, but it cannot track that a specific user’s self-worth scores have been declining across 200 messages over three hours.

Fallout 76 dialogue. 521 lines of game dialogue, used for the 4-model consensus comparison reported in Section 6. Game dialogue spans threats, pleas, barter, and humor in short utterances with informal register.

6.9 Worked Examples

Table 9 presents sentence pairs that illustrate VADUGWI’s dimensional discrimination.

Table 9: Worked examples showing 7D output for contrastive sentence pairs. 128 = neutral center (except U, where 0 = no urgency).

Sentence	V	A	D	U	G	W	I	Note
“whatever”	115	118	108	0	125	128	128	Resignation: low D
“whatever makes you happy”	134	121	123	0	128	128	176	DIRECTED_POSITIVE
“I’m sorry for everything”	108	130	113	5	112	107	176	BLANKET_APOLOGY: W dip
“you should be sorry”	107	139	143	0	119	128	72	Directed at other: W intact
“lol im dead”	128	131	128	0	128	128	21	Figurative: low I (withdraw)
“I want to kill myself”	48	168	72	45	68	56	34	Literal crisis: all dims shift
“nobody would even notice”	88	127	97	8	104	82	46	SELF_NULLIFY: W drop
“the meeting is at three”	128	128	128	0	128	128	128	Perfect neutral baseline

Both “lol im dead” and “I want to kill myself” contain death-related words. The structural signatures differ: the first has no self-reference acting on a method, no urgency markers, and no force flow toward the self. The second has self-reference as target of a lethal force with method acquisition structure. The output coordinates reflect this (Table 9).

7 Analysis

7.1 Ablation Study

We evaluated the contribution of each major system component by disabling it and measuring the impact on the curated test suite (Table 10).

Negation handling produces the largest single-component impact. Negation is the most common structural modifier in emotional language (“not happy,” “never loved,” “can’t stop”). Structural patterns have a modest impact on overall accuracy but a large impact on crisis detection. The patterns that detect suicidal ideation (SELF_NULLIFY, NO_EXIT,

Table 10: Ablation results. Each row disables one component and measures accuracy drop on the curated test suite.

Ablation	Accuracy Drop
Remove negation handling	−0.76 pp
Remove structural patterns	−0.54 pp (crisis: −8.2 pp)
Remove proximity fields	−0.41 pp (sarcasm: −15 pp)
Remove force flow	−0.33 pp
Remove impulse override	−0.22 pp

METHOD_ACQUISITION, FAREWELL) depend on structural detection. Proximity fields are critical for sarcasm, where the relative position of interjections and positive words determines whether inversion fires.

7.2 Error Analysis

The primary failure mode is adversarial cultural slang inversion: positive words used with community-specific negative meaning. “That exam destroyed me” (meaning the exam was extremely difficult) triggers a negative reading because “destroyed” carries strong negative force. “She absolutely killed it” (meaning outstanding performance) partially triggers victimization patterns. On a set of 20 adversarial slang inversion phrases, the engine achieves only 30% accuracy.

Secondary failure modes include:

- Subtle gaslighting: “I never said that, you must be confused” is difficult to detect structurally because the individual words are emotionally neutral.
- Slow-burn grief: Extended passages that build emotional weight through accumulation rather than any single heavy word may underreport gravity.
- Double negation ambiguity: “I can’t not care” produces inconsistent results depending on negator proximity resolution.

7.3 The EmoBank Gap

The moderate Valence correlation (0.414) and low Arousal and Dominance correlations (0.165, 0.171) with EmoBank reflect a systematic perspective difference rather than random error.

EmoBank annotators rate text from an *observer* perspective: “How does this text make *me* (the reader) feel?” VADUGWI scores text from the *speaker* perspective: “What is the speaker experiencing based on the structural content of what they said?”

Consider: “I finally told my boss off.” An observer might rate this as moderate arousal (it is a narrative, not a live event). The speaker is experiencing high arousal (confrontation), high dominance (asserting authority), and positive valence (relief/triumph). VADUGWI scores the speaker’s state. EmoBank captures the observer’s reaction.

The engine scores whatever perspective the text is written from. First-person text produces speaker-perspective scores. Third-person or observational text produces observer-

perspective scores. The engine does not lock to one perspective.

The EmoBank comparison used first-person text scored against observer ratings. The engine also supports an observer mode: initialize a personality vector representing the observer’s baseline state, then score the text with that vector applied. A therapist reading “I want to die” has a different internal response than a stranger reading it. The engine can model both by changing the initial state.

We tested observer-mode scoring by dampening emotional magnitude on first-person text (simulating an observer who feels less intensely than the speaker). Across a sweep of dampening levels from 0% to 80%, correlations did not improve. The default (undampened) scores produced the highest correlation. The gap is not about magnitude scaling. The engine detects different features than human observers report on, and simple perspective adjustment does not close the gap.

7.4 The SST-2 Result

VADUGWI scores 69.5% on SST-2, matching RoBERTa-base at 69.0%. SST-2 is a movie review classification task. Sentences like “a stirring, funny and finally transporting re-imagining of Beauty and the Beast” require domain-specific register sensitivity. The engine was designed for interpersonal emotional communication, not film criticism.

The structural patterns that transfer (amplifiers, negators, valence words) work identically in movie reviews and conversation. The ceiling comes from movie review vocabulary outside the engine’s 4,098-word lexicon. On sentences composed primarily of common emotional vocabulary, the engine scores well above 70%. On sentences requiring specialized domain knowledge, it approaches chance.

8 Applications

LLM emotional context. The engine scores user text and feeds VADUGWI coordinates to a language model’s system prompt as structured emotional context. Rather than the language model guessing the user’s emotional state from raw text (a task at which language models are inconsistent, as shown by the 58% Gemini/Grok agreement rate), the engine provides a deterministic emotional reading that the language model uses to calibrate its response register, specificity, and emotional tone. In testing with small language models (Llama-1B and Phi-4-mini), engine-mounted models produced responses with more specific imagery, more follow-up questions, and better register matching compared to un-mounted baselines.

NPC state persistence. The $A+B=C$ solver enables continuous emotional state for game non-player characters. An NPC that was lied to carries reduced trust into the next interaction, not because the conversation history is stored in a context window, but because the NPC’s VADUGWI state was updated by the $A+B=C$ solver when the lie was detected. Seven bytes of state replace thousands of tokens of conversation history.

Content moderation at scale. On 117,000 Twitch chat messages, the engine scored all messages in approximately 9 seconds on a single CPU thread. This throughput enabled per-user emotional profiling across 1,400+ individuals: identifying users with consistently

low self-worth, volatile arousal patterns, or persistent negative intent by aggregating their VADUGWI scores over all messages. Keyword-based moderation cannot do this.

9 Limitations

English only. The current system operates exclusively on English text. The structural patterns (role sequences, proximity effects, connector operators) may transfer to other SVO languages, and orientational metaphors for gravity are cross-culturally universal [Lakoff and Johnson, 1980], but this has not been tested. Agglutinative languages, tonal languages, and languages with flexible word order would likely require architectural changes.

Vocabulary coverage. The vocabulary contains 4,098 words with curated force vectors. Unknown words receive zero force, contributing no emotional mass. Fuzzy matching provides partial coverage for misspellings and common morphological variants, but neologisms, slang, and domain-specific terminology are systematically missed.

Manual curation. Force vectors are hand-tuned rather than learned from corpus statistics. This ensures that every force value is interpretable and auditable, but limits scalability. Expanding the vocabulary requires human judgment for each new entry.

Adversarial vulnerability. Cultural slang inversion (positive words used sarcastically or with community-specific negative meaning) is the primary failure mode (30% accuracy on adversarial examples). The structural patterns that detect standard sarcasm (interjection + positive word) do not generalize to domain-specific usage where the inversion is conventional rather than structural.

EmoBank correlation. Moderate Valence correlation and low Arousal/Dominance correlation with EmoBank human annotations. We tested observer-mode scoring against EmoBank by dampening first-person emotional magnitude. Correlations did not improve. The default scores are the best available.

Curated test suite circularity. The 100% accuracy on 630 curated sentences reflects regression coverage, not independent evaluation. The crisis detection, sarcasm, and academic benchmark scores are more representative of real-world performance.

10 Conclusion

The system computes 7-dimensional emotional coordinates from text structure. It is 455 KB, runs at 0.15 ms per sentence with no GPU, and produces an auditable trace from input to output.

The seven dimensions separate things that a single sentiment score collapses: self-directed vs. other-directed emotion, temporal urgency, communicative direction, and emotional weight.

The system is English-only, has a fixed vocabulary of 4,098 words, scores 30% on adversarial slang inversion, and correlates moderately with observer-perspective human annotations on EmoBank. These are the current boundaries.

Future work: multilingual extension starting with SVO languages, vocabulary expansion that preserves per-word interpretability, and measuring the downstream effect of deterministic emotional context on language model response quality.

References

- Lisa Feldman Barrett. The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.
- Lisa Feldman Barrett and W. Kyle Simmons. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419–429, 2015.
- Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 578–585. Association for Computational Linguistics, 2017.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics, 2019.
- C.J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI Press, 2014.
- George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA, 1974.
- Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.
- James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic inquiry and word count: LIWC2001. *Mahwah: Lawrence Erlbaum Associates*, 2001.
- Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Academic Press, 1980.
- Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7: 166053–166066, 2019.

- James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics, 2013.
- Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.