Semantic Entropy and Tariff Policy Uncertainty Discourse on Reddit: Integrating Message-Level and Network-Level Optimal Information Theory

James A Danowski *

Posted Date: 19 May 2025

doi: 10.20944/preprints202505.1388.v1

Keywords: semantic entropy; optimal information theory; uncertainty; Reddit; tariff policy; market volatility; VIX; text analysis; bigram analysis; computational social science; linguistic novelty; financial communication

 doi:10.20944/preprints202505.1388.v1

Article

Semantic Entropy and Tariff Policy Uncertainty Discourse on Reddit: Integrating Message-Level and Network-Level Optimal Information Theory

James A. Danowski

Department of Communication, University of Illinois at Chicago; jdanowski@gmail.com

Abstract: This study examines how structural linguistic novelty (bigram-based semantic entropy) and explicit uncertainty markers co-vary and predict market volatility in Reddit discussions of U.S. tariff policy. Analyzing 101 days of Reddit posts (February–May 2025), we computed daily aggregate entropy scores and uncertainty term frequencies, evaluating their relationship with the CBOE VIX index using 1–7 day lags. Bigram entropy moderately correlated with uncertainty word counts ($r=0.47, p<0.001$) and strongly with post volume ($r=0.98, p<0.001$). A seven-day ahead model demonstrated the highest explanatory power in multivariate regressions controlling for volume, uncertainty language, fear words, and sentiment ($R2=0.42$). Higher log-entropy robustly predicted increased future VIX ($\beta=146.14, p<0.001$), while log-uncertainty words also contributed positively ($\beta=20.76, p<0.001$). Fear terms and VADER sentiment were nonsignificant. These findings extend Danowski's Optimal Information Theory (OIT), showing that message-level novelty (entropy) and redundancy (uncertainty cues) co-regulate in online discourse and differentially influence market responses over a week-long horizon. Policy implications include monitoring semantic entropy as an early-warning indicator for market instability and pairing novel announcements with clear guidance to mitigate volatility.

Keywords: semantic entropy; optimal information theory; uncertainty; Reddit; tariff policy; market volatility; VIX; text analysis; bigram analysis; computational social science; linguistic novelty; financial communication

1. Introduction

Digital platforms like Reddit are now pivotal arenas for policy debate, shaping public understanding and influencing financial markets [1,2]. Their rapid dissemination of diverse information and opinions creates a high-velocity interplay between online conversations and economic outcomes, challenging traditional information gatekeepers. This necessitates more nuanced analytical tools to understand how these emergent online narratives, especially concerning significant economic policies, translate into market behavior—a critical area of inquiry.

Existing lexicon-based uncertainty measures [3], while valuable for capturing explicit doubt, overlook structural linguistic novelty. This novelty—quantifiable as semantic entropy via Shannon's information theory [4,5]—reflects linguistic unpredictability and complexity. High entropy can signal emerging ideas, cognitive effort with unfamiliar information, or departures from established narratives, all crucial to public sensemaking. Thus, reliance on explicit markers alone offers an incomplete picture of public apprehension and the processing of new policy implications.

This study examines the interplay of structural linguistic signals (bigram-based semantic entropy) and lexical signals (explicit uncertainty term frequencies) in Reddit discussions of U.S. tariff policy. Our objectives are to: 1) investigate the co-variance of semantic entropy and explicit uncertainty in this discourse, and 2) determine their collective power in predicting future financial market volatility (CBOE Volatility Index, VIX). The complexities and economic impact of tariff policy

 doi:10.20944/preprints202505.1388.v1

and Reddit's extensive user-generated content provide a rich context and ideal data source for this analysis.

By examining explicit lexical uncertainty and structural semantic entropy, this research addresses a key gap in understanding how online public discourse impacts financial markets, moving beyond simpler sentiment or keyword analyses to probe deeper informational characteristics of language. The findings aim to advance financial communication and information processing theories, showing how distinct linguistic features in online conversations can serve as early indicators of market instability. Ultimately, this study seeks to foster more sophisticated, linguistically informed approaches for monitoring and interpreting the economic impact of digital-age policy discourse.

2. Literature Review

The intersection of online discourse, policy uncertainty, and market dynamics presents a fertile ground for research. This review examines existing literature in three key areas: the measurement and impact of uncertainty in economic and political communication, the application of semantic entropy in linguistics and communication studies, and the role of narratives and sentiment in shaping economic behavior. We identify the current study's conceptual space and methodological opportunities by synthesizing these areas.

2.1. Uncertainty in Economic and Political Communication: Beyond Explicit Markers

The study of uncertainty and its effects on economic and political landscapes has a rich history. A prominent approach to quantifying such ambiguity involves lexicon-based indices, exemplified by the Economic Policy Uncertainty (EPU) index developed by Baker, Bloom, and Davis [3]. These indices typically track the frequency of predefined "uncertainty" terms (e.g., 'uncertain,' 'risk,' 'instability') within large textual corpora, often news media. The EPU index, for instance, has demonstrated robust correlations with various macroeconomic indicators, providing valuable insights into how explicit articulations of uncertainty in formal communication channels can foreshadow or coincide with economic shifts [3].

However, while powerful, these lexicon-dependent methods capture *explicit* manifestations of doubt or ambiguity. They may not fully encompass the more subtle, *implicit* structural cues within public discourse that can also signal underlying collective unease, cognitive load when processing novel or complex information, or the emergence of new, unsettling themes [6]. When individuals and groups grapple with new policy pronouncements or rapidly evolving situations, the very structure of their language—its predictability, complexity, and novelty—can change, even before explicit uncertainty terms become prevalent. This cognitive load, inherent in sensemaking during times of change or informational influx related to concepts in Weick's work on sensemaking, suggests that a deeper understanding of uncertainty requires looking beyond mere word counts. The linguistic patterns, such as an increase in unusual word pairings or convoluted phrasing, might indicate that a collective struggles to assimilate new information, thereby reflecting latent uncertainty. This limitation of purely lexicon-based approaches highlights the need for complementary methods to discern these subtler, structural linguistic patterns to gain a more holistic view of public uncertainty and its potential market impacts.

2.2. Semantic Entropy: Quantifying Novelty and Unpredictability in Discourse

To capture these structural linguistic characteristics, the concept of semantic entropy, rooted in Shannon's foundational information theory [4], offers a powerful analytical lens. Shannon originally defined entropy as a measure of unpredictability or the average information content within a message or system [4]. When applied to linguistics, semantic entropy quantifies the "surprise" value or average information content inherent in sequences of words or symbols [7]. A linguistic corpus with lower entropy is characterized by predictable, common, and perhaps redundant language patterns. Conversely, higher semantic entropy signifies greater novelty, structural complexity, and less

 doi:10.20944/preprints202505.1388.v1

predictable word sequences, indicating that the discourse contains more unexpected or freshly combined ideas.

N-gram models are commonly employed to operationalize linguistic entropy. These models assess the probability of a word occurring given its preceding context (i.e., the preceding n-1 words). Bigram models (n=2), which analyze sequences of two words, are frequently utilized as they offer a practical balance for measuring discourse novelty, particularly in dynamic and often noisy online environments like Reddit. Bigrams embody a link between two co-occurring terms and are thus the basic building blocks of larger semantic networks. They capture more local syntactic and semantic structure than unigram (single-word) models, which primarily reflect topic distribution. Simultaneously, bigrams are less prone to the data sparsity issues that can plague higher-order n-grams (e.g., trigrams, quadgrams) when analyzing diverse and rapidly changing corpora where many longer word sequences may appear too infrequently for stable probability estimation [cf. Petersen et al., 7]. By measuring bigram-based semantic entropy, this study aims to quantify the degree of linguistic novelty and structural unpredictability in online policy discussions, indicating how much new or complex information is being processed or generated by the community.

2.3. Narrative Economics, Sentiment, and the Structure of Discourse

The importance of public discourse, particularly its evolving narratives, in shaping economic decisions and market movements is central to Shiller's theory of Narrative Economics [2]. Shiller posits that popular stories and explanations—which can embody elements of novelty, hope, fear, or uncertainty—can spread contagiously, influencing collective behavior and thereby impacting economic outcomes, from investment decisions to consumer confidence [2]. Understanding the structural aspects of these narratives as they unfold in real-time on platforms like Reddit—such as the emergence of novel themes (high entropy) or the articulation of collective doubt—is crucial for grasping how online conversations might translate into tangible market effects.

While narrative economics provides a broad framework, computational sentiment analysis tools like VADER (Valence Aware Dictionary and sEntiment Reasoner) offer methods to quantify text's affective tone or emotional valence. These tools can identify whether discourse is predominantly positive, negative, or neutral. However, sentiment, while an important dimension of communication, is conceptually distinct from a message's linguistic novelty (unpredictability) or its structural redundancy (predictability). A message can be novel and positive, or novel and negative; similarly, it can be redundant and positive, or redundant and negative. Semantic entropy, therefore, measures a different characteristic of the text—its inherent informational surprise and structural complexity—which may not be captured by sentiment scores alone. For instance, a stream of highly predictable, low-entropy anxious statements might yield a strong negative sentiment score but low novelty. Conversely, a discussion involving the innovative framing of a policy, even if affectively neutral, could exhibit high entropy. Distinguishing between these linguistic properties is essential for a nuanced understanding of how different facets of online discourse contribute to public awareness and market responses.

By integrating insights from these areas—the limitations of traditional uncertainty measures, the potential of semantic entropy to capture linguistic novelty, and the influence of narratives and sentiment—this study seeks to explore how the structural and lexical characteristics of online policy discussions on Reddit co-vary and jointly predict market volatility. This approach allows for a more comprehensive examination of the mechanisms through which public discourse translates into financial market behavior, addressing a gap in the current understanding of these complex interactions.

3. Theoretical Framework: Optimal Information Theory (OIT)

Navigating and making sense of the torrent of information in digital environments, particularly concerning complex topics like tariff policy, presents significant cognitive challenges for individuals and collectives. This study employs Danowski's Optimal Information Theory (OIT) [9–11] as its

 doi:10.20944/preprints202505.1388.v1

primary theoretical lens to understand how linguistic features of online discourse reflect and influence these sensemaking processes and their broader market implications. OIT posits that for communication to be effective—that is, for information to be successfully processed and understood by an audience—communication systems inherently strive for a balance between informational novelty (which can be quantified by measures like semantic entropy) and informational redundancy (which includes clarity cues and explicit markers of uncertainty) [9].

The core tenet of OIT is that human information processing operates within an optimal range. Excessive novelty, such as highly complex, unpredictable, or entirely unfamiliar information (high entropy), can lead to cognitive overload, confusion, and an inability to integrate the new information effectively. Conversely, excessive redundancy, characterized by overly simplistic, repetitive, or predictable information (low entropy, high explicit redundancy), can result in disengagement, boredom, and a failure to capture attention or convey new insights. OIT therefore suggests that optimal communication involves presenting new or complex information in a digestible manner, often by accompanying novel elements with cues that acknowledge complexity or guide interpretation.

OIT proposes mechanisms through which this balance is sought at distinct but interconnected levels:

- **Message-Level Dynamics: Co-regulation of Novelty and Clarity Cues:** At the message level, OIT predicts a functional co-occurrence between the introduction of novel or complex information (characterized by high semantic entropy) and the use of explicit linguistic markers that signal this complexity or provide context. In this study, such markers include terms of uncertainty (e.g., "it is unclear," "potentially," "risk"). From an OIT perspective, these explicit uncertainty markers are not merely indicators of doubt; they can also function as a form of adaptive redundancy or as "clarity cues." By explicitly acknowledging that information is difficult, new, or not fully resolved, speakers or writers help manage the audience's cognitive load. Such cues prepare the receiver for more effortful processing, frame the novel information appropriately, and can guide comprehension by highlighting areas that require careful consideration. Including these cues prevents the cognitive system from being overwhelmed by sheer novelty. This study directly tests this message-level proposition by hypothesizing a positive correlation between the daily aggregate semantic entropy of Reddit discourse on tariffs and the daily frequency of explicit uncertainty words. This co-occurrence would suggest a communicative strategy, whether conscious or emergent, to buffer the impact of linguistic novelty.
- **Network-Level Dynamics: Diffusion, Distributed Redundancy, and Collective Sensemaking:** Beyond individual messages, OIT considers how complex and novel information diffuses and is processed within broader communication networks, such as the extensive discussions on a platform like Reddit. As information with high semantic entropy permeates a social system, OIT suggests that mechanisms for distributing redundancy will emerge to ensure wider comprehension and facilitate collective sensemaking. This distributed redundancy might manifest in various ways: through subsequent posts that rephrase, simplify, or elaborate on initially novel or confusing statements; through question-and-answer exchanges that clarify ambiguities; or through the emergence of summary threads or interpretations offered by influential users. The dynamic interplay between the injection of novel information (entropy) and the responsive generation of clarifying discourse (uncertainty language, explanations) over time shapes the collective understanding and sentiment within the network. The current study views the aggregate daily measures of entropy and uncertainty language, and their evolving relationship with market volatility (VIX) over different time lags (1–7 days), through this network-level lens. The aim is to understand how these collective, platform-level discourse dynamics, reflecting the balance (or imbalance) of novelty and clarity, might translate into systemic financial market responses.

 doi:10.20944/preprints202505.1388.v1

This study operationalizes OIT by systematically measuring message novelty through daily aggregate bigram-based semantic entropy in Reddit posts concerning U.S. tariff policy. Simultaneously, it quantifies explicit uncertainty through the frequency of specific uncertainty terms. By examining how these two linguistic dimensions (1) co-vary, and (2) jointly predict future financial market volatility (VIX), while controlling for other factors like post volume and sentiment, the research empirically tests OIT's core propositions in a highly relevant, real-world context of financial communication. OIT thus provides a nuanced framework for understanding not just *that* online discourse might affect markets, but *how* specific structural and lexical properties of that discourse, reflecting fundamental principles of information processing, contribute to such effects

4. Materials and Methods

4.1. Corpus Collection and Preprocessing

We collected ~200,000 English language- Reddit posts and comments referencing "tariff OR tariffs" over 101 days (1 February 2025–11 May 2025) via Meltwater. Reddit was chosen for its widespread use in news discussion and public data. Daily post volume for tariff-related content varied significantly. For days where post volume exceeded 20,000, a random sample of 20,000 posts was extracted due to download restrictions; this applied to most weekdays.

Preprocessing involved: (1) removing exact duplicates; (2) excluding known bot posts; (3) eliminating emojis/special characters; (4) filtering posts <5 words. Text was lowercased. NLTK (Python) was used for tokenization. Near-duplicates were also removed to ensure entropy calculations reflected genuine discourse variation.

4.2. Semantic Entropy Calculation

We computed bigram entropy in two steps: first, we estimated a bigram-based language model over the entire corpus and then used that model to score each post.

Estimating Bigram Entropy

To quantify the linguistic novelty of Reddit posts, we calculated semantic entropy based on bigram (two-word sequence) probabilities across the entire 101-day corpus. We began by counting how many times each word appeared in the corpus (called a "unigram" count). These unigram counts served as the base for estimating the likelihood of a word following another.

Next, we counted how often each consecutive word pair appeared (called a "bigram"). For example, we counted how often "tariff increase" appeared as a sequence. We then estimated the probability of a word given its predecessor by dividing each bigram's count by the first word's count in that pair. The likelihood of word B following word A equals the number of times "A B" appears, divided by the total number of times "A" appears. For bigrams that were not observed in the data, we assigned a very small fallback probability (such as 0.00000001) to avoid undefined calculations when taking logarithms.

Each Reddit post was then evaluated for how unpredictable its word sequences were. After splitting the text into individual words (tokens), we scanned through all consecutive pairs and calculated each bigram probability's negative log base 2. We averaged these values across the post to produce a single entropy score. A higher score indicated more unexpected or novel phrasing.

We then summed the entropy scores of all daily posts to create a daily aggregate entropy value. This daily total was log-transformed to reduce skew and normalize the distribution for regression models.

This resulting measure represents how surprising or diverse the language was on a given day. Higher daily entropy indicates that Reddit users discussing tariffs used more varied and less predictable word combinations, suggesting greater novelty or complexity in public discourse.

4.3. Lexical Indices and Sentiment Analysis

 doi:10.20944/preprints202505.1388.v1

Daily counts of uncertainty (240-term lexicon, e.g., "risk," "unclear") and fear (144-term lexicon, e.g., "panic," "crisis") words were extracted using WORDij software's [12] include list option. These lexicons were based on prior work [3,13]. Daily counts were log-transformed. VADER [8] provided daily mean compound sentiment scores (-1 to +1) for overall affective tone.

4.4. Market Volatility Data Acquisition

Daily closing values of the CBOE Volatility Index (VIX), a measure of expected 30-day S&P 500 volatility, were obtained. Higher VIX indicates greater expected instability. VIX data were lagged 1–7 days (VIXt+k, k=1…7) to assess discourse's predictive power on future volatility.

4.5. Statistical Analysis Procedures

Pearson correlations assessed pairwise associations between daily log-transformed semantic entropy (Hday), log-uncertainty counts, log-fear counts, log-post volume, VADER scores, and unlagged VIX. Ordinary Least Squares (OLS) regression models predicted VIXt+k for each lag. Predictors included log-entropy, log-uncertainty, log-fear, log-volume (control), and sentiment. Models with insufficient observations due to lagging were omitted. Significance was $p<0.05$. Adjusted R2 assessed model fit.

5. Results

Table 1 presents the descriptive statistics for the primary daily linguistic and market variables analyzed in this study. The average daily log-transformed bigram entropy, which serves as a proxy for structural linguistic novelty in Reddit discussions, was 8.84 (SD = 0.10), with values ranging from 8.59 to 9.10. The mean log-transformed uncertainty word count was 4.36 (SD = 0.92), while the log-transformed fear word count averaged 3.55 (SD = 0.66), reflecting lower overall levels of these affective expressions relative to structural novelty. Based on the VADER compound score, the average daily sentiment was slightly positive at 0.06 (SD = 0.04), ranging from mildly negative to moderately positive. The total daily word count, log-transformed, had a mean of 14.59 (SD = 0.64), indicating a high level of variation in posting volume. The VIX index, used to represent market volatility, had a mean value of 16.66 (SD = 1.72), with observed values ranging from 14.05 to 22.52 during the 101-day observation window.

These figures provide a foundational overview of the data, highlighting moderate variability in linguistic complexity, uncertainty markers, and more stable sentiment expressions. They also contextualize the subsequent analysis of how these features relate to shifts in market volatility.

Table 1. Descriptive Statistics.

Variable	Mean	SD	Min	Max	N
total_entropy	67872.22	9700.74	23339.58	76645.74	101
total_bigram_entropy	116330.13	17287.46	37843.76	128434.47	101
avg_vader_sentiment	0.02	0.01	-0.01	0.04	101
total_words_all	371141.55	51360.46	135692.0	430788.0	101
post_count	16644.4	2471.15	5422.0	18431.0	101
total_uncertainty_words	7781.4	1901.45	3647.0	14022.0	100

 doi:10.20944/preprints202505.1388.v1

total_fear_words	11714.99	2906.07	5053.0	20447.0	100
VIX	23.69	7.8	14.77	52.33	68

5.1. Correlation Analysis

Table 2 reports the Pearson correlation coefficients among the key daily variables: log-transformed bigram entropy, uncertainty word counts, fear word counts, average sentiment scores, total word counts, and the VIX index. Several notable patterns emerged.

First, log-transformed bigram entropy showed a strong positive correlation with total word count ($r = 0.98$, $p < 0.001$), indicating that days with higher discussion volume on Reddit also exhibited greater linguistic novelty. This underscores the importance of controlling for volume effects when interpreting entropy. Bigram entropy also showed a moderate positive correlation with uncertainty word counts ($r = 0.47$, $p < 0.001$), suggesting that as discourse becomes more structurally novel, it tends to contain more explicit expressions of uncertainty—consistent with the predictions of Optimal Information Theory.

Correlations between bigram entropy and the VIX were more modest ($r = 0.38$), suggesting a potential link between discourse novelty and market volatility, though weaker in simple bivariate terms than in the full multivariate model. Uncertainty and fear word counts were moderately correlated ($r = 0.37$), yet each showed weak associations with the VIX ($r = 0.18$ and $r = 0.24$, respectively). Sentiment scores demonstrated weak or negligible correlations with all other variables, including the VIX ($r = –0.14$), indicating that average sentiment alone may not capture these discussions' epistemic or market-relevant dimensions.

Together, the correlation results provide preliminary support for the co-regulation of novelty and uncertainty in public discourse and motivate the need for multivariate modeling to disentangle these effects in predicting market outcomes.

Table 2. Correlation Matrix.

	tot_entropy	tot_bigram_entropy	avg_vader_sentiment	tot_words_all	post_count	tot_uncert_words	tot_fear_words	VIX
total_entropy	1.0	0.996	-0.002	0.992	0.996	0.488	0.523	0.33
total_bigram_entropy	0.996	1.0	-0.033	0.977	1.0	0.466	0.507	0.376
avg_vader_sentiment	-0.002	-0.033	1.0	0.051	-0.039	0.315	-0.104	-0.139
total_words_all	0.992	0.977	0.051	1.0	0.977	0.522	0.539	0.284
post_count	0.996	1.0	-0.039	0.977	1.0	0.466	0.511	0.374

 doi:10.20944/preprints202505.1388.v1

total_uncertainty_words	0.488	0.466	0.315	0.522	0.466	1.0	0.371	0.177
total_fear_words	0.523	0.507	-0.104	0.539	0.511	0.371	1.0	0.24
VIX	0.33	0.376	-0.139	0.284	0.374	0.177	0.24	1.0

5.2. Regression Outcomes for Predicting Market Volatility

To assess the predictive power of linguistic indicators on future market volatility, we conducted a series of multiple linear regression models with the CBOE VIX as the dependent variable, lagged from one to seven days ahead. Each model included the following predictors: log-transformed daily bigram entropy, log-transformed counts of uncertainty-related words, log-transformed counts of fear-related words, average daily VADER sentiment scores, and log-transformed total word counts (as a control for overall discussion volume). Of these, the seven-day-ahead model yielded the strongest results.

Table 3. OLS Regression Predicting VIX (7-Day Lag).

Variable	Coefficient	Std. Error	t	P>|t|	[0.025	0.975]
Constant	264.60	127.96	2.07	0.043	8.26	520.94
log_total_bigram_entropy	146.14	29.42	4.97	<0.001	87.21	205.07
log_total_uncertainty_words	20.76	4.82	4.31	<0.001	11.10	30.41
log_total_fear_words	0.48	4.08	0.12	0.907	-7.69	8.65
log_total_words_all	-166.61	35.43	-4.70	<0.001	-237.59	-95.62
avg_vader_sentiment	107.90	104.06	1.04	0.304	-100.54	316.35

In the optimal seven-day lag model (n = 62), log bigram entropy emerged as a highly significant predictor of increased market volatility ($\beta = 146.14$, $p < 0.001$), suggesting that greater structural novelty in Reddit discussions on tariffs forecasts heightened investor uncertainty approximately one week later. Similarly, the frequency of uncertainty-related words was a significant positive predictor ($\beta = 20.76$, $p < 0.001$), indicating that explicit expressions of doubt or ambiguity in discourse also contribute to anticipated volatility.

Interestingly, total word count was negatively associated with VIX ($\beta = -166.61$, $p < 0.001$), implying that high-volume discussions—when not marked by structural novelty—may reflect a more familiar or processed information environment, dampening volatility. Neither the frequency of fear-related terms nor average sentiment scores significantly predicted VIX in lagged models. This reinforces the view that epistemic rather than affective discourse features hold stronger explanatory value in this context.

Overall, the model explained 42% of the variance in VIX values seven days later (adjusted $R^2 = 0.42$), a substantial level of predictive power in financial modeling. These findings highlight the importance of monitoring both structural and lexical dimensions of public discourse for anticipating shifts in market sentiment.

The OLS regression model predicting VIXt+7 (seven days ahead) demonstrated the highest explanatory power (Adjusted R2=0.42,n=62).

Key findings for the VIXt+7 model:

- Log-transformed semantic entropy: significant positive predictor (β=146.14,p<0.001).
- Log-transformed uncertainty word counts: significant positive predictor (β=20.76,p<0.001).

 doi:10.20944/preprints202505.1388.v1

- Log-transformed total post volume: significant negative predictor (β=–166.6,$p<0.001$).
- Log-transformed fear word counts and VADER sentiment scores were not significant predictors.

Models for other lags showed lower R2 values and less consistent predictor significance.

6. Discussion

This study provides novel insights into the complex interplay between the linguistic structures of online public discourse, particularly on the Reddit platform, and financial market dynamics, specifically market volatility as measured by the CBOE VIX. Interpreted through the lens of Danowski's Optimal Information Theory (OIT) [9,10], the findings illuminate how both structural linguistic novelty (semantic entropy) and explicit expressions of uncertainty co-vary and collectively offer predictive power for market fluctuations.

6.1. Co-Regulation of Novelty and Explicit Uncertainty: An OIT Perspective

A key finding is the moderate positive correlation ($r=0.47,p<0.001$) between daily aggregate semantic entropy and the frequency of uncertainty-related terms in Reddit discussions on U.S. tariff policy. This observation lends empirical support to a core proposition of Optimal Information Theory [9]: that communication systems often exhibit a co-regulation of informational novelty and cues that help manage the cognitive load associated with that novelty. As discourse becomes more structurally complex or unpredictable (higher bigram entropy), indicating the introduction or processing of new perspectives or intricate details, a concurrent increase in language explicitly acknowledging ambiguity or risk.

This co-regulation suggests an adaptive mechanism within the discourse community. When participants introduce or grapple with novel aspects of tariff policy—perhaps new regulations, unforeseen consequences, or complex economic arguments—the simultaneous rise in uncertainty language can be seen as a collective effort to signal the challenging nature of the information, thereby priming readers for more effortful processing or inviting clarification. This likely occurs organically at the individual message level, where users might hedge novel statements with cautious phrasing, and at the aggregate discourse level, where the overall conversation reflects this blend.

Furthermore, the joint significance of both log-entropy ($\beta=146.14,p<0.001$) and log-uncertainty words ($\beta=20.76,p<0.001$) in predicting future VIX in the seven-day ahead model is particularly revealing. It implies that while related, semantic novelty and explicit uncertainty markers capture distinct and additive dimensions of the discourse that contribute to market participants' perception of future risk. High entropy might reflect the emergence of fundamentally new information or frameworks that challenge existing market assumptions. In contrast, high uncertainty language might reflect unresolved questions and a lack of clear consensus, even around more familiar topics. The fact that both contribute independently underscores the multifaceted nature of how online discourse can signal impending volatility.

6.2. Temporal Dynamics and Market Information Channels: The Seven-Day Horizon

The emergence of the seven-day ahead model for VIX prediction as the one with the highest explanatory power (Adjusted R2=0.42) is a significant finding, suggesting a characteristic lag in how diffuse public discourse on platforms like Reddit translates into broader market volatility. While official news releases or policy announcements from recognized authorities are often assumed to be priced into markets almost instantaneously, the signals from decentralized online conversations appear to follow a different, more gradual impact pathway.

This approximately one-week cycle could reflect several underlying processes. It may represent the time it takes for nascent themes, novel linguistic framings, or shifts in collective uncertainty within a large online community to diffuse sufficiently to reach and influence a critical mass of market participants, including institutional investors or influential financial analysts. Alternatively, this timeframe might align with weekly economic reporting cycles or when market sentiment gradually

 doi:10.20944/preprints202505.1388.v1

consolidates in response to sustained online discussion trends. The finding that semantic entropy, a measure of linguistic novelty, has predictive power seven days out suggests that the very structure of public conversation can serve as a leading indicator of market stress, well before such stress is universally acknowledged or reflected in traditional news sentiment. This contrasts sharply with the often more immediate impact of formal information channels and points to the unique role of social media in incubating and signaling future market conditions.

6.3. Implications for Policy Communication and Market Monitoring

The results of this study carry several important practical implications for policymakers, communication strategists, and market analysts:

- **Developing Early-Warning Systems for Market Instability:** The robust predictive relationship between semantic entropy in Reddit discourse and future VIX, particularly at a seven-day lead, suggests its utility as a component in an early-warning system for market stress. Real-time monitoring of linguistic novelty within relevant online policy discussions could provide a valuable, data-driven indicator for proactive risk assessment and financial surveillance. However, implementing such a system would require addressing challenges related to data filtering, contextual understanding, and distinguishing signal from noise in vast social media data streams.
- **Informing Strategic Policy Communication:** To mitigate potentially destabilizing market volatility arising from new or complex policy announcements, policymakers should consider the principles of OIT. This involves strategically pairing novel information with clear, unambiguous, and sufficiently redundant explanations. Practices such as providing detailed FAQs, utilizing consistent messaging across multiple channels, and proactively addressing potential areas of confusion can help reduce the perceived linguistic novelty and manage public uncertainty more effectively, potentially dampening adverse market reactions.
- **Facilitating Constructive Network-Level Outreach:** Given the influence of online discourse, engaging with online communities like Reddit may become increasingly important. Identifying and interacting with key "bridging actors" [13] or influential community members could help to contextualize complex policy information accurately. Such engagement could aid in collective sensemaking processes within these forums, fostering more informed discussion and mitigating extreme market reactions driven by unclarified novelty or unmanaged uncertainty.
- **Enhancing Financial Literacy and Market Analysis:** For financial market participants and analysts, these findings highlight the value of looking beyond traditional news sources and incorporating signals from social media discourse. Understanding measures like semantic entropy could offer an additional tool for gauging public interpretation of policy and anticipating potential market shifts.

6.4. Limitations and Future Research

While this study offers valuable insights, its limitations must be acknowledged to contextualize the findings and guide future research:

- **Platform and Topic Specificity:** The focus on Reddit discussions related to U.S. tariff policy means the findings may not be directly generalizable to other social media platforms (e.g., Twitter, Facebook, specialized financial forums), which have different user demographics, communication norms, and information diffusion patterns. Similarly, the dynamics observed for tariff policy might differ for other types of economic or political news (e.g., monetary policy, elections).
- **Single Volatility Measure:** While VIX is a widely recognized measure of expected market volatility, it reflects expectations for the S&P 500. Future research could explore impacts on other financial indicators, such as specific industry stock indices, bond yields, or trading volumes.

 doi:10.20944/preprints202505.1388.v1

- **Operationalization of Entropy:** Bigram-based semantic entropy, while well-justified for its balance of structural insight and computational feasibility, is one of many ways to quantify linguistic properties. Future work could explore higher-order n-grams (with larger datasets), syntactic complexity measures, semantic coherence, or topic-specific entropy variations.
- **Correlational Nature:** The study identifies strong correlations and predictive relationships but cannot establish causality due to its observational nature. It is possible that underlying economic conditions or unobserved variables simultaneously influence the nature of online discourse and market volatility. Nevertheless, the lag results establish two of three necessary conditions for causality, association and time-order.
- **Data Sampling:** The necessity of sampling posts on high-volume days due to download restrictions, while randomized, could potentially introduce biases if the sampled posts are not fully representative of the entirety of the discourse on those specific days. To obtain the full loads of posts on Meltwater requires accounts costing tens of thousands of dollars annually. Our account was a low-cost, monthly $31 community one.

Building on this research, several avenues for future inquiry emerge:

- **Cross-Platform and Cross-Topic Analyses:** Comparative studies across different social media platforms and diverse policy domains are needed to assess the robustness and generalizability of these findings.
- **Event-Based Studies:** Investigating the precise dynamics of semantic entropy and uncertainty language immediately before, during, and after specific, major policy announcements or economic shocks could provide finer-grained insights into the information assimilation process.
- **Experimental Research:** Experimental designs could be employed to address causality. For example, researchers could expose participants to messages about fictitious policies, varying the levels of semantic entropy and explicit uncertainty cues, and then measure participants' market expectations, perceived risk, or simulated investment decisions.
- **Advanced Linguistic Features:** Incorporating more sophisticated natural language processing techniques to analyze argument structure, stance detection, emotional tone beyond simple sentiment, and the propagation of specific narratives could offer a richer understanding of discourse dynamics.
- **Longitudinal and Regime-Change Studies:** Examining these relationships over extended periods, including different market regimes (e.g., bull vs. bear markets, high vs. low baseline volatility), could reveal how the predictive power of discourse features might change.

By addressing these limitations and exploring these future directions, the field can continue to build a more comprehensive understanding of the intricate connections between public discourse in the digital age and the functioning of financial markets.

7. Conclusions

Bigram-based semantic entropy (structural linguistic novelty) and explicit uncertainty language co-regulate in Reddit tariff discourse and jointly predict market volatility (CBOE VIX) with optimal power at a seven-day horizon. Higher entropy and uncertainty independently forecasted higher volatility, controlling for volume, fear, and sentiment. These findings empirically extend Danowski's Optimal Information Theory to digital policy forums and their market connections, highlighting how information structure and framing in online conversations impact financial outcomes. This research supports developing linguistically-informed monitoring tools for economic policy management, positioning semantic entropy as a potential early-warning indicator for market turbulence.

Author Contributions: J.A.D. conceived and designed the study, performed data collection and analysis, and authored the manuscript.

Funding: This research received no external funding.

 doi:10.20944/preprints202505.1388.v1

Institutional Review Board Statement: Ethical review and approval were waived for this study, as it relied exclusively on publicly available, anonymized data from Reddit, in compliance with the platform's terms of service.

Informed Consent Statement: Not applicable as the study used publicly available, anonymized data.

Data Availability Statement: Aggregated daily data supporting reported results are available from the author upon reasonable request. Raw post data cannot be shared directly due to Reddit's Terms of Service, but can be independently collected using the described methodology.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Castells, M. *Communication Power*; Oxford University Press: Oxford, UK, 2013.
2. Shiller, R.J. Narrative Economics. *Am. Econ. Rev.* **2017**, *107*, 967–1004.
3. Baker, S.R.; Bloom, N.; Davis, S.J. Measuring Economic Policy Uncertainty. *Q. J. Econ.* **2016**, *131*, 1593–1636.
4. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
5. Danowski, J.A. An Information Theory of Communication Functions. Ph.D. Thesis, Michigan State University, East Lansing, MI, USA, 1975.
6. Danowski, J.A.; Edison-Swift, P. Crisis effects on intraorganizational computer-based communication. *Commun. Res.* **1985**, *12*, 251–270.
7. Petersen, A.M.; Tenenbaum, J.N.; Havlin, S.; Stanley, H.E.; Perc, M. Characterizing the Shannon Entropy of Natural Language. *J. Stat. Mech. Theory Exp.* **2012**, *2012*, P04010.
8. Weick, K.E. *Sensemaking in Organizations*; Sage Publications: Thousand Oaks, CA, USA, 1995.
9. Danowski, J.A. An emerging macro-level theory of organizational communication: Organizations as virtual reality management systems. In4 *Emerging Perspectives in Organizational Communication*; Thayer, L., Barnett, G., Eds.; Ablex Publishing: Norwood, NJ, USA, 1993; pp. 141–174.
10. Danowski, J.A. WORDij Software; University of Illinois Chicago: Chicago, IL, USA, 2013.
11. Loughran, T.; McDonald, B. When is a liability not a liability? *J. Financ.* **2011**, *66*, 35–65.
12. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, Ann Arbor, MI, USA, 1–4 June 2014;6 AAAI Press: Palo Alto, CA, USA, 2014; pp. 216–225.
13. Burt, R.S. *Brokerage and Closure: An Introduction to Social Capital*; Oxford University Press: Oxford, UK, 2005.