Sentium: Sentiment Evaluation through Neurosymbolic Taxonomy - an Interpretable and Understandable Model

Anonymous ACL submission

Abstract

Sentiment analysis has seen rapid progress driven by deep learning, but the opaque black-box nature of these models hinders trustworthy deployment in high-stakes domains where interpretability is crucial. We propose **Sentium** (**S**entiment **E**valuation through **N**eurosymbolic **T**axonomy, an **I**nterpretable and **U**nderstandable **M**odel), a cognitively-inspired architecture that closely emulates human sentiment comprehension processes. Sentium takes a hybrid approach by combining structured sentiment knowledge with neural models, achieving state-of-the-art performance while maintaining transparency through explicit compositional reasoning over semantic propositions. Compared to state-of-the-art financial language models, Sentium showed substantially lower misclassification rates for predicting true negatives as positive (Sentium=1.97%; FLANG-BERT (Shah et al., 2022) =6.78%, FinBERT (Araci, 2019) =10.17%). The code are available at: https://github.com/anonymous-submission

1 Introduction

Sentiment analysis aims to bridge the gap between human and machine capabilities in analysing sentiment (Yusof et al., 2018). This objective can be interpreted through two complementary lenses following Gobet and Lane (2010): (i) An *engineering approach* that narrows the performance disparity, harnessing computer science techniques to create intelligent artifacts achieving human-level outcomes. (ii) A *cognitive modeling* approach that aligns the underlying processes, developing computational architectures that closely emulate human behavior for interpretable simulations.

To reach state-of-the-art performance, the field has extensively leveraged deep neural networks for natural language processing tasks (Chen et al., 2023). Indeed, sentiment analysis has transitioned

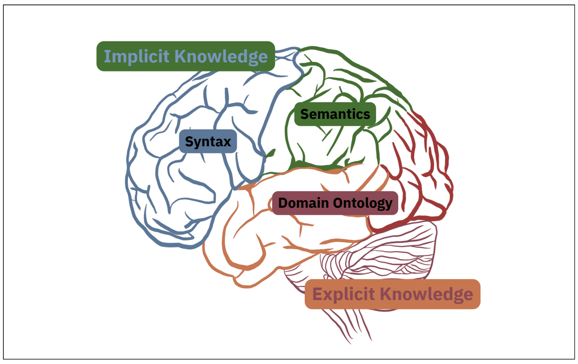

Figure 1: Cognitive Architecture of Sentium. This diagram illustrates Sentium's hybrid approach, unifying implicit knowledge (semantics/syntax captured by neural models) and explicit knowledge (encoded rule-based domain ontology).

from traditional rule-based and lexicon-based models (Stone et al., 1962; Bradley and Lang, 1999; Hu and Liu, 2004; Esuli and Sebastiani, 2006; Nielsen, 2011; Taboada et al., 2011; Hutto and Gilbert, 2014; Cambria et al., 2022) to transformer-based approaches like Small Language Models (SLMs) (Araci, 2019; Alaparthi and Mishra, 2021; Prottasha et al., 2022; Shah et al., 2022; Cho et al., 2023), and more recently, Large Language Models (LLMs) (Nadi et al., 2023; Kheiri and Karimi, 2023). This transition was inevitable, as lexicon-based methods remained below acceptable performance levels (Muhammad et al., 2016), typically achieving 55-85% accuracy compared to deep learning models' 70-95% range (Al-Qablan et al., 2023).

However, this pursuit of performance gains has given rise to profound challenges. While traditional deep learning drawbacks like substantial data, computational resource, and training time requirements (Muhammad et al., 2016; Schouten et al., 2017; Sarker, 2021) have been relatively mitigated through fine-tuning (Talaei Khoei et al., 2023; Wojciuk et al., 2024), a fundamental issue persists – the inherent lack of interpretability in these black-box neural architectures.

Despite extensive exploration of four common interpretation methods (Chen et al., 2023), the true model interpretability remains unresolved. Post-hoc techniques like LIME (Ribeiro et al., 2016) offer local approximations but fail to capture the global logic encoded within model parameters. Even for LLMs, methods like sparse autoencoders (Templeton et al., 2024) and chain-of-thought reasoning (Turpin et al., 2024) provide limited post-hoc justifications rather than intrinsic interpretability. After all, if these interpretations faithfully mirrored the original model, the explanation would equal the model itself, rendering the original redundant (Rudin, 2019).

This lack of transparency significantly hinders the trustworthy and responsible deployment of deep learning for sentiment analysis, especially in high-stakes domains where decision rationales profoundly impact businesses, investments, and lives (Rudin, 2019; Rudin et al., 2022; Oh, 2024). Opaque black-box predictions, while accurate, offer little insight into the reasoning behind sentiment derivations – an untenable predicament given the real-world consequences.

In contrast to opaque black-box models, we take a step forward towards interpretable and understandable sentiment analysis through cognitive modelling. By uniting structured domain knowledge with neural architectures in a cognitively-plausible manner, our approach achieves state-of-the-art performance while maintaining full interpretability. Predictions are firmly grounded in an intuitive sentiment ontology, enabling comprehensive rationale generation through explicit compositional reasoning over human-readable semantic propositions.

This human-inspired interpretability bridges a crucial gap in current black-box methods. Rather than inscrutable mappings from inputs to outputs, Sentium offers a transparent window into its inner workings, closely emulating the cognitive processes underlying human semantic comprehension. Stakeholders can intuitively audit and verify the evidence chain driving each sentiment prediction, fostering accountability and trust.

As the complexity of AI systems increases, embedding interpretability as a core architectural principle becomes vital. Sentium represents a tangible step in this critical direction, establishing human-centred transparency without compromising state-of-the-art performance.

The main contributions of this work are three-fold:

1. *Demonstrating that models need not be opaque end-to-end black boxes.* Our rule-based approach matches and even exceeds the performance of deep learning models, yet with the additional benefit of intuitive interpretability – a capability previously highlighted as advantageous by Hutto and Gilbert (2014).

2. *Proposing Ontological Sentiment Labelling Framework (OSLF)* – a machine-readable and human-interpretable knowledge base that captures the compositional semantics of how sentiment expressions interact with real-world concepts and aspects. OSLF enables more elaborate analysis of opinions on specific topics.

3. *Introducing a cognitively-inspired neural architecture that closely approximates human sentiment comprehension and reasoning processes* – an area receiving relatively less attention compared to the performance-driven engineering approaches in AI.

Through these contributions, Sentium paves the way towards developing trustworthy, accountable, and transparently-aligned systems that can be robustly deployed in high-stakes real-world domains. Rather than pursuing a broad cross-domain approach, we concentrate our efforts on showcasing Sentium's capabilities for the financial domain.

2 Sentium

Sentium is composed of three major modules inspired by theories of language comprehension from cognitive psychology (Kintsch and Van Dijk, 1978; Fodor, 1983; Anderson, 2000). These theories posit that comprehension involves several distinct yet interconnected stages. Fodor (1983) proposed a modular view, where a dedicated linguistic module first analyses the incoming language before passing its output to general cognition. Similarly, Kintsch and Van Dijk (1978) assumed an initial parsing stage that transforms the text into a set of propositions, which are then further processed.

Anderson (2000) outlined three key stages: 1) *Perceptual* encoding of the textual input, 2) *Parsing*, which involves syntactic and semantic analysis to derive a coherent mental representation of meaning, acting as an interface between low-level encoding and higher-level cognition, and 3) *Utilisation*, where this mental representation is used for tasks like reasoning and decision-making. This three-stage pipeline directly inspires the modular design of Sentium.

Sentium's modular architecture directly mirrors this systematic progression from perception to parsing to cognitive utilisation.

2.1 `tag.pos` replicates perceptual encoding by annotating the input text with low-level linguistic features like parts-of-speech, dependencies and lemmas.

2.2 `parse.aspect` models the parsing stage by extracting key semantic representations like entities and phrases, leveraging the annotated linguistic knowledge.

2.3 `evaluate.senti` captures utilisation by performing the target task – sentiment evaluation – grounded in the previous analyses and explicit domain knowledge.

A key contribution that advances the field of traditional rule-based sentiment analysis methods is how the `evaluate.senti` module incorporates explicit structured knowledge from the financial sentiment ontology, enabling interpretable reasoning. This maps to the distinction between implicit and explicit cognitive processes (Anderson, 2000). While `tag.pos` and `parse.aspect` rely on implicit learned representations, `evaluate.senti` combines these with explicit ontological knowledge to produce human-intelligible sentiment prediction rationales.

By systematic modelling of both implicit learned representations and explicit structured knowledge in a cognitively-plausible architecture, Sentium achieves a powerful synthesis: the predictive accuracy of neural models with the intuitive interpretability of human-like reasoning grounded in real-world finance knowledge. This synergy addresses key limitations of existing black-box sentiment analysis methods.

2.1 `tag.pos`

Humans possess an innate *linguistic competence* (Chomsky, 2014) - an implicit, abstract knowledge of language that allows intuitive judgments about syntactic structure, despite the infinite possible utterances (Anderson, 2000). We internalise thousands of subtle grammatical rules without being able to explicitly articulate them.

Sentium's `tag.pos` module aims to computationally capture this implicit low-level linguistic knowledge by leveraging neural models from `spaCy` (Honnibal et al., 2020). The input text is

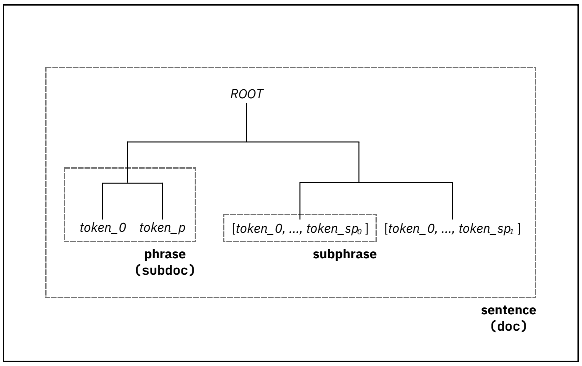

Figure 2: Sentence Subtree Representation. Subphrases are processed from phrases iff `len(phrase)>15`, segmented based on hierarchical subtree structure of such a phrase.

encoded with linguistic annotations like parts-of-speech tags, dependencies, and lemmas, producing sentence-level `doc` objects and phrase-level `subdoc` objects.

The hierarchical division of sentences into phrases is a core component of parsing and interpretation (Anderson, 2000) (Figure 2). As demonstrated by Graf and Torrey (1966), identifying constituent phrase structure is crucial for sentence comprehension. Sentium emulates this process by first segmenting sentences based on punctuation boundaries, following evidence that humans naturally pause at clause boundaries when reading (Aaronson and Scarborough, 1977). Coordinating conjunctions like "but" and subordinating conjunctions like "while", which link phrases and convey relationships (Gleitman, 1965), then guide further subdivision.

To handle long phrases that may require simplification, phrases exceeding 15 tokens are split into sub-phrases sharing a common parent node within the dependency parse subtree. This 15 token threshold aligns with typical readability guidelines and automatic simplification targets (DuBay, 2004).

Both doc and `subdoc` objects in Sentium encapsulate the encoded linguistic features, mirroring the perceptual process of syntactic analysis in human cognition (Anderson, 2000). While concatenating the `subdoc` (phrase) objects to construct `doc` (sentence) representations, or passing multiple `subdocs` to subsequent modules may seem cognitively plausible, Sentium deliberately avoids these approaches. A simplistic concatenation risks failing to accurately capture the syntactic structure and compositional semantics of sentences, as emphasised by compositional semantics theories (Partee, 2007). A sentence's meaning does not merely arise from combining its constituent phrases (Dankers and Lucas, 2023) – it emerges through nuanced composition rules governing how phrase meanings systematically interact[1].

Passing the complete, structured doc representation is not only more cognitively plausible by better approximating human-level composition abilities, but also computationally more efficient. By allowing subsequent modules to analyse a single doc object that encapsulates the full sentential context, rather than operating over multiple disconnected subdoc phrases, Sentium can construct more holistic and contextualised sentence interpretations.

2.2 `prase.aspect`

Building upon the syntactically-informed doc representations from `tag.pos`, the `parse.aspect` module aims to derive semantic interpretations more aligned with human language comprehension. This involves two core capabilities.

1. Extracting rich noun phrases by leveraging the encoded universal dependency parse structures (Manning, 2015; De Marneffe et al., 2021) within each doc object. While basic noun chunks provide a foundational starting point, `parse.aspect` goes further by capturing crucial prepositional modifier relationships. Prepositions like "in", "of", and "at" link nouns and noun phrases, expressing specific semantic relationships between the connected concepts. By modelling these dependency structures where one noun modifies another via a prepositional link, `parse.aspect` identifies semantically richer noun phrases than simple chunks alone.

2. In parallel, dedicated neural Named Entity Recognition (NER) models are employed to classify mentions of real-world entities like organisations and persons based on contextualised semantic representations. This separable semantics pathway accounts for how syntax alone cannot reliably disambiguate meanings – for instance, whether "Apple" refers to the fruit or technology company. Currently, apart from `spaCy`, three additional NER models pre-trained on diverse datasets like OntoNotes, Reuters corpus, WikiNews and Wikipedia by Aarsen are avaliable.

The architectural separation of these syntax-guided phrase extraction and neural entity recognition functionalities is grounded in cognitive psychology findings. Electrophysiological studies unveil distinct neural signatures for syntactic and semantic processing through dissociable N400 and P600 event-related potential components (Kutas and Hillyard, 1980a,b; Osterhout and Holcomb, 1993, cited in Anderson, 2000). This empirical evidence motivates modelling syntax and semantics as separable yet interacting mechanisms.

2.3 `evaluate.senti`

Cognitive systems gain the ability to predict – expectation about the concept – by categorising the concept, and because of this ability, categories give us great economy in representation and communication (Anderson, 2000, p.151). Traditional sentiment analysis methods have attempted to operationalise this by manually associating sentiment lexicons with conceptual representations. For instance, Henry (2008) examined the context of each lexicon's occurrence by calculating collocation percentages with desirable financial terms like "revenue" versus undesirable ones like "expenses" to categorise words as positive or negative.

While pioneering, such dictionary-based approaches have inherent limitations. Henry's (2008) lexicon achieved 80.12% accuracy in our **3 Experiment** – impressive yet insufficient for real-world robustness. Even "increased", positively used 66% of the time (Henry, 2008, p.33), carries a 34% chance of being neutral or negative. This variability arises from failing to account for the compositional effects of combining sentiment expressions with different semantic contexts.

To address these shortcomings, this paper proposes the *Ontological Sentiment Labelling Framework* (OSLF). The "ontology" refers to an expertly-curated, structured knowledge base defining relevant concepts and their interrelationships (Bandari and Bulusu, 2020; Kontopoulos et al., 2013). For example, in finance, representing "sales", "profit", and "loss" as distinct aspects, with modelled associations to sentiment-bearing expressions like "increase" or "decrease". Grounding sentiment analysis in such a rich, domain-adapted ontology enables interpretable rule-based propositional analysis over the input text. By mapping linguistic entities to ontological concepts, and expressions to sentiment variables, `evaluate.senti` later instantiates intuitive propositions capturing how each sentiment contributor interacts with the referenced aspect.

[1] It is important to note that the doc object primarily encodes syntactic representations, while largely abstracting away the compositional and non-compositional semantic nuances that contribute to a sentence's full meaning.

2.3.1 Ontological Sentiment Labelling Framework (OSLF)

The framework is centred around curating financial sentiment ontology, inspired by ontology construction methods (Kontopoulos et al., 2013; Schouten et al., 2017). The central idea was to systematically group key financial constructs and explicitly represent the relationships between them as an intuitive cross-table taxonomy (Table 1), later translated into a structured dictionary format below.

Ontological Sentiment Labelling Framework

---

domain_ontology ← {
  "*Synset*": {
    "*Target*": *Sentiment*
  }
}

finance_ontology ← {
  "increase": {
    "positive_financial_metrics": 1,
    "negative_financial_metrics": -1,
    "market_consensus": 1
  },
  "decrease": {
    "positive_financial_metrics": -1,
    "negative_financial_metrics": 1,
    "market_consensus": -1
  },
  "strength": {
    "positive_financial_metrics": 1,
    "strategic_partnerships": 1
  },
  "warn": {
    "positive_financial_metrics": -1,
    "negative_financial_metrics": -1,
    "performance_indicators": -1
  }
  ...
}

The ontology coherently models three core components essential for nuanced financial sentiment analysis:

1. *Targets* represent important aspects in the financial domain, acting as overarching concepts. Each *Target* encompasses a set of related aspects exhibiting superordinate-subordinate conceptual relationships. For example, "positive_financial_metrics" is a broad *Target* under which more specific metrics like "sales," "revenue," and "profit" are subsumed as subordinate terms. In total, 12 hierarchically-organised *Targets* were manually curated by experts.

2. *Synsets* represent sentiment-laden expressions prevalent in financial discourse, encapsulating collections of synonymous lemmas that convey analogous meanings. For instance, the "increase" synset comprises 23 lemmas, encompassing terms like "expand" and "rocket" that articulate a connotation of growth and positive trajectory. In total, 23 *Synset* categories were systematically adapted to the financial domain by leveraging semi-automatic methods (Hu and Liu, 2004; Strapparava and Valitutti, 2004; Esuli and Sebastiani, 2006). This involved iteratively expanding initial seed lists using WordNet (Miller, 1995; Fellbaum, 1998), followed by manual filtering to retain only expressions highly relevant to the financial news genre, accounting for genre variations noted by Pennebaker et al. (2015) and domain-specific language needs highlighted by Loughran and McDonald (2011).

3. *Sentiments* precisely capture the contextual sentiment polarity associated with each (*Target*, *Synset*) pair in the taxonomy. This models how the same sentiment expression can convey opposite polarities depending on the financial aspect referenced – a key challenge in this domain. For example, the "increase" *Synset* conveys positive sentiment for "positive_financial_metrics" *Target* like higher sales or revenue. However, it indicates negative sentiment when used with "negative_financial_metrics" *Target* such as rising costs or losses, capturing how the same expression can flip polarity across financial aspects.

The hallmark of this ontological framework is explicitly representing the nuanced, many-to-many relationships between *Targets* and *Synsets* in an intuitive yet comprehensive taxonomy manually curated by experts. This structured knowledge modelling enables highly precise, domain-specific sentiment analysis grounded in finance knowledge, while maintaining crucial transparency and audibility often lacking in opaque neural models.

2.3.2 Rule-based Propositional Analysis

A core capability of Sentium's `evaluate.senti` module is performing rule-based propositional analysis grounded in the financial sentiment ontology. This approach models the semantic composition of sentiments by systematically mapping linguistic inputs to propositions representing coherent units of sentiment-bearing knowledge.

		pfm	nfm	pi	ca	mc	div	sp	op	stf	tc	par	nar
Directional	increase	1	-1			1							
	decrease	-1	1			-1							
	higher	1	-1	1		1							
	lower	-1	1	-1		-1							
Performance	win				1								
	beat					1							
	reach	1											
	continue	1		1		1							
	strength	1						1					
Action	generate	1	-1										
	cause	1	-1										
	protect	1	-1	1									
	turn	1		1									
	propose						1						
	equip										1		
	improve	1		1									
	expect	1	-1										
	recommend											1	-1
Temporal	faster								1				
	slower								-1				
Negative	warn	-1	-1	-1									
	lose									-1			
	slip	-1	-1	-1									

Table 1: Cross-table taxonomy (OSLF). This cross-table taxonomy systematically organises key financial constructs (*Targets* and *Synsets*) and explicitly represents their relationships. The grouping is based on the semantic meanings and contexts in which these *Synsets* are typically used in the financial domain. For example, the Directional group captures *Synsets* that describe the upward or downward movement of financial metrics, while the Performance group encompasses *Synsets* that represent the results or achievements of financial entities or activities. Abbreviations: pfm (positive_financial_metrics), nfm (negative_financial_metrics), pi (performance_indicators), ca (contractual_agreements), mc (market_consensus), div (dividend), sp (strategic_partnership), op (operation_process), stf (staff), tc (technological_capabilities), par (positive_analyst_recommendation), nar (negative_analyst_recommendation)

The key idea, inspired by theories from Kintsch (1974), is to represent the smallest units of knowledge that can be evaluated as true or false sentiment propositions. Specifically, `evaluate.senti` identifies dependencies between ontological *Targets* (e.g. "positive_financial_metrics") and sentiment-bearing *Synset* expressions (e.g. "own", "lose") in the input text. When a valid (*Synset*, *Target*) mapping is detected based on the ontology, a corresponding proposition is instantiated.

However, unlike Kintsch's (1974) propositions containing arguments like entities and objects, Sentium's propositions focus solely on the (*Synset*, *Target*) relations that convey sentiment polarity. This abstract semantic structure aligns with how humans conceptualise sentiment, facilitating intuitive modelling and interpretability.

Analysing sentiment through propositions also accounts for how different semantic scope interpretations can lead to divergent annotations, even among expert human labellers, as observed in Malo et al. (2014). Such variability likely arises from backward inferencing processes and differing proposition weighting strategies employed by each annotator.

To illustrate, consider "NVIDIA, owning 80% of the $65.3B GPU market, is slowly losing share to AMD". One annotator may label this negatively by prioritising the "lose" proposition, which could be structurally represented as {"own": {"market_share": +1}, "lose": {"market_share": -2}}. Another may view it as positive, giving more weight to the "own" proposition about NVIDIA's large market share, represented as {"own": {"market_share": +2}, "lose": {"market_share": -1}}. OSLF allows the adaptive combination of propositions into personalised ontologies mapping *Synset* to *Target* polarity weights, akin to human subjectivity. Representing each (*Synset*, *Target*) mapping as an interpretable proposition enables capturing and examining these distinct reasoning paths.

To empirically extract robust dependency patterns between ontological *Targets* and sentiment-bearing *Synsets*, we applied a 50-50 split on the Financial Phrasebank dataset (Malo et al., 2014) instead of the traditional 80-20. While using more data could increase accuracy, the aim was not to exhaustively cover all possible dependencies. Rather, it is sought to derive a representative set of high-confidence rules capturing common sentiment composition phenomena in this domain.

Through this data-driven analysis, 61 dependency patterns were identified between *Targets* like "positive_financial_metrics" and *Synsets* like "increase" and "decrease". For example, such *Targets* frequently depend on were objects of these *Synsets* (nsubj/dobj dependency relations), as in "Profit increased this quarter".

Importantly, news headlines exhibit their own grammar structures for concisely conveying key information (Salih and Abdulla, 2012), unlike noisier social media text (Kontopoulos et al., 2013). To handle this, we uncovered 9 common grammatical templates like "versus" comparisons (e.g. "pre-tax profit of $100M versus a loss of $50M") and "up/down" framing (e.g. "operating profit totalled $7.2M, up from a $4.0M loss year-on-year").

This empirical pattern mining approach allows Sentium to robustly capture the diverse linguistic constructions used to express financial sentiment, beyond just simplistic word co-occurrences. The extracted dependency rules systematically map natural language to proposition-like semantic representations grounded in the ontology. This tight coupling of data-driven patterns with structured knowledge facilitates precise sentiment composition modeling.

For example, analysing "revenue increased 5% over projections" involves accessing the ontology {"increase": {"positive_financial_metrics": +1}} based on matching the dependency $revenue(Target) \rightarrow increased(Synset)$. In contrast, "costs increased unexpectedly" would yield {"increase": "negative_financial_metrics": -1}} – the same *Synset* flips polarity for a different *Target* concept.

By deriving these rich semantic parses in an automated yet interpretable, reasoning-driven manner, Sentium can provide reliable sentiment predictions along with rationale explanations auditable by humans. This combination of empirical pattern coverage and cognitive modeling of compositional semantics allows our approach to achieve new levels of accuracy and transparency for sentiment analysis.

3 Experiment

To conduct an evaluation, we leverage the remaining 50% test set of the Financial Phrasebank dataset (Malo et al., 2014), compared against four benchmark models. We include two dictionary-based bag-of-words approaches, Henry (Henry, 2008) and MASTER (Loughran and McDonald, 2011), accessed through the `sentibank` library (Oh, 2024) (under CC-BY-NC-SA-4.0 license). Additionally, we consider two state-of-the-art financial language models, FinBERT (Araci, 2019) and FLANG-BERT (Shah et al., 2022), leveraging the HuggingFace `transformers` library (Wolf et al., 2020) (under Apache 2.0 license).

The Henry dictionary, designed explicitly for analysing tones in earnings press releases, comprises 189 unigram entries selected based on contextual analysis, with a focus on directional collocates. The MASTER dictionary targets sentiment expressions commonly encountered in financial regulatory filings, such as 10-K reports. With 3,876 domain-specific affect terms, this lexicon has demonstrated a statistically significant negative correlation with file date excess returns, underscoring its applicability. Both dictionaries underwent a manual labelling process by the authors.

Both FinBERT and FLANG-BERT are state-of-the-art language models based on the BERT architecture (Devlin et al., 2018). While both models were originally pre-trained on the Financial Phrasebank dataset, to ensure optimal performance, we further fine-tuned these models using the 50% training set, aligning them with the task-specific data distribution.

4 Results

The accuracy results demonstrate Sentium consistently outperforming traditional dictionary-based approaches (Henry=80.12%; MASTER=58.83%) while achieving highly competitive results compared to state-of-the-art language models (FinBERT=96.03%; FLANG-BERT=97.35%) with an accuracy of 92.05% (Table 2).

Model	Accuracy	Precision	F1
Henry *(Henry, 2008)*	0.8012	0.7985	0.7976
MASTER *(Loughran and McDonald, 2011)*	0.5883	0.6171	0.5666
FinBERT *(Araci, 2019)*	0.9603	0.9604	0.9602
FLANG-BERT *(Shah et al., 2022)*	0.9735	0.9738	0.9736
Sentium	0.9205	0.9228	0.9191

Table 2: Performance comparison of Sentium against benchmarks on financial sentiment analysis task.

While the overall accuracy is impressive by itself, Sentium's true strength lies in its *precision* - a

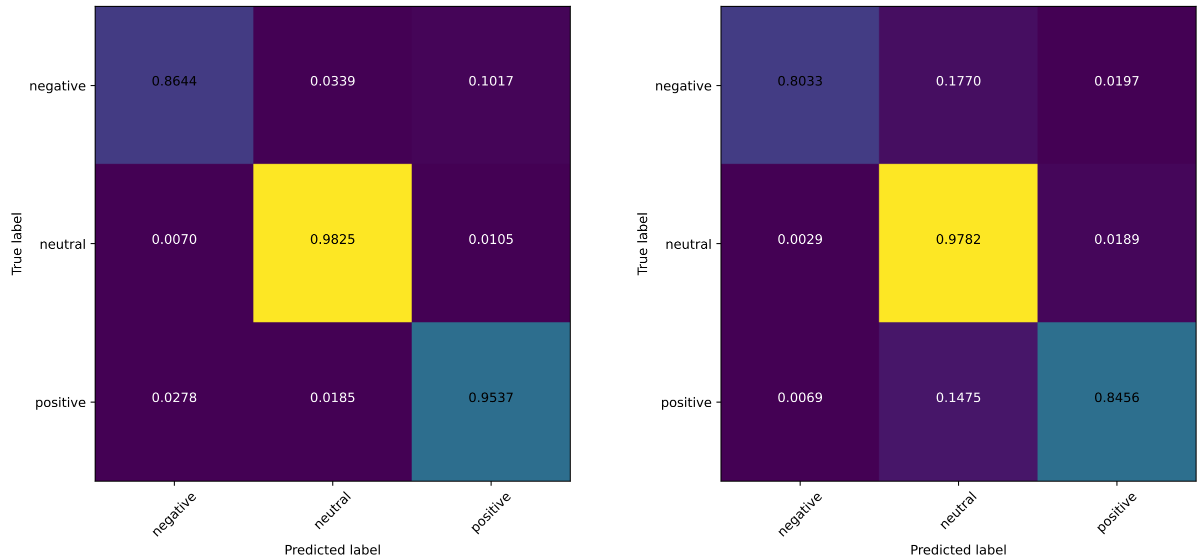

Figure 3: Confusion Matrix Analysis (Left=FinBERT; Right=Sentium). The comparison highlights Sentium's strength in precision compared to FinBERT's (Araci, 2019) baseline.

crucial capability for financial sentiment analysis. Both FLANG-BERT (Negative=93.22%; 6.78% misclassified as positive) and FinBERT (Negative=86.44%; 10.17% misclassified as positive) exhibit non-trivial error rates in misclassifying true negatives as positive sentiment.

In contrast, Sentium demonstrates a substantially lower 1.97% misclassification rate for true negatives predicted as positive - a 3.5x and 5x reduction compared to FLANG-BERT and FinBERT respectively. Additionally, unlike FinBERT which misclassified 2.78% of true positives as negative, Sentium's error rate is a mere 0.69% for this type of egregious polarity reversal (Figure 3).

The implications are clear: **Sentium excels at reliably distinguishing positive and negative sentiments, a critical requirement in a domain where misinterpreting pessimistic or optimistic signals can have severe consequences.** While FLANG-BERT and FinBERT achieve higher overall accuracy on this dataset, their error profiles are considerably more skewed towards costly polarity confusions between positive and negative classes.

5 Conclusion

Sentium represents a significant stride towards developing transparent, interpretable and understandable sentiment analysis systems. By uniting structured knowledge from the financial domain with neural models under a cognitively-inspired framework, it achieves state-of-the-art performance while maintaining crucial interpretability. Sentium's explicit compositional reasoning over semantic propositions grounded in an intuitive ontology enables comprehensive rationale generation, fostering trust and auditability in high-stakes decision-making scenarios. This human-centred approach bridges a critical gap in existing opaque black-box methods, paving the way for the responsible deployment of AI in sentiment analysis and allied domains where decision transparency is paramount.

6 Limitation

While Sentium demonstrates impressive performance and interpretability, certain limitations should be acknowledged. First, the financial sentiment ontology currently focuses exclusively on the financial domain, potentially constraining its applicability across diverse domains. Extending the ontology to capture sentiment nuances in other sectors would be a valuable future endeavor. Additionally, the ontology construction process, although grounded in empirical data analysis, still involves manual curation by domain experts, introducing potential human biases. Exploring semi-automated or fully automated ontology learning methods could alleviate this limitation. Finally, Sentium's modular architecture, while cognitively inspired, may not fully capture the complex, parallel processing dynamics of human language comprehension, suggesting opportunities for further refinement.

References

Doris Aaronson and Hollis Shapiro Scarborough. 1977. Performance theories for sentence coding: Some quantitative models. *Journal of Verbal Learning and Verbal Behavior*, 16(3):277–303.

Tom Aarsen. SpanMarker.

Tamara Amjad Al-Qablan, Mohd Halim Mohd Noor, Mohammed Azmi Al-Betar, and Ahamad Tajudin Khader. 2023. A survey on sentiment analysis and its applications. *Neural Computing and Applications*, 35(29):21567–21601.

Shivaji Alaparthi and Manit Mishra. 2021. Bert: A sentiment analysis odyssey. *Journal of Marketing Analytics*, 9(2):118–126.

John R Anderson. 2000. Cognitive psychology and its implications.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Sumalatha Bandari and Vishnu Vardhan Bulusu. 2020. Survey on ontology-based sentiment analysis of customer reviews for products and services. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19*, pages 91–101. Springer.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical Report 1, Technical report C-1, the center for research in psychophysiology, University of Florida.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Z Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Zhuo Chen, Chengyue Jiang, and Kewei Tu. 2023. Using interpretation methods for model enhancement. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ikhyun Cho, Yoonhwa Jung, and Julia Hockenmaier. 2023. Sir-absc: Incorporating syntax into roberta-based sentiment analysis models with a special aggregator token. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Verna Dankers and Christopher G Lucas. 2023. Non-compositionality in sentiment: New data and analyses. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William H DuBay. 2004. The principles of readability. *Online Submission*.

Andrea Esuli and Fabrizio Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. In *11th Conference of the European chapter of the association for computational linguistics*, pages 193–200.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Jerry A Fodor. 1983. *The modularity of mind*. MIT press.

Lila R Gleitman. 1965. Coordinating conjunctions in english. *Language*, 41(2):260–293.

Fernand Gobet and Peter CR Lane. 2010. The chrest architecture of cognition: The role of perception in general intelligence. In *Procs 3rd Conf on Artificial General Intelligence: AGI-2010*. Atlantis Press.

Richard Graf and Jane W Torrey. 1966. Perception of phrase structure in written language. In *American Psychological Association Convention Proceedings*, volume 83, page 84.

Elaine Henry. 2008. Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, 45(4):363–407.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

Walter Kintsch. 1974. The representation of meaning in memory.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. 2013. Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074.

Marta Kutas and Steven A Hillyard. 1980a. Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, 11(2):99–116.

Marta Kutas and Steven A Hillyard. 1980b. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Christopher D Manning. 2015. The case for universal dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, volume 1.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Aminu Muhammad, Nirmalie Wiratunga, and Robert Lothian. 2016. Contextual sentiment analysis for social media genres. *Knowledge-based systems*, 108:92–101.

Farhad Nadi, Hadi Naghavipour, Tahir Mehmood, Alliesya Binti Azman, Jeetha A/P Nagantheran, Kezia Sim Kui Ting, Nor Muhammad Ilman Bin Nor Adnan, Roshene A/P Sivarajan, Suita A/P Veerah, and Romi Fadillah Rahmat. 2023. Sentiment analysis using large language models: A case study of gpt-3.5. In *The International Conference on Data Science and Emerging Technologies*, pages 161–168. Springer.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nick Oh. 2024. sentibank: A unified resource of sentiment lexicons and dictionaries. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2003–2013.

Lee Osterhout and Phillip J Holcomb. 1993. Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, 8(4):413–437.

Barbara Partee. 2007. Compositionality and coercion in semantics: The dynamics of adjective meaning. *Cognitive foundations of interpretation*, 2007:145–161.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud, and Mohammed Baz. 2022. Transfer learning for sentiment analysis using bert based supervised fine-tuning. *Sensors*, 22(11):4157.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Younis Mehdi Salih and Q Abdulla. 2012. Linguistic features of newspaper headlines. *Journal of Al-Anbar University for Language and Literature*, 7:192–213.

Iqbal H Sarker. 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420.

Kim Schouten, Flavius Frasincar, and Franciska de Jong. 2017. Ontology-enhanced aspect-based sentiment analysis. In *Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings 17*, pages 302–320. Springer.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484.

Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tala Talaei Khoei, Hadjar Ould Slimane, and Naima Kaabouch. 2023. Deep learning: Systematic review, models, challenges, and research directions. *Neural Computing and Applications*, 35(31):23103–23124.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Mikolaj Wojciuk, Zaneta Swiderska-Chadaj, Krzysztof Siwek, and Arkadiusz Gertych. 2024. Improving classification accuracy of fine-tuned cnn models: Impact of hyperparameter optimization. *Heliyon*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Nor Nadiah Yusof, Azlinah Mohamed, and Shuzlina Abdul-Rahman. 2018. A review of contextual information for context-based approach in sentiment analysis. *International Journal of Machine Learning and Computing*, 8(4):399–403.