Faculty of Media Engineering and Technology
German University in Cairo

LLM for Multi-Source Financial Sentiment

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

By

Ali Ussama Elsayed Mohamed Youssef

Supervised by

Prof. Dr. Walaa Khaled Ibn Elwaleed Ibrahim Gad

June 6, 2026

Faculty of Media Engineering and Technology
German University in Cairo

LLM for Multi-Source Financial Sentiment

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

By

Ali Ussama Elsayed Mohamed Youssef

Supervised by

Prof. Dr. Walaa Khaled Ibn Elwaleed Ibrahim Gad

June 6, 2026

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree of Science (B.Sc.) at the German University in Cairo (GUC),

(ii) due acknowledgment has been made in the text to all other material used

Ali Ussama Elsayed Mohamed Youssef
June 6, 2026

Acknowledgments

First and foremost, all praise and gratitude are due to God, whose guidance and grace made this work possible and carried me through the most difficult moments of this journey.

I would like to sincerely thank my supervisor, Prof. Dr. Walaa Khaled Ibn Elwaleed Ibrahim Gad, for her guidance, patience, and support throughout this thesis. Her belief in the topic when its direction was still taking shape gave me the confidence to pursue it with conviction, and her availability whenever I needed counsel, even after many late-night updates sent at 4 a.m., made the process far less difficult than it would otherwise have been. Thank you.

To my mom and dad: no words in any acknowledgment section can do justice to what you have given me. Your patience, sacrifice, and steadfast support across every stage of this degree are woven into every page of this thesis, even where your names do not appear. I hope this work is something you can be proud of.

To my brothers: thank you for tolerating me through the late nights, the stress, and the moments when I was far less pleasant to be around than I should have been. Your presence made the hard stretches manageable.

To the Future and Beyond

Abstract

Financial markets receive information through sources that differ in institutional accountability, including regulatory filings, professional news, and social media discussion. Existing financial sentiment analysis pipelines often collapse these sources into a single sentiment estimate, which can obscure disagreement among sources describing the same corporate event. Our thesis presents Sentinel, an LLM-based framework that treats cross-source sentiment variation as an analytical signal rather than as noise to be averaged away.

Sentinel introduces two main analytical components. The Sentiment Disagreement Index (SDI) measures event-level disagreement among source documents using a bounded pairwise distance formulation with source-type credibility weights. The Aspect Divergence Index (ADI), within the Multi-Aspect Sentiment Divergence (MASD) framework, decomposes disagreement across financial aspects and records directional differences in a source-aspect sentiment matrix. The implemented case-study pipeline uses FinBERT to score event-aligned documents from professional news, Reddit discussion, and Securities and Exchange Commission (SEC) filings. A separate open-weight scoring engine, Sentinel-LLaMA3-FPB75, is trained using Quantized Low-Rank Adaptation (QLoRA), evaluated on Financial PhraseBank (FPB), and applied to the Tesla corpus through token log-probability extraction.

In the Tesla Q1 2024 primary case study, the Level 2 normalised SDI is 0.35172, placing the event in the Partial Disagreement zone. Revenue is the most contested keyword-based aspect, and the dominant source-pair contribution comes from professional news versus the SEC filing. Level 3 Monte Carlo Dropout (MC Dropout) uncertainty adjustment preserves the same zone classification. The AppLovin Q4/FY2025 exploratory case also falls in the Partial Disagreement zone, with a Level 2 normalised SDI of 0.248746. Sentinel-LLaMA3-FPB75 achieves macro F1 0.97526 on the FPB 75%-agreement benchmark, while lower-agreement audits show reduced robustness. These results establish Sentinel as a proof-of-concept framework for measuring and explaining cross-source disagreement in financial event data, with calibration, broader source coverage, and predictive validation left to future multi-event studies.

Contents

List of Abbreviations

ABSA	Aspect-Based Sentiment Analysis
ADI	Aspect Divergence Index
AI	Artificial Intelligence
API	Application Programming Interface
ARVol	Abnormal Return Volatility
BERT	Bidirectional Encoder Representations from Transformers
CFA	Chartered Financial Analyst Institute
ECE	Expected Calibration Error
EDGAR	Electronic Data Gathering, Analysis, and Retrieval
EU AI Act	European Union Artificial Intelligence Act
FinBERT	Financial BERT
FinEntity	Financial Entity-Level Sentiment Dataset
FinGPT	Financial GPT
FiQA	Financial Question Answering
FLANG	Financial Language Model
FLUE	Financial Language Understanding Evaluation
FNSPID	Financial News Stock Price Integration Dataset
FP16	16-bit Floating Point
FP32	32-bit Floating Point
FPB	Financial PhraseBank
FSA	Financial Sentiment Analysis
FSD	Full Self-Driving
GB	Gigabyte
GPU	Graphics Processing Unit
HAD	Heterogeneous Agent Discussion
LIME	Local Interpretable Model-Agnostic Explanations
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MASD	Multi-Aspect Sentiment Divergence
MC Dropout	Monte Carlo Dropout

MSFSA	Multi-Source Financial Sentiment Analysis
NF4	NormalFloat 4-bit
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QLoRA	Quantized Low-Rank Adaptation
RAMAS	Retrieval-Augmented Multi-Agent Sentiment System
SDI	Sentiment Disagreement Index
SEC	Securities and Exchange Commission
Sentinel	an LLM-based multi-source financial sentiment analysis framework
SEntFiN	Sentiment Entity in Finance
SHAP	SHapley Additive Explanations
SR 11-7	Federal Reserve Supervisory Guidance on Model Risk Management
VIX	Volatility Index
VRAM	Video Random Access Memory
XAI	Explainable Artificial Intelligence

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Financial Markets as Information-Processing Systems

Financial markets respond not only to numerical disclosures, but also to the language through which events are described and interpreted. Beliefs about an asset's future cash flows are shaped by regulatory filings, professional news, analyst commentary, earnings calls, and social media discussion. These sources already inform practical financial systems: news-derived sentiment signals are used in algorithmic trading, textual indicators contribute to credit-risk assessment, disclosure language is monitored for irregularities, and portfolio managers use sentiment summaries when evaluating market narratives. Financial Sentiment Analysis (FSA) is therefore concerned with a problem that is both methodological and economically relevant. Its outputs must be accurate, robust, and interpretable because they may inform financially consequential decisions.

The same financial event can produce different sentiment signals across source types. Regulatory filings are written under legal liability and tend to use restrained language; professional news may emphasise consequence; and social media permits more direct and variable reactions. Our thesis represents these source differences through an accountability spectrum, ranging from SEC filings and scripted earnings calls to professional news and retail social media. Aggregating these sources without preserving their identity may conceal economically relevant disagreement.

1.2 Progress and Persistent Limits of Financial Sentiment Analysis

FSA has progressed from lexicon-based methods to domain-adapted transformer and LLM-based approaches. Early methods relied on expert-curated dictionaries, most notably the Loughran–McDonald financial sentiment dictionary [1], which showed that financial terms such as "liability" and "risk" cannot be interpreted reliably using general-purpose sentiment lexicons. Although dictionary-based approaches are transparent and computationally inexpensive, they have limited ability to represent contextual polarity, negation, and changing language use. Subsequent supervised models used handcrafted features with classifiers such as support vector machines and logistic regression, while universal language model fine-tuning [2] and domain-adaptive pretraining [3] enabled stronger adaptation to financial text.

Domain-specific transformer models further improved financial sentiment classification. Financial BERT (FinBERT) [4, 5] achieves approximately 88% macro F1 (on the FPB 75%-agreement subset; the all-agree subset yields higher results, as reported in Section 4.2) on the Financial PhraseBank (FPB) benchmark [6] and remains a widely adopted baseline. Other domain-adapted models, including Financial Language Model (FLANG) [7] and BloombergGPT [8], have broadened financial language modelling at different scales. More recent approaches include GPT-4o mini, reported by Zhang et al. [9] at 0.87090 macro F1 on FPB; FinSentLLM [9], reported at approximately 95% macro F1; and Heterogeneous Agent Discussion (HAD) [10], which applies multi-agent deliberation to FSA. Together, these studies indicate that sentiment classification on established financial benchmarks is well developed.

A limitation remains: major FSA benchmarks do not preserve source-type structure. FPB consists of financial news sentences [6]; Financial Question Answering (FiQA) combines headlines and microblogs without source-type disaggregation [11]; Sentiment Entity in Finance (SEntFiN) [12] focuses on news; and Financial Entity-Level Sentiment Dataset (FinEntity) [13] extends entity-level annotation without addressing source-type disagreement. Du et al. [14] survey methods applied across several financial text sources, but do not propose a framework for comparing their sentiment signals within the same event. Existing work therefore provides strong tools for sentiment classification, but limited support for measuring disagreement across sources describing a shared corporate event.

1.3 Problem Statement

Existing FSA research largely evaluates sentiment within individual source types, although evidence suggests that differences across sources are economically relevant. Garcia [15], analysing 1,823 U.S. firms from 2015 to 2021, found that sentiment divergence between social media and traditional news predicts financial distress, with a one standard deviation increase in divergence volatility associated with a 46-basis-point increase in the one-year probability of default. Cookson et al. [16] similarly showed that social media platforms may exhibit correlated attention without corresponding agreement in sentiment. These studies motivate the treatment of cross-source variation as a measurable signal.

Two methodological gaps remain. Single-source sentiment classifiers do not measure disagreement among sources reporting on the same event, while methods such as FinSentLLM [9] compare classifier outputs on identical text rather than source-specific accounts of a shared event. Moreover, an aggregate disagreement score cannot indicate whether divergence arises from Revenue, Management, Risk, or Market interpretation.

The Tesla Q1 2024 earnings event provides the case-study setting for examining these gaps. Average sentiment scores from the company's Form 8-K filing, Guardian news coverage, and Reddit discussion span 0.420 units on the $[-1, +1]$ scale. The filing is near-neutral, whereas professional news is strongly negative. This event-aligned setting allows Sentinel to measure source-level disagreement and examine its aspect-level structure.

Sentinel combines source-level disagreement measurement with aspect-level decomposition. The current case-study implementation executes unweighted and static credibility-weighted disagreement measurement, while model-uncertainty treatment remains part of the specified framework rather than the reported case-study results. To the best of our knowledge, no published framework combines these elements for cross-source financial sentiment analysis.

1.4 Proposed Sentinel Framework

Our thesis presents Sentinel, an end-to-end framework for Multi-Source Financial Sentiment Analysis (MSFSA) that measures how professional news, social media discussion, and regulatory filings differ in their characterisation of the same corporate event. The framework produces continuous signed sentiment scores for each source document and preserves cross-source variation for analysis rather than collapsing it into one aggregate score.

Sentinel has two principal analytical components. The Sentiment Disagreement Index (SDI) produces a normalised event-level disagreement score in $[0, 1]$ and attributes that disagreement to contributing source pairs. The Multi-Aspect Sentiment Divergence (MASD) framework extends the analysis through the Aspect Divergence Index (ADI), which identifies the financial aspects on which sources diverge and records the direction of that divergence in a source-aspect

sentiment matrix. The full SDI specification includes credibility and uncertainty-aware levels, while the current case-study implementation evaluates the unweighted and static credibility-weighted levels.

The case-study analysis and the scoring-engine evaluation serve different purposes in our thesis. FinBERT is used to compute SDI and ADI for the Tesla Q1 2024 earnings event, whereas Sentinel-LLaMA3 is evaluated separately on the FPB benchmark as an open-weight financial sentiment classifier and applied to the Tesla corpus via token log-probability extraction (Section 4.4.5). Tesla Q1 2024 provides a suitable setting for the primary case study because the selected sources offer different accounts of the same earnings, safety, strategy, and market developments. The same pipeline is also applied to AppLovin Q4/FY2025 as an exploratory secondary case.

1.4.1 Scope of Our Current Thesis Implementation

The implementation evaluated in our thesis covers the parts of Sentinel required for the reported case studies and benchmark experiments. It includes three-source event collection, FinBERT-based case-study scoring, Level 1 and Level 2 SDI, keyword-based MASD/ADI, source-pair contribution reporting, Sentinel-LLaMA3 evaluation on FPB, a scoring engine comparison on the Tesla Q1 2024 corpus reported in Section 4.4.5, and a sentence-ID LLM MASD sensitivity audit for Tesla Q1 2024. Other components remain part of the broader framework specification but are not evaluated here: source-type calibration gates, dynamic credibility weighting, earnings-call and analyst-report ingestion, production-scale sentence-ID or span-grounded LLM aspect extraction across events, and multi-event SDI-to-market-outcome regression. They are therefore presented as future extensions, not as findings of our current study.

1.5 Research Questions

Four research questions guide the design and evaluation of Sentinel.

RQ1: Can SDI, in its implemented unweighted and static credibility-weighted forms, measure sentiment disagreement across heterogeneous sources reporting on a shared financial event?

RQ2: Can ADI-based aspect decomposition identify the financial dimensions and sentiment directions underlying observed cross-source disagreement?

RQ3: Can a QLoRA fine-tuned open-weight LLM (Sentinel-LLaMA3) match or exceed proprietary model performance on standard financial sentiment benchmarks while remaining open-weight, locally deployable, and reproducible from public components?

RQ4: In the Tesla Q1 2024 case study, is observed cross-source disagreement associated primarily with source-type editorial constraints or with social media noise?

1.6 Contributions

Our thesis makes four principal contributions.

1. **Sentiment Disagreement Index (SDI).** Our thesis introduces SDI as a formal measure of sentiment disagreement among source types reporting on the same company-event instance. The metric is based on a pairwise root-mean-square distance and produces a normalised scalar disagreement score in $[0, 1]$. The SDI admits an $\mathcal{O}(n)$ computation via the standard algebraic identity for weighted pairwise squared differences; the contribution lies in its credibility-weighted structure, interpretation ladder, and source-pair decomposition, rather than in the identity itself. Its purpose is to retain disagreement across source types as an analytical output rather than remove it through aggregation. In addition to the scalar score, the SDI output record reports an interpretation zone classification, a per-source-pair weighted contribution breakdown, and metadata recording source sentiment signs and applied credibility weights. Accordingly, the novelty of the SDI is methodological rather than algebraic: it combines a credibility-weighted source-type structure, an interpretation ladder, and source-pair attribution outputs within an auditable disagreement framework, while its underlying pairwise distance formula remains a standard weighted statistical distance.

   The full SDI framework is specified as a four-level experimental ladder. Level 1 provides an unweighted baseline, while Level 2 incorporates static source-type credibility priors. Level 3 adds instance-level uncertainty adjustment through Monte Carlo Dropout (MC Dropout), and Level 4 specifies dynamic credibility weighting conditioned on event-level evasion scoring, historical source track record, and market regime. In our current evaluation, Levels 1 and 2 are reported for both case studies, and Level 3 is reported for the Tesla Q1 2024 case study. Level 4 remains a specified extension and is not evaluated empirically in this thesis.

2. **Multi-Aspect Sentiment Divergence (MASD) framework and Aspect Divergence Index (ADI).** The second contribution extends disagreement measurement from the document level to specific financial aspects. The MASD framework examines whether observed cross-source divergence is associated with Revenue, Management, Risk, or Market interpretation, rather than reporting only an aggregate event-level score. Within this framework, ADI is computed as the population variance of per-source aspect sentiment scores. The resulting divergence vector identifies which dimensions account for the greatest observed disagreement.

   The source-aspect sentiment matrix complements the ADI vector by recording the signed

sentiment assigned by each source type to each aspect. This makes it possible to distinguish differences in sentiment magnitude from directional splits, where sources assign opposing sentiment orientations to the same financial dimension.

3. **Sentinel-LLaMA3, a QLoRA fine-tuned scoring engine.** Sentinel-LLaMA3-FPB75 is a 4-bit quantised Large Language Model Meta AI (LLaMA) 3.1 8B Instruct model with low-rank adapters [17, 18], trained on the FPB 75%-agreement subset. On the held-out test set ($n = 691$), the model achieves macro F1 0.97526. This score exceeds the literature-reported GPT-4o mini macro F1 of 0.87090 [9] and the approximate FinSentLLM ensemble score of 0.95 [9], subject to the caveat that these values are reported under different evaluation protocols rather than obtained from an identical controlled test set.

   Our thesis further evaluates the limits of this benchmark result through leakage-safe robustness audits. Applying the trained FPB75 adapter to non-overlapping lower-agreement subsets yields macro F1 0.72713 on the FPB66-exclusive set ($n = 763$) and 0.56338 on the FPB50-exclusive set ($n = 627$), showing substantial degradation as annotator agreement decreases. A separate Sentinel-LLaMA3-Robust-FPB50 adapter, trained on the full FPB 50%-agreement subset under a duplicate-safe stratified protocol, achieves macro F1 0.87113 on its 973-example test split. This additional experiment shows that dedicated training on lower-agreement data improves performance under that separate evaluation protocol.

4. **End-to-end application to real financial events.** Our thesis implements a reproducible case-study pipeline spanning FinBERT baseline validation, multi-source data collection from the Guardian Open Platform, the Arctic Shift Reddit archive, and SEC EDGAR via edgartools, FinBERT-based case-study scoring, SDI and ADI computation, and unified result synthesis. In the primary Tesla Q1 2024 case study, the Level 2 normalised SDI is 0.35172, placing the event in the Partial Disagreement zone. The News–SEC pair is the dominant divergence contributor, with a weighted contribution of 0.124.

   Aspect-level decomposition yields a divergence vector of $(0.199, 0.135, 0.097, 0.013)$ for Revenue, Management, Market, and Risk, respectively. Under keyword-based extraction, Revenue is the most contested dimension and includes a Reddit–SEC sentiment gap of 1.091 units on the $[-1, +1]$ scale, while Management is the only aspect exhibiting a potential directional split. This observation is based on a single Reddit-extracted sentence for the Management aspect and should be treated as illustrative rather than validated. The near identity of SDI Level 1 and Level 2, with a normalised delta of 0.0046, suggests that the observed disagreement in this event is associated more strongly with differences between professional news and regulatory disclosure than with social media noise. The same pipeline is also applied to the AppLovin Q4/FY2025 earnings event as an exploratory secondary case in a different company and event setting.

1.7 Thesis Organisation

The remainder of our thesis is organised as follows.

Chapter 2 establishes the literature basis for our study. It reports the PRISMA-guided screening process used to identify the reviewed studies and organises the resulting literature across ten analytical sections. These sections examine the progression of FSA, domain-adapted financial language models, heterogeneous financial source types, multi-source fusion, aspect-level sentiment analysis, disagreement measurement, source credibility, uncertainty, and explainability. The chapter concludes by identifying the combination of requirements addressed by Sentinel, which was not found in a single framework within the reviewed literature.

Chapter 3 defines the Sentinel framework and the scope of its current implementation. It introduces the source-accountability rationale, data collection and preprocessing pipeline, sentiment-score construction, and the calibration limitations governing cross-source comparison. It then develops the SDI computation ladder and the MASD/ADI aspect-level decomposition, followed by the explainability outputs and end-to-end workflow. The chapter also presents the Sentinel-LLaMA3 QLoRA fine-tuning methodology and makes explicit that the reported case-study scores are produced using FinBERT, while Sentinel-LLaMA3 is evaluated separately on FPB and applied to the Tesla corpus via token log-probability extraction (Section 4.4.5).

Chapter 4 presents two related but distinct evaluation tracks. The first reports the FinBERT-based event analysis: baseline validation, the Tesla Q1 2024 primary case study across sentiment scoring, SDI, and ADI, and the exploratory AppLovin Q4/FY2025 secondary application. The second evaluates Sentinel-LLaMA3-FPB75 on the FPB benchmark and reports the lower-agreement robustness audits and the separate Sentinel-LLaMA3-Robust-FPB50 experiment. The chapter concludes by interpreting the results across phases and consolidating the reported outputs.

Chapter 5 answers the four research questions within the evidential scope of our thesis. It discusses the implications of the case-study and benchmark findings, states the limits of the conclusions that may be drawn from them, and presents eight limitations of the current implementation together with their corresponding extension paths.

Chapter 6 develops the future work arising from those limitations. It focuses on multi-event validation and downstream SDI-to-volatility testing, dynamic credibility weighting, temporal propagation modelling, broader source coverage, and source-type-aware fine-tuning with cross-benchmark evaluation of Sentinel-LLaMA3.

Chapter 2

Literature Review

Financial sentiment analysis has moved from isolated-text classification toward methods that account for how domain and source type shape sentiment expression. Recent LLM-based methods support richer semantic extraction from financial language, but most still operate at the level of individual documents or individual source types. The reviewed literature did not identify a published framework that combines multi-source data alignment, aspect-level decomposition, continuous LLM-derived sentiment scores, and a formal measure of source-level divergence. This chapter reviews the relevant literature across ten analytical sections, showing how each strand informs Sentinel and how the research gap addressed in this thesis is defined. The review includes 36 studies selected through the PRISMA screening process described in Section 2.1 and concludes with an explicit positioning matrix in Section 2.11.

2.1 Screening Methodology

2.1.1 Search Protocol and Inclusion Criteria

The literature reviewed in this chapter was selected through a PRISMA-guided screening process covering seven search sources: EBSCO, ScienceDirect, Google Scholar, IEEE Xplore, arXiv, Semantic Scholar, and the ACL Anthology. The search identified 7,245 records in total before deduplication, with 5,816 remaining after duplicate removal. Title and abstract screening excluded 5,651 records outside the review scope, leaving 165 records for full-text assessment. Of these, 129 were excluded because they lacked sufficient methodological detail, did not analyse text, were not relevant to sentiment conflict, or had been superseded by a later version. The final review corpus contains 36 studies, with two additional methodological references cited for supporting background discussion. Table 2.1 reports the screening counts for each phase. The IEEE Xplore search returned 173 results spanning 2019–2025, consistent with the emergence of transformer-based methods in financial Natural Language Processing (NLP) fol-

lowing the introduction of Bidirectional Encoder Representations from Transformers (BERT) in late 2018.

Table 2.1: PRISMA-guided literature review screening process.

Phase	Stage	Breakdown / Details	n
Identification	Database Searching	EBSCO: 2,253; ScienceDirect: 4,218; Google Scholar: 412; IEEE Xplore: 173. Searches used full-text, peer-reviewed, English-language, and 2015–2025 filters where supported. IEEE Xplore returned 173 records under this protocol, all from 2019–2025; no records were returned for 2015–2018.	7,056
	Additional Sources	arXiv: 87; citation tracking: 64; Semantic Scholar: 22; ACL Anthology: 16.	189
	Total Before Deduplication		**7,245**
Screening	After Duplicate Removal	Duplicates removed: 1,429, approximately 20% duplication rate.	5,816
	Excluded at Title/Abstract	Not financial-text sentiment: 3,218; numerical-only: 987; prediction without NLP: 681; non-English: 512; other/insufficient relevance: 253.	5,651
	Passed to Full-Text		**165**
Eligibility	Full-Text Articles Assessed		165
	Full-Text Excluded	Insufficient detail: 37; no text analysis: 31; not relevant to conflict: 38; superseded: 23.	129
	Passed to Inclusion		**36**
Inclusion	**Final Studies Included**	**T1** LLMs/Transformers: ≈11 **T2** domain-specific LMs: ≈6 **T3** multi-source: ≈6 **T4** sentiment divergence: ≈6 **T5** aspect-based: ≈5 **T6** explainability: ≈2	**36**

The 36 included studies are classified by primary analytical focus into six themes, as summarised in Table 2.1: LLM and transformer architectures (T1), domain-specific language models (T2), multi-source fusion (T3), sentiment divergence (T4), aspect-based analysis (T5), and explainability (T6). The chapter uses these themes as an organising basis rather than as a fixed section structure. Several themes contain distinct lines of work that are clearer when discussed separately, so the review is developed across ten analytical sections while retaining the same overall thematic coverage.

2.2 Foundations of Financial Sentiment Analysis

2.2.1 The Single-Source Paradigm

Most FSA research is conducted within a single-source classification setting. A corpus is selected from one source type, annotated for sentiment, used to train or fine-tune a classifier, and divided into training and held-out evaluation partitions. This design supports reproducible model comparison and has enabled substantial progress in financial sentiment classification. However, it does not examine whether the source through which a financial event is communicated affects the sentiment signal assigned to that event.

Du et al. [14], in their survey of FSA methods in *ACM Computing Surveys*, review techniques applied to corporate disclosures, earnings calls, financial news, and social media. Although these source types are all represented in the surveyed literature, they are considered as separate application settings rather than as aligned channels reporting on the same event. The principal annotated FSA benchmarks, including FPB [6], FiQA [11], and SEntFiN [12], follow the same single-source design. They therefore support evaluation of sentiment classification performance, but not measurement of disagreement across sources describing a shared corporate event.

This distinction motivates our thesis. A model evaluated on news data alone cannot reveal whether regulatory filings or Reddit discussions of the same event produce different sentiment scores. If outputs from these sources are aggregated before their differences are examined, source-specific variation is lost. Sentinel begins from the premise that this variation should be retained and measured because it may reflect differences in how institutional channels frame the same financial event.

2.2.2 Lexicon-Based and Pre-Neural Approaches

Before transformer-based models became established in NLP, FSA was commonly performed using sentiment lexicons, bag-of-words representations, and shallow classifiers. Applying general-purpose lexicons such as LIWC and the Harvard-IV dictionary to financial text exposed an important domain mismatch: terms such as “liability,” “tax,” and “risk,” which often carry negative connotations in ordinary language, may function as neutral technical vocabulary in regulatory filings and earnings reports. Loughran and McDonald [1] addressed this mismatch by developing a finance-specific dictionary based on the language of SEC filings, distinguishing evaluative terms from routine financial terminology. The resulting dictionary remains valuable for its interpretability and domain relevance, although neural approaches have achieved stronger classification performance on held-out benchmarks.

Subsequent advances in transfer learning shifted financial sentiment modelling toward neural language models. Howard [2] showed through ULMFiT that general language-model pretrain-

ing followed by task-specific fine-tuning improves text classification performance relative to feature-engineering approaches. Gururangan et al. [3] later demonstrated that further pretraining on domain-specific corpora can improve an already pretrained model. These developments established the methodological basis for financial language models that begin with a general pretrained model, adapt it to financial text where appropriate, and fine-tune it on labelled task data.

2.3 Transformer-Based Financial Language Models

2.3.1 Encoder-Based Domain Adaptation

Domain-adapted encoder models based on Bidirectional Encoder Representations from Transformers (BERT) substantially improved benchmark performance in financial NLP. FinBERT [4, 5], which continues BERT-base pretraining on financial corpora, achieves approximately 88% macro F1 on the FPB 75%-agreement subset and is widely used as a domain-specific baseline for financial sentiment classification. Its performance relative to general-purpose BERT is consistent with the broader finding of Gururangan et al. [3] that continued pretraining on domain-relevant text improves downstream task performance. FLANG [7], developed alongside the Financial Language Understanding Evaluation (FLUE) benchmark, extends this evidence beyond BERT by reporting gains from an ELECTRA-based domain-adapted encoder on financial NLP tasks including sequence labelling and multi-class classification.

A limitation of domain-adapted encoder models for cross-source divergence measurement is the form of their output. These models generally produce class labels or probability distributions over positive, neutral, and negative sentiment, whereas divergence measurement requires continuous signed scores on a common scale. A signed score can be derived from the posterior distribution as $s_i = p_{\text{positive}} - p_{\text{negative}}$, but its interpretation across source types depends on calibration. Scores derived from different source domains are comparable only if the model's probability outputs are comparably calibrated across those domains. Since modern neural networks may be systematically overconfident [19], this requirement cannot be assumed without evaluation. Chapter 3 therefore specifies a calibration protocol for cross-source comparison.

2.3.2 Large-Scale Pre-trained Models

BloombergGPT [8] demonstrates the potential of large-scale financial-domain pretraining. The model contains 50 billion parameters and was trained on 708 billion tokens drawn from financial and general-purpose text, achieving state-of-the-art performance on several financial NLP tasks at the time of publication. Its methodological value for reproducible experimentation is nevertheless constrained by its proprietary status: the model is not available for independent evaluation, reproduction, or task-specific fine-tuning.

This constraint highlights a practical distinction between reported large-scale performance and reproducible model development. Researchers seeking locally deployable financial language models require an alternative that can be inspected, adapted, and evaluated without dependence on proprietary infrastructure or API access. Parameter-efficient fine-tuning of open-weight models provides such an alternative and motivates the approach discussed in Section 2.4.4 and the model selection presented in Chapter 3.

2.4 Large Language Models for Financial Reasoning

2.4.1 Few-Shot and Zero-Shot Prompting

Brown et al. [20] showed that GPT-3 could perform diverse language tasks in few-shot settings without task-specific parameter updates, establishing in-context prompting as an alternative to fine-tuning for general NLP applications. Evidence of transfer to financial sentiment classification is provided by Zhang et al. [9], who report a zero-shot macro F1 of 0.87090 for GPT-4o mini on FPB. This result indicates that a general-purpose instruction-following model can produce competitive financial sentiment predictions without domain-specific fine-tuning, subject to the reported evaluation protocol. Wei et al. [21] demonstrate that prompts incorporating intermediate reasoning steps can improve performance on reasoning-intensive tasks. Although their study does not evaluate financial sentiment analysis, it provides a basis for considering structured prompting in financial text extraction tasks that require semantic interpretation beyond a single label.

A reproducible disagreement-measurement pipeline requires greater control over inference than proprietary API-based models typically provide. Their outputs may vary following model updates, posterior probabilities may be unavailable, and repeated processing of large corpora introduces additional cost. The scoring engine intended for Sentinel is therefore locally deployable and open-weight. Prompt-based methods remain relevant for the separate task of extracting aspect-level sentiment from financial text.

2.4.2 Open-Source Financial LLMs

Financial GPT (FinGPT) [22] develops an open-source financial LLM framework in which data collection and adaptation are central design components. Although the framework draws on multiple types of financial text, its objective is to provide broader training material for a financial language model rather than to compare how different sources describe the same event. This distinction is central to our thesis. A model may be trained on news, social media, and financial documents without producing an explicit measure of disagreement among those sources. Sentinel, by contrast, retains source identity at inference time and compares sentiment scores for texts aligned to a shared corporate event.

The LLaMA family [23] provides the open-weight model basis required for this type of reproducible experimentation. In particular, LLaMA 3.1 8B Instruct offers a model that can be adapted and evaluated locally at a scale compatible with the parameter-efficient fine-tuning methodology used in our thesis.

2.4.3 Multi-Agent Financial Reasoning Systems

Multi-agent approaches in financial sentiment analysis use several reasoning roles to improve the classification of a given text. In HAD, Xing [10] assigns specialised roles to LLM agents and aggregates their discussion into a final sentiment assessment. The reported results show benefits from heterogeneous agent reasoning, but the framework remains document-centred: all agents evaluate the same input text. This differs from the problem addressed in our thesis, where disagreement arises among separate source documents describing a common corporate event.

Retrieval-Augmented Multi-Agent Sentiment System (RAMAS) [24] follows a related direction by combining retrieval augmentation with generator, discriminator, and arbitrator agents. Its reported gains concern classification on established FSA benchmarks, using retrieval and agent interaction to improve the assessment of each input document. The scope of both systems remains document-level sentiment assessment rather than comparison of event-aligned sentiment signals across institutional source types. Sentinel addresses this latter problem by preserving source identity throughout the analysis.

2.4.4 Parameter-Efficient Fine-Tuning

Adapting multi-billion-parameter LLMs through full fine-tuning is computationally demanding because training must retain gradients and optimiser states for the entire parameter set. For models of the scale used in our thesis, this requirement is difficult to satisfy on single-Graphics Processing Unit (GPU) research hardware. Parameter-efficient fine-tuning provides a practical alternative by leaving the pretrained model weights fixed and learning a comparatively small task-specific update.

LoRA [18] represents this update through a low-rank decomposition. Given a frozen weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the adapted matrix is

$$W' = W_0 + \frac{\alpha}{r} BA,$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable matrices and $r \ll \min(d, k)$. This formulation reduces the number of parameters updated during training and permits the learned update to be merged into the base model for inference. Hu et al. [18] report that LoRA achieves perfor-

mance comparable to full fine-tuning across several evaluated language tasks without additional inference latency.

QLoRA [17] reduces memory requirements further by combining low-rank adaptation with a frozen base model quantised to 4-bit NormalFloat 4-bit (NF4) precision. The method also uses double quantisation and paged optimisers to reduce memory consumption and manage training-time memory spikes. Dettmers et al. [17] demonstrate that QLoRA can fine-tune a 65-billion-parameter model on a single 48 Gigabyte (GB) GPU while preserving full 16-bit fine-tuning task performance in their evaluation. These properties motivate its use for adapting LLaMA 3.1 8B Instruct to financial sentiment classification in Chapter 3.

2.5 Source-Type Characteristics in Financial NLP

2.5.1 Financial News

Financial news differs from regulatory filings and social media not only in style, but also in the conditions under which it is produced. Regulatory disclosures are governed by formal reporting obligations, whereas social media commentary is largely unconstrained by editorial review. Professional news is prepared within editorial and attribution standards, while allowing journalists to determine which consequences of an event warrant emphasis. This combination makes financial news an appropriate comparison source in a study of how institutional channels characterise the same corporate event.

A substantial portion of the FSA literature is built on financial-news data. FPB [6] contains 4,846 sentences annotated by 16 domain experts and is available at the all-agree, 75%, 66%, and 50% agreement thresholds, allowing performance to be examined under differing levels of annotation agreement. SEntFiN [12] moves from sentence-level to entity-level annotation in financial headlines, so that sentiment can be associated with the relevant firm rather than only with the headline as a whole. Outside the 36-study PRISMA-screened corpus, Financial News Stock Price Integration Dataset (FNSPID) [25] is included as contextual evidence of the scale of available news data, with 15.7 million time-aligned financial news records for studies combining textual and market information.

Aspect-oriented resources also inform the present framework. The FiQA benchmark [11] includes sentiment scores and categories such as Corporate/Sales, Stock/Price Action, and Market/Volatility for financial headlines and microblog posts. These categories inform the construction of the aspect taxonomy in Sentinel's MASD framework. However, the available datasets do not align financial news with regulatory and social-media texts concerning the same event. They can therefore support sentiment modelling and aspect definition, but not the cross-source aspect-level comparison examined in our thesis. Section 2.7 develops this point further.

2.5.2 Social Media and Retail Investor Sentiment

Social media differs from professional news and regulatory disclosure in both its production setting and the distribution of its sentiment signals. Unlike edited news reports or legally constrained filings, social-media posts reflect decentralised commentary from users whose information sets, interpretations, and communication practices may vary considerably. Cookson et al. [16] show that attention is correlated across major social-media platforms, while sentiment is much less closely aligned. This result cautions against representing social-media sentiment as a single platform-independent signal.

Related evidence indicates that disagreement within social-media discussion may be economically relevant. Using data from StockTwits, an investor-focused social-media platform, Cookson and Niessner [26] distinguish disagreement arising from different information sets from disagreement arising from different interpretations, finding that both contribute to overall disagreement. Siganos et al. [27] report that sentiment divergence measured from Facebook data is positively associated with trading volume and stock-price volatility. Although these studies examine disagreement within social-media environments rather than across institutional source types, they motivate the decision to retain sentiment variation rather than remove it through aggregation.

Reddit financial communities such as r/wallstreetbets and r/investing provide a high-volume channel for retail-investor commentary. Their distinct community identities and discussion norms make subreddit identity relevant when variation in the collected Reddit evidence is interpreted. Reddit serves as the social-media source in the present case-study implementation, but our thesis does not assume that it fully represents retail sentiment or that its coverage of financial aspects is uniform. Instead, Reddit documents are analysed as one available source channel, with aspect coverage recorded explicitly in the ADI evaluation. This allows sparse coverage in the collected corpus to be treated as an empirical limitation of the case study rather than as a general claim about social-media communication.

2.5.3 Regulatory Filings

Regulatory filings provide a legally mandated source of company-disclosed financial information. Through EDGAR, the SEC makes available annual reports on Form 10-K, quarterly reports on Form 10-Q, and current reports on Form 8-K. These documents differ from professional news and social-media commentary because they are produced by the reporting company within a regulated disclosure process and are structured around formal reporting requirements.

The linguistic specificity of regulatory disclosure is established by Loughran and McDonald [1], who show that general-purpose sentiment dictionaries frequently misclassify routine financial terms, including “tax,” “liability,” and “risk,” when applied to 10-K filings. This domain mismatch is relevant to the interpretation of Sentinel’s outputs. The near-neutral score obtained

for the SEC filing in the present case study may reflect the language of formal disclosure rather than a failure of the sentiment model. The result is therefore interpreted within the case-study setting only; whether the same pattern appears across a wider filing corpus remains to be tested.

2.5.4 Earnings Calls and Analyst Reports

Earnings calls constitute a distinct form of corporate financial communication. In a typical call, management presents quarterly results through prepared remarks and subsequently responds to analyst questions. The prepared portion commonly addresses financial dimensions such as revenue, margins, operating expenses, and forward guidance, while the question-and-answer portion permits analysts to focus on issues requiring further explanation or scrutiny.

This organisation is relevant to aspect-level sentiment analysis because the same financial dimension may be framed in prepared disclosure and examined differently during analyst questioning. Regulation FD, adopted by the SEC in 2000, limits selective disclosure of material nonpublic information to certain market participants unless the information is also publicly disclosed in accordance with the rule. Within the FSA literature, Du et al. [14] identify earnings calls as one of the principal financial text sources studied for sentiment analysis. Earnings-call transcripts thus provide a source type that is institutionally distinct from both regulatory filings and external news reporting.

Analyst reports are similarly distinct from the sources used in the present case study. They communicate professional investment analysis through narrative evaluation, recommendations, and price targets, but they are not issuer-filed disclosures. Although earnings calls and analyst reports are not analysed in the implemented Sentinel case studies, they indicate how the framework could later be extended to additional professionally mediated sources.

2.6 Multi-Source Sentiment Fusion

Research on sentiment fusion in finance has primarily examined whether combining textual signals improves prediction, rather than whether disagreement among sources is itself informative. Passalis et al. [28] combine sentiment information from multiple online sources in a deep-learning framework for Bitcoin price-change prediction. Their work establishes the feasibility of incorporating heterogeneous financial text signals into a predictive model, but its output is a market-prediction signal rather than a measure of disagreement among aligned sources.

Liu et al. [29] approach integration at a different level. Their Chinese bond-market framework combines firm-specific and industry-specific sentiment with temporal smoothing to produce a daily composite sentiment index for credit-spread forecasting. The study therefore cap-

tures variation across analytical levels and time, but does not compare source-specific interpretations of the same event or decompose such differences by financial aspect.

A related aggregation problem is examined by Wang and Ma [30]. MANA-Net assigns dynamic weights to news sentiment representations for market prediction in order to reduce the information loss associated with equal-weight aggregation of multiple news items. Although this approach preserves differential relevance during aggregation, it operates within financial news rather than across distinct institutional source types.

Taken together, these studies show how multiple sentiment inputs can be combined to support financial prediction. Their focus is the construction of an improved aggregate signal, rather than the analysis of differences among sources. Sentinel addresses the latter problem by aligning source documents to a common event and measuring the resulting source-specific divergence.

2.7 Aspect-Based Sentiment Analysis in Finance

The financial sentiment literature includes several approaches that move beyond document-level polarity, although these approaches are generally evaluated within a single corpus or source setting. FiGAS [31] assigns fine-grained sentiment scores in the range $[-1, +1]$ to economic and financial topics of interest using an unsupervised lexicon-based method. This provides an interpretable form of topic-specific sentiment measurement, but it does not compare how separate institutional sources describe the same event.

The FiQA benchmark [11] offers aspect-oriented annotations for financial headlines and microblog posts, including Corporate/Sales, Stock/Price Action, and Market/Volatility. These categories provide a useful basis for the aspect taxonomy used in Sentinel's MASD framework. Nevertheless, the benchmark was not designed to align regulatory, professional-news, and social-media texts to shared corporate events, which is necessary for measuring disagreement among source-specific aspect scores.

Yan and Qin [32] extend this line of work through a FinBERT-BiGRU-CNN model applied to Chinese financial media commentary, reporting an F1 score of 0.873. Their study confirms the feasibility of neural aspect-level sentiment analysis in a financial setting, while also noting limitations in the availability of authoritative datasets for financial Aspect-Based Sentiment Analysis (ABSA). Its scope remains within financial media commentary.

FinEntity [13] addresses a neighbouring problem by annotating entities and their sentiment in financial news. This level of annotation supports sentiment attribution to a named firm where multiple entities appear in the same text, but it is not equivalent to aspect-level annotation and does not enable cross-source aspect comparison.

The remaining gap is therefore not the absence of fine-grained financial sentiment resources, but the absence within the reviewed studies of an event-aligned method for comparing aspect-

level sentiment across source types. In MASD, aspect scores are obtained separately for each source document, and the ADI is then used to measure divergence for each aspect.

2.8 Opinion Divergence and Investor Disagreement

2.8.1 Finance Theory of Disagreement

The economic relevance of cross-source divergence is grounded in finance theory rather than in NLP alone. Miller [33] showed that, under short-sale constraints, disagreement among investors can lead to systematic overvaluation because pessimistic views are less easily reflected in market prices. Hong and Stein [34] developed this account further by identifying heterogeneous priors and limited attention as sources of disagreement and by showing how gradual information diffusion can produce momentum over short horizons and reversals over longer horizons. This mechanism is pertinent to multi-source sentiment analysis: different source types may reach different investor audiences at different times, meaning that changes in cross-source sentiment divergence may reflect the sequence through which an event is interpreted across market participants.

Diether et al. [35] provided an empirical operationalisation of disagreement through analyst earnings forecast dispersion, demonstrating that this measure predicts subsequent stock returns independently of the level of analyst consensus. Their approach establishes that dispersion in structured financial signals can contain information beyond average expectations. Our present thesis examines an analogous possibility in textual data: narrative sentiment dispersion across institutional source types may contain related predictive information. The SDI is designed as a computable measure of this form of narrative disagreement.

2.8.2 Computational Measurement of Sentiment Divergence

Garcia [15] provides relevant computational motivation for the SDI. In that study, sentiment divergence is measured as the absolute difference between quarterly Twitter and Bloomberg news sentiment scores across 1,823 U.S. firms from 2015 to 2021. Garcia reports that divergence predicts financial distress, with an effect of approximately 46 basis points per standard deviation of divergence, even after accounting for the sentiment level in either source. This finding motivates the treatment of disagreement across sources as a signal in its own right.

The Sentinel framework extends this approach in four respects. Garcia's measure compares exactly two sources, whereas the present framework is designed for multiple source types across the accountability spectrum. It also uses absolute difference rather than a weighted pairwise distance measure, and it does not incorporate either source credibility weighting or model uncertainty quantification. As a result, calibration artefacts cannot be separated from genuine narrative disagreement within that measure. Finally, Garcia assumes sentiment scores

as inputs, while Sentinel specifies an NLP pipeline for producing and comparing source-level scores from raw text. The SDI therefore develops the same underlying idea of divergence as the variable of interest in a broader computational setting.

FinSentLLM [9] is related to this discussion but examines a different source of divergence. Its meta-classifier combines FinBERT and RoBERTa-sentiment posteriors generated from the same text, so the resulting disagreement occurs between models rather than between information sources. This distinction matters for our thesis. Model-level disagreement indicates that classifiers interpret the same input differently, which is relevant to annotation ambiguity and model calibration. Source-level disagreement indicates that institutional channels describe a shared event differently, which is relevant to framing and information ecology. The SDI measures the second form of disagreement.

2.9 Source Credibility, Uncertainty, and Calibration

Compared with the large body of single-source FSA research, relatively little work has examined how credibility should affect the weighting of financial text sources. Huang et al. [36] study this problem for financial users on Twitter by deriving user-level trust scores from account history, follower-network properties, and content consistency. They find that trust-weighted sentiment improves downstream stock prediction accuracy relative to unweighted sentiment. Their study shows that credibility can be informative, although its setting is confined to one platform and does not address credibility across different source types.

Kengmegni [37] provides a further reason to treat credibility weights cautiously. In the tested setting, sources commonly viewed as less credible in financial discourse, including Benzinga and Zacks, sometimes yield more stable and consistent sentiment signals than traditional newspaper-of-record sources. Institutional reputation alone is therefore not a sufficient basis for fixing credibility priors. For Sentinel, this finding motivates evaluating the SDI under alternative credibility weight configurations, so that any observed divergence can be assessed for sensitivity to the chosen priors.

A related issue is uncertainty in the sentiment model itself. MC Dropout [38] estimates model uncertainty by retaining dropout during repeated inference passes and examining the resulting distribution of predictions. This procedure provides an uncertainty estimate without changing the model architecture or retraining it. Guo et al. [19] show that modern deep neural networks can be systematically overconfident, especially among predictions assigned high confidence. This matters in a cross-source setting because the same model may be differently calibrated on regulatory filings and social-media text. If predictions for SEC filings are more overconfident than predictions for social media, some of the observed difference in $s_i = p_{\text{positive}} - p_{\text{negative}}$ may arise from calibration rather than source-level disagreement. For this reason, the full Sentinel design includes Expected Calibration Error (ECE) measurement

and probability calibration as requirements for reliable cross-source comparison; their empirical application remains outside the current case-study implementation.

2.10 Explainability in Financial NLP

Financial NLP studies have applied several Explainable Artificial Intelligence (XAI) methods to explain sentiment predictions, but these methods remain focused mainly on individual model outputs rather than on the needs of different financial users. SHapley Additive Explanations (SHAP)-based approaches have been used in sentiment lexicon construction to identify token contributions to classification decisions and to produce explainable lexicon extensions that outperform the Loughran-McDonald dictionary on held-out text. Local Interpretable Model-Agnostic Explanations (LIME) and Anchors have also been used in entity-level financial sentiment architectures to generate faithful local explanations. Such methods can clarify why a model assigns a particular sentiment label to a specific text, but they do not show why two source types characterise the same event differently. That distinction is central to cross-source sentiment analysis.

The Chartered Financial Analyst Institute (CFA) describes a related difficulty in financial Artificial Intelligence (AI) explainability: the form of explanation required by a compliance officer, portfolio manager, regulator, or quantitative analyst may differ substantially. Attribution maps developed for machine-learning practitioners may therefore require interpretation before they become useful in financial decision or oversight contexts. The European Union Artificial Intelligence Act (EU AI Act) and Federal Reserve Supervisory Guidance on Model Risk Management (SR 11-7) also point towards greater transparency requirements for AI systems that influence investment or credit decisions, although neither specifies how explanations should be presented for financial sentiment systems.

Sentinel approaches this problem through outputs that arise directly from its divergence measures. The per-source-pair contribution breakdown shows how each pairwise distance contributes to the aggregate SDI. The directional divergence vector indicates whether sources differ in the direction of sentiment rather than only in its magnitude. The per-aspect ADI matrix identifies which financial dimensions contribute to the observed disagreement. Together, these outputs provide an interpretable account of cross-source divergence for financial audiences and can also be represented in machine-readable form for downstream risk monitoring or regulatory reporting.

2.11 Research Gap and Thesis Positioning

Four gaps in the surveyed literature motivate the design of Sentinel. The first concerns the data required for cross-source evaluation. The major benchmarks considered in our thesis [6, 11, 12] are single-source datasets and do not align multiple textual sources to the same corporate event for sentiment annotation. They therefore cannot be used for retrospective evaluation of cross-source disagreement. The second concerns calibration. Although modern neural networks are known to exhibit systematic overconfidence [19], the reviewed frameworks do not make probability calibration a precondition for comparing sentiment posteriors across source domains. The third concerns the treatment of credibility and uncertainty: existing work studies source credibility weighting and model uncertainty quantification separately rather than combining them within a divergence metric. The fourth concerns interpretability at the aspect level. In the reviewed multi-source sentiment frameworks, document-level aggregation does not identify the financial dimensions that contribute to cross-source disagreement.

2.11.1 Design Requirements Derived from the Literature

The identified gaps lead to six methodological requirements for our thesis. Textual sources must first be aligned to the same corporate event, since comparing unrelated corpora would not measure disagreement about a shared information setting. Sentiment outputs must then be expressed on a continuous signed scale so that differences between sources can be measured as distances rather than only as categorical mismatches. The framework must also retain source identity throughout the pipeline, because its purpose is to examine source-level disagreement rather than produce an aggregate sentiment estimate. Credibility priors and calibration metadata must remain explicit, including where calibration has not yet been empirically applied. In addition, a document-level disagreement score must be decomposable into pairwise source-type contributions and aspect-level components. Finally, the output must be auditable, allowing the resulting score to be traced to the source types, aspects, and signed sentiment values that produced it.

Chapter 3 implements these requirements through event-window alignment, signed sentiment scoring, source-type priors, the SDI, the MASD/ADI decomposition, and a structured output record. Our thesis does not claim that every requirement has been completed at deployment level. Calibration gates are specified within the methodology and retained as output metadata; their empirical execution is reserved for future multi-event validation. MC Dropout uncertainty estimation (Level 3) has been implemented and is reported in Section 4.4.3.

Taken together, the surveyed studies address parts of this problem, but not its full combination of requirements. Multi-source sentiment research does not decompose disagreement at the aspect level, while aspect-level sentiment research does not compare institutional source types. Research on divergent signals generally measures disagreement between models rather than be-

tween sources, and work using LLM-based extraction typically considers individual documents without cross-source comparison. Garcia [15] measures source-level divergence but does not include NLP infrastructure or aspect decomposition. FinSentLLM [9] measures model-level divergence with LLM-derived features but is not multi-source in the institutional sense. Yan and Qin [32] perform aspect-level analysis within a single source type. Table 2.2 positions ten representative works against the four dimensions required for cross-source aspect-level sentiment disagreement analysis.

Table 2.2: Positioning of representative works against four dimensions of cross-source aspect-level sentiment disagreement analysis.

Work	Multi-Source Analysis	Aspect-Level Analysis	LLM-Based Scoring	Cross-Source Divergence
Garcia [15]	Yes, two sources	No	No	Absolute difference
FinSentLLM [9]	No	No	Ensemble	Model-level
HAD [10]	No	Limited	Yes	No
RAMAS [24]	No	No	Yes	No
FinGPT [22]	Training data only	No	Yes	No
Passalis et al. [28]	Yes	No	Hybrid / partial	No
Liu et al. [29]	Yes	No	Hybrid / partial	No
FiGAS [31]	No	Yes	No	No
Yan & Qin [32]	No	Yes	Hybrid / partial	No
Sentinel (our thesis)	Yes	Yes	Specified*	Source-level

Note: "Training data only" denotes studies in which multiple source types are used for model development but are not compared as aligned inputs for the same event. "Hybrid / partial" denotes studies that include an LLM or neural language-model component without employing it as the complete mechanism for scoring cross-source divergence. "Specified*" means that LLM-based scoring is specified at the framework level and evaluated through Sentinel-LLaMA3 on FPB, whereas the reported case-study SDI/ADI values are computed using FinBERT.

Among the 36 sources reviewed, no prior published work was identified that addresses all four dimensions together. This observation defines the gap addressed by our thesis: a formal, computable measure of source-level sentiment disagreement based on continuous LLM-derived scores, decomposed to the aspect level, implemented with a reproducible open-weight scoring engine, and demonstrated on real multi-source financial events.

2.11.2 Thesis Proposition

Our thesis treats cross-source financial sentiment disagreement as a potentially meaningful signal in its own right, rather than as variation that should automatically be removed by aggregation. Previous studies provide the foundations for multi-source analysis, aspect-level sentiment analysis, and credibility-aware weighting, but these elements remain largely separate in the literature. Sentinel combines them to examine whether regulatory filings, professional news, and social media characterise the same financial event differently. The central proposition is that a credibility-weighted disagreement metric, specified with uncertainty handling and aspect-level decomposition, can make such source-level differences observable where aggregate FSA alone cannot.

2.11.3 Thesis Objective

The objective of our thesis is to develop and evaluate Sentinel as an end-to-end framework for MSFSA. Sentinel is designed to generate continuous source-specific sentiment scores for aligned financial events, measure document-level cross-source disagreement using the SDI, and decompose that disagreement through the MASD framework and ADI. The framework is then demonstrated on real financial data. In this way, our thesis brings together four dimensions that remain separately addressed in the literature: multi-source analysis, aspect-level sentiment, LLM-based extraction with a fine-tuned open-weight scorer, and cross-source divergence measurement.

Chapter 3

Methodology

This chapter describes the Sentinel framework and the methodology used in our thesis. The framework treats cross-source variation in financial sentiment as an object of analysis rather than as variation to be removed by averaging. It therefore follows the full analytical path from event definition and data collection through preprocessing, sentiment inference, disagreement measurement, and aspect-level decomposition. The chapter also describes the QLoRA fine-tuning procedure used for the Sentinel-LLaMA3 scoring engine. Where the broader framework extends beyond the present evaluation, we distinguish clearly between implemented components and components reserved for future work.

3.1 Framework Overview and Design Principles

Sentinel is structured as a five-layer framework that preserves source identity throughout the analysis. Layer 1 assembles the multi-source text corpus from documents located at different points on the institutional accountability spectrum. Layer 2 defines the calibration and quality-gating procedures required for cross-source comparison, and Layer 3 converts the processed documents into signed sentiment scores. At Layer 4, the framework separates into two analytical branches: the SDI measures document-level disagreement across the event corpus, while the MASD/ADI branch identifies the aspects on which sources diverge. Layer 5 consolidates these results into the explainability output and final structured record. In the main Tesla and AppLovin Level 1–2 analyses, the case-study scores are uncalibrated FinBERT outputs combined with static type-prior weights. The Tesla Level 3 analysis, reported separately in Chapter 4, additionally evaluates MC Dropout-based uncertainty adjustment.

Figure 3.1 presents the system architecture, tracing the information path from three heterogeneous source types through a common preprocessing and inference layer to the parallel SDI and ADI computation branches and the explainability output. Table 3.1 summarises the input, output, and key component of each of the five operational phases that traverse this architecture.

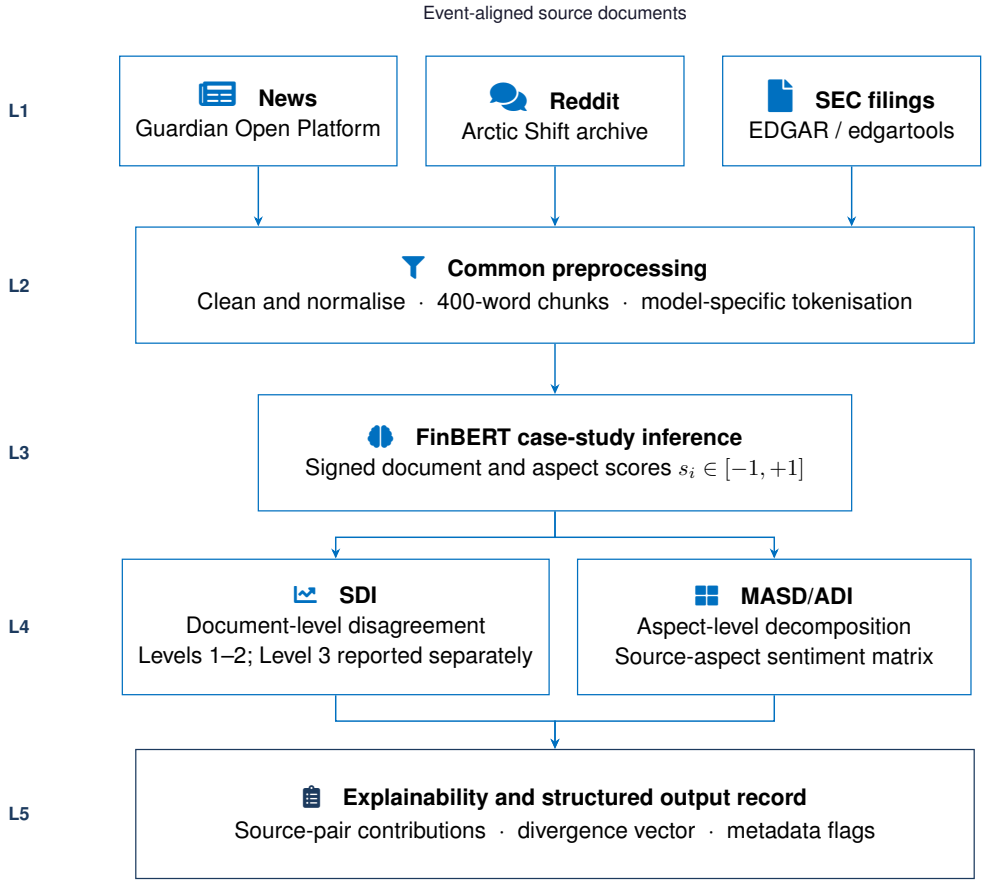

Figure 3.1: Sentinel case-study architecture. The three source types are processed through a shared preprocessing and FinBERT inference path, after which the framework computes document-level SDI and aspect-level ADI in parallel. Level 3 MC Dropout uncertainty estimation is reported separately in Section 4.4.3. Sentinel-LLaMA3-FPB75 is evaluated separately on Financial PhraseBank and does not generate the reported Tesla SDI/ADI values.

Table 3.1: Sentinel five-phase pipeline: input, output, and key component per phase.

Phase	Name	Input	Output	Key Component
1	Baseline Validation	FPB subsets	Scoring-engine benchmark metrics	FinBERT or Sentinel-LLaMA3 inference
2	Multi-Source Collection	News, Reddit, SEC filings	Event-aligned document corpus	Source APIs and chunk scoring
3	SDI Computation	Per-document signed scores	Cross-source disagreement scalar	Credibility-weighted SDI
4	MASD/ADI	Per-source aspect scores	Divergence vector	Aspect Divergence Index
5	Output and Reporting	SDI, ADI, benchmark metrics	Structured output record	Explainability layer

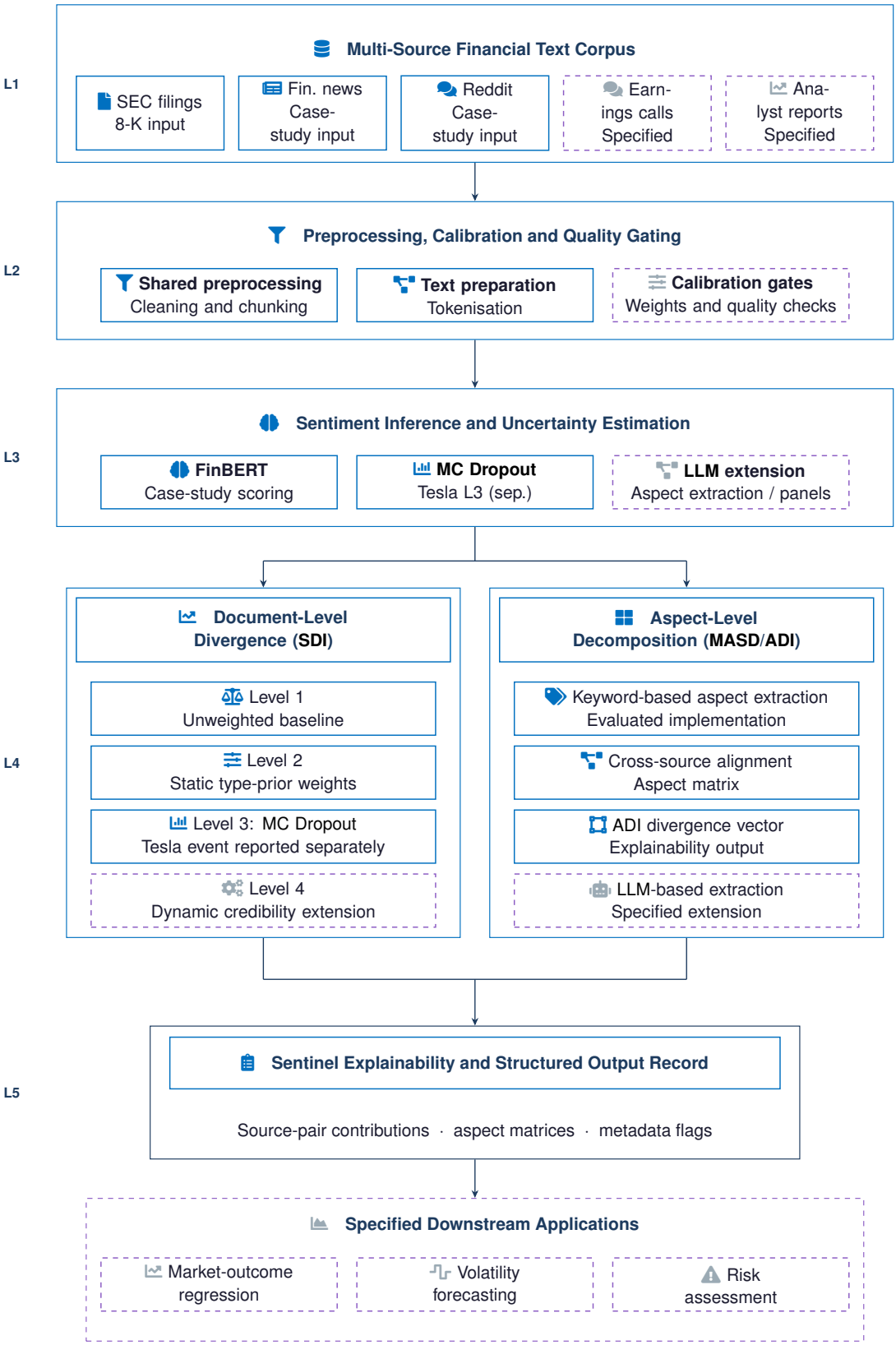

Figure 3.2: Conceptual full Sentinel architecture. Solid nodes identify evaluated components: news, Reddit, and SEC inputs; FinBERT case-study scoring; Level 1–2 SDI; keyword-based MASD/ADI; and Level 3 MC Dropout, which is reported separately in Section 4.4.3. Dashed nodes identify specified extensions: additional source types, calibration gates, LLM-based extraction, and downstream applications.

3.1.1 Implemented Components and Specified Extensions

The framework is intentionally broader than the version evaluated in this thesis. The implemented subset covers the three-source ingestion path for professional news, Reddit social media, and SEC filings; FinBERT-based document scoring for the Tesla Q1 2024 and AppLovin Q4/FY2025 case studies; Level 1 and Level 2 SDI; keyword-based MASD/ADI; source-pair contribution reporting; Level 3 MC Dropout uncertainty adjustment reported in Section 4.4.3; Sentinel-LLaMA3 benchmark evaluation on FPB and scoring engine comparison on the Tesla Q1 2024 corpus (Section 4.4.5); and sentence-ID LLM aspect extraction implemented as an extractive MASD sensitivity audit for Tesla Q1 2024. Other components are specified as part of the framework but are not executed in the present evaluation, including source-type calibration gates, dynamic credibility weighting, earnings-call and analyst-report ingestion, production-scale sentence-ID or span-grounded LLM aspect extraction across events, and downstream SDI-to-market-outcome regression.

This separation matters for the interpretation of the results. The Tesla and AppLovin SDI/ADI values are proof-of-concept disagreement measurements produced with uncalibrated FinBERT scores and static credibility priors. The Sentinel-LLaMA3 results establish the feasibility of an open-weight fine-tuned scoring engine on FPB; a separate scoring engine comparison using token log-probability extraction is reported in Section 4.4.5 and does not replace FinBERT as the primary case-study scorer. Components that are specified mathematically or architecturally are therefore treated as future implementation requirements unless they are explicitly reported in Chapter 4.

3.2 Ethical Considerations

The Sentinel framework raises several ethical considerations. We address them at the design level and state them explicitly here to clarify the intended research scope.

Data source selection and privacy. The case studies use only publicly accessible regulatory documents and public-facing platform text. SEC EDGAR filings are statutory public disclosures, Guardian articles are retrieved through the publisher's Open Platform API for non-commercial research access, and Reddit posts are obtained from public finance-oriented communities through the Arctic Shift archive. The pipeline does not collect, store, or analyse private messages, user profile data, or non-public account information, and it does not construct user-level profiles. The document is the unit of analysis.

Potential for misuse. A sentiment disagreement signal produced by Sentinel could in principle be misused in ways that conflict with market integrity requirements. For example, coordinated artificial social media content could inflate the SDI score for a target company and create the appearance of stronger cross-source conflict. The framework does not include adversarial-robustness controls at the data-collection layer, and the evasion-scoring component identified as

a Level 4 extension in Section 3.8.6 is only a partial structural mitigation. Users who integrate Sentinel outputs into operational decision systems are therefore responsible for upstream data quality checks and source-validation procedures.

Fairness and representational bias. The MASD keyword lists (Appendix A.1) and source-type credibility priors (Section 3.3.3) embed design choices with distributional implications. The English-language, financially conventional keyword lists may under-detect aspect mentions in non-standard registers or other languages. The lower prior assigned to social media reflects documented sentiment variability [37], but may also reduce signal from communities that rely primarily on social channels. These limitations should be revisited in multilingual or demographically broader applications.

Intended use boundaries. Sentinel is a research instrument for measuring and decomposing cross-source sentiment disagreement. It produces disagreement scores and aspect-level divergence vectors, not investment recommendations, credit decisions, or regulatory determinations. Its outputs should not be used in consequential decision contexts without further validation, human oversight, and documented governance controls consistent with the EU AI Act and SR 11-7 frameworks discussed in Chapter 2.

3.3 Event Definition and Source Selection

3.3.1 Event Window

A corporate event e is defined as a company-specific occurrence that generates commentary across multiple institutional source types. The event is bounded by a temporal collection window:

$$W(e) = [t_0 - \delta_{\text{pre}},\, t_0 + \delta_{\text{post}}], \tag{3.1}$$

where t_0 is the event trigger timestamp, and δ_{pre} and δ_{post} define the event-type-specific look-back and look-ahead periods. The recommended default windows are reported in Table 3.2. Only documents published within $W(e)$ are included in the SDI computation for event e. This temporal alignment ensures that cross-source comparisons refer to a shared corporate event rather than to different underlying information states.

Table 3.2: Default event-window parameters by event type

Event type	δ_{pre}	δ_{post}
Earnings release	24 hours	48 hours
Regulatory filing event	0 hours	24 hours
Major news event	6 hours	24 hours

3.3.2 Source Type Taxonomy and Accountability Spectrum

The framework recognises five source types along an institutional accountability spectrum in the financial information environment. Each source type differs in editorial mandate, regulatory constraint, audience, and temporal cadence. These characteristics shape expected sentiment expression patterns and guide the credibility assignment.

Table 3.3: Source-type taxonomy and accountability characteristics

Source type	Position	Defining characteristics
Regulatory filings (SEC Forms 10-K, 10-Q, and 8-K)	Highest	Legally reviewed disclosures subject to statutory liability under the Securities Exchange Act.
Earnings call transcripts	High	Legally reviewed prepared remarks with more spontaneous analyst question-and-answer segments.
Analyst reports	Professional	Written by credentialled professionals constrained by reputation and regulatory obligations.
Professional news	Intermediate	Governed by commercial editorial standards and an intermediate temporal cadence.
Social media (Reddit, X, and StockTwits)	Low	No editorial filter and high variation in content quality and intent.

3.3.3 Credibility Type-Prior Weights

Each source type is assigned a static type-prior credibility weight π_τ, interpreted as the credibility ceiling for that source type under perfect model confidence. Table 3.4 reports the recommended weights. These priors encode relative institutional accountability and are treated as hyperparameters, with sensitivity analysis discussed in Section 3.8. The values $\{1.0, 0.9, 0.8, 0.7, 0.4\}$ are author-defined operationalisation choices, not parameters estimated from data. The ordering is grounded in source-credibility literature, including trust-based weighting approaches [36] and evidence that unconventional sources exhibit higher sentiment variability than institutional channels [37]. Data-driven estimation of π_τ from held-out event corpora is reserved for the Level 4 dynamic credibility extension in Section 3.8.6.

Table 3.4: Recommended credibility type-prior weights π_τ by source type.

Source Type	π_τ	Rationale
SEC filing (8-K, 10-K/Q)	**1.0**	Legally mandated disclosure; highest accountability.
Earnings call transcript	0.9	Legal review with partially spontaneous Q&A.
Analyst report	0.8	Professional standards and institutional reputation.
Wire / major news	0.7	Editorial standards; intermediate accountability.
Social media (Reddit, X)	0.4	No editorial filter; high content variability.

Table 3.5: SDI Level 2 sensitivity to credibility-weight perturbations for the Tesla Q1 2024 earnings event ($n = 11$). $\pi_{\text{SEC}} = 1.0$ in all scenarios.

Scenario	π_{news}	π_{Reddit}	$\text{SDI}_{\text{L2}}^{\text{norm}}$	Zone	Dominant pair
Base (current)	0.7	0.4	0.35172	Partial Disagreement	News vs. SEC, 0.42000
Social-upweighted	0.7	0.6	0.35209	Partial Disagreement	News vs. SEC, 0.42000
News-downweighted	0.5	0.4	0.34764	Partial Disagreement	News vs. SEC, 0.42000
Equalised	1.0	1.0	0.35628	Partial Disagreement	News vs. SEC, 0.42000

Table 3.5 reports SDI Level 2 under four credibility weight configurations. Across these scenarios, the normalised SDI ranges from 0.34764 to 0.35628, yielding a total spread of 0.00864, or less than one percentage point on the $[0, 1]$ scale. The Partial Disagreement zone classification and the dominant source pair (News vs. SEC 8-K, pairwise distance $= 0.42000$) remain unchanged. This stability occurs because the principal disagreement lies between professional news and the regulatory filing, both of which carry high credibility priors. Consequently, varying the Reddit weight has little effect on the aggregate disagreement score. The Tesla Q1 2024 interpretation is therefore robust to the tested credibility perturbations.

3.3.4 Current Instantiation and Secondary Case Study

Our thesis uses three of the five source types: professional news via the Guardian Open Platform, social media via the Arctic Shift Reddit archive, and regulatory filings via SEC EDGAR. Earnings call transcripts and analyst reports are supported by the framework architecture but excluded from the current evaluation due to the lack of pre-annotated neural sentiment corpora and the need for commercial data subscriptions (a limitation acknowledged in Section 3.14).

The primary case study is the Tesla Q1 2024 earnings event, analysed in Chapter 4. The AppLovin Q4/FY2025 event serves as a secondary case study to assess whether the unmodified pipeline transfers across different corporate events and source configurations. Both case studies use the three-source instantiation described above.

3.4 Data Collection

3.4.1 Collection Principles

For each source type, the data collection layer performs three operations: retrieval within the event window $W(e)$, metadata capture, and provenance persistence. Metadata include timestamps, document identifiers, and source URLs or accession numbers. Provenance information is stored in a structured record to support downstream reproducibility. Each document is tagged with its source type before being passed to the preprocessing layer.

3.4.2 News Collection

Professional news is collected through the Guardian Open Platform API. The company name is used as the query term, and the `show-fields=bodyText` parameter is used to retrieve full article bodies, headlines, publication dates, and URLs. Results are ordered by relevance within the event window, and the top-ranked articles are retained. Most Guardian articles fit within three or fewer 400-word preprocessing chunks.

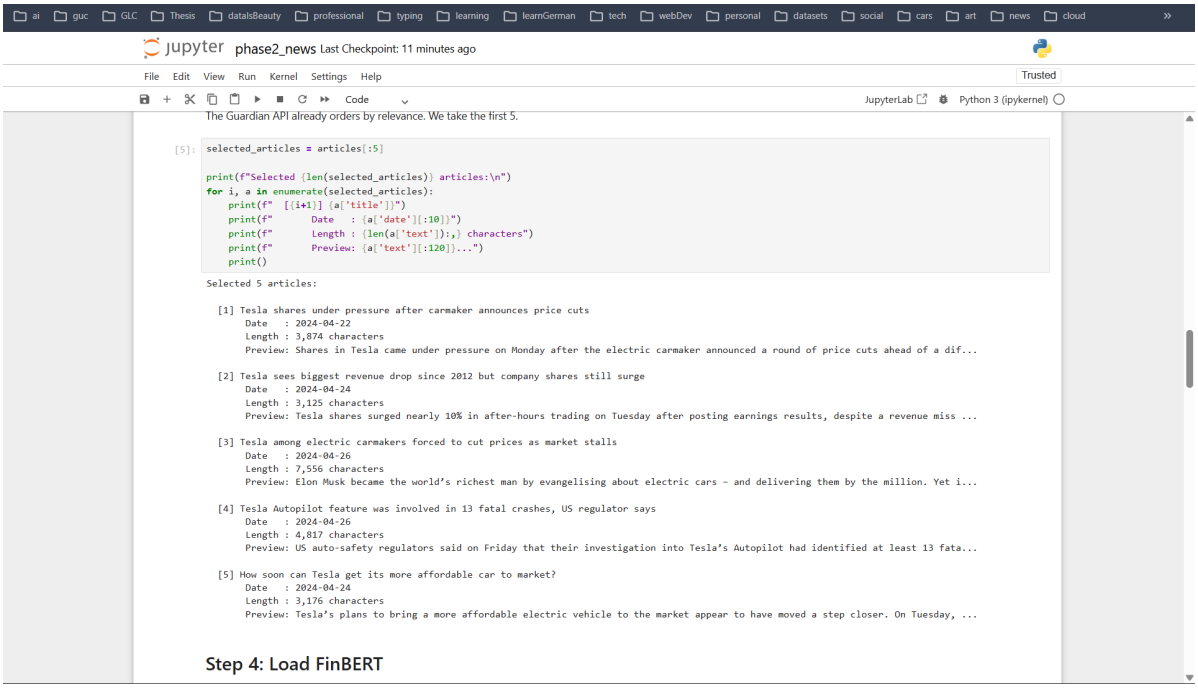

Figure 3.3: Guardian Open Platform API output for the Tesla Q1 2024 event window (April 21–26, 2024): five relevance-ranked articles returned by the `q='tesla'` query with `show-fields=bodyText`. Titles, publication dates, and character lengths correspond to the five news documents used in the primary case study (Table 3.6).

3.4.3 Social Media Collection

Social media documents are collected from the Arctic Shift public Reddit archive through pageable subreddit-level feed queries. The pipeline queries finance-oriented subreddits and applies a case-insensitive keyword filter to the company name and ticker symbol in the post title or body. Matching posts are ranked by upvote score, and the highest-ranked posts within the event window are retained. Unix epoch timestamps are converted to datetime format before event-window comparison.

3.4.4 Regulatory Filing Collection

SEC EDGAR filings are retrieved by accession number using the `edgartools` Python library. For earnings events, Form 8-K is the primary filing type because it contains the earnings press release and key financial data. Long-form filings are extracted as full text and passed to the preprocessing layer, where filings of approximately 6,000 to 7,000 words are split into multiple chunks.

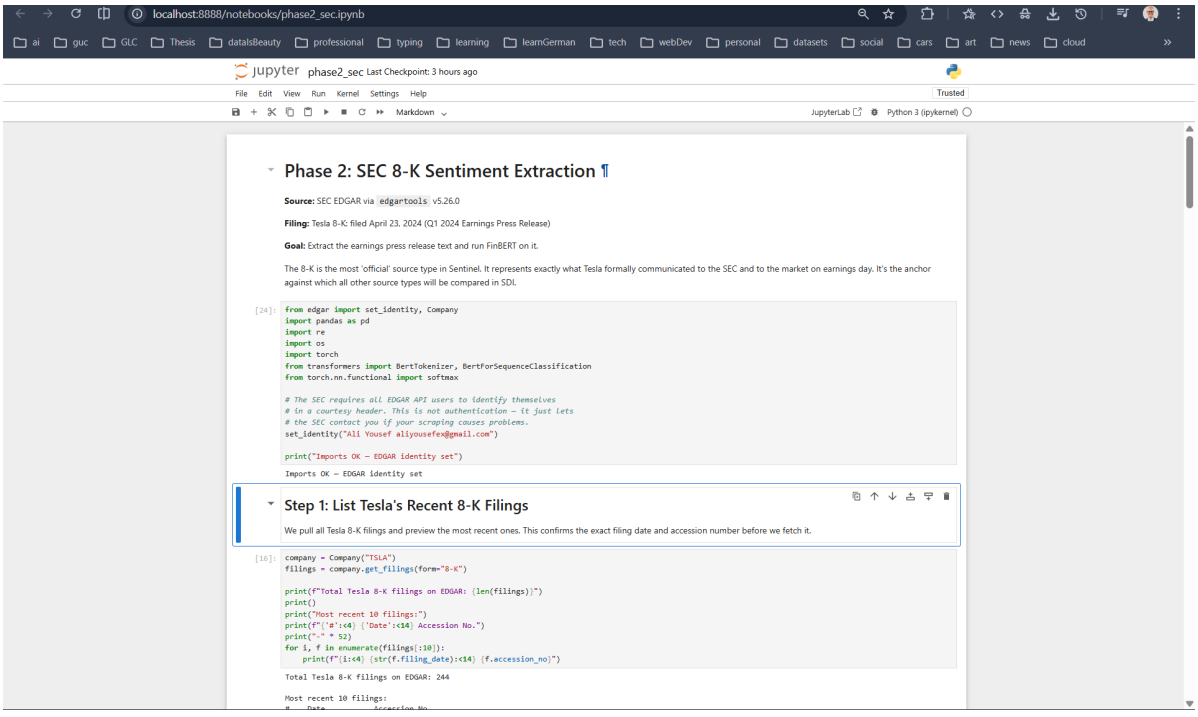

Figure 3.4: `edgartools` library initialisation and the SEC EDGAR 8-K filing history query for Tesla, confirming access to the regulatory filing pipeline. The filing listed for April 23, 2024 is the Q1 2024 Earnings press release (accession number `0000950170-24-046895`) used in the primary case study.

3.4.5 Collection Summary

Table 3.6 summarises how each source type is collected for the Tesla Q1 2024 primary case study, including the collection method, event window, and number of retained documents.

Table 3.6: Multi-source data collection configuration for the Tesla Q1 2024 primary case study.

Source Type	Collection Method	Event Window	n
News (Guardian)	Guardian Open Platform API, relevance-ranked	Apr. 21–26, 2024	5
Reddit (Social Media)	Arctic Shift archive, upvote-ranked, keyword-filtered	Apr. 21–26, 2024	5
SEC 8-K (Regulatory)	edgartools, accession 0000950170-24-046895	Apr. 23, 2024	1
Total			**11**

3.5 Preprocessing

The preprocessing layer uses the same core workflow for all source types. After any source-specific adjustments, each document is cleaned and normalised, split into 400-word chunks, and tokenised with the model-specific tokeniser. Figure 3.5 summarises this document-level workflow.

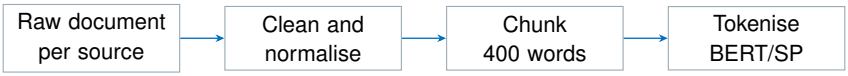

Figure 3.5: Preprocessing workflow. Each document passes through cleaning and normalisation, 400-word chunking, and model-specific tokenisation.

Cleaning and Normalisation.

Cleaning removes HTML residues, normalises unicode quotation characters, and removes boilerplate content. Numerical tokens and currency symbols are retained because they may determine financial sentence polarity. Source-specific cleaning removes subscription prompts and navigation fragments from news articles, excludes URL shorteners and bot-detection markers from Reddit posts, and preserves structural item markers in regulatory filings.

Chunking Strategy. FinBERT inherits BERT's 512-token context limit. Documents are split into non-overlapping chunks of at most 400 words to allow for subword tokenisation overhead and the required `[CLS]` and `[SEP]` tokens. Word count is used because it is model-independent and provides a consistent boundary across source types. A Tesla SEC 8-K of approximately 6,400 words is therefore split into 16 preprocessing chunks.

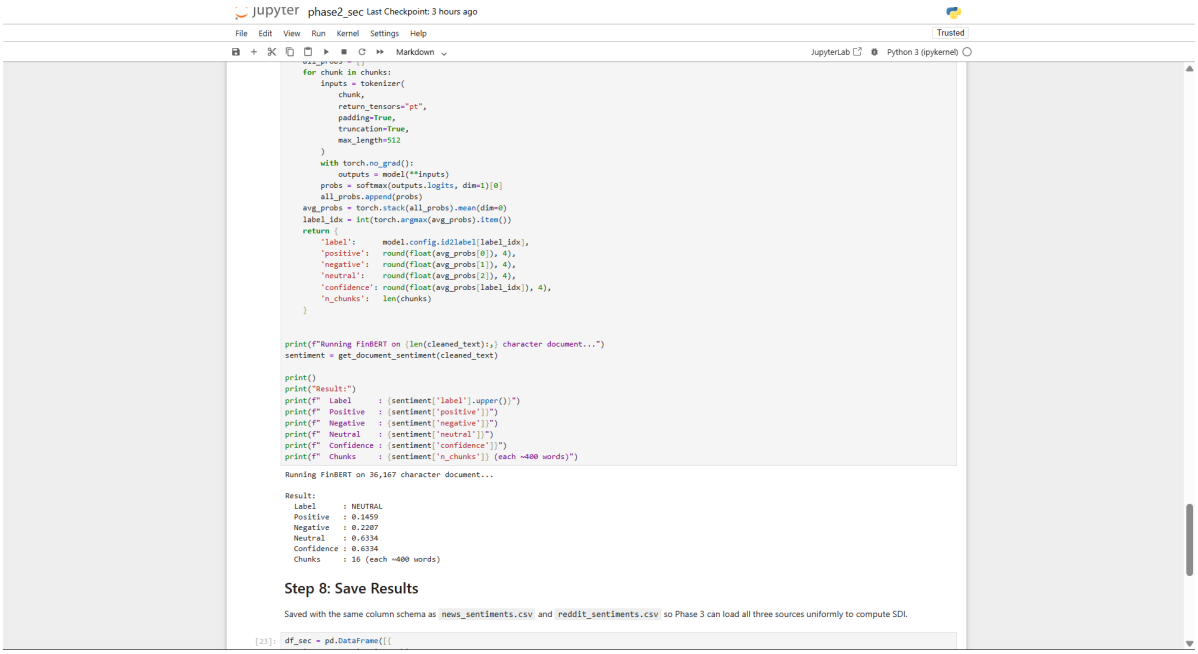

Figure 3.6: FinBERT inference output for the Tesla Q1 2024 Form 8-K (36,167 characters). The output confirms the 16-chunk segmentation described in this section and produces $p_{\text{pos}} = 0.1459$, $p_{\text{neg}} = 0.2207$, yielding signed score $s_i = -0.0748$. This value is used in the SDI computation in Chapter 4.

Tokenisation. Each chunk is tokenised before inference using the relevant model-specific tokeniser. FinBERT uses the WordPiece tokeniser inherited from BERT-base-uncased, with a 30,522-token vocabulary. Sentinel-LLaMA3 uses the 128,000-token LLaMA 3.1 tokeniser, reducing fragmentation for specialised financial terms, ticker symbols, and compound expressions relative to the BERT vocabulary.

3.6 Sentiment Inference and Score Production

3.6.1 Scoring Engine Assignment by Experiment

Our thesis uses separate scoring engines for baseline validation, case-study inference, and open-weight benchmark evaluation. Table 3.7 records these assignments to distinguish benchmark validation from the scoring configuration used for the reported SDI/ADI case-study results.

Table 3.7: Scoring engine assignment by experiment.

Experiment	Scoring engine	Role in thesis
FinBERT baseline validation	FinBERT	Validates the baseline scorer on the FPB all-agree subset.
Tesla Q1 2024 case study	FinBERT	Produces document and aspect signed scores used for reported Tesla SDI/ADI.
AppLovin Q4/FY2025 case study	FinBERT	Produces document and aspect signed scores used for the secondary exploratory SDI/ADI.
Sentinel-LLaMA3-FPB75 benchmark	Sentinel-LLaMA3-FPB75	Evaluates the fine-tuned open-weight scorer on the FPB 75%-agreement split.
Robust-FPB50 benchmark	Sentinel-LLaMA3-Robust-FPB50	Evaluates a separate duplicate-safe lower-agreement adapter under an FPB 50% protocol.

The reported Tesla and AppLovin disagreement results are therefore produced under the FinBERT scoring instantiation. The Sentinel-LLaMA3 experiments address a separate scoring-engine research question: whether an open-weight QLoRA-fine-tuned model can support reproducible financial sentiment classification, evaluated on FPB and additionally through a Tesla scoring-engine comparison in Section 4.4.5.

3.6.2 Per-Chunk Inference

Each chunk is passed independently through the sentiment model, yielding an uncalibrated three-class softmax probability distribution $[p_{\text{pos}}, p_{\text{neg}}, p_{\text{neu}}]$, where $p_{\text{pos}} + p_{\text{neg}} + p_{\text{neu}} = 1$. This distribution is retained as the raw model output for the chunk. Figure 3.7 shows how chunk-level probabilities are aggregated into a signed document-level sentiment score.

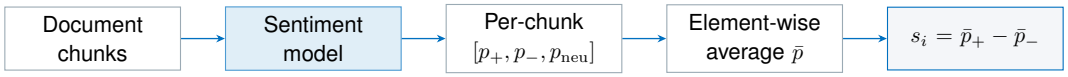

Figure 3.7: Sentiment inference pipeline. Each document chunk is passed independently through the sentiment model to produce a three-class probability vector. Averaging these vectors element-wise across all chunks preserves the full distributional information from every chunk before collapsing to the scalar signed score $s_i = \bar{p}_+ - \bar{p}_-$.

3.6.3 Document-Level Score: Chunk Aggregation

Let document d be split into C chunks. Each chunk c produces a probability vector $[p^{(c)}_{\text{pos}}, p^{(c)}_{\text{neg}}, p^{(c)}_{\text{neu}}]$. The document-level probability vector is the element-wise mean across all chunks:

$$\bar{\mathbf{p}}_d = \frac{1}{C} \sum_{c=1}^{C} [p^{(c)}_{\text{pos}}, p^{(c)}_{\text{neg}}, p^{(c)}_{\text{neu}}]. \tag{3.2}$$

Averaging the probability vectors element-wise preserves class-level confidence information that would otherwise be discarded if only the predicted labels were averaged. The argmax of $\bar{\mathbf{p}}_d$ gives the document-level class label for classification tasks.

Definition: Signed Sentiment Score. Let $\bar{p}_{\text{pos}}$ and $\bar{p}_{\text{neg}}$ denote the document-level mean positive and negative class probabilities from Equation (3.2). The signed sentiment score s_i for document i is defined as

$$s_i = \bar{p}_{\text{pos}} - \bar{p}_{\text{neg}}. \tag{3.3}$$

At the extremes, a document scored as entirely positive yields $s_i = +1$ and one scored as entirely negative yields $s_i = -1$; a document with no directional signal yields $s_i = 0$. The neutral probability does not appear in the score because it encodes the absence of directional signal rather than a point on the positive-negative axis. All signed sentiment scores satisfy $s_i \in [-1, +1]$ by construction, because the class probabilities are non-negative and sum to one. The $[-1, +1]$ range is required for the pairwise distance computation in the SDI: scores from different source types must be on the same scale before their differences carry interpretable meaning.

3.6.4 Observation Design: Individual Documents as SDI Inputs

The SDI is computed over document-level observations. In Sentinel, each document is an observation i with its own signed sentiment score s_i; n counts documents across all source types, not source types.

For the Tesla Q1 2024 case study, $n = 11$: five news articles, five Reddit posts, and one SEC filing, yielding $|P| = n(n-1)/2 = 55$ unordered document pairs. Each observation is tagged with source type $\tau(i)$, which determines weight $w_i = \pi_{\tau(i)} g(\sigma_i)$. Same-source documents share the same source-type weight but retain separate sentiment scores, so both within-type and cross-type pairs enter the SDI. Pair-level contributions are aggregated by source-type pair for reporting.

This prevents within-source variation from being averaged out before disagreement is measured.

3.7 Source Calibration Protocol

3.7.1 Calibration Specification

Cross-source comparison is meaningful only when model scores are calibrated across domains. For example, a FinBERT score of $+0.6$ on an SEC filing and a score of $+0.6$ on a Reddit post must represent the same sentiment level for their difference to be interpretable. If not, the comparison captures a calibration gap rather than real disagreement. Domain shift can create this problem by making the model more or less confident on one source type than another, independently of the underlying sentiment. Guo et al. [19] demonstrated that modern deep neural networks are systematically overconfident; in the cross-source SDI setting, this means that apparent disagreement between source types may reflect model miscalibration rather than genuine sentiment divergence.

Calibration Function. For each source type τ, the framework specifies a post-hoc calibration function g_τ that maps raw model outputs to calibrated sentiment scores:

$$s_i^{\text{cal}} = g_{\tau(i)}(s_i^{\text{raw}}). \tag{3.4}$$

The recommended default is isotonic regression calibration, fitted separately for each source type on a held-out calibration set where human-annotated sentiment serves as ground truth. Isotonic regression is non-parametric and requires no distributional assumptions. When only limited calibration data are available, Platt scaling provides a two-parameter parametric alternative:

$$g_\tau(s) = \frac{2}{1 + \exp(-a_\tau s - b_\tau)} - 1, \tag{3.5}$$

where a_τ and b_τ are learned per source type and the outer transform maps the logistic output to $[-1, +1]$.

Expected Calibration Error. Calibration quality is assessed via the ECE computed per source type. Predicted scores are binned into B equally spaced intervals, and the ECE is

$$\text{ECE}_\tau = \sum_{b=1}^{B} \frac{|n_b|}{N_\tau} \left| \bar{s}_b^{\text{pred}} - \bar{s}_b^{\text{true}} \right|, \tag{3.6}$$

Table 3.8 defines the notation in Equation (3.6).

Table 3.8: Notation for Equation (3.6).

Symbol	Definition
n_b	Number of samples in bin b
N_τ	Total number of samples for source type τ
$\bar{s}_b^{\text{pred}}$	Mean predicted sentiment score in bin b
$\bar{s}_b^{\text{true}}$	Mean human-annotated sentiment score in bin b

Dual-Gate Filtering. The framework specifies two quality gates before observations enter the SDI. The calibration gate requires re-calibration or exclusion when source-type calibration error exceeds a pre-specified tolerance. The uncertainty gate flags the event-level SDI as high-uncertainty when mean epistemic uncertainty exceeds $\sigma_{\max}$. Individual observations remain included through w_i, but the output metadata records the reliability warning. The numerical gate tolerances are implementation hyperparameters to be calibrated on a larger held-out multi-source corpus; they are not estimated or applied in the current case-study evaluation.

3.7.2 Calibration Status in the Current Evaluation

The calibration and uncertainty gates define how the full framework would handle source-level reliability checks once calibration data are available. In the current case-study implementation, these gates are retained as framework specifications rather than executed filtering steps. The evaluation therefore uses the available model scores for all source types and records the calibration status in the output metadata. The resulting SDI values should be read as proof-of-concept disagreement measurements within the case-study setting, not as calibrated population-level estimates.

3.8 The Sentiment Disagreement Index

The SDI is designed as both a disagreement metric and a diagnostic output. It measures how strongly sources disagree about a financial event, while also identifying which source-type pairs contribute most to that disagreement. Formally, $\mathrm{SDI}(e) = 0$ when all source observations agree on sentiment, and $\mathrm{SDI}(e)$ increases monotonically as observations diverge, with greater weight assigned to cases where highly credible, high-confidence sources disagree. The event-level output includes the scalar score, interpretation zone, dominant source-type-pair contribution, and per-source sentiment signs, allowing the disagreement to be traced to the source structure that produced it.

3.8.1 Computation Ladder: Level 1 to Level 4

The SDI is organised as a four-level computation ladder. Level 1 serves as the unweighted baseline, with every document observation contributing equally. Level 2 adds static source-type credibility priors π_τ and is the highest SDI level executed across both current case studies. Level 3 extends this formulation by adding instance-level epistemic uncertainty through MC Dropout, so that observations with higher uncertainty have reduced influence. Level 4 further specifies dynamic credibility terms for source behaviour, historical track record, and market regime. In this thesis, Level 1 and Level 2 are reported for both case studies, while Level 3 is reported only for the Tesla Q1 2024 event. Level 4 remains a specified extension for future empirical evaluation on a larger multi-event corpus.

3.8.2 Notation

Table 3.9 provides the complete symbol table, and Figure 3.8 summarises the computational flow from scores and weights to the normalised SDI.

Table 3.9: Symbol table for SDI notation.

Symbol	Meaning
e	A company-event instance
n	Total number of source documents across all source types for e
s_i	Signed sentiment score for document i, uncalibrated in the current implementation; $s_i \in [-1, +1]$
σ_i	Epistemic uncertainty of s_i, estimated via MC Dropout ($N = 30$ stochastic forward passes); Level 3 values reported in Section 4.4.3
$g(\sigma_i)$	Uncertainty penalty function, $g(\sigma_i) = 1/(1 + \sigma_i)$
$\tau(i)$	Source type of document i
π_τ	Type-prior credibility weight for source type τ
w_i	Composite weight, $w_i = \pi_{\tau(i)} g(\sigma_i)$
P	Set of all unordered document pairs, $\lvert P\rvert = n(n-1)/2$
Z	Pairwise normalisation term; $Z = \sum_{(i,j)\in P} w_i w_j$, equivalently $Z = (W^2 - \sum_i w_i^2)/2$

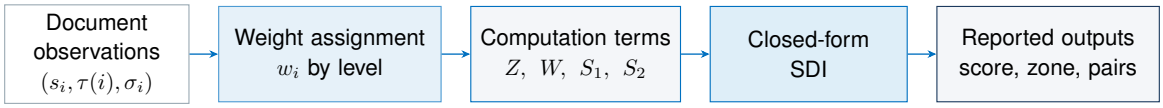

Figure 3.8: SDI computation and reporting flow. Document observations are assigned level-dependent weights, converted into the pairwise and aggregate terms required by the closed-form computation, and reported as a normalised score with an interpretation zone and source-pair contribution breakdown.

3.8.3 Level 1: Unweighted Baseline SDI

Level 1 defines SDI as the root-mean-square sentiment distance over all unordered document pairs:

$$\mathrm{SDI}(e) = \sqrt{\frac{1}{|P|} \sum_{(i,j)\in P} (s_i - s_j)^2}. \tag{3.7}$$

It represents the typical pairwise sentiment gap in the event corpus and provides a credibility-free baseline for evaluating the added effect of Level 2 weighting.

3.8.4 Level 2: Credibility-Weighted SDI

Level 2 introduces source-type credibility into the SDI calculation. In Level 2, each document receives a weight based on its source-type prior; in the uncertainty-aware form, this weight is also adjusted for model uncertainty:

$$w_i = \pi_{\tau(i)} g(\sigma_i), \qquad g(\sigma_i) = \frac{1}{1+\sigma_i}. \tag{3.8}$$

The penalty function $g(\sigma_i)$ decreases as epistemic uncertainty σ_i increases. A lower-confidence document therefore has less influence on the SDI than a higher-confidence document from the same source type. The Level 3 implementation applies MC Dropout with $N = 30$ stochastic forward passes; per-document σ_i values and the resulting Level 3 SDI are reported in Section 4.4.3.

Definition: Credibility-Weighted SDI. The pairwise normalisation term is defined as $Z = \sum_{(i,j)\in P} w_i w_j$, equivalently $Z = (W^2 - \sum_i w_i^2)/2$ where $W = \sum_i w_i$. The credibility-weighted SDI is

$$\mathrm{SDI}_w(e) = \sqrt{\frac{1}{Z} \sum_{(i,j)\in P} w_i w_j (s_i - s_j)^2}. \tag{3.9}$$

Normalising by Z allows events with different source mixes to be compared on the same scale. If every source-type prior is set to one, so that $\pi_{\tau(i)} = 1$ for all i, the expression collapses to the Level 1 baseline. Level 2 therefore adds credibility weighting without changing the underlying pairwise-distance structure.

Source-Type-Pair Contribution Attribution. For each unordered document pair $(i, j) \in P$, the normalised contribution of that pair to aggregate disagreement is

$$C_{ij}(e) = \frac{w_i w_j (s_i - s_j)^2}{\sum_{(p,q)\in P} w_p w_q (s_p - s_q)^2}. \tag{3.10}$$

When the SDI numerator is non-zero, these contributions sum to one across all unordered pairs. For interpretation, they are then aggregated by source-type pair (τ, τ'): all document pairs (i, j) with $\tau(i) = \tau$ and $\tau(j) = \tau'$ are summed into a single source-type-pair contribution. This makes the source-channel structure of the disagreement visible in the output.

3.8.5 Level 3: Uncertainty-Aware Credibility SDI

Level 3 extends Level 2 by estimating document-level epistemic uncertainty σ_i using MC Dropout during inference. For a document chunk d_i, dropout remains enabled across N stochastic forward passes (recommended $N = 30$), yielding signed-score samples $\{s_i^{(1)}, \ldots, s_i^{(N)}\}$. The mean of these samples is used as the point estimate, and their population standard deviation is used as the uncertainty estimate:

$$s_i = \frac{1}{N} \sum_{k=1}^{N} s_i^{(k)}, \qquad \sigma_i = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(s_i^{(k)} - s_i \right)^2}. \tag{3.11}$$

This Level 3 formulation has been implemented and evaluated on the Tesla Q1 2024 event. The per-document σ_i values and resulting Level 3 normalised SDI are reported in Section 4.4.3.

3.8.6 Level 4: Dynamic Credibility Extension

Level 4 extends the Level 3 weight by adding an event- and time-conditional credibility modifier:

$$w_i(t) = \pi_{\tau(i)} g(\sigma_i) f(ev_i, tr_i, reg(t)), \tag{3.12}$$

The modifier f is a learned multiplicative scalar with three inputs: an evasion score ev_i for earnings call question-and-answer pairs, a rolling track-record indicator tr_i comparing past source sentiment with realised outcomes, and a market regime indicator $reg(t)$, such as a volatility-state proxy derived from the Volatility Index (VIX). In our thesis, Level 4 remains a formal specification rather than an implemented component. Its empirical estimation requires a multi-event validation corpus, and no fixed market-regime threshold is estimated or applied in the present evaluation.

3.8.7 Efficient Closed-Form Computation

Computing Equation (3.9) directly requires all unordered document pairs and therefore scales as $\mathcal{O}(n^2)$. For implementation, the same quantity can be computed in $\mathcal{O}(n)$ using a standard identity for weighted squared differences. We include the derivation to make the calculation explicit, not to claim a new algorithmic contribution. The closed form depends on three weighted aggregates: $W = \sum_i w_i$, $S_1 = \sum_i w_i s_i$, and $S_2 = \sum_i w_i s_i^2$.

Identity.

$$\sum_{(i,j)\in P} w_i w_j (s_i - s_j)^2 = W S_2 - S_1^2, \tag{3.13}$$

with $Z = (W^2 - \sum_i w_i^2)/2$. The efficient closed-form expression is therefore

$$\mathrm{SDI}_w(e) = \sqrt{\frac{W S_2 - S_1^2}{Z}}. \tag{3.14}$$

Proof. Expanding the squared difference:

$$\sum_{(i,j)\in P} w_i w_j (s_i - s_j)^2 = \sum_{(i,j)\in P} w_i w_j s_i^2 + \sum_{(i,j)\in P} w_i w_j s_j^2 - 2 \sum_{(i,j)\in P} w_i w_j s_i s_j. \tag{3.15}$$

Because the pairs are unordered, the first two sums are symmetric and have the same value, $\frac{1}{2}(W S_2 - \sum_i w_i^2 s_i^2)$. The cross term can be written as $\sum_{(i,j)\in P} w_i w_j s_i s_j = \frac{1}{2}(S_1^2 - \sum_i w_i^2 s_i^2)$. Combining these terms cancels $\sum_i w_i^2 s_i^2$ and gives $W S_2 - S_1^2$. The result matches the explicit pairwise computation while avoiding enumeration of all document pairs.

3.8.8 Normalisation and Interpretation Scale

The normalised SDI lies in $[0, 1]$:

$$\mathrm{SDI}_{w,\mathrm{norm}}(e) = \frac{\mathrm{SDI}_w(e)}{2}, \tag{3.16}$$

because the maximum absolute difference between any two sentiment scores on $[-1, +1]$ is 2, achieved when one source scores $+1$ and another scores -1. We interpret the normalised score using three zones. Scores below 0.15 are treated as *Aligned*, meaning that sources characterise the event consistently. Scores from 0.15 up to, but not including, 0.45 are treated as *Partial Disagreement*, where sources broadly point in the same direction but differ in intensity or magnitude. Scores of 0.45 or above are treated as *High Conflict*, where sources contradict one another in direction or magnitude.

Algorithm 1 Credibility-weighted SDI computation for one event

Require: Source documents $\{(s_i, \tau(i), \sigma_i)\}_{i=1}^{n}$ for event e; source-type priors π_τ

Ensure: $\mathrm{SDI}_{w,\mathrm{norm}}(e)$, interpretation zone, and dominant source-type-pair contribution

1: Apply calibration gate when calibration metadata are available; otherwise flag uncalibrated source types.
2: **for all** included observations i **do**
3: Compute $w_i = \pi_{\tau(i)}/(1 + \sigma_i)$.
4: **end for**
5: Compute $W = \sum_i w_i$, $S_1 = \sum_i w_i s_i$, $S_2 = \sum_i w_i s_i^2$, and $Z = (W^2 - \sum_i w_i^2)/2$.
6: Compute $\mathrm{SDI}_w(e) = \sqrt{(WS_2 - S_1^2)/Z}$ and $\mathrm{SDI}_{w,\mathrm{norm}}(e) = \mathrm{SDI}_w(e)/2$.
7: Assign interpretation zone: Aligned, Partial Disagreement, or High Conflict.
8: For attribution, compute $C_{ij}(e)$ for document pairs using Equation (3.10).
9: Sum C_{ij} values by source-type pair $(\tau(i), \tau(j))$ to obtain source-type-pair contributions.
10: Return normalised SDI, interpretation zone, dominant source-type pair contribution, and metadata flags.

A complete numerical worked example of this computation is provided in Appendix A.6.

3.9 The Multi-Aspect Sentiment Divergence Framework

The MASD framework extends the document-level SDI by locating disagreement within specific financial dimensions and identifying its direction. Whereas the SDI measures whether sources diverge on an event, MASD specifies what the divergence concerns. This decomposition is needed because aggregate disagreement may obscure aspect-level directional splits: two sources can show moderate event-level disagreement while taking opposite positions on a specific financial dimension, which a document-level score alone cannot distinguish.

The SDI and MASD are therefore complementary rather than redundant. The SDI operates over individual document observations and detects overall cross-source conflict for an event. The ADI operates over source-type-level aspect scores and localises that conflict to financial dimensions such as revenue performance, management credibility, regulatory risk, or market reaction. The MASD framework follows four sequential stages.

3.9.1 Stage 1: Aspect Taxonomy and Keyword Extraction

The aspect taxonomy is derived from financial named-entity-recognition categories and the FiQA aspect categories [11]. The full taxonomy covers six financial dimensions: Performance (Revenue/Sales, Profitability, Earnings, Growth); Financial Health (Debt/Leverage, Liquidity, Cash Flow); Valuation (Stock Price Action, Market Capitalisation); Strategy (Mergers and Acquisitions, Partnerships, Product Launches); Risk (Regulatory, Legal, Competitive Pressure); and Management (Leadership Quality, Governance, Guidance Credibility).

For the Tesla Q1 2024 case study, we operationalise this broader taxonomy as four aspect groups that are directly represented in the collected corpus. The four aspects are selected from the full six-dimension taxonomy as the most directly implicated dimensions for an earnings and operational risk event: Revenue subsumes Performance and Financial Health signals relevant to the quarterly results; Market captures the Valuation response to the post-earnings share price movement; and Strategy and standalone Valuation are excluded because the Tesla Q1 2024 event does not feature a material merger, acquisition, product launch, or market capitalisation announcement. Table 3.10 lists the keyword sets used to identify aspect-relevant sentences for each group.

Table 3.10: Aspect keyword sets used for Tesla Q1 2024 MASD implementation.

Aspect Group	Keywords
Revenue	revenue, sales, income, earnings, profit, margin, growth, decline, miss, beat, forecast, guidance, quarter, annual, fiscal
Management	CEO, management, executive, leadership, board, director, governance, strategy, restructur, layoff, workforce, travel, Musk, guidance, credibility, decision
Risk	recall, crash, fatality, safety, NHTSA, regulation, legal, liability, lawsuit, penalty, investigation, compliance, risk, hazard, audit
Market	stock, share, price, market, valuation, surge, drop, rally, sell-off, trading, volume, investor, capitalization, analyst

Aspect-specific sentiment is extracted through sentence-level filtering. Each document is first split into sentences using regular-expression tokenisation on sentence-terminal punctuation. A sentence is retained for an aspect group if it contains at least one keyword from that group, and sentences shorter than four words are excluded as insufficiently informative. Since the groups can overlap, the same sentence may be retained for more than one aspect group. This keyword-based approach is used because it is transparent and computationally efficient; it is not intended to replace neural aspect extraction methods. The exact same keyword lists are reproduced in Appendix A.1 for auditability.

3.9.2 Aspect Keyword Mapping to AppLovin Event Labels

The AppLovin Q4/FY2025 case study uses the same four-group taxonomy, but assigns event-specific labels that reflect AppLovin's business context. Revenue is labelled as Financial Performance, covering earnings beats, revenue growth, and forward guidance in AppLovin's advertising technology segment. Management is labelled as AI Platform Growth, reflecting the prominence of the AXON AI engine and platform strategy in management commentary. Risk is labelled as Regulatory Data Risk, corresponding to data-privacy scrutiny and app-store compliance exposure discussed across source types. Market is labelled as Market Valuation, covering share-price reactions, analyst price-target revisions, and institutional trading volume. The keyword lists remain unchanged; only the interpretive labels differ, preserving cross-case comparability. The full keyword-to-label mapping is included in Appendix A.1.

3.9.3 Stage 2: Cross-Source Aspect Alignment

Source types often describe the same financial dimension with different vocabulary and levels of formality. An SEC filing may refer to "operating margin compression"; a news article may report "cost increases"; a Reddit post may state that the company "is not making money." Aspect alignment maps these semantically related expressions to the same taxonomy node before divergence is computed. In our current keyword-based implementation, this alignment is handled implicitly through the keyword lists in Table 3.10, which include multiple surface expressions for each aspect.

3.9.4 Stage 3: Aspect Divergence Index

After aspect alignment, we compute the ADI for each covered aspect within each event.

Definition: Aspect Divergence Index. For company c, aspect a, and event window t, let $D_{k,a,t}$ denote the set of aspect-relevant scored sentences from source type k. The aspect sentiment for source type k is

$$s_{c,a,t}^{(k)} = \frac{1}{|D_{k,a,t}|} \sum_{d \in D_{k,a,t}} s_{d,a}, \qquad |D_{k,a,t}| > 0, \tag{3.17}$$

where $s_{d,a}$ is the signed sentiment score of an aspect-relevant sentence or chunk d. If a source type contains no evidence for aspect a in the event window, it is left out of the ADI calculation for that aspect rather than filled with an imputed value. Let K denote the number of source types

covering aspect a; at least two source types ($K \geq 2$) are required for the ADI to be meaningful. The ADI is computed as the population variance of the per-source aspect sentiments:

$$\mathrm{ADI}(c, a, t) = \frac{1}{K} \sum_{k=1}^{K} \left(s_{c,a,t}^{(k)} - \bar{s}_{c,a,t} \right)^2, \tag{3.18}$$

where $\bar{s}_{c,a,t} = \frac{1}{K} \sum_{k=1}^{K} s_{c,a,t}^{(k)}$ is the cross-source mean aspect sentiment. We use denominator K because the calculation is taken over all source types available for that aspect in the dataset, not over a sampled subset. A high ADI for a specific aspect indicates strong cross-source disagreement on that financial dimension. The ADI is computed only when at least two source types cover the aspect; aspects covered by a single source type are excluded from the divergence analysis.

The ADI is intended as a descriptive measure of cross-source disagreement magnitude rather than an inferential statistical test. In our three-source setting ($K = 3$), one atypical or weakly supported source-level observation can materially affect the reported variance. Each ADI value is therefore interpreted together with the signed source-aspect sentiment matrix and the corresponding sentence coverage counts. Aspects supported by fewer than approximately five source-sentences per source type are reported as low-reliability estimates requiring additional evidence.

3.9.5 Coverage Rules and Undefined Aspects

We record aspect coverage separately from aspect sentiment because missing evidence is not the same as neutral sentiment. If a source type contributes no retained sentences for aspect a, we leave that source-aspect cell undefined rather than imputing a value. If fewer than two source types cover the aspect, the ADI for that aspect is undefined and the aspect is excluded from the directional divergence vector for that event. When two or more source types cover the aspect, we compute the ADI over the covered source types only and report the coverage count K alongside the result.

This rule matters when coverage is sparse, especially for social-media sources. In the Tesla Q1 2024 case study, Reddit contributes no Risk-relevant sentences and therefore does not enter the Risk ADI. Under keyword-based extraction, the resulting low Risk ADI should be read as near-consensus among the covering sources, not as evidence that all three source types fully covered the Risk aspect. In the AppLovin Q4/FY2025 secondary case, all four reported aspects have three-source coverage, so the aspect-level comparisons are more balanced but remain exploratory because the event corpus is small.

Directional Divergence Vector. The directional divergence vector aggregates ADI values across all M aspects in the taxonomy:

$$\vec{D}(c,t) = (\mathrm{ADI}(c,a_1,t), \mathrm{ADI}(c,a_2,t), \ldots, \mathrm{ADI}(c,a_M,t)) \,. \tag{3.19}$$

This vector gives a multi-dimensional sentiment divergence profile for a company-event instance. High values in specific components of $\vec{D}$ identify the financial aspects that require targeted investigation. The term "directional" reflects that the per-source scores $s_{c,a,t}^{(k)}$ retain sign information, positive or negative on each aspect, which is shown in the source-aspect matrix defined below.

Source-Aspect Sentiment Matrix. For deeper analysis, we construct a source-aspect sentiment matrix with rows indexed by source type and columns indexed by aspect. Each cell contains the signed sentiment score of source type k on aspect a. The matrix makes the direction of disagreement explicit by showing which source types are positive and which are negative on each financial dimension. A positive score from one source and a negative score from another on the same aspect indicates a directional split that the scalar ADI summarises.

Figure 3.9 illustrates the ADI computation flow from per-source documents to aspect extraction, taxonomy alignment, per-aspect sentiment scoring, and population-variance aggregation.

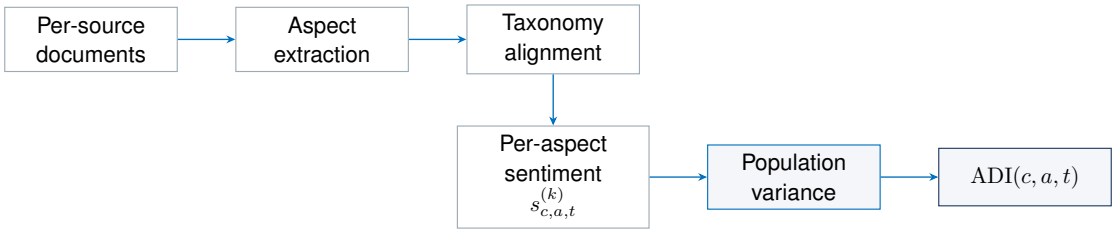

Figure 3.9: ADI computation flow. Aspect-relevant source documents are mapped to the taxonomy, converted into per-source aspect sentiment scores, and summarised through population variance to produce $\mathrm{ADI}(c,a,t)$.

Algorithm 2 MASD/ADI aspect-level decomposition for one event

Require: Event documents grouped by source type; aspect taxonomy with keyword sets; signed sentiment scorer

Ensure: Aspect divergence vector $\vec{D}(c,t)$ and source-aspect sentiment matrix

for all source types k and aspect groups a_m **do**
  Split documents into sentences and retain the sentences associated with aspect a_m by keyword matching.
  if at least one sentence is retained **then**
    Score each retained sentence using the signed sentiment scorer to obtain s_{d,a_m}.
    Compute $s^{(k)}_{c,a_m,t}$ using Equation (3.17).
  else
    Leave the corresponding source-aspect cell undefined.
  end if
end for
for all aspect groups a_m covered by at least two source types **do**
  Compute $\mathrm{ADI}(c,a_m,t)$ using Equation (3.18).
end for
Assemble the source-aspect sentiment matrix and $\vec{D}(c,t)$.

3.10 Explainability Layer

Sentinel's explainability layer reports three artefacts that are derived directly from the metric construction, without relying on post-hoc attribution methods such as SHAP or LIME. The first is the source-type-pair contribution breakdown: for each source-type pair (τ, τ'), the aggregated weighted contribution from all document pairs of that type is computed from Equation (3.10), showing which institutional channel combination contributes most to aggregate disagreement. The second is the directional divergence vector $\vec{D}$ from Equation (3.19), which decomposes document-level disagreement into per-aspect ADI values. The third is the source-aspect sentiment matrix, which records the signed sentiment of each source type on each aspect and makes visible the directional splits summarised by the scalar ADI.

For the Tesla Q1 2024 case study, these three artefacts are: the three-element source-type-pair contribution breakdown over News-SEC, News-Reddit, and Reddit-SEC; the four-element ADI vector over Revenue, Management, Market, and Risk; and the three-by-four source-aspect sentiment matrix with rows for News, Reddit, and SEC and columns for the four aspect groups. These artefacts are human-interpretable for financial audiences without requiring attribution expertise and are machine-readable for downstream integration. Dedicated explanation layers for compliance officers, portfolio managers, and regulators remain future work; the current

artefacts nonetheless support inspectability and provide a limited transparency layer relevant to the EU AI Act and SR 11-7.

3.11 End-to-End Computational Workflow

Algorithm 3 summarises the implemented Sentinel workflow at the event level. We state the workflow at the framework level so that the signed sentiment model can be FinBERT, as in the Tesla and AppLovin case studies, or Sentinel-LLaMA3 via token log-probability extraction, as evaluated in Section 4.4.5.

Algorithm 3 Sentinel multi-source sentiment analysis workflow

Require: Event e; source documents $\mathcal{D}$; source-type map τ; source-type priors π_τ; aspect taxonomy $\mathcal{A}$; keyword sets $\{K_a\}$; signed sentiment model

Ensure: Normalised SDI, source-pair contribution summary, ADI vector, source-aspect matrix, and metadata flags

Align all documents to the event window $W(e)$ and retain the supported source types.
for all documents $d_i \in \mathcal{D}$ **do**
    Clean, normalise, chunk, and tokenise d_i.
    Run the assigned sentiment model on each chunk and aggregate chunk probabilities.
    Compute the signed document score $s_i = p_{\text{pos}} - p_{\text{neg}}$.
    Attach source type $\tau(i)$, credibility prior $\pi_{\tau(i)}$, calibration status, and uncertainty status.
end for
Compute Level 1 and Level 2 SDI; compute Level 3 when uncertainty estimates are available, while leaving Level 4 as a future dynamic-credibility extension.
Aggregate pairwise SDI contributions by source-type pair.
for all aspects $a \in \mathcal{A}$ **do**
    Filter aspect-relevant sentences using K_a and compute source-aspect sentiment scores.
    if at least two source types cover aspect a **then**
        Compute $\mathrm{ADI}(a)$ and append it to $\vec{D}(e)$.
    else
        Mark $\mathrm{ADI}(a)$ as undefined and report the coverage limitation.
    end if
end for
Return the structured output record for interpretation and downstream validation.

3.12 Sentinel-LLaMA3 Fine-Tuning Methodology

Sentinel-LLaMA3 is our intended open-weight scoring model for the Sentinel framework. It is parameter-efficiently fine-tuned and evaluated on the FPB benchmark. In our thesis, Sentinel-LLaMA3 is validated on FPB, while the Tesla Q1 2024 SDI and ADI case-study run uses FinBERT as the scoring engine. A scoring engine comparison using Sentinel-LLaMA3 token log-probability extraction is reported in Section 4.4.5; full pipeline integration as the primary scorer remains a next implementation step.

3.12.1 Limitations of FinBERT as a Scoring Engine

Although FinBERT is a well-validated baseline for financial sentiment classification, it has three structural constraints that motivate our move toward a fine-tuned instruction-following LLM. First, FinBERT inherits BERT's 512-token context limit, so long financial documents must be split into chunks and recombined through chunk-level probability averaging. A Tesla 8-K of approximately 6,400 words is processed in 16 independent chunks under this setup. This averaging treats sentiment as if it were distributed uniformly across the document, which is a poor fit for regulatory filings with structured item-level organisation.

Second, FinBERT produces categorical class labels or three-way probability distributions; the continuous signed scores $s_i \in [-1, +1]$ required for SDI computation must be derived indirectly via $s_i = p_{\text{positive}} - p_{\text{negative}}$, without guarantee that the derived scores are comparably calibrated across source types with different text distributions. Third, FinBERT cannot be prompted to identify the financial aspect under discussion within a given sentence, a capability required by Stage 1 of the MASD pipeline. An instruction-following LLM can be conditioned on a structured prompt that simultaneously requests aspect identification and sentiment labelling in a single inference pass.

3.12.2 Model Selection

We selected LLaMA 3.1 8B Instruct [23] as the base model for four reasons: its open weights support reproducibility without proprietary API access; its instruction-following performance is stronger than prior-generation open-weight models of comparable parameter count; it can be run at 4-bit quantised precision, requiring approximately 6 GB of Video Random Access Memory (VRAM) on a 16 GB GPU; and Meta AI's licence permits research use and adaptation.

Full fine-tuning an 8-billion-parameter model would require updating all parameters and maintaining the corresponding gradient and optimiser states, which exceeds the VRAM available on single-GPU research hardware at full precision. QLoRA [17] addresses this constraint through quantisation and low-rank adaptation.

3.12.3 QLoRA Methodology

QLoRA combines quantisation with low-rank adaptation. The base model weights are stored in 4-bit NF4 representation using the BitsAndBytes library with double quantisation enabled, reducing memory consumption by approximately eight times relative to 32-bit Floating Point (FP32). During computation, the quantised values are upcast to 16-bit Floating Point (FP16) before matrix multiplication. Trainable low-rank adapter matrices are then injected at selected linear layers according to the LoRA formulation [18]:

$$W' = W_0 + \frac{\alpha}{r} BA, \tag{3.20}$$

where W_0 denotes the frozen quantised base weight matrix, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are the trainable adapter matrices with rank $r \ll \min(d, k)$, and α is a scaling hyperparameter. Only A and B receive gradient updates; W_0 remains frozen. The weight update $\Delta W = (\alpha/r)BA$ is therefore constrained to a low-rank matrix, substantially reducing the number of trainable parameters.

LoRA adapters are applied to seven linear layer types in the LLaMA 3.1 attention and feed-forward blocks: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. With rank $r = 16$ and $\alpha = 32$, this configuration introduces 41,943,040 trainable parameters out of 8,072,204,288 total, representing approximately 0.52% of the model. The NF4 quantisation reduces the base model's memory footprint from approximately 32 GB at FP32 to approximately 6 GB, enabling training on a single 16 GB GPU.

3.12.4 Training Configuration

Table 3.11 reports the full training configuration for Sentinel-LLaMA3-FPB75. We format each training example using the Alpaca instruction-tuning convention. The system prompt casts the model as a financial sentiment analyst, and the instruction restricts the response to exactly one of `positive`, `neutral`, or `negative` after the `Sentiment:` prefix. During inference, we use greedy decoding with `do_sample=false` and a maximum of 5 new tokens per prediction, producing a short deterministic sentiment label without free-form explanation.

Table 3.11: Sentinel-LLaMA3-FPB75 training configuration.

Parameter	Value
Base model	meta-llama/Llama-3.1-8B-Instruct
Fine-tuning method	QLoRA (NF4 4-bit, double quantisation enabled)
LoRA rank r	16
LoRA alpha α	32
LoRA dropout	0.05
LoRA target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Maximum sequence length	256 tokens
Training epochs	3
Per-device batch size	1
Gradient accumulation steps	8 (effective batch size: 8)
Learning rate	2×10^{-4}
Warmup ratio	0.03
Weight decay	0.01
Gradient checkpointing	enabled
Final training loss	0.9818
Hardware	NVIDIA GeForce RTX 4090 (CUDA enabled)

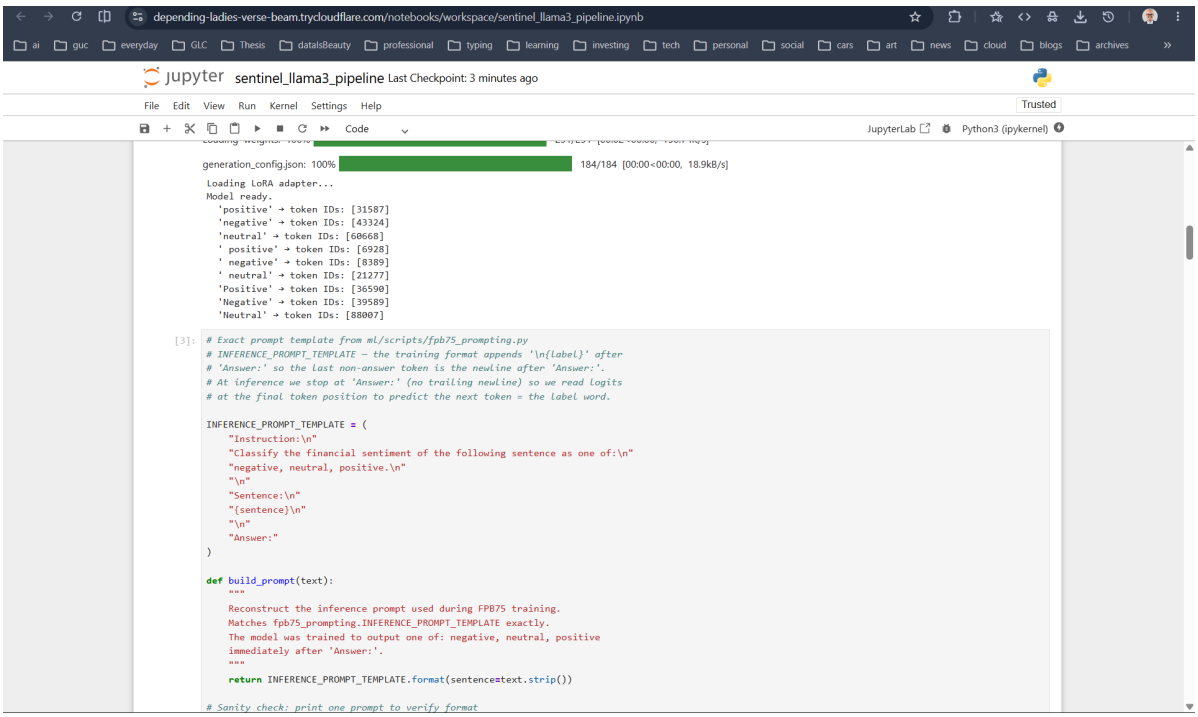

Figure 3.10: Sentinel-LLaMA3-FPB75 inference pipeline: LoRA adapter loading confirmation, label token ID mappings for all case and spacing variants (`positive`→31587, `negative`→43324, `neutral`→60668), and the Alpaca-format inference prompt template. The three primary token IDs are the targets of the three-way softmax that produces the signed score $s_i = p_{\text{pos}} - p_{\text{neg}}$ used in the scoring engine comparison (Section 4.4.5).

3.12.5 Positioning Relative to FinSentLLM

Sentinel-LLaMA3 and FinSentLLM [9] differ in what their disagreement signal represents. FinSentLLM measures model-level disagreement by applying multiple classifier models (FinBERT and a RoBERTa-based sentiment model) to the same input text and comparing their predicted posteriors. Sentinel-LLaMA3, by contrast, is a single scoring model applied to documents from heterogeneous institutional source types. The SDI therefore captures disagreement between source-level signals, not disagreement between model predictions on identical text. Model-level disagreement reflects uncertainty or variation within a classifier ensemble, whereas source-level disagreement reflects variation in the information environment around a shared event.

3.13 Framework Output Specification

For each company-event instance, the framework returns a structured output record that combines metric values, diagnostic decompositions, and metadata. The record includes per-source-type average sentiment and dominant direction; raw and normalised Level 1 and Level 2 SDI values; the dominant source-type pair and its aggregated weighted contribution; the ADI vector over the aspect taxonomy; the source-aspect sentiment matrix; and metadata flags for calibration status, uncertainty-adjustment status, source-type coverage, and scorer identity. The scorer identity flag is included because the Tesla Q1 2024 and AppLovin Q4/FY2025 records are FinBERT-based case-study outputs, whereas Sentinel-LLaMA3-FPB75 is evaluated on FPB and applied to the Tesla corpus in the scoring-engine comparison (Section 4.4.5). The case-study records are presented and interpreted in Chapter 4.

Chapter 6 details the statistical validation design for the SDI-to-volatility hypothesis. It presents the nested two-stage regression structure and corpus-scale targets as the main extension of the framework.

3.14 Methodology Assumptions and Scope Limitations

The current implementation relies on six assumptions that set the boundaries of the proof-of-concept evaluation.

Source representativeness. The Guardian Open Platform provides a single-publisher proxy for professional news; a multi-publisher collection would give a more stable news sentiment estimate. Arctic Shift Reddit data covers finance-oriented subreddits but not the wider social-media environment. The SEC 8-K represents one regulatory filing per event. Within this proof-of-concept evaluation, we treat these sources as representative samples of their respective institutional categories.

Calibration deferral. Sentiment scores from FinBERT are used without post-hoc calibration across source types. As established in Section 3.7, systematic differences in model confidence across text domains can produce calibration artefacts that inflate apparent disagreement. This limitation is flagged in all output metadata and acknowledged as a first-order extension in Chapter 6.

Uncertainty estimates. The Level 3 implementation applies MC Dropout with $N = 30$ stochastic forward passes per document; per-document σ_i values and the resulting Level 3 SDI are reported in Section 4.4.3.

Keyword-based aspect extraction. The MASD implementation uses the keyword lists in Table 3.10 for sentence-level aspect filtering. This approach may retain sentences where a keyword appears in a context unrelated to the target aspect (false positives) and may miss aspect mentions expressed without any listed keyword (false negatives). The keyword lists were constructed to cover the primary lexical expressions for each aspect in English financial text and are sufficient for a proof-of-concept case study, but they would require validation and expansion for deployment across a broader event corpus.

As a methodological sensitivity check, sentence-ID LLM aspect extraction was implemented as an extractive MASD sensitivity audit for Tesla Q1 2024 and compared against the verified keyword-based MASD baseline. This audit did not replace the keyword-based pipeline. Instead, each source document was segmented into numbered candidate sentences, and the LLM was prompted to select sentence IDs only for each aspect rather than generate new text. Selected IDs were then mapped back to the original source sentences before FinBERT scoring. This preserves extractive grounding while testing whether the evidence-selection layer changes the resulting ADI values.

Public-data and model-risk boundaries. The current case studies use public or publicly accessible textual sources: professional news retrieved through publisher or archive access routes, Reddit posts from public finance-oriented communities, and SEC filings from EDGAR. Our thesis does not attempt user-level profiling, private-message analysis, or investment recommendation generation. The framework outputs are research artefacts for measuring disagreement and should not be interpreted as trading advice, credit decisions, or regulatory determinations without additional validation, governance review, and documented model-risk controls.

Case-study scope and generalisability. The Tesla Q1 2024 event was selected as a stress test for cross-source disagreement analysis because it contains multiple partially conflicting signals: the largest quarterly revenue decline since 2012, an Autopilot safety recall linked to formal NHTSA review, concurrent Full Self-Driving (FSD) technology licensing negotiations framed as strategic validation, and a post-announcement share price surge of approximately 12%. This combination of signals is expected to maximise cross-source divergence; it is not representative of a typical corporate event. The resulting SDI values establish that the metric can surface meaningful disagreement on a high-conflict event; they do not establish what SDI

level is typical across the distribution of corporate events. Multi-event validation is the first-order extension required before the interpretation zones can be treated as calibrated thresholds.

With the framework specification, pipeline design, and scope boundaries established, Chapter 4 presents the empirical application of this pipeline to the Tesla Q1 2024 primary and AppLovin Q4/FY2025 secondary events, reporting case-study results across all four phases alongside the independent Sentinel-LLaMA3 benchmark evaluation on FPB.

Chapter 4

Results

This chapter reports the experimental results from the benchmark validation, robustness audits, and multi-source case-study evaluations. In each section, results are stated before they are interpreted, and interpretation is kept concise and tied to the specific evidence at hand.

4.1 Experimental Setup Summary

The evaluation draws on two data types: FPB benchmark data for scoring-engine validation and robustness auditing, and two multi-source corporate event corpora for SDI and ADI case-study evaluation. Table 4.1 summarises each.

Table 4.1: Datasets used in the Sentinel evaluation.

Dataset	Source	Access	Role
FPB all-agree	Financial news	HuggingFace `takala/financial_phrasebank`	FinBERT baseline validation
FPB 75%-agree	Financial news	HuggingFace `takala/financial_phrasebank`	Sentinel-LLaMA3-FPB75 training and test
FPB 66%-exclusive	Financial news	Raw PhraseBank files	Leakage-safe robustness audit
FPB 50%-exclusive	Financial news	Raw PhraseBank files	Leakage-safe robustness audit
FPB 50%-minus-75%	Financial news	Raw PhraseBank files	Broader non-overlap audit
FPB 50%-full	Financial news	Raw PhraseBank files	Sentinel-LLaMA3-Robust-FPB50 training and test
Tesla Q1 2024	Multi-source event	Guardian / Arctic Shift / edgartools	SDI and ADI primary case-study evaluation
AppLovin Q4/FY2025	Multi-source event	Independent news / social-retail discussion / SEC filing	Secondary exploratory SDI/ADI case study

The chapter is organised around two parallel evaluation tracks. The case-study track ap-

plies FinBERT to the Tesla Q1 2024 primary and AppLovin Q4/FY2025 secondary corporate event corpora, covering per-source sentiment scoring, SDI computation, and ADI decomposition across Phases 1–4. The scoring-engine track evaluates Sentinel-LLaMA3-FPB75 on the FPB benchmark and additionally applies token log-probability extraction to the Tesla Q1 2024 corpus, reported in Section 4.4.5. The two tracks address distinct research questions and are reported in sequence: case-study phases first, scoring-engine benchmark after Phase 4.

4.2 Phase 1: FinBERT Baseline Validation

FinBERT achieves 97.17% accuracy and macro F1 0.96000 on the FPB all-agree subset, with $n = 2{,}264$ sentences and neutral-class support of 1,391. Both results align with published high-agreement benchmarks and confirm FinBERT as an appropriate baseline for the case-study pipeline phases. A separately fine-tuned variant, evaluated on a 227-example split, achieves 98.24% accuracy and macro F1 0.97050; it is included for reference only and does not contribute to the Sentinel-LLaMA3 benchmark.

4.3 Phase 2: Per-Source Sentiment Scoring

All three source types produced net-negative sentiment signals for the Tesla Q1 2024 event within the five-day collection window from 21–26 April 2024, but at markedly different intensities. This window extends beyond the framework default of 24 hours pre-event and 48 hours post-event to accommodate the Guardian API's relevance-ranked article retrieval, which surfaces contextual coverage from the days preceding the announcement alongside immediate post-earnings reporting; the wider window ensures that all five relevance-ranked articles fall within the collection scope. Table 4.2 presents the source-type prior, the average signed sentiment score, and the dominant direction for each source.

Table 4.2: Phase 2 multi-source sentiment summary, Tesla Q1 2024 earnings event.

Source	π_τ	**Avg.** s_i	**Direction**
News (Guardian)	0.7	-0.495	Negative
Reddit (Social)	0.4	-0.346	Negative
SEC 8-K (Regulatory)	1.0	-0.075	Neutral

The spread of average scores across source types is 0.420 units, from -0.495 for news to -0.075 for the SEC 8-K, indicating marked cross-source variation before any formal disagreement measurement is applied. The SEC 8-K's near-neutral average score is expected given the structural constraints of regulatory disclosure language: even a filing reporting the largest quarterly revenue decline in the company's history since 2012 produces a near-neutral aggregate sentiment score under FinBERT inference, because obligatory hedging and qualification

suppress evaluative language. The 0.420-unit spread is therefore treated as a structural feature of the informational ecology, not merely as a measurement artefact.

4.4 Phase 3: SDI Computation

4.4.1 Aggregated Statistics

The corpus comprises $n = 11$ source observations: five news articles, five Reddit posts, and one SEC 8-K filing, giving $|P| = n(n-1)/2 = 55$ unique source pairs. The Level 2 weighted statistics are $W = 6.50$, $S_1 = -2.499$, $S_2 = 2.407$, and $Z = 19.00$.

4.4.2 SDI Results

Both SDI levels fall in the Partial Disagreement zone, $[0.15, 0.45)$. Table 4.3 places the result on the interpretation scale, and Figure 4.1 shows the Tesla Q1 2024 placement.

Table 4.3: SDI normalised score interpretation scale with Tesla Q1 2024 result.

$\mathrm{SDI}_{\mathrm{norm}}$ **range**	**Interpretation**	**Tesla Q1 2024**
$[0.00, 0.15)$	Aligned: sources characterise the event consistently	
$[0.15, 0.45)$	Partial Disagreement: same sign, varying degree	**0.35172**
$[0.45, 1.00]$	High Conflict: sources contradict each other	

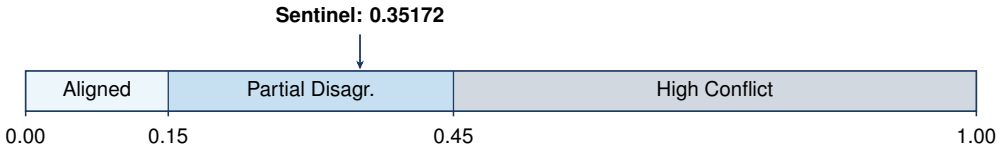

Figure 4.1: Normalised SDI interpretation scale with the Tesla Q1 2024 Level 2 result, 0.35172, placed in the Partial Disagreement zone.

The Level 2 credibility-weighted normalised score of 0.35172 is $\Delta = 0.0046$ below the Level 1 unweighted score of 0.3563. Credibility weighting therefore had a negligible effect on the aggregate disagreement for this event.

4.4.3 Level 3: MC Dropout Uncertainty Estimates

Table 4.4: MC Dropout uncertainty estimates for the eleven Tesla Q1 2024 source documents ($N = 30$ stochastic forward passes). σ_i is the population standard deviation of signed scores across passes; $g(\sigma_i) = 1/(1 + \sigma_i)$ is the uncertainty discount; $w_{i,\mathrm{L3}} = \pi_\tau \cdot g(\sigma_i)$ is the Level 3 composite weight.

Document	Source type	Score	σ_i	$g(\sigma_i)$	π_τ	$w_{i,\mathrm{L3}}$
news_1	News	-0.947	0.008	0.992	0.7	0.694
news_2	News	-0.853	0.041	0.961	0.7	0.672
news_3	News	-0.523	0.049	0.953	0.7	0.667
news_4	News	-0.431	0.055	0.948	0.7	0.663
news_5	News	$+0.353$	0.049	0.953	0.7	0.667
reddit_1	Reddit	$+0.369$	0.140	0.877	0.4	0.351
reddit_2	Reddit	-0.758	0.100	0.909	0.4	0.364
reddit_3	Reddit	-0.433	0.121	0.892	0.4	0.357
reddit_4	Reddit	-0.457	0.124	0.890	0.4	0.356
reddit_5	Reddit	-0.048	0.057	0.946	0.4	0.378
SEC 8-K	SEC	-0.031	0.014	0.986	1.0	0.986

Table 4.4 shows a clear uncertainty gradient across source types that tracks the structural accountability spectrum. The SEC 8-K filing, scored over 16 chunks of formal regulatory language, exhibits the lowest uncertainty ($\sigma_i = 0.014$): the high redundancy of legal disclosure text produces stable signed scores across stochastic inference passes. News articles, scored over two to four chunks, exhibit moderate uncertainty ($\sigma_i = 0.008$–0.055). Reddit posts, each represented as a single short-text chunk, exhibit the highest uncertainty ($\sigma_i = 0.057$–0.140), reflecting their sensitivity to individual token dropout in the absence of cross-chunk averaging. Longer, institutionally constrained documents produce more stable sentiment estimates under stochastic inference, consistent with the structural accountability spectrum.

The Level 3 uncertainty adjustment follows the expected direction. Reddit documents receive the largest $g(\sigma_i)$ discount (minimum 0.877), reducing their effective weights relative to news ($g(\sigma_i) \geq 0.948$) and the SEC 8-K ($g(\sigma_i) = 0.986$). Despite this reweighting, the Level 3 normalised SDI (0.32323) differs from the Level 2 value computed over the same MC Dropout mean scores (0.32251) by only $+0.00071$. All three SDI levels remain in the Partial Disagreement zone, and the dominant source pair (News vs. SEC 8-K) is unchanged. The MC Dropout uncertainty estimates therefore confirm rather than revise the Level 2 finding: the observed cross-source disagreement in the Tesla Q1 2024 event is robust to instance-level uncertainty adjustment.

A note on MC Dropout score convergence: the MC Dropout mean scores used in Level 3 differ from the original deterministic Level 2 scores (e.g., Level 2 normalised SDI = 0.35172 vs. Level 3 baseline = 0.32251) because stochastic dropout-enabled forward passes produce different activations than the deterministic `model.eval()` pass. Both values remain in the Partial

Disagreement zone. The key comparative result is the minimal L3–L2 delta of $+0.00071$; it is computed consistently from MC Dropout mean scores and is not affected by this difference.

4.4.4 Source-Pair Contributions

The most divergent source pair is News versus the SEC 8-K, with a pairwise distance of $d(\text{News}, \text{SEC}) = 0.420$ and a weighted pairwise contribution of 0.124. By contrast, the News–Reddit pair contributes 0.006 and the Reddit–SEC pair contributes 0.030. This dominance reflects the weight structure: the News–SEC pair carries a combined weight of $w_{\text{News}} w_{\text{SEC}} = 0.7 \times 1.0 = 0.70$, higher than the News–Reddit pair's $0.7 \times 0.4 = 0.28$ and the Reddit–SEC pair's $0.4 \times 1.0 = 0.40$.

4.4.5 Scoring Engine Comparison: FinBERT versus Sentinel-LLaMA3

To evaluate whether the SDI zone classification is robust to scoring engine choice, the Tesla Q1 2024 corpus was re-scored using Sentinel-LLaMA3-FPB75 via token log-probability extraction. Rather than generating a label token, the model produces a signed score by extracting the logits for the three label tokens (*positive*, *negative*, *neutral*) at the final prompt position and applying a three-way softmax, yielding $s_i = p_{\text{pos}} - p_{\text{neg}} \in [-1, +1]$ for each document chunk. This approach requires no retraining and integrates Sentinel-LLaMA3 directly into the SDI pipeline. Because edgartools was unavailable in the inference environment, the SEC 8-K was scored using a constructed near-neutral summary; its score of -0.016 is close to the deterministic FinBERT value of -0.075, and the impact on SDI is minimal.

Table 4.5: Per-document signed scores and SDI Level 2 results for FinBERT and Sentinel-LLaMA3 scoring engines, Tesla Q1 2024 earnings event. Sentinel-LLaMA3 scores are derived from token log-probability extraction without label generation.

Document	Source	FinBERT score	LLaMA3 score	Δ
news_1	News	−0.960	−0.540	+0.420
news_2	News	−0.877	+0.214	+1.091
news_3	News	−0.535	+0.013	+0.548
news_4	News	−0.511	+0.063	+0.574
news_5	News	+0.410	−0.384	−0.794
reddit_1	Reddit	+0.473	+0.005	−0.469
reddit_2	Reddit	−0.839	+0.233	+1.071
reddit_3	Reddit	−0.627	−0.729	−0.101
reddit_4	Reddit	−0.683	−0.937	−0.254
reddit_5	Reddit	−0.055	−0.255	−0.200
SEC 8-K	SEC	−0.075	−0.016	+0.059
SDI Level 2 (normalised)		0.35172	0.25434	−0.09738
Zone classification		Partial Disag.	Partial Disag.	—
Dominant source pair		News vs. SEC	Reddit vs. SEC	—

Both scoring engines place the Tesla Q1 2024 event in the Partial Disagreement zone, but diverge substantially at the document level. Sentinel-LLaMA3 assigns neutral or positive scores to several news articles that FinBERT scores negatively, reflecting its tendency to weight positive framing, such as the post-earnings share price increase in news_2, alongside negative signals; this shifts the dominant source pair from News versus the SEC filing under FinBERT to Reddit, which shows greater polarity dispersion under Sentinel-LLaMA3. We find this contrast analytically useful: the SDI zone classification characterises event-level disagreement more stably than individual document scores, which are sensitive to architectural and training differences between encoder-based and instruction-tuned models. Between-model score differences reach 1.091 units for individual documents in several cases, exceeding the within-model cross-source distance of 0.420 units for FinBERT, confirming that scoring engine choice is a material source of upstream variation that multi-event validation should treat as a controlled variable.

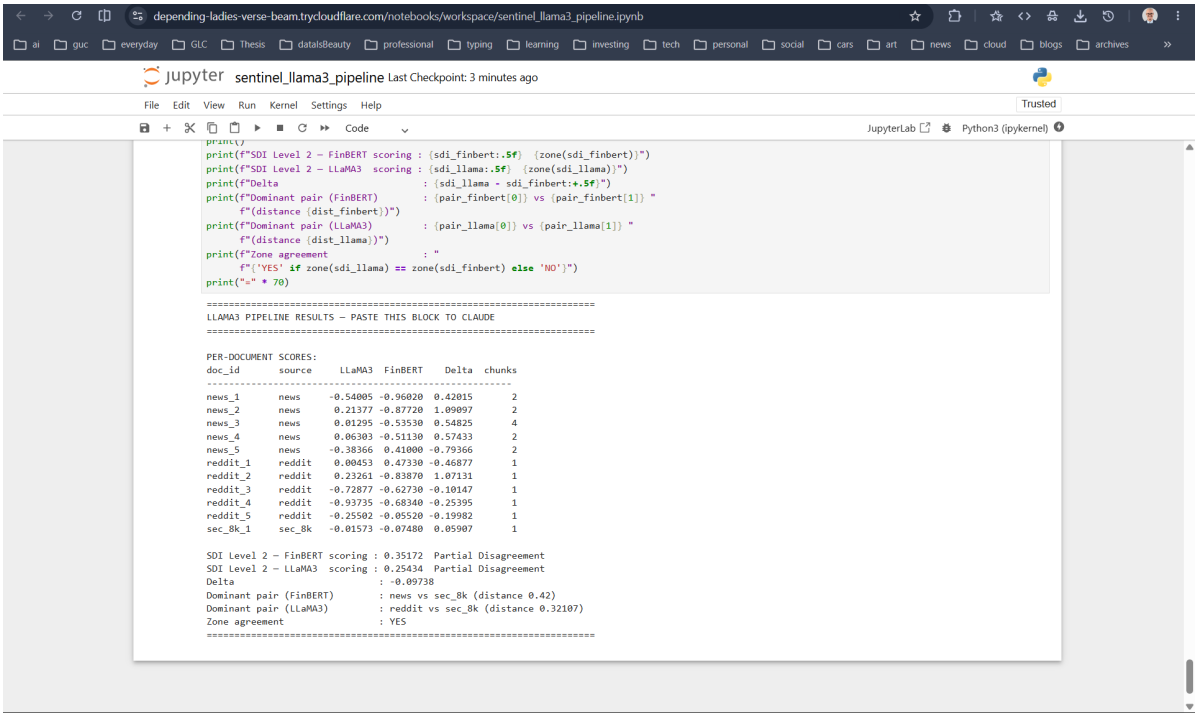

Figure 4.2: Console output from the Sentinel-LLaMA3-FPB75 scoring engine comparison on the Tesla Q1 2024 corpus (11 source documents). Per-document signed scores confirm the values in Table 4.5. The aggregate SDI Level 2 values (FinBERT: 0.35172; Sentinel-LLaMA3: 0.25434) and the zone agreement (`Zone agreement: YES`) confirm Partial Disagreement under both scoring engines.

4.5 Phase 4: ADI Decomposition

4.5.1 Aspect Sentence Coverage

Table 4.6 reports aspect sentence counts per source document; Figure 4.3 aggregates the same counts by source type. Coverage is highly uneven across source types: the SEC 8-K and news articles dominate, while Reddit posts are structurally sparse and contribute zero Risk-relevant sentences across all five posts, leaving Risk ADI with only 2/3 source coverage.

Table 4.6: Aspect sentence coverage counts per source document.

Document (truncated)	Rev.	Mgt.	Risk	Mkt.
News: shares under pressure	11	11	6	15
News: EV makers cut prices	23	5	2	41
News: Autopilot fatal crashes	10	3	24	9
News: affordable car market	11	9	1	14
News: biggest revenue drop	12	6	6	13
Reddit: FSD licensing talks	0	1	0	0
Reddit: TSLA up 12%	1	0	0	1
Reddit: META earnings selloff	1	0	0	1
Reddit: FSD transfer policy	0	0	0	0
Reddit: shorted TSLA	0	0	0	1
SEC 8-K (Q1 2024 filing)	72	23	9	46
Total	**141**	**58**	**48**	**141**

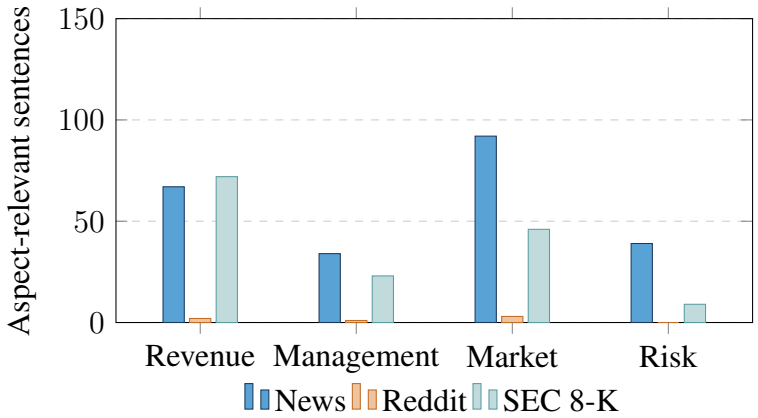

Figure 4.3: Aspect-relevant sentence counts by source type, aggregated across all documents. Reddit bars are near-zero across all aspects (range 0–3), reflecting the structural sparseness of short-form social media posts relative to news articles and the SEC 8-K filing.

4.5.2 ADI Results

Table 4.7 reports the per-source aspect-level signed sentiment scores. Figure 4.4 ranks the resulting keyword-based ADI values, and Figure 4.5 shows the directional source-aspect sentiment matrix. The keyword-based directional divergence vector is

$$\vec{D}(\text{Tesla}, \text{Q1-2024}) = (0.199,\ 0.135,\ 0.097,\ 0.013), \tag{4.1}$$

for Revenue, Management, Market, and Risk respectively.

Table 4.7: Per-source aspect-level sentiment scores and ADI values, Tesla Q1 2024.

Aspect	**News** ($\pi = 0.7$)	**Reddit** ($\pi = 0.4$)	**SEC 8-K** ($\pi = 1.0$)	**ADI**	**Coverage**
Revenue	-0.354	-0.948	+0.143	**0.199**	3/3
Management	-0.334	+0.565	+0.134	0.135	3/3
Market	-0.357	-0.628	+0.126	0.097	3/3
Risk	-0.304	N/A	-0.073	0.013	2/3

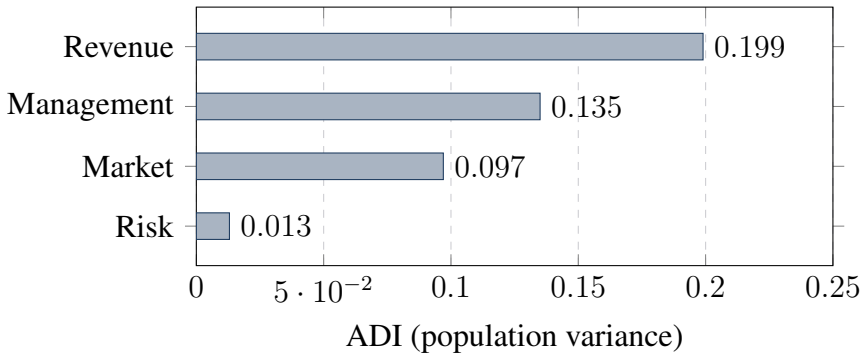

Figure 4.4: ADI values per aspect, Tesla Q1 2024.

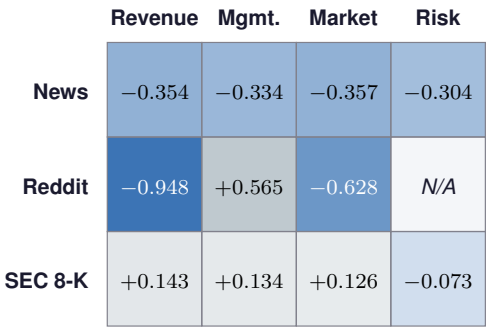

Figure 4.5: Source-aspect sentiment heatmap, Tesla Q1 2024. Cell labels report signed aspect sentiment values. The sign indicates direction, and darker saturation indicates larger magnitude.

4.5.3 Key Aspect Findings

Finding F-A1: Under keyword-based extraction, Revenue is the most contested aspect. Revenue has the highest keyword-based ADI, at 0.199. Reddit scores Revenue at -0.948, reflecting crowd concern about the earnings miss, while the SEC 8-K scores Revenue at $+0.143$, reflecting measured regulatory disclosure language. The 1.091-unit gap between these values is the largest source-pair spread in the dataset. The result shows that the same financial fact, a quarterly revenue decline of unprecedented magnitude for the company, can receive near-opposite sentiment characterisations depending on source type. This is consistent with Garcia's account [15], where cross-source sentiment divergence is linked to differences in institutional mandate rather than to different underlying facts.

Finding F-A2: Management is the only aspect exhibiting a potential directional split in the current dataset. Management has a keyword-based ADI of 0.135. Reddit scores Management at $+0.565$, mainly because the relevant Reddit post frames the FSD licensing announcement as strategic validation and evidence of management foresight. News scores Management at -0.334, reflecting coverage of chief executive travel cancellations, workforce restructuring announcements, and leadership uncertainty. The SEC 8-K scores Management near-neutrally at $+0.134$. These signals show why aspect-level analysis is useful: the document-level scores register moderate disagreement, while the aspect-level matrix exposes a possible directional split. This interpretation is bounded by coverage. The Reddit Management score is based on a single aspect-relevant sentence across all five Reddit posts, so the finding is best read as an illustration of directional-split detection rather than as a validated empirical result. Subject to that caveat, the pattern is consistent with Hong and Stein [34], where heterogeneous information-processing constraints can produce divergent interpretations of the same event.

Finding F-A3: Under keyword-based extraction, Risk approaches consensus across covering sources. Risk has a keyword-based ADI of 0.013. News scores Risk at -0.304 and the SEC 8-K at -0.073, while Reddit contributes zero Risk-relevant sentences. The Autopilot recall, linked to 13 fatal crashes and subject to a formal NHTSA ruling, leaves limited room for positive framing under either news or regulatory disclosure conventions. Under keyword-based extraction, the low ADI indicates that the sources covering Risk assign it a similar negative interpretation; the sentence-ID LLM audit in Table 4.8 later shows that Risk divergence increases under semantic extractive selection. This distinguishes measured agreement from missing coverage and shows that ADI can identify convergence as well as divergence. The analyst-dispersion literature [35] similarly treats low dispersion as evidence that a signal is strong enough to dominate heterogeneous prior beliefs.

Finding F-A4: Market exhibits moderate divergence. Market has a keyword-based ADI of 0.097. News scores Market at -0.357 and Reddit at -0.628, both negative, while the SEC 8-K scores Market slightly positively at $+0.126$. The divergence reflects the counterintuitive post-earnings share surge of approximately 12%. Reddit emphasises the contradiction of a market advance despite a significant earnings miss, whereas the regulatory filing presents stock price action in non-evaluative disclosure language aligned with the observed market movement. Under keyword extraction, Market therefore falls between the near-consensus Risk aspect and the highly divergent Revenue aspect. This supports the interpretation of ADI as a graded measure of aspect-level disagreement, consistent with Miller's [33] account that disagreement varies with event type and with the scope for legitimate interpretive variance.

4.5.4 Extraction-Method Sensitivity Analysis

Before comparing extraction strategies, the keyword ADI baseline was reproduced for the Tesla Q1 2024 case study. Sentence-ID LLM aspect extraction was implemented as an extractive MASD sensitivity audit and compared against the verified keyword-based MASD baseline. In the audit, the model selected sentence IDs from numbered source sentences; only valid IDs that mapped back to original source sentences were passed to FinBERT for scoring.

Table 4.8: Keyword and Sentence-ID LLM ADI Comparison for Tesla Q1 2024

Aspect	Keyword ADI	Sentence-ID LLM-ADI	Change	Coverage
Revenue	0.198831	0.098466	Decrease	3/3
Management	0.135021	0.030068	Decrease	2/3
Market	0.097599	0.067127	Decrease	3/3
Risk	0.013820	0.073896	Increase	2/3

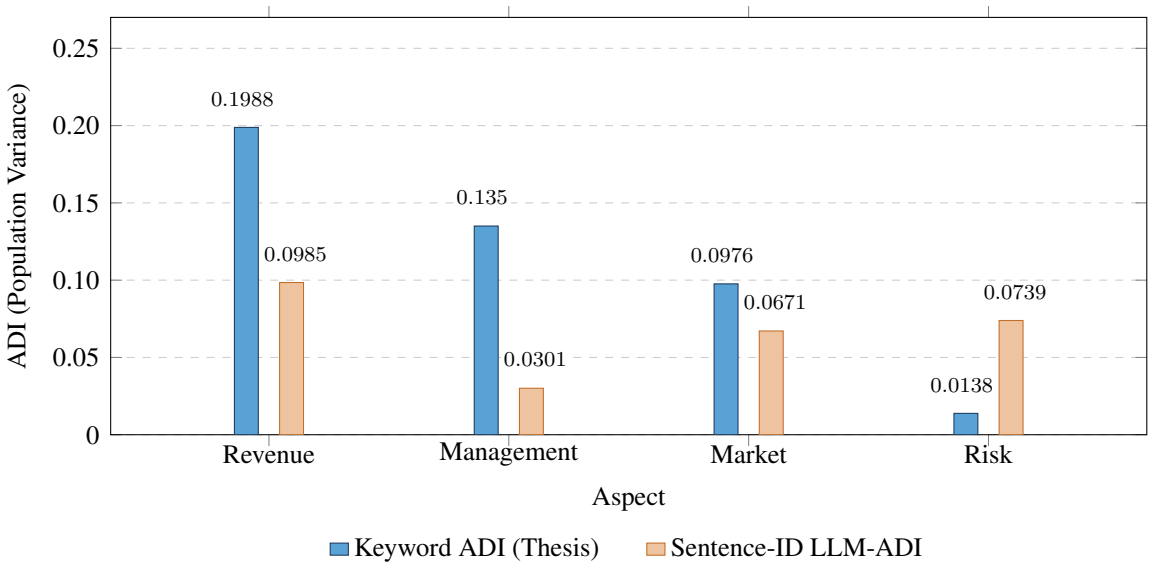

Figure 4.6: Keyword-based and sentence-ID LLM-based ADI values for the Tesla Q1 2024 case study. Sentence-ID extraction reduces Revenue, Management, and Market divergence estimates while increasing Risk, indicating that aspect-level disagreement is sensitive to the extraction method.

The audited sentence-ID experiment covered 11 documents with 0 fetch errors and 0 label failures. Exact-source validation found that 100.0% of selected sentences mapped back to source text. The run still produced 52 format violations and 30 invalid ID outputs; these invalid IDs were discarded, and only valid IDs mapped to original source sentences were scored. The result is therefore extractively grounded, while still indicating that instruction-following robustness remains an implementation concern.

Sentence-ID LLM extraction produced lower ADI values for Revenue, Management, and Market, but a higher ADI value for Risk. Therefore, aspect-level disagreement is sensitive to the extraction strategy. Under keyword extraction, Revenue remains the highest ADI aspect. Under sentence-ID LLM extraction, Risk becomes more salient. This does not invalidate MASD; it clarifies that MASD depends on the evidence-selection layer. In this thesis, keyword MASD is therefore best read as a transparent and reproducible baseline, while the implemented sentence-ID LLM audit evaluates a more semantically expressive and extractively grounded aspect-selection method for MASD.

4.6 Sentinel-LLaMA3 Benchmark Comparison

4.6.1 Protocol Caveats and Comparability

This section uses published benchmark results to place Sentinel-LLaMA3-FPB75 in context, but the rows are not directly comparable under a single shared protocol. Sentinel-LLaMA3-FPB75 is evaluated on the verified FPB 75%-agreement held-out split used in our thesis. The GPT-4o mini and FinSentLLM values are reported by Zhang et al. [9], while the HAD and FinBERT values come from their original evaluation protocols. The comparison should therefore be read as evidence of competitive in-domain benchmark performance, not as a controlled leaderboard across all systems.

4.6.2 Main Result

Sentinel-LLaMA3-FPB75 achieves macro F1 0.97526 on the FPB 75%-agreement test set, with $n = 691$. Relative to literature-reported results, this is 0.10436 above the GPT-4o mini score of 0.87090 reported by Zhang et al. [9] and 0.02526 above the approximate FinSentLLM macro F1 of 0.95, subject to protocol differences. Table 4.9 presents the benchmark comparison, and Figure 4.7 visualises the same macro-F1 comparison across model families.

Table 4.9: Macro-F1 benchmark comparison on FPB; protocol differences are discussed in Section 4.6.1.

Model	Macro F1	Type
FinBERT [5]	∼0.88	Fine-tuned BERT
GPT-4o mini [9]	0.87090	Proprietary LLM (zero-shot)
HAD / Xing [10]	∼0.93	Multi-agent LLM
FinSentLLM [9]	∼0.95	Training-free ensemble
Sentinel-LLaMA3-FPB75 (ours)	**0.97526**	Fine-tuned open-weight LLM

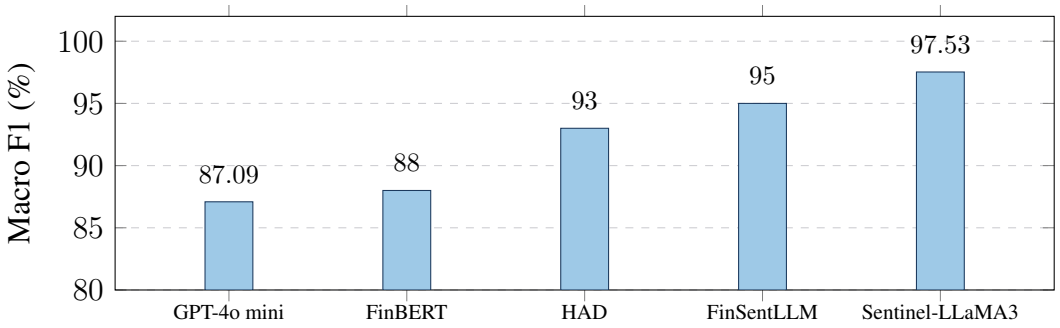

Figure 4.7: Benchmark comparison on FPB macro F1. Sentinel-LLaMA3-FPB75 is evaluated on the verified 75%-agreement split; other values are literature-reported under their original protocols.

Table 4.10: Sentinel-LLaMA3-FPB75 main evaluation metrics on the FPB 75%-agreement held-out test set.

Panel A: Evaluation protocol	
Model	Sentinel-LLaMA3-FPB75
Dataset	Financial PhraseBank
Subset	75%-agreement
Split	Stratified 80/20 train/test, seed 42
Test size	691

Panel B: Performance metrics	
Accuracy	0.97685
Macro precision	0.98672
Macro recall	0.96485
Macro F1	0.97526
Weighted precision	0.97744
Weighted recall	0.97685
Weighted F1	0.97669

Table 4.11 reports the audited lower-agreement evaluation sets and the resulting robustness metrics for Sentinel-LLaMA3-FPB75.

Table 4.11: Audited lower-agreement robustness evaluation for Sentinel-LLaMA3-FPB75.

Evaluation set	**Definition**	**Audit summary**
FPB66-exclusive	`Sentences_66Agree` \ `Sentences_75Agree`	$n = 763$; FPB75 overlap: 0 Correct / Incorrect: 561 / 202
FPB50-exclusive	`Sentences_50Agree` \ `Sentences_66Agree`	$n = 627$; FPB75 overlap: 0 Correct / Incorrect: 382 / 245
FPB50 – FPB75	`Sentences_50Agree` \ `Sentences_75Agree`	$n = 1{,}392$; FPB75 overlap: 0 Correct / Incorrect: 945 / 447

Metric	**FPB66-exclusive**	**FPB50-exclusive**	**FPB50 – FPB75**
Accuracy	0.73526	0.60925	0.67888
Macro precision	0.76648	0.63899	0.71461
Macro recall	0.70790	0.53491	0.63004
Macro F1	**0.72713**	**0.56338**	**0.65714**
Weighted F1	0.73060	0.59188	0.66977

When overlap with the FPB75 test setting is removed, evaluation on lower-agreement subsets shows a clear degradation in performance. This suggests that lower annotator agreement introduces additional ambiguity and label noise, and it does not support the earlier claim of 90% macro-F1 robustness at lower agreement thresholds. Table 4.12 reports the separate duplicate-safe Robust-FPB50 experiment used to test performance under a protocol designed for lower-agreement data.

Table 4.12: Evaluation metrics for Sentinel-LLaMA3-Robust-FPB50 under the duplicate-safe FPB 50% train/dev/test protocol. The corresponding protocol details and split counts are reported in Table 4.13.

Panel A: Model and protocol	
Model	Sentinel-LLaMA3-Robust-FPB50
Protocol	Duplicate-safe FPB50; stratified 70/10/20 split, seed 42
Test set size (n)	973
Panel B: Classification metrics	
Accuracy	0.87667
Macro precision	0.89331
Macro recall	0.85397
Macro F1	**0.87113**
Weighted precision	0.87769
Weighted recall	0.87667
Weighted F1	0.87478

Because the evaluation sets are not identical, we interpret this result as a protocol-specific improvement experiment, not as a direct paired comparison on the same test set. The audited Robust-FPB50 artifacts are stored under `ml/artifacts/fpb50_robust_llama3`, and the adapter is stored under `ml/checkpoints/fpb50_robust_llama3/adapter`.

Table 4.13: Duplicate-safe train/dev/test split counts for Sentinel-LLaMA3-Robust-FPB50.

Split	Rows	Negative	Neutral	Positive
Train	3,388	423	2,012	953
Dev	485	60	289	136
Test	973	121	578	274

Table 4.14: Duplicate-safety and training summary for Sentinel-LLaMA3-Robust-FPB50.

Item	Value
Total rows	4,846
Total normalised groups	4,838
Duplicate normalised sentence groups	8
Cross-split duplicate group count	0
Conflicting-label duplicate groups	2
Epochs	3.0
Final train loss	1.07985

Table 4.15 gives the corresponding Robust-FPB50 confusion matrix on the duplicate-safe test split.

Table 4.15: Confusion matrix for Sentinel-LLaMA3-Robust-FPB50 on the duplicate-safe FPB50 test split. Rows denote true labels and columns denote predicted labels.

Actual \ Predicted	Negative	Neutral	Positive	Support
Negative	106	15	0	121
Neutral	7	542	29	578
Positive	0	69	205	274
Predicted total	113	626	234	973

Figure 4.8 visualises the macro-F1 drop from the audited FPB75 test set to the non-overlapping lower-agreement evaluation sets.

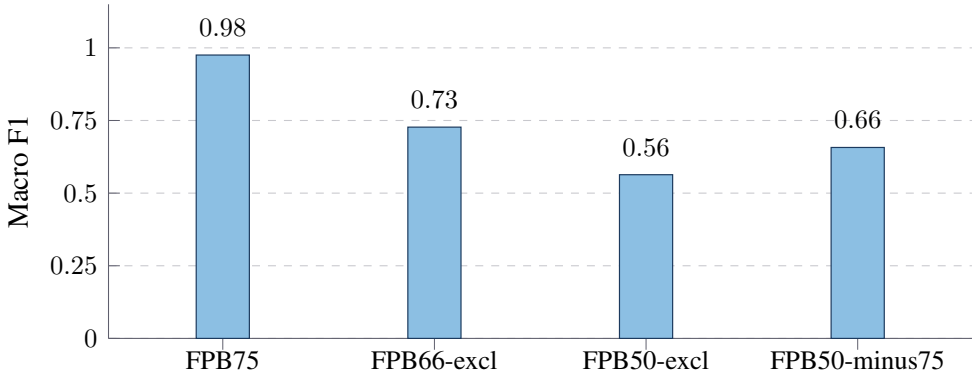

Figure 4.8: Macro F1 of Sentinel-LLaMA3-FPB75 on the audited FPB75 test set and non-overlapping lower-agreement evaluation sets.

4.6.3 Per-Class Evaluation

Table 4.16 reports the per-class precision, recall, and F1 scores for Sentinel-LLaMA3-FPB75 on the FPB 75% test set. Table 4.17 reports the corresponding confusion matrix.

Table 4.16: Sentinel-LLaMA3-FPB75 per-class evaluation on the FPB 75% test set, with overall accuracy of 0.97685.

Class	Precision	Recall	F1	Support
Negative	1.00000	0.96429	0.98182	84
Neutral	0.96614	0.99767	0.98165	429
Positive	0.99401	0.93258	0.96232	178
Macro avg.	**0.98672**	**0.96485**	**0.97526**	691
Weighted avg.	0.97744	0.97685	0.97669	691

Table 4.17: Confusion matrix for Sentinel-LLaMA3-FPB75 on the FPB 75%-agreement test set. Rows correspond to actual labels, and columns correspond to predicted labels.

Actual \ Predicted	Negative	Neutral	Positive	Support
Negative	81	3	0	84
Neutral	0	428	1	429
Positive	0	12	166	178
Predicted total	81	443	167	691

The dominant error mode in the confusion matrix is positive-to-neutral slippage, occurring in 12 of the 691 test examples, or approximately 1.7% of the full test set. This pattern is consistent with hedged positive language in financial reporting, where optimistic projections are often qualified by conditional clauses and risk disclosures.

4.7 Secondary Case Study: AppLovin Q4/FY2025

The AppLovin Q4/FY2025 case study applies the same FinBERT-based Sentinel pipeline to a second corporate event, allowing the framework to be examined in a setting distinct from the primary Tesla Q1 2024 case. AppLovin was selected for three reasons. First, it belongs to a different industry: ad-technology and AI-platform software rather than automotive manufacturing, which helps check whether the primary results are tied to the Tesla sector setting. Second, the Q4/FY2025 period combined strong financial performance with simultaneous regulatory scrutiny of the company's data practices, producing competing narrative threads across source types. These threads are analogous to, but not identical with, the earnings-miss and safety-recall dynamic in the Tesla case. Third, the AppLovin corpus includes a larger social/retail

layer (six documents) relative to the news layer (three documents), providing a different source-count composition from the Tesla corpus's balanced five-five-one structure. The case study is therefore used as a secondary exploratory application that tests the pipeline in a different event setting, with its results interpreted within that bounded scope.

4.7.1 Source Inventory

The AppLovin event corpus comprises ten documents: one SEC filing, three independent news items, and six social/retail discussion documents. All documents were processed using the same preprocessing, chunking, and FinBERT-based inference pipeline applied in the Tesla case study, without pipeline modification.

4.7.2 Per-Source Sentiment Scoring

The three source types differ in their average sentiment within the event window: news is the most negative, social/retail discussion is near-neutral, and the SEC disclosure is mildly positive. Table 4.18 reports the source-type prior, mean signed sentiment score, and dominant direction for each source type. The spread across the three averages is 0.625900 units on the $[-1, +1]$ scale, compared with 0.420 units in the Tesla Q1 2024 case. The larger spread reflects the difference between negative professional-news framing and mildly positive company disclosure language.

Table 4.18: AppLovin Q4/FY2025 source-type sentiment summary.

Source	π_τ	Mean signed score	Direction
News	0.7	-0.513705	Negative
Social/retail discussion	0.4	-0.081376	Negative (near neutral)
SEC disclosure	1.0	$+0.112195$	Positive (mildly)

4.7.3 SDI Computation

Table 4.19 presents the Level 1 and Level 2 SDI results for AppLovin Q4/FY2025. Both levels are classified as Partial Disagreement.

Table 4.19: AppLovin Q4/FY2025 SDI results by computation level.

SDI Level	Normalised score	Zone
Level 1 (unweighted)	0.226592	Partial Disagreement
Level 2 (credibility-weighted)	0.248746	Partial Disagreement

The Level 2 normalised score increases from 0.226592 to 0.248746 after credibility weighting, unlike the Tesla case, where the weighted score decreased only marginally. In the Ap-

pLovin corpus, this increase is driven primarily by disagreement between the news layer and the SEC disclosure.

4.7.4 ADI Decomposition

The AppLovin MASD extraction uses an event-specific aspect taxonomy reflecting four dominant narrative dimensions in the Q4/FY2025 event: Financial Performance, AI Platform Growth, Regulatory Data Risk, and Market Valuation. Table 4.20 summarises aspect coverage across the news, social/retail, and SEC source types.

Table 4.20: AppLovin Q4/FY2025 aspect coverage by source type.

Aspect	News	Social/retail	SEC	Coverage
Financial Performance	Covered	Covered	Covered	3/3
AI Platform Growth	Covered	Covered	Covered	3/3
Regulatory Data Risk	Covered	Covered	Covered	3/3
Market Valuation	Covered	Covered	Covered	3/3

Table 4.21 reports the per-source aspect-level signed sentiment scores together with the corresponding ADI values. Figure 4.9 visualises the same source-aspect sentiment matrix as a colour-coded heatmap. The directional divergence vector for the AppLovin Q4/FY2025 case is

$$\vec{D}(\text{AppLovin}, \text{Q4-FY2025}) = (0.029014,\ 0.056362,\ 0.010678,\ 0.012742), \tag{4.2}$$

where the entries correspond to Financial Performance, AI Platform Growth, Regulatory Data Risk, and Market Valuation, respectively.

Table 4.21: AppLovin Q4/FY2025 per-source aspect sentiment scores and ADI values.

Aspect	News ($\pi = 0.7$)	Social/retail ($\pi = 0.4$)	SEC disclosure ($\pi = 1.0$)	ADI	Coverage
Financial Performance	+0.583331	+0.197642	+0.252661	0.029014	3/3
AI Platform Growth	−0.375277	−0.032940	+0.202993	**0.056362**	3/3
Regulatory Data Risk	−0.207545	−0.160893	+0.031234	0.010678	3/3
Market Valuation	+0.057798	−0.052535	−0.216937	0.012742	3/3

Figure 4.9: AppLovin Q4/FY2025 source-aspect sentiment heatmap. Cell labels report signed values on the $[-1, +1]$ scale; darker saturation indicates larger magnitude.

Finding F-B1: AI Platform Growth is the most contested aspect. AI Platform Growth has the highest ADI in the AppLovin corpus, at 0.056362. The SEC disclosure scores this aspect positively, at $+0.202993$, aligning with its presentation of the AI advertising platform as a strategic growth driver. The news layer scores the same aspect negatively, at -0.375277, while the social layer is near neutral at -0.032940. This pattern indicates that retail investor commentary is less sharply directional on AI platform growth than either professional news or regulatory disclosure language.

Finding F-B2: Financial Performance is the only aspect with directional consensus across all source types. Financial Performance is the only aspect with positive sentiment across all three source types, with scores of $+0.583331$ for news, $+0.197642$ for social, and $+0.252661$ for the SEC disclosure. The resulting ADI of 0.029014 reflects differences in intensity among positive scores, not a directional split. This matters because the news layer is the most negative source type at the document level, with an average score of -0.513705. The aspect-level extraction therefore separates recognition of strong financial results from the broader negative framing of the event.

4.7.5 Comparison with the Tesla Primary Case

Two compositional features distinguish the AppLovin case from Tesla Q1 2024. First, the maximum ADI is 0.056362, substantially below the Tesla maximum of 0.199, indicating weaker aspect-level divergence despite the larger source sentiment spread. Second, credibility weighting raises the normalised SDI from 0.226592 to 0.248746 in the AppLovin case, whereas it produces a negligible decrease from 0.3563 to 0.35172 in the Tesla case. Both patterns trace to the larger social layer in the AppLovin corpus. Near-neutral social sentiment moderates the pairwise numerator under unweighted computation, while Level 2 down-weighting of this layer shifts relative weight toward the news-versus-SEC pair, which has both the highest pairwise weight and the largest pairwise distance in the corpus.

4.7.6 Exploratory Assessment

The AppLovin Q4/FY2025 case study shows that the Sentinel pipeline can be applied to a second corporate event without pipeline modification while still producing interpretable cross-source sentiment outputs. Its Level 2 normalised SDI of 0.248746 falls in the Partial Disagreement zone, consistent with the Tesla result, and the dominant source pair in both cases is professional news versus SEC disclosure. These parallels are best treated as a pattern for future testing in a larger multi-event corpus, rather than as evidence of a stable property of the financial information ecology. The AppLovin case adds a useful secondary check on the framework while retaining the scope boundaries of the current implementation. The social/retail layer is small and uneven, and the source inventory is intentionally limited because the case examines methodological transfer to a different event setting rather than population-level performance. Calibration has not been applied, and aspect extraction remains keyword-based, consistent with the current implementation described in Chapter 3. The case therefore supports the portability of the pipeline at proof-of-concept scale, with broader performance claims left to the multi-event validation study identified in Chapter 6.

4.8 Cross-Phase Interpretation

The unified results show that the framework's descriptive contribution and scoring-engine contribution are connected but analytically distinct. The Tesla case-study phases show that source-type variation can be preserved, measured, and decomposed: all three source types are negative on average, yet they differ substantially in intensity, source-pair contribution, and aspect-level structure. The Sentinel-LLaMA3 benchmark phases show that an open-weight fine-tuned scorer can achieve strong in-domain FPB performance while revealing lower-agreement robustness limits. The two strands therefore support different parts of the thesis. The case-study SDI/ADI results evaluate the framework's measurement logic on real multi-source events, while the Sentinel-LLaMA3 results establish benchmark performance on FPB and demonstrate pipeline integration through the Tesla scoring-engine comparison reported in Section 4.4.5.

The evidence supports a scoped conclusion: Sentinel surfaces interpretable disagreement on the selected Tesla event and produces coherent exploratory results on AppLovin. Broader claims about population-level SDI thresholds, predictive validity for market outcomes, calibrated cross-source score comparability, and general performance across corporate events remain reserved for the multi-event validation and calibration work identified in Chapter 6.

Chapter 5 interprets the findings from both evaluation tracks. The discussion addresses each research question using these consolidated results, examines what cross-source divergence implies about the structure of financial information, and situates the thesis contributions within the limitations of the current proof-of-concept implementation.

4.9 Unified Results Summary

Table 4.22 reports the combined results from all five pipeline phases, which support six key findings.

Table 4.22: Unified Sentinel evaluation results: Tesla Q1 2024 earnings event, 23 April 2024.

Phase	Metric	Value	Zone / Interpretation
Phase 1	FinBERT accuracy / macro F1 (FPB all-agree)	97.17% / 0.96000	Validated baseline
Phase 2	News avg. sentiment	−0.495	Negative
	Reddit avg. sentiment	−0.346	Negative
	SEC 8-K avg. sentiment	−0.075	Neutral
	Source spread	0.420 units	Pre-SDI cross-source variation
Phase 3	SDI Level 1 raw / normalised	0.7126 / 0.3563	Partial Disagreement
	SDI Level 2 raw / normalised	0.7034 / **0.35172**	Partial Disagreement
	SDI Level 3 (uncertainty-weighted) normalised	0.32323[a]	Partial Disagreement
	SDI L1–L2 delta	Raw 0.0092; Normalised 0.0046	Negligible credibility weighting effect
	Most divergent pair	News vs. SEC 8-K	Dist. 0.420; weighted contrib. 0.124
Sentinel	Sentinel-LLaMA3-FPB75 macro F1 (FPB 75%)	**0.97526**	+0.10436 vs. GPT-4o mini; +0.02526 vs. FinSentLLM
	FPB75-trained model macro F1 (FPB66-exclusive)	0.72713	Leakage-safe non-overlap audit
	FPB75-trained model macro F1 (FPB50-exclusive)	0.56338	Leakage-safe non-overlap audit
	Sentinel-LLaMA3-Robust-FPB50 macro F1	0.87113	Duplicate-safe FPB50 protocol
Phase 4	Revenue ADI	**0.199**	Highest keyword-based divergence; gap 1.091 units
	Management ADI	0.135	Potential directional split; low Reddit coverage (one sentence)
	Market ADI	0.097	Moderate divergence
	Risk ADI	0.013	Keyword-based near-consensus; 2/3 source coverage

[a]Level 3 uses MC Dropout mean scores; the appropriate L2-to-L3 comparison therefore uses the MC Dropout-baseline SDI Level 2 of 0.32251, rather than the deterministic Level 2 of 0.35172.

Chapter 5

Discussion and Conclusion

5.1 Discussion

The findings support Sentinel's central claim within the Tesla Q1 2024 case study: the type of information channel systematically shapes the sentiment characterisation of a financial event, and this variation can be quantified and decomposed. In this section, we address each research question in sequence before synthesising the results into a scoped structural conclusion.

5.1.1 Evidence Boundaries

The conclusions in this chapter are bounded by the evidence produced in our thesis. Direct empirical evidence comes from the FinBERT-based Tesla Q1 2024 primary case study, the FinBERT-based AppLovin Q4/FY2025 exploratory secondary case, and the Sentinel-LLaMA3 FPB benchmark experiments. Theoretical support for the economic relevance of disagreement comes from the literature reviewed in Chapter 2. Since the thesis establishes the framework at proof-of-concept scale, tasks such as testing whether SDI predicts future volatility, applying calibration gates and MC Dropout gate thresholds, and estimating population-level thresholds are left to future work. The claims below are therefore treated as conclusions about the case studies, benchmark experiments, and framework design, rather than as evidence for market-wide generalisation.

5.1.2 RQ1: Can SDI Quantify Meaningful Cross-Source Disagreement?

The Tesla Q1 2024 case study answers RQ1 affirmatively. The 0.420-unit spread in average sentiment across channels is quantified as a structured feature of the event's information environment rather than treated as random noise. The SDI formalises this variation as a scalar disagreement score, yielding a Level 2 normalised value of 0.35172 (Partial Disagreement), which is bounded, theoretically comparable across events, and decomposable into source-pair contributions.

The primary value of the SDI lies less in the isolated numerical result, which requires a multi-event distribution for full interpretation, and more in its ability to preserve cross-source variation. Conventional single-score FSA would reduce heterogeneous source evidence to one aggregate score. By contrast, Sentinel keeps the underlying source-level divergence measurable and available for interpretation.

Level 1 and Level 2 SDI give the same interpretation in this case: both fall in the Partial Disagreement zone. This suggests that the observed disagreement is stable under the implemented weighting scheme, but it should not be read as evidence that credibility weighting is unnecessary in general. The effect of weighting remains event-specific. As shown in Table 3.5, varying π_{news} and π_{Reddit} (with $\pi_{\text{SEC}} = 1.0$) keeps the normalised SDI within $[0.34764, 0.35628]$. This supports zone-level stability under reasonable prior perturbations, while broader grid sensitivity testing remains future work for multi-event evaluations.

5.1.3 RQ2: Aspect-Level Structure

RQ2 is answered affirmatively, with an important sensitivity qualification. The Management aspect illustrates how ADI can identify directional splits, although the Reddit estimate for this aspect has low reliability because it is based on one extracted sentence; see Section 5.2.5. At the document level, Reddit's overall sentiment (-0.346) and news overall sentiment (-0.495) are both negative, while the SEC 8-K (-0.075) is near-neutral. The SDI therefore records these sources as differing mainly in sentiment intensity. At the aspect level, however, Reddit and news move in opposite directions on Management. Reddit scores Management at $+0.565$, reflecting a Reddit narrative centred on FSD licensing as evidence that management strategy was being validated by external partners. News scores the same aspect at -0.334, reflecting coverage of chief executive travel cancellations and workforce restructuring as evidence of leadership instability. The Management result should therefore be treated as illustrative due to low Reddit coverage and extraction-method sensitivity, but it shows why ADI is useful: it can reveal aspect-level directional disagreement that a document-level SDI score summarises only as cross-source difference.

The Revenue finding provides the largest source-pair spread in the dataset, at 1.091 units between Reddit (-0.948) and the SEC 8-K ($+0.143$). This contrast shows how source type affects aspect-level sentiment even when the underlying financial fact, a record quarterly revenue decline, is the same across sources. Reddit frames the earnings miss in strongly negative terms, whereas the SEC 8-K presents the same revenue figures through item-by-item factual disclosure and forward-looking-statement disclaimers. The model therefore assigns the filing a slightly positive score, partly because its language places the decline within a longer-term growth narrative. The 1.091-unit spread is the largest source-pair contrast observed in the evaluation and would be lost under a single-score sentiment framework.

The Risk aspect shows that low disagreement can also be informative under keyword extrac-

tion. Its near-zero keyword-based ADI of 0.013 indicates that the evaluated keyword-selected evidence characterises this dimension in broadly similar terms across covering sources. However, the sentence-ID LLM audit in Section 4.5.4 increases Risk ADI to 0.073896 and makes Risk more salient under semantic extractive selection. This does not undermine ADI/MASD; it clarifies that the metric is conditional on the evidence-selection layer and should not be interpreted as method-invariant at proof-of-concept scale.

5.1.4 RQ3: Sentinel-LLaMA3 as a Scoring Engine

RQ3 is answered affirmatively for the in-domain FPB 75% benchmark. Sentinel-LLaMA3-FPB75 achieves macro F1 0.97526, which is 0.10436 macro-F1 points above the literature-reported GPT-4o mini zero-shot score, subject to the caveat that the two values do not come from a single controlled same-test leaderboard experiment. The result is consistent with domain-adaptation findings from FinBERT [5] and FLANG [7]: supervised training on in-domain financial text can improve domain-specialised classification performance in ways that general pretraining scale alone does not guarantee.

Second, the audited non-overlap robustness evaluation shows a clear performance drop on lower-agreement examples that were not present in the FPB 75% training data: macro F1 0.72713 on FPB 66%-exclusive, 0.56338 on FPB 50%-exclusive, and 0.65714 on FPB 50%-minus-FPB 75. These results replace the earlier unsupported 0.90 robustness claim with a more conservative finding: annotator ambiguity remains a substantive limitation. Third, the separate Sentinel-LLaMA3-Robust-FPB50 experiment achieves macro F1 0.87113 on its duplicate-safe FPB 50% test split, indicating that training on lower-agreement examples improves performance under a protocol designed for that setting. The resulting 160 MB adapter file also avoids reliance on proprietary API access, reducing reproducibility and deployment barriers relative to GPT-4o mini-based baselines.

Sentinel-LLaMA3 is incorporated into the case-study scoring pipeline through token log-probability extraction, as reported in Section 4.4.5. Rather than generating a label token, the model extracts logits for the three label tokens at the final prompt position and applies a three-way softmax, yielding the signed score $s_i = p_{\text{pos}} - p_{\text{neg}} \in [-1, +1]$ without any additional fine-tuning beyond the original QLoRA training. The resulting SDI is 0.25434, compared with the FinBERT baseline value of 0.35172, and both values remain in the Partial Disagreement zone. This suggests zone-level stability under the scoring-engine comparison performed in our thesis. Direct regression over $[-1, +1]$ remains a separate integration path for future implementations that require continuous score outputs without log-probability extraction.

The scientific contribution of Sentinel-LLaMA3-FPB75 is therefore threefold: it demonstrates that an open-weight locally-deployable model can match proprietary zero-shot baselines on domain-specific benchmarks after parameter-efficient fine-tuning; it establishes reproducibility without proprietary API access; and the scoring-engine comparison in Section 4.4.5

suggests zone-level stability in this comparison between FinBERT and Sentinel-LLaMA3 as the scoring engine. The primary Tesla and AppLovin SDI/ADI values use FinBERT because its chunk-averaging pipeline was validated in Phase 1 and its three-class probability vector is natively compatible with the signed score formula $s_i = p_{\text{pos}} - p_{\text{neg}}$. Full replacement of FinBERT as the primary case-study scorer is the immediate next implementation step.

5.1.5 RQ4: Editorial Constraints or Social Media Noise?

RQ4 produces one of the main interpretive results of the Tesla Q1 2024 evaluation. The near-identity of SDI Level 1 and Level 2, with a normalised decrease of 0.0046, suggests that the observed disagreement is not primarily driven by Reddit noise. Credibility weighting is designed to attenuate disagreement attributable to lower-credibility sources: assigning Reddit $\pi_{\text{Reddit}} = 0.4$ reduces its effective influence on the aggregate SDI. If social media noise were the main driver of disagreement in this case study, this attenuation should produce a larger reduction in the Level 2 SDI relative to Level 1. Instead, the pairwise contribution breakdown shows that the dominant disagreeing pair is News, with $\pi = 0.7$, versus the SEC 8-K, with $\pi = 1.0$. Their combined pairwise weight is 0.70, so this pair dominates both SDI levels and Reddit's reduced prior has negligible effect on the aggregate score.

The interpretation is structural: the disagreement is not primarily explained by a high-noise, low-credibility social media source producing an outlier signal. Instead, it is associated with two institutionally constrained source types, professional news and a regulatory filing. News coverage tends to foreground consequence, market reaction, and narrative salience; it scored the Tesla Q1 2024 event at -0.495, reflecting the earnings miss, the safety recall, and the workforce restructuring as the dominant narrative threads. The regulatory filing is legally constrained to present material facts in measured, liability-aware language; it scored the same event at -0.075, reflecting the qualification of negative statements and the compression of sentiment in formal disclosure language.

Neither source is necessarily inaccurate. Rather, the sources are constrained by different communicative mandates, and those mandate differences explain the observed cross-source sentiment disagreement in this case study. For sentiment-based financial risk modelling, this means that source type is part of the signal, not only a source of noise. Aggregating heterogeneous sources into a single score may therefore remove variation that reflects institutional mandate rather than random informational noise.

5.1.6 Implications for Financial Risk Modelling and Sentiment Analysis

The results collectively have four implications for the broader FSA community.

For investors and risk managers. The SDI and ADI distinguish negative sentiment from cross-source disagreement. A low aggregate score suggests uniform negativity, whereas a high SDI indicates that sources interpret the event differently. For portfolio managers and credit-risk analysts, this reveals whether an event is unambiguously negative or contested across channels. The ADI then isolates where contestation occurs. For example, a high Management ADI indicates a divided leadership narrative, while a low Risk ADI suggests consensus. Together, these metrics structure event-level disagreement and its aspect concentration.

For financial NLP researchers. The framework supports event-study designs in which cross-source variation is treated as the variable of interest rather than as noise to be removed. The downstream validation design specified in Chapter 3, where SDI is regressed on abnormal return volatility, makes the relationship between informational heterogeneity and market outcomes empirically testable. This design builds on Garcia [15], who showed that sentiment divergence between news and financial-platform sources predicts financial distress at a 46-basis-point effect size. Sentinel extends this line of work by moving from a two-source divergence setting to a multi-source, aspect-decomposed, and credibility-weighted framework. It therefore provides a basis for studying not only whether sources disagree, but also which source types diverge and on which financial aspects.

For practitioners building production sentiment systems. The case-study result suggests a practical design requirement for financial NLP pipelines: source type should be retained during sentiment computation rather than collapsed at the preprocessing stage. Aggregating regulatory filings, professional news, and social media into a single corpus implicitly assumes that these channels are comparable sentiment producers. The Tesla Q1 2024 case study shows why this assumption is problematic. The 0.420-unit spread across source-type averages reflects systematic differences associated with institutional mandate, not only random variation. Preserving source identity and reporting disagreement alongside the aggregate score can therefore give practitioners a clearer view of the event's information environment without requiring additional data collection.

For regulatory-grade financial intelligence. These implications are research directions, not production-readiness claims. In the current proof-of-concept, the SDI and ADI outputs make disagreement inspectable by showing which source types and financial aspects drive it. This is relevant to the transparency concerns discussed in Chapter 2, but deployment in risk-sensitive settings would still require the validation steps described in Chapter 6.

5.2 Limitations

The current implementation is intentionally bounded in several respects. These boundaries define the scope of the present evaluation rather than limitations of the underlying approach, and each has a corresponding extension path. Table 5.1 maps each boundary to the current evaluation status and the corresponding extension path.

Table 5.1: Scope boundaries of the current Sentinel implementation and corresponding extension paths.

Scope boundary	Current evaluation	Extension path
Single-event scope	Tesla Q1 2024 primary; AppLovin Q4/FY2025 exploratory	Multi-event validation corpus spanning diverse industries, company sizes, and event types
Source calibration	Unadjusted FinBERT scores; cross-source calibration deferred	Isotonic regression calibration fitted on a held-out source-type calibration set
Aspect extraction	Keyword-based filtering plus Tesla Q1 2024 sentence-ID LLM sensitivity audit	Sentence-ID or span-grounded LLM extraction integrated across events
Data coverage	Single-publisher news sampling (Guardian); restricted Reddit body-text availability	Multi-publisher news collection; full post body persistence at collection time
Predictive and cross-benchmark validation	Scoring evaluation bounded to FPB; market regression deferred to future work	Downstream event-study regression; cross-benchmark evaluation on FiQA-SA and SEntFiN

5.2.1 Single-Event Scope

The primary SDI and ADI results derive from a single corporate event, supplemented by AppLovin as an exploratory secondary application. The Tesla Q1 2024 earnings announcement was selected specifically as a high-disagreement stress test. Consequently, the observed Level 2 normalised SDI of 0.35172 establishes a proof of concept, but its relative magnitude requires an empirical distribution for full interpretation. Future multi-event validation will contextualise this disagreement and enable the SDI-to-volatility regression that forms the framework's primary falsifiable hypothesis. A validated corpus spanning diverse industries, company sizes, and event types is the foundation for this extension.

5.2.2 Source Calibration Not Applied

The current implementation defers the isotonic regression calibration specified in Chapter 3. As a result, the absolute SDI magnitude rests on the controlled assumption that FinBERT's signed score distributions are comparable across source types. If the model's baseline calibration differs systematically between formal regulatory filings and informal social media text, some observed divergence may reflect calibration artefacts rather than true sentiment disagreement. Applying the specified calibration protocol on a held-out set provides the necessary extension to test this assumption and strengthen future absolute SDI magnitude claims.

5.2.3 MC Dropout Uncertainty Estimates

The current implementation executes the Level 3 SDI formulation, incorporating instance-level uncertainty estimation via MC Dropout (Section 4.4.3). The applied procedure uses $N = 30$ stochastic forward passes per document, yielding uncertainty estimates σ_i that range from 0.00796 (news article, 2 chunks) to 0.14021 (Reddit post, 1 chunk). Consistent with its institutional constraints, the SEC 8-K filing exhibits the lowest uncertainty ($\sigma_i = 0.01445$) across its 16 chunks. The resulting Level 3 normalised SDI is 0.32323, which remains in the Partial Disagreement zone. Because the current phase uses a fixed N, sensitivity analysis of this parameter ($N \in \{10, 20, 30, 50\}$) is deferred to future work.

5.2.4 Keyword-Based Aspect Extraction

Aspect-specific sentiment in Phase 4 is extracted using keyword-based sentence filtering as the primary reported MASD baseline. This approach has three principal limitations. First, it misses implicit or indirect aspect mentions. Second, it produces false positives when keywords appear in irrelevant contexts. Third, it cannot distinguish incidental aspect mentions from those that drive the primary sentiment. Sentence-ID LLM aspect extraction was implemented as an extractive MASD sensitivity audit for Tesla Q1 2024 and compared against the verified keyword-based MASD baseline (Section 4.5.4), showing that semantic extractive selection can materially change aspect rankings while preserving source-sentence grounding. Scaling this implemented audit into a default sentence-ID or span-grounded LLM MASD pipeline across events remains future work, including stronger instruction-following controls and validation across source types.

5.2.5 Reddit Post Body Unavailability

In Phase 4, the Reddit corpus was constrained by restricted body-text availability during the initial data collection. This sparsity resulted in zero Risk-relevant sentences and a Reddit Management score derived from a single extracted sentence, rendering these specific aspect estimates less robust than their news and SEC 8-K counterparts. Resolving this constraint by persisting full post bodies at collection time is a direct structural extension. In the current evaluation, the Reddit Management score of $+0.565$ serves as an illustrative low-sample estimate, and the resulting directional split should therefore be interpreted with appropriate caution.

5.2.6 Single-Publisher News Approximation

The current implementation represents the professional news source type exclusively through Guardian coverage. Because different publishers may exhibit systematic framing differences, expanding this single-publisher sample is necessary to measure within-news-type variation. A multi-publisher corpus, incorporating sources such as Reuters, the Associated Press, and the Financial Times, is the designated extension path. This expansion will provide a broader sentiment estimate and enable direct within-source-type consistency analysis.

5.2.7 Benchmark Generalisability of Sentinel-LLaMA3

In the current implementation, Sentinel-LLaMA3 evaluation is scoped exclusively to FPB subsets. Consequently, cross-benchmark evaluation on FiQA-SA, SEntFiN, and FinEntity remains a necessary extension. This broader testing will determine whether the performance improvements achieved on FPB generalise reliably to financial sentiment benchmarks that employ different annotation schemes, domain coverage, and class distributions.

5.2.8 Absence of Multi-Event Regression Evidence

The framework's core falsifiable hypothesis posits that SDI is a statistically significant predictor of short-term Abnormal Return Volatility (ARVol). Because a single-event case study cannot provide statistical validation, testing this hypothesis is deferred to future work. Consequently, the predictive value of SDI currently rests on theoretical motivation and established empirical literature [15, 27]. Executing the nested regression framework (Chapter 3) across a multi-event corpus remains the primary experimental step required to generate direct evidence for this hypothesis.

Crucially, these scope boundaries are paired with specific extension paths and do not diminish the framework's structural claims within its intended proof-of-concept parameters. The concluding remarks assess these structural claims directly.

5.3 Concluding Remarks

In this thesis, we presented Sentinel, an end-to-end framework for MSFSA that formalises, computes, and decomposes cross-source sentiment disagreement. The framework's primary contributions span both disagreement measurement and aspect-level decomposition. First, the SDI introduces a credibility-weighted, uncertainty-aware scalar measure of cross-source divergence. Supported by a rigorous mathematical formulation and an $\mathcal{O}(n)$ algebraic closed form, the SDI establishes a four-level methodological progression from unweighted to dynamic credibility-aware computation. Second, the MASD framework decomposes document-level disagreement into per-aspect divergence via the ADI, yielding a directional vector that constitutes a multi-dimensional sentiment portrait of a financial event.

In its current evaluation, the implementation successfully executes SDI Levels 1 through 3. The Level 3 computation, incorporating MC Dropout uncertainty estimation, yields a normalised SDI of 0.32323 (Section 4.4.3), placing the primary case study clearly within the Partial Disagreement zone. The Level 4 dynamic weighting mechanism is deferred to future work.

Methodologically, Sentinel-LLaMA3-FPB75 demonstrates that QLoRA fine-tuning of an open-weight 8B model can surpass the literature-reported GPT-4o mini zero-shot baseline on the FPB 75% benchmark, achieving high performance while remaining locally deployable and independent of proprietary APIs. This scoring capability is explicitly bounded to high-agreement settings, as evidenced by material degradation on lower-agreement subsets.

Empirically, applying the full pipeline to the Tesla Q1 2024 event yields a Partial Disagreement SDI of 0.35172. The resulting keyword-based directional ADI vector identifies Revenue as the most contested aspect dimension. Management emerges as the sole aspect exhibiting a potential directional split under the keyword baseline, though this specific vector requires future validation against a higher-coverage Reddit text corpus and the extraction-method sensitivity identified in the sentence-ID LLM audit.

Theoretically, these findings support a scoped structural interpretation: the observed cross-source disagreement stems from institutional mandates and editorial framing rather than mere social media noise. Professional news and regulatory filings are both credible sources describing the same event; they diverge because they operate under fundamentally different communicative obligations. These findings caution against treating cross-source variation in multi-source sentiment modelling solely as calibration noise or retail noise. Instead, sentiment variation itself, when measured and decomposed explicitly, carries economically meaningful information that warrants testing across a larger event corpus.

Our review of 36 studies identified a clear gap: no prior published work occupies the intersection of multi-source analysis, aspect-level sentiment decomposition, and formal divergence measurement. Sentinel addresses this gap at a proof-of-concept scale through a mathematically grounded, reproducible pipeline evaluated on real financial data. In doing so, the framework

provides a methodological foundation for further financial NLP work to treat disagreement structure as an analytical object. In this paradigm, the structure of disagreement, defining precisely which source types diverge on which aspects and in what direction, is elevated from a preprocessing nuisance to primary empirical evidence.

Chapter 6

Future Work

This thesis presents Sentinel as a mathematically grounded framework for MSFSA. Having evaluated the core analytical components in Chapter 5, including the SDI formulation, the MASD protocol, and the Sentinel-LLaMA3 scoring engine, the framework is now ready for broader empirical testing. The following research directions outline how these components can be scaled from proof-of-concept case studies toward broader empirical validation.

6.1 Multi-Event Validation and Regression Design

With the mathematical formulation and single-event evaluation of the SDI now established, the natural progression is to scale the framework for broader market analysis. The Tesla case study demonstrates the metric's computability by yielding a Level 2 normalised SDI of 0.35172. By constructing a validated multi-event corpus, future research can establish a population-level distribution for this metric. This empirical baseline would allow future studies to benchmark corporate events and examine whether particular SDI levels are statistically unusual.

A practical future corpus should span approximately 200 to 500 company-event instances drawn from at least three heterogeneous source channels per event. This dataset requires stratification across event types, including earnings announcements, regulatory actions, executive leadership changes, and material corporate transactions. Data collection can leverage the existing pipeline components: SEC EDGAR for regulatory filings, Arctic Shift for Reddit social media, and multi-publisher professional news (Reuters, the Associated Press, the Financial Times, and the Guardian). To guard against look-ahead contamination, the study must employ a temporal holdout split, training on earlier events and evaluating on later ones.

The structured output records produced by this corpus will serve as the primary input for testing the framework's core falsifiable hypothesis: that the normalised SDI ($\mathrm{SDI}_{w,\mathrm{norm}}$) serves as a statistically significant predictor of short-term Abnormal Return Volatility (ARVol). This

conjecture is theoretically grounded in the investor disagreement literature, particularly the work of Miller [33], and in the empirical finding of Garcia [15] that sentiment divergence predicts financial distress. The validation study follows a nested regression design. In the first stage, ARVol is regressed on $\mathrm{SDI}_{w,\mathrm{norm}}$ alongside standard event-study covariates, including event type, market capitalisation, and a volatility proxy such as the VIX. In the second stage, the ADI component vector $\vec{D}(c,t)$ is added as a block of additional predictors, enabling a test of whether aspect-level divergence provides incremental explanatory power over the document-level disagreement score.

Successful validation requires a statistically significant positive coefficient on $\mathrm{SDI}_{w,\mathrm{norm}}$ in the first stage, followed by a significant incremental R^2 from the ADI block in the second stage. Notably, a null result in the first stage would be equally informative, indicating that cross-source divergence is structurally descriptive rather than independently market-predictive. Rooted in established empirical evidence [15, 27], the direct validation of this hypothesis aims to produce an operational disagreement signal. If validated, this would move Sentinel beyond descriptive analysis toward a predictive research instrument for event-driven risk analysis.

6.2 Dynamic Credibility Weighting

The Level 4 SDI formulation is specified in Chapter 3 and represents a natural extension of the current framework. It addresses the primary limitation of static prior weighting: a universal π_τ value cannot distinguish between professional news outlets with varying editorial standards, nor can it correct for a specific source exhibiting systematic bias on certain event types. To resolve this, Level 4 introduces evasion scoring, historical track records, and market regime awareness as dynamic credibility modifiers.

These three modifiers adapt the credibility weights to empirical source behaviour. Evasion detection operationalises the ev_i component by measuring the deviation of a source's current sentiment from its historical baseline on comparable events; sources that systematically diverge receive elevated evasion penalties. Track-record computation leverages the multi-event corpus (Section 6.1) to calculate the Pearson correlation between a source's historical SDI contributions and subsequent ARVol, scaling the credibility prior proportionally. Finally, regime-switching detection applies separate adjustment curves for low, medium, and high market-volatility states, capturing the empirical reality that credibility differentials often widen during high-stress market periods.

The recommended prerequisite for full Level 4 implementation is a sensitivity analysis on static π_τ variation, isolating the baseline contribution of fixed priors before introducing dynamic complexity. A successful Level 4 deployment requires a statistically significant reduction in SDI prediction error during volatility regression relative to the Level 2 baseline. If

successful, this extension would produce adaptive disagreement scores informed by historical source behaviour rather than fixed expert-assigned priors.

6.3 Sentence-ID and Span-Grounded Aspect Extraction

The Tesla Q1 2024 sentence-ID LLM audit shows that MASD is sensitive to the evidence-selection layer: Revenue remains most divergent under keyword filtering, while Risk becomes more salient under semantic extractive selection. Future work should therefore scale this implemented audit into the default MASD pipeline across events, using sentence-ID or span-grounded LLM aspect extraction in place of keyword filtering. The extraction model should select only source-grounded sentence IDs or explicit text spans, which are then mapped back to original documents before sentiment scoring.

This extension must also improve instruction-following robustness. The audited sentence-ID run preserved exact-source grounding after invalid IDs were discarded, but its format violations and invalid-ID outputs show that constrained decoding, schema validation, retry logic, and conservative discard rules are necessary before production use. Validation should compare keyword, sentence-ID, and span-grounded extraction across more events and source types, including professional news, regulatory filings, Reddit, earnings calls, and analyst reports. The goal is not to make ADI method-invariant, but to document how extraction choices affect aspect rankings and to select the most reliable evidence-selection protocol for future multi-event studies.

6.4 Temporal Propagation Modelling

The current framework treats all documents within a five-day event window as temporally simultaneous, compressing the ordered diffusion of sentiment signals across source types into a single aggregated SDI score. This is a deliberate simplification. In reality, SEC filings, earnings call transcripts, professional news articles, and social media posts arrive in a structured sequence: regulatory disclosures typically precede formal news coverage, which subsequently overlaps with peak social media response.

To capture this sequential dynamics, future work should model this diffusion as a directed temporal graph. In this graph, nodes represent per-source sentiment time series and directed edges represent statistically significant lead-lag relationships. The input is a rolling SDI computed over sliding six-hour windows within the event period, generating time series from which Granger causality tests can estimate directed propagation coefficients. To ensure validity, the sensitivity of these coefficients to window size variations (from two to twenty-four hours) must be evaluated as a core robustness check.

The analytical goal of this temporal modelling is to establish whether institutional sentiment consistently leads retail sentiment across specific event types. A finding of consistent institutional-to-retail Granger causality would empirically validate this propagation hypothesis. Conversely, detecting event types where social media sentiment leads institutional coverage would identify critical scenarios where retail channels anticipate the formal news cycle. The potential value of this extension is temporal interpretation: it could show whether institutional disagreement tends to precede retail sentiment, or whether retail discussion sometimes anticipates formal coverage.

6.5 Extended Data Source Coverage

To improve the external validity of the Sentinel framework, future work should expand the current three-source inventory to include earnings call transcripts and sell-side analyst reports. Earnings calls occupy a unique hybrid register: their legally reviewed opening remarks mirror SEC filings, while their unscripted question-and-answer segments resemble professional news. This allows the SDI to test for intra-source register variation. Conversely, analyst reports offer forward-looking interpretations and explicit valuation targets, producing sentiment that closely tracks market price formation. Acquiring these sources is feasible in principle; earnings calls are accessible as 8-K exhibits via the existing `edgartools` pipeline, while analyst reports are available through academic databases and commercial terminals (e.g., Bloomberg).

Incorporating these channels requires the empirical calibration of new credibility priors. Earnings calls necessitate a prior reflecting their dual-register nature, while analyst reports require a penalised prior to account for documented sell-side optimism bias. To validate this five-source extension, future implementations must demonstrate a measurable increase in SDI discriminative validity on high-controversy events, confirmed by a non-redundancy test ensuring the new channels contribute mathematically independent signals.

6.6 Source-Type-Aware Fine-Tuning and Cross-Benchmark Evaluation

Because Sentinel-LLaMA3 was fine-tuned exclusively on the FPB news corpus, its application to SEC filings and Reddit posts exposes the pipeline to register-induced domain shift. Regulatory text relies on liability hedging, while social media employs sarcasm and intensity amplification; neither is represented in the FPB. To mitigate this, future work must develop a source-type-aware fine-tuning regime. The simplest methodology conditions the instruction prompt with a register prefix (e.g., `[SOURCE: SEC_FILING]`). A more rigorous architecture would employ source-specific QLoRA adapters, applied conditionally during inference to

prevent cross-register gradient contamination. Both approaches require assembling a multi-register training corpus.

To validate this register-aware engine, it must undergo cross-benchmark evaluation on FiQA-SA, SEntFiN, and FinEntity to probe distinct generalisation axes. Success requires a consistent macro F1 improvement across these benchmarks compared to the FPB-only baseline, alongside reduced isotonic calibration error on held-out SEC and Reddit data. If successful, this would support a scoring engine with more comparable sentiment outputs across source types, reducing reliance on post-hoc calibration and strengthening the SDI computation.

Appendix

A.1 Aspect Keyword Lists

Keyword-based sentence filtering in Phase 4 aspect extraction used the lists reported in Table A.1. Each sentence was assigned to an aspect when it contained at least one keyword from the corresponding list, with matching performed case-insensitively. The same four lists were used for both the Tesla Q1 2024 and AppLovin Q4/FY2025 case studies. For AppLovin, the event-specific labels map directly to the shared lists: Financial Performance to Revenue, AI Platform Growth to Management, Regulatory Data Risk to Risk, and Market Valuation to Market.

Table A.1: Aspect keyword lists used for Phase 4 keyword-based sentence filtering.

Aspect	Keywords
Revenue	revenue, sales, income, earnings, profit, margin, growth, decline, miss, beat, forecast, guidance, quarter, annual, fiscal
Mgmt.	CEO, management, executive, leadership, board, director, governance, strategy, restructur, layoff, workforce, travel, Musk, guidance, credibility, decision
Risk	recall, crash, fatality, safety, NHTSA, regulation, legal, liability, lawsuit, penalty, investigation, compliance, risk, hazard, audit
Market	stock, share, price, market, valuation, surge, drop, rally, sell-off, trading, volume, investor, capitalization, analyst

A.2 API Collection Details

This appendix documents the source-collection routes used for the primary Tesla Q1 2024 case study. These details are reported to support reproducibility.

Guardian news. Five Guardian Open Platform articles matching the query term `tesla` were retrieved within the 21–26 April 2024 event window. The retrieved fields included article body text, headline, publication date, and URL. The retained items were ranked by relevance within the collection route.

Reddit social media. Five Reddit posts were retrieved from the Arctic Shift Reddit archive across finance-oriented subreddits using `tesla` or `tsla` as query terms. Candidate posts were filtered to the 21–26 April 2024 window and ranked by upvote score. As noted in the limitations, full post body text was not persisted for all Phase 4 aspect extraction inputs.

SEC EDGAR. The Tesla Q1 2024 Form 8-K was retrieved from SEC EDGAR using accession number `0000950170-24-046895`, filed on 23 April 2024. The filing was extracted with `edgartools` and split into 16 chunks for FinBERT-compatible inference.

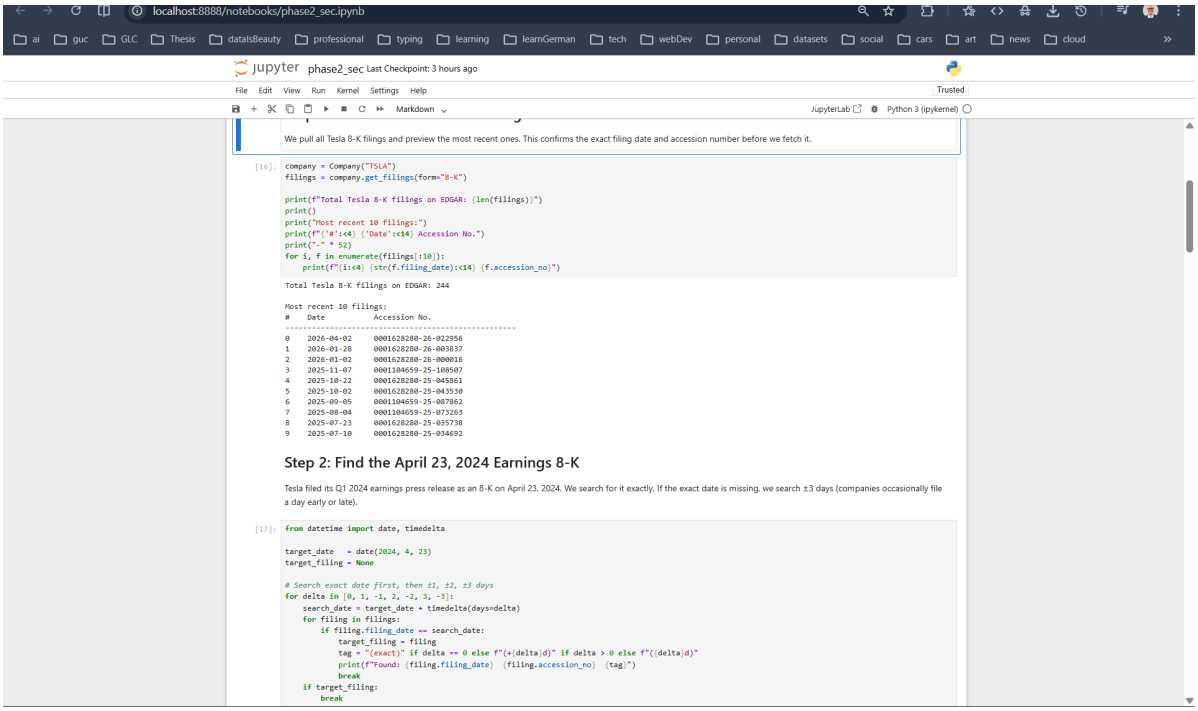

Figure A.1: Tesla 8-K filing history via `edgartools`. The 23 April 2024 filing is the Q1 2024 earnings press release used in the primary case study (Table 3.6).

A.3 Extended Training Environment

This appendix documents the software and hardware environment used for Sentinel-LLaMA3-FPB75 training. The values are reported as reproducibility metadata for the benchmark run and should be treated as environment documentation rather than as additional experimental results.

Table A.2 summarises the extended Sentinel-LLaMA3-FPB75 training environment.

Table A.2: Extended training environment for Sentinel-LLaMA3-FPB75.

Item	Recorded value
Hardware	NVIDIA GeForce RTX 4090
CUDA availability	true
PyTorch	`2.6.0+cu124`
Transformers	`4.45.2`
Datasets	`2.18.0`
PEFT	`0.13.2`
TRL	`0.8.6`
BitsAndBytes	`0.49.2`
Accelerate	`0.34.2`
Logging and saving	logging every 10 steps; save strategy by epoch
Optimiser	AdamW fused

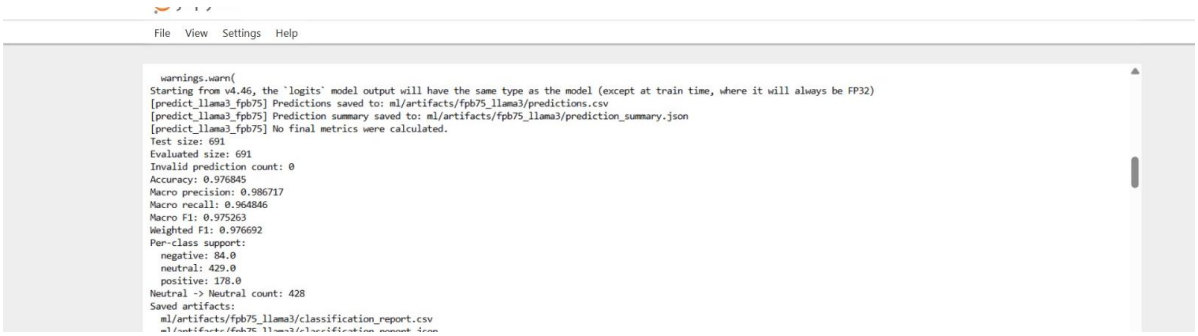

Figure A.2: Evaluation terminal output ($n = 691$). Confirms Table 4.10.

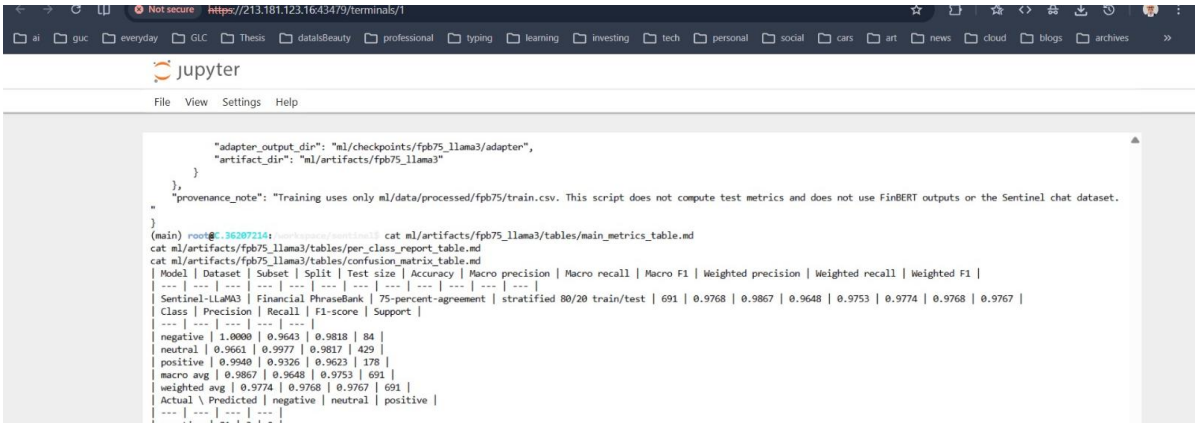

Figure A.3: Per-class metrics and partial confusion-matrix output. Confirms Table 4.16; the complete confusion matrix is reported in Table 4.17.

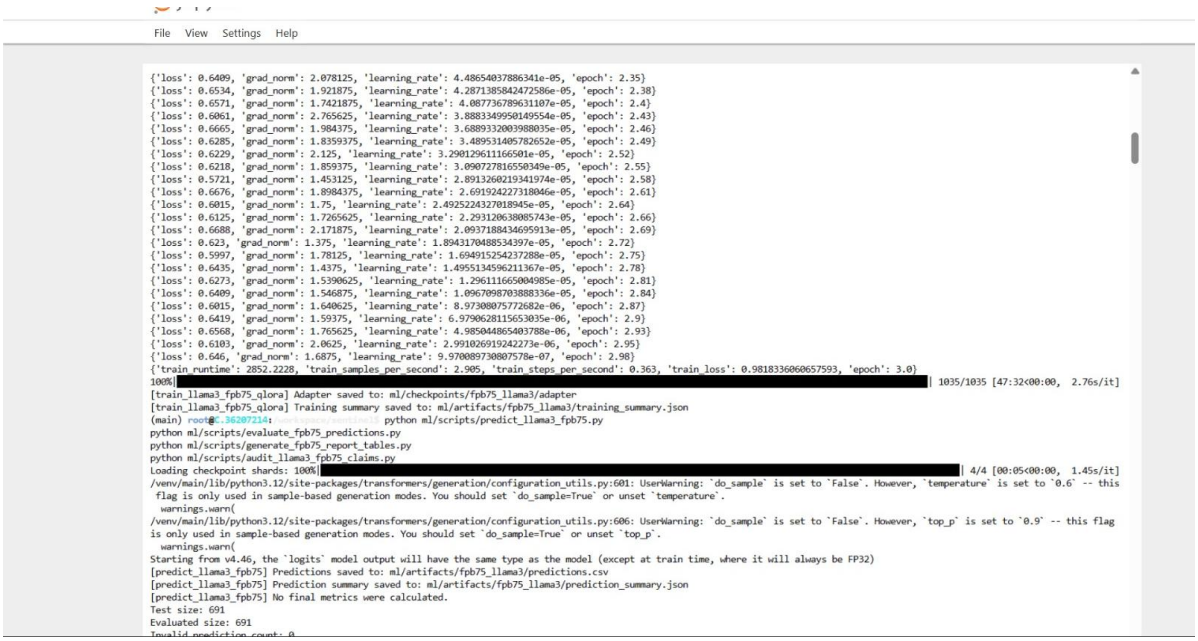

Figure A.4: Sentinel-LLaMA3-FPB75 training completion output.

A.4 AppLovin Secondary Case Study Supporting Outputs

The AppLovin Q4/FY2025 case study is reported in Chapter 4 as a secondary exploratory validation. Tables A.3 and A.4 provide the corresponding source-aspect sentiment matrix and ADI ranking for reproducibility. Chapter 4 discusses the derived findings rather than repeating these raw supporting values.

Table A.3: AppLovin secondary case source-aspect matrix.

Aspect	News	Social	SEC	Coverage
Financial Performance	0.583331	0.197642	0.252661	3
AI Platform Growth	-0.375277	-0.032940	0.202993	3
Regulatory Data Risk	-0.207545	-0.160893	0.031234	3
Market Valuation	0.057798	-0.052535	-0.216937	3

Table A.4: AppLovin secondary case ADI ranking.

Aspect	ADI
AI Platform Growth	0.056362
Financial Performance	0.029014
Market Valuation	0.012742
Regulatory Data Risk	0.010678

A.5 Phase 3–4 Pipeline Execution Evidence

The following figures record the runtime output of the Phase 3 SDI and Phase 4 ADI computations on the Tesla Q1 2024 event corpus, confirming that the values in Tables 4.2, 4.3, and 4.7 were produced by executed code.

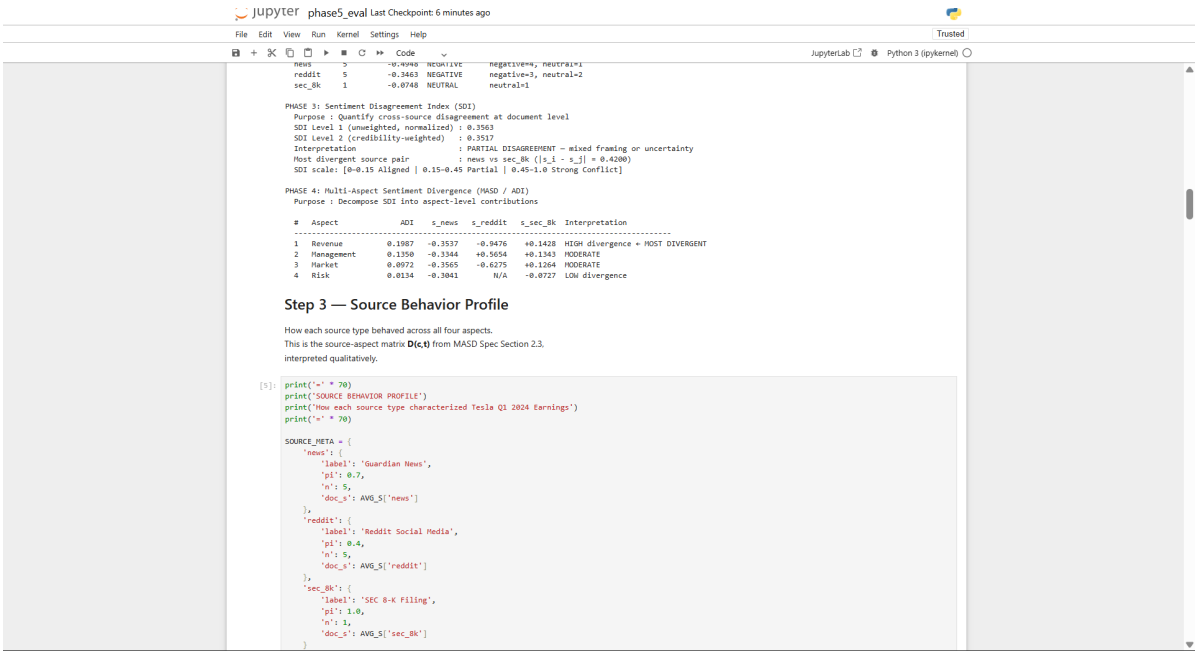

Figure A.5: Phase 5 notebook: Tesla Q1 2024 execution summary. Phase 2-4 scores confirm Tables 4.2, 4.3, and 4.7.

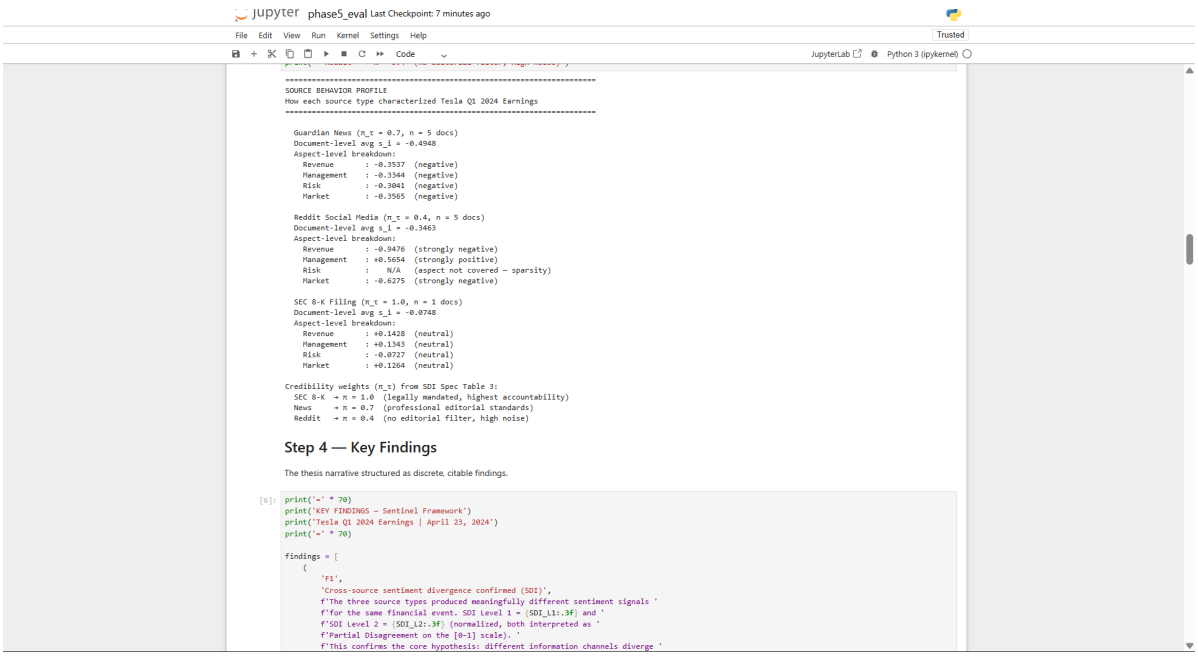

Figure A.6: Source behaviour profile. `N/A` confirms Section 3.9.5 coverage limitation. Priors π_τ correspond to Table 3.4.

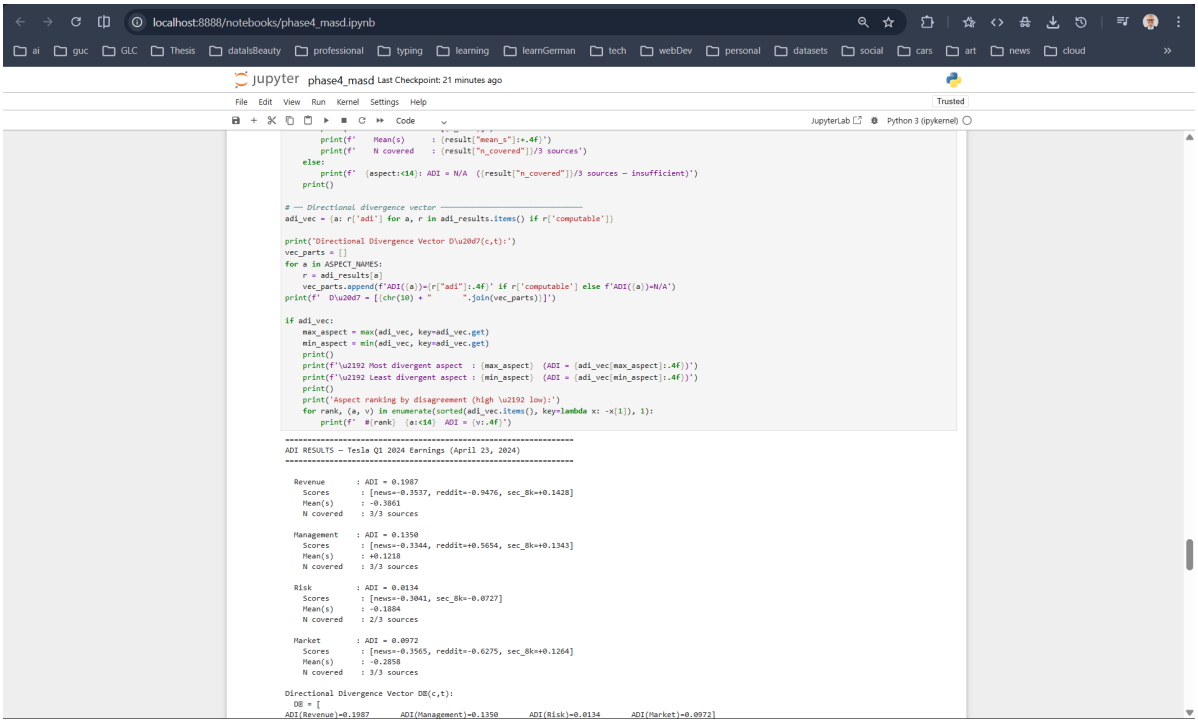

Figure A.7: Phase 4 ADI terminal output for Tesla Q1 2024. The divergence vector corresponds to Table 4.7 and Eq. (4.1).

A.6 SDI Computation Worked Example

Figure A.8 summarises the SDI computation pipeline. As a worked example, consider a hypothetical company-event instance with five sources: an SEC filing with $s = +0.60$ and $\pi = 1.0$, an earnings call transcript with $s = +0.70$ and $\pi = 0.9$, news headlines with $s = -0.20$ and $\pi = 0.7$, social media with $s = -0.80$ and $\pi = 0.4$, and an analyst report with $s = +0.30$ and $\pi = 0.8$. With $\sigma_i = 0$ for all sources, corresponding to Level 2 computation, the weights are

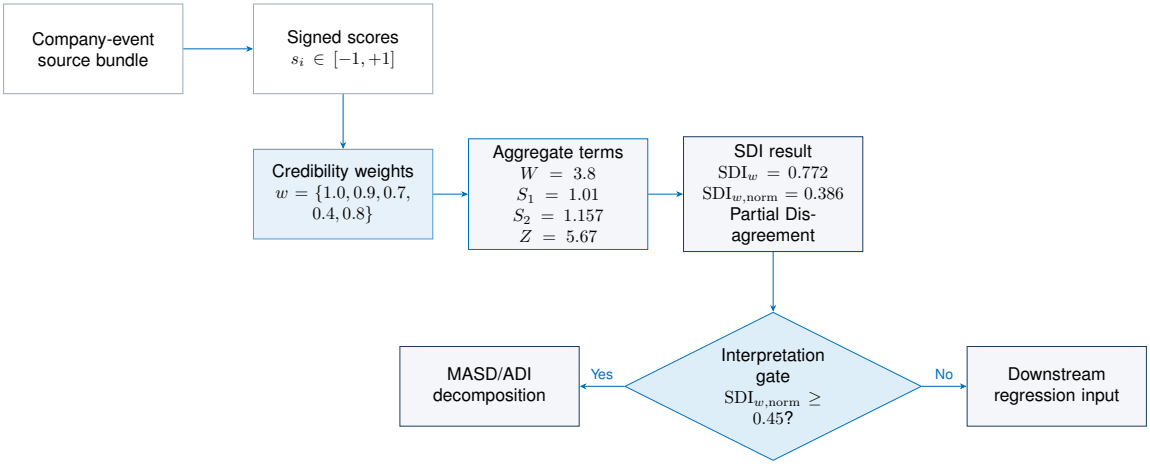

Figure A.8: SDI computation walkthrough using a hypothetical five-source company-event instance. The diagram shows source scoring, credibility-weighted aggregation, SDI scoring, and routing to either MASD/ADI decomposition or downstream regression input.

$$w = \{1.0, 0.9, 0.7, 0.4, 0.8\}.$$

The aggregated weighted statistics are

$$W = 1.0 + 0.9 + 0.7 + 0.4 + 0.8 = 3.8, \tag{1}$$

$$\begin{aligned} S_1 &= 1.0(0.60) + 0.9(0.70) + 0.7(-0.20) + 0.4(-0.80) + 0.8(0.30) \\ &= 0.60 + 0.63 - 0.14 - 0.32 + 0.24 = 1.01, \end{aligned} \tag{2}$$

$$\begin{aligned} S_2 &= 1.0(0.36) + 0.9(0.49) + 0.7(0.04) + 0.4(0.64) + 0.8(0.09) \\ &= 0.36 + 0.441 + 0.028 + 0.256 + 0.072 = 1.157. \end{aligned} \tag{3}$$

The pairwise normalisation term is

$$\begin{aligned} Z &= \frac{3.8^2 - (1.0^2 + 0.9^2 + 0.7^2 + 0.4^2 + 0.8^2)}{2} \\ &= \frac{14.44 - (1.0 + 0.81 + 0.49 + 0.16 + 0.64)}{2} \\ &= \frac{14.44 - 3.10}{2} = 5.67. \end{aligned} \tag{4}$$

The credibility-weighted SDI is therefore

$$\begin{aligned} \mathrm{SDI}_w &= \sqrt{\frac{WS_2 - S_1^2}{Z}} \\ &= \sqrt{\frac{3.8(1.157) - 1.01^2}{5.67}} \\ &\approx \sqrt{\frac{4.397 - 1.020}{5.67}} \\ &\approx \sqrt{\frac{3.377}{5.67}} \\ &\approx \sqrt{0.596} \approx 0.772. \end{aligned} \tag{5}$$

The normalised value is

$$\mathrm{SDI}_{w,\mathrm{norm}} = \frac{0.772}{2} = 0.386.$$

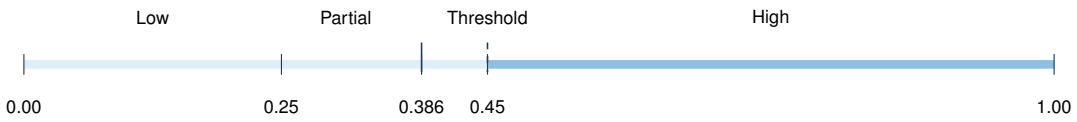

Figure A.9: Worked-example SDI interpretation scale: $\mathrm{SDI}_{w,\mathrm{norm}} = 0.386 < 0.45$.

A.7 SDI Pipeline and Framework Design Diagrams

This section provides two supporting diagrams for the SDI worked example in Appendix A.6.

Figure A.10 traces the full SDI computation pipeline for the hypothetical AAPL Q1 2025 worked example. It illustrates the sequence of operations from initial multi-source event collection and per-source calibration, through credibility-weighted aggregation, to the final threshold routing decisions and downstream MASD/ADI decomposition.

Figure A.11 maps the four theoretical design pillars of the Sentinel Framework: temporal propagation, source credibility, audit-ready explainability, and conflict detection. By linking these pillars to specific measurement approaches and the limitations of prior work, the diagram provides the conceptual grounding for the computation architecture developed in Chapter 3.

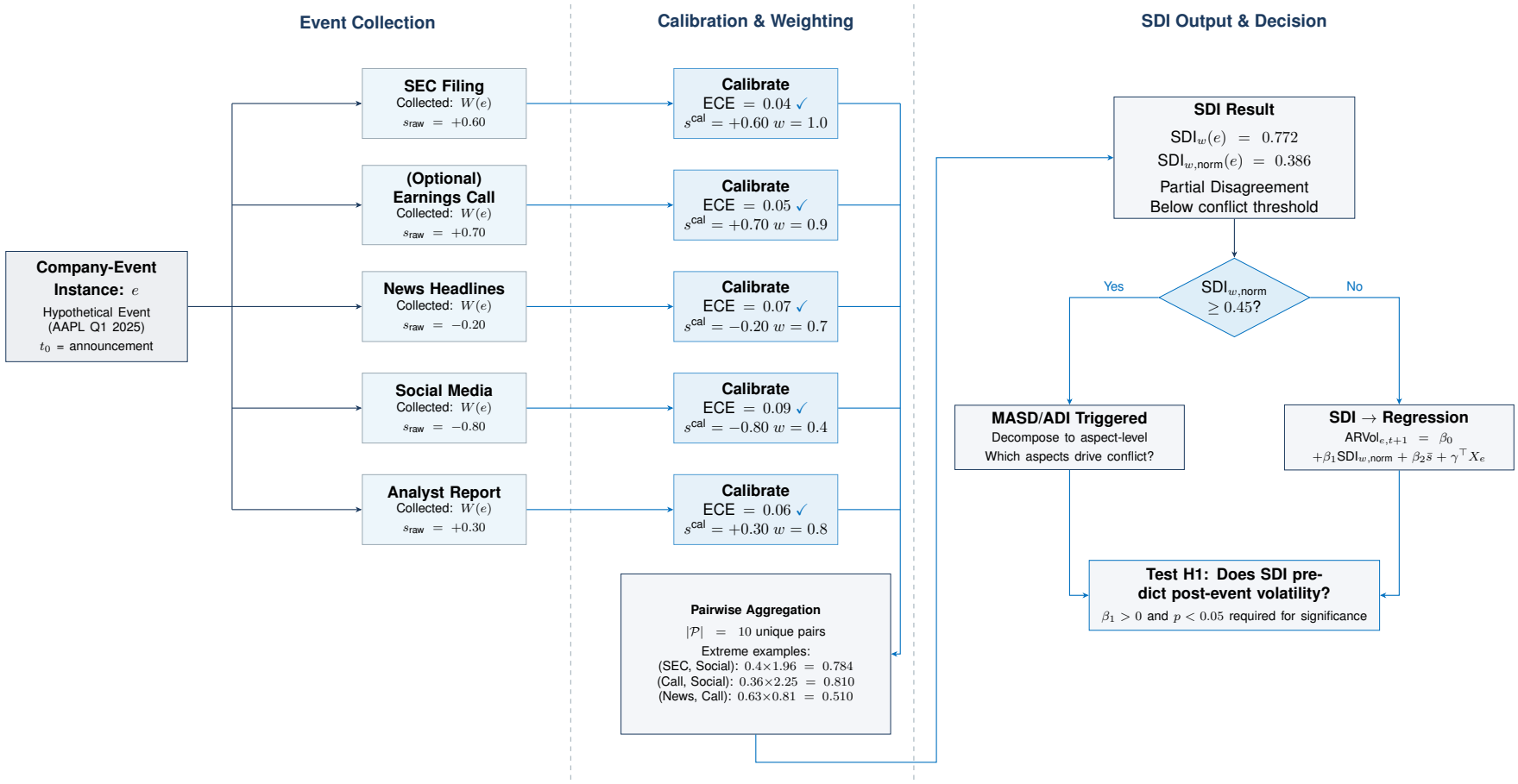

Figure A.10: SDI computation and routing pipeline for the hypothetical AAPL Q1 2025 worked example.

Figure A.11: Sentinel Framework conceptual structure: four design pillars with their analytical or regulatory grounding (dark branch) and identified gaps from prior work (grey branch).

A.8 Supplementary Implementation Evidence

This section provides supplementary computational evidence supporting the implemented pipeline components described in Chapter 3 and evaluated in Chapter 4. The figures document corpus loading outputs, inference configuration, and evaluation results as produced by the implemented pipeline notebooks.

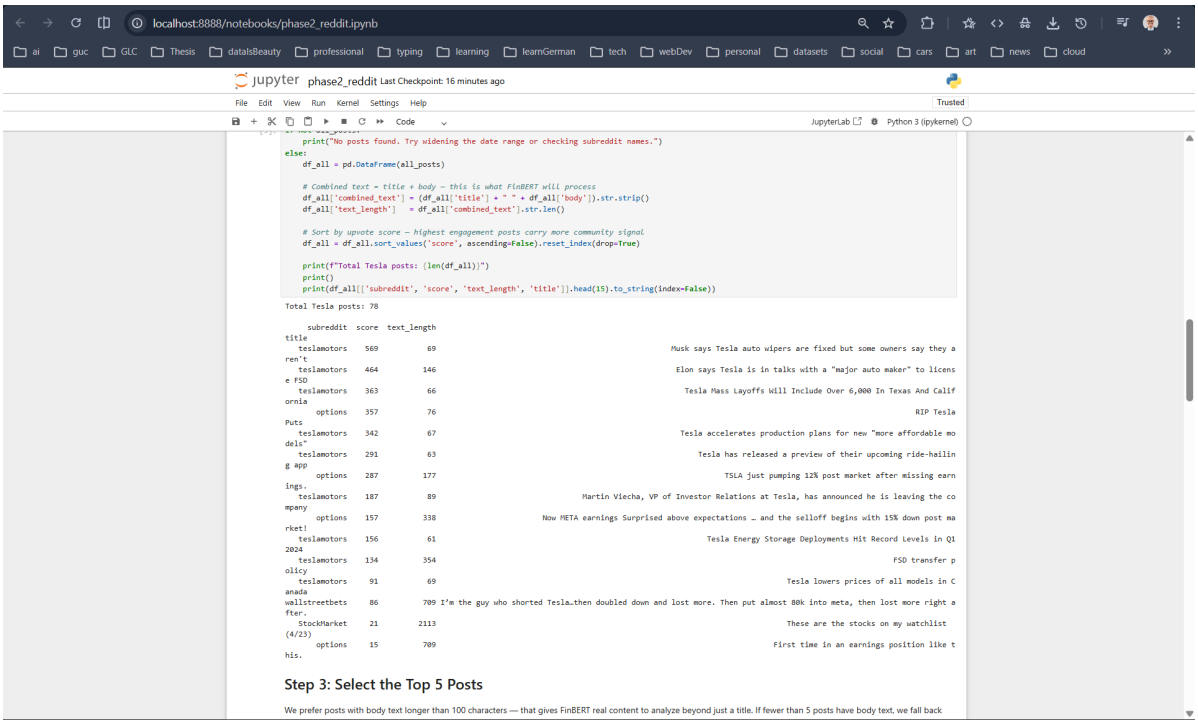

Figure A.12: Reddit post ranking output from Phase 2 social media collection: top-ranked Tesla-related posts from finance-oriented subreddits within the April 21–26, 2024 event window, ranked by upvote score. Post topics (FSD licensing, TSLA 12% surge, META earnings comparison, FSD transfer policy, short position) correspond to the five Reddit documents in Table 4.6.

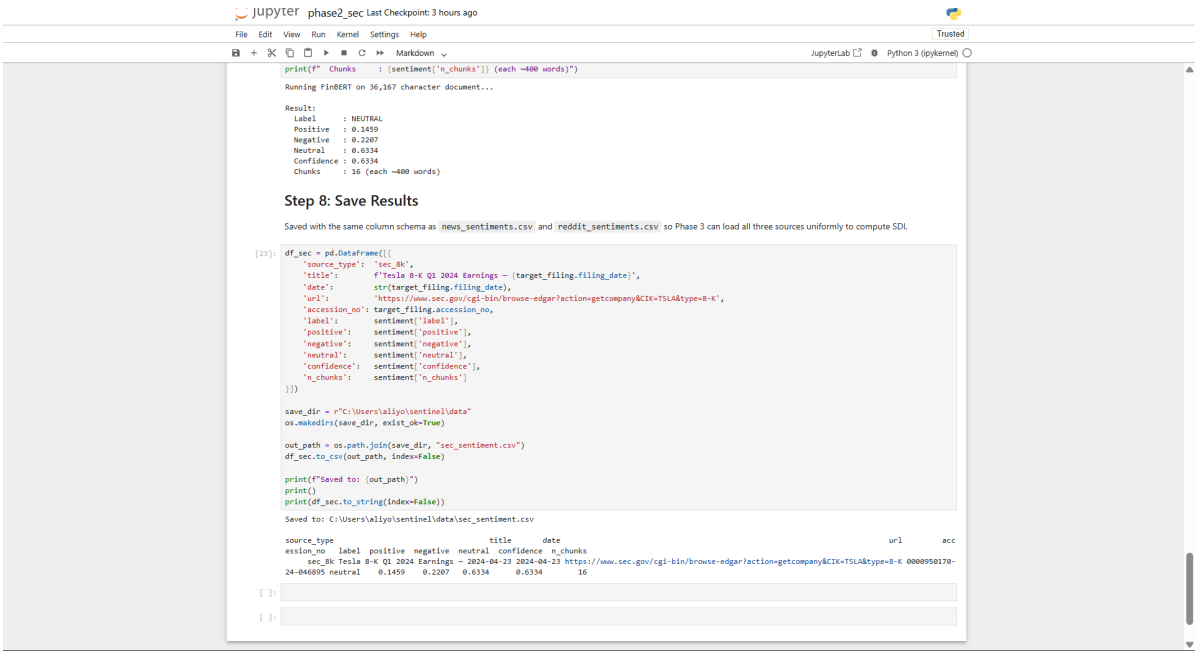

Figure A.13: Phase 2 SEC collection Step 8 save output: the constructed DataFrame row for the Tesla Q1 2024 Form 8-K. Accession number `0000950170-24-046895`, sentiment label `neutral`, and chunk count of 16 confirm the values cited in Appendix A.2 and Table 3.6.

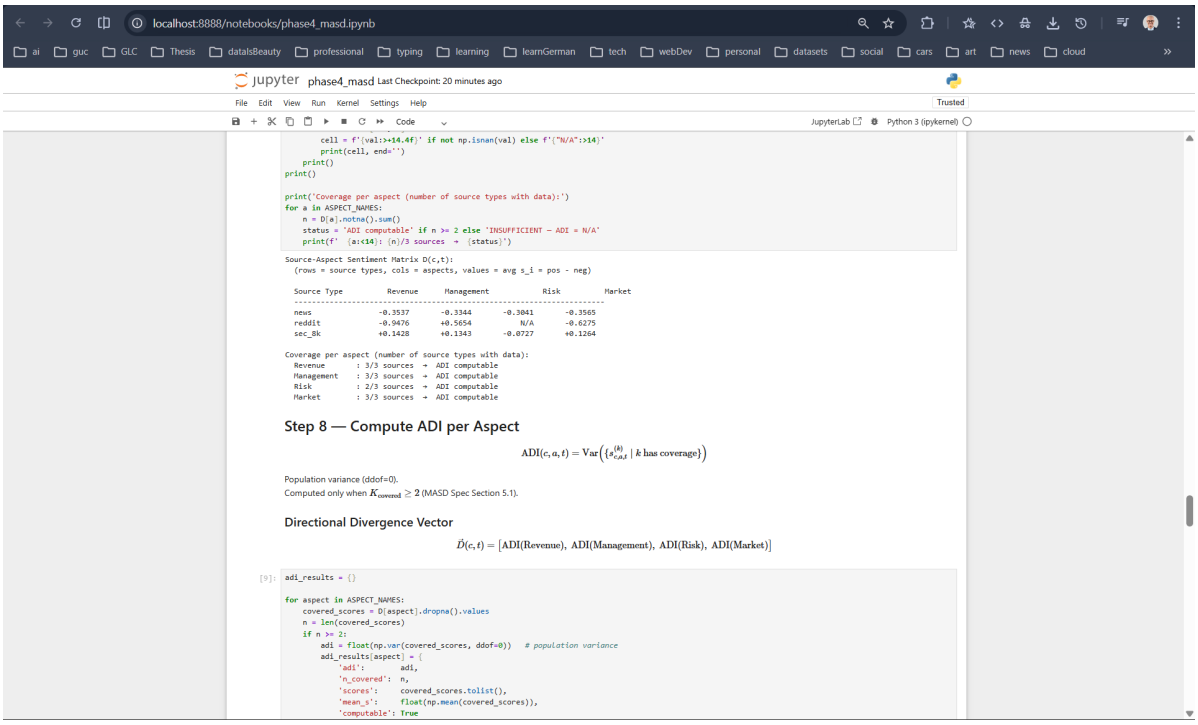

Figure A.14: Phase 4 intermediate source-aspect sentiment matrix before ADI aggregation. Per-source signed scores for all four aspect groups are shown alongside coverage fractions (3/3 for Revenue, Management, and Market; 2/3 for Risk). Values correspond to Table 4.7.

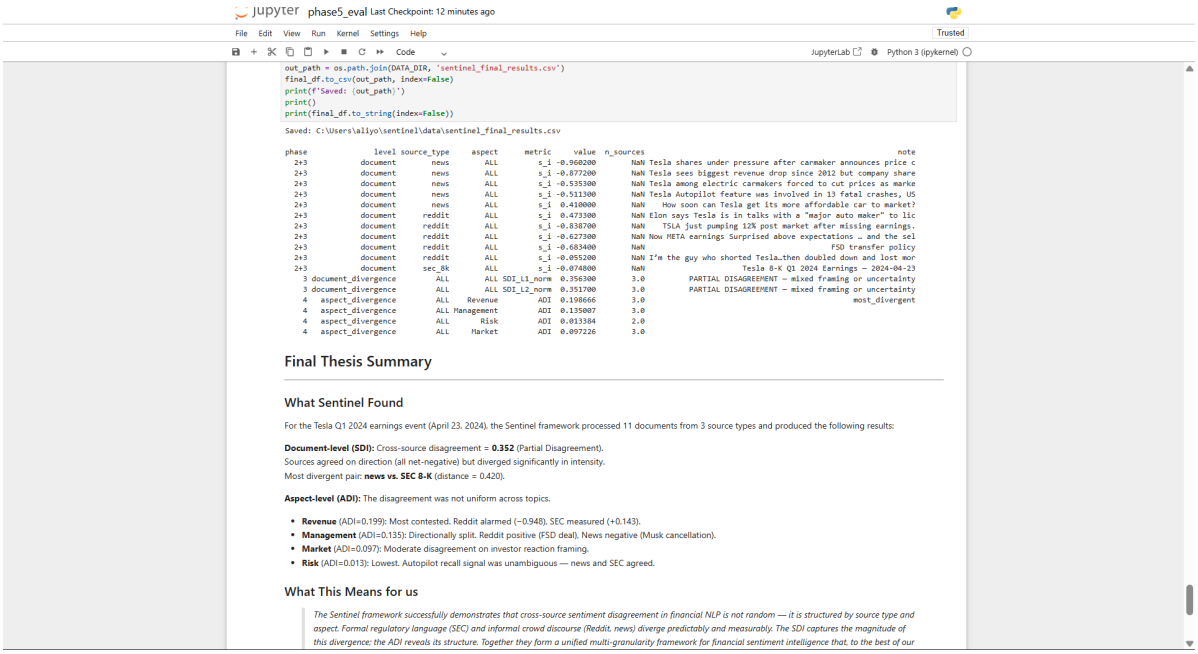

Figure A.15: Phase 5 final evaluation summary notebook cell: per-document signed scores for all 11 source documents alongside aspect labels, and a structured narrative summary of key findings. Keyword-based Revenue ADI 0.199 (highest divergence) and the Management directional split are highlighted in the output.

Bibliography

[1] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

[2] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics.

[3] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, 2020. Association for Computational Linguistics.

[4] Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[5] Allen H. Huang, Hui Wang, and Yi Yang. FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023.

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

[7] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When FLUE meets FLANG: Benchmarks and large pre-trained language model for the financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2322–2335, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[8] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann.

BloombergGPT: A large language model for finance, 2023. arXiv preprint, revised Dec. 2023.

[9] Zijian Zhang, Rong Fu, Yangfan He, Xinze Shen, Yanlong Wang, Xiaojing Du, Haochen You, Jiazhao Shi, and Simon Fong. FinSentLLM: Multi-LLM and structured semantic signals for enhanced financial sentiment forecasting. In *ICASSP 2026 — 2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 17682–17686, Barcelona, Spain, 2026.

[10] Frank Xing. Designing heterogeneous LLM agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*, 16(1):1–24, 2025.

[11] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of The Web Conference 2018 (WWW '18 Companion)*, pages 1941–1942, Lyon, France, April 2018. ACM / International World Wide Web Conferences Steering Committee.

[12] Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335, 2022.

[13] Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore, 2023. Association for Computational Linguistics.

[14] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42, April 2024.

[15] John Garcia. Beyond the headlines: Sentiment divergence and financial distress. *Global Finance Journal*, 66:101126, 2025.

[16] J. Anthony Cookson, Runjing Lu, William Mullins, and Marina Niessner. The social signal. *Journal of Financial Economics*, 158:103870, 2024.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 10088–10115. Curran Associates, Inc., 2023.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th*

International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, Sydney, Australia, 2017. PMLR.

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[22] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. In *FinLLM Symposium at IJCAI 2023*, Macao, 2023.

[23] Llama Team, AI @ Meta. The Llama 3 herd of models, 2024.

[24] Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. A retrieval-augmented multiagent system for financial sentiment analysis. *IEEE Intelligent Systems*, 40(2):15–22, 2025.

[25] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4918–4927. Association for Computing Machinery, 2024.

[26] J. Anthony Cookson and Marina Niessner. Why don't we agree? Evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228, 2020.

[27] Antonios Siganos, Evangelos Vagenas-Nanos, and Patrick Verwijmeren. Divergence of sentiment and stock market trading. *Journal of Banking and Finance*, 78:130–141, 2017.

[28] Nikolaos Passalis, Loukia Avramelou, Solon Seficha, Avraam Tsantekidis, Stavros Doropoulos, Giorgos Makris, and Anastasios Tefas. Multisource financial sentiment analysis for detecting Bitcoin price change indications using deep learning. *Neural Computing and Applications*, 34(22):19441–19452, 2022.

[29] Yiwei Liu, Junbo Wang, Lei Long, Xin Li, Ruiting Ma, Yuankai Wu, and Xuebin Chen. A multi-level sentiment analysis framework for financial texts, 2025.

[30] Mengyu Wang and Tiejun Ma. MANA-Net: Mitigating aggregated sentiment homogenization with news weighting for enhanced market prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*, pages 2379–2389, Boise, ID, USA, 2024. Association for Computing Machinery.

[31] Sergio Consoli, Luca Barbaglia, and Sebastiano Manzan. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, 247:108781, 2022. Earlier version circulated as SSRN Working Paper No. 3766194.

[32] Yiling Yan and Qiuli Qin. Sentiment analysis of financial media commentary text based on FinBERT-BiGRU-CNN. In *Proceedings of the 2024 International Conference on Smart City and Information System (ICSCIS '24)*, Kuala Lumpur, Malaysia, August 2024. Association for Computing Machinery.

[33] Edward M. Miller. Risk, uncertainty, and divergence of opinion. *The Journal of Finance*, 32(4):1151–1168, September 1977.

[34] Harrison Hong and Jeremy C. Stein. Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2):109–128, 2007.

[35] Karl B. Diether, Christopher J. Malloy, and Anna Scherbina. Differences of opinion and the cross section of stock returns. *The Journal of Finance*, 57(5):2113–2141, October 2002.

[36] Teng-Chieh Huang, Razieh Nokhbeh Zaeem, and K. Suzanne Barber. It is an equal failing to trust everybody and to trust nobody: Stock price prediction using trust filters and enhanced user sentiment on Twitter. *ACM Transactions on Internet Technology*, 19(4):1–20, 2019.

[37] Gael Kengmegni. Limitations of news sentiment analysis in short-term stock return prediction: A multi-level approach. SSRN Working Paper 5086825, Social Science Research Network (SSRN), November 2024.

[38] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, NY, USA, 2016. PMLR.