Information Sharing for Validation: Evidence from Social Networks*

Minji Yoo

University of Miami

January 2026

Abstract

This paper examines how investors' communication incentives on social networks shape information processing in financial markets. I distinguish between influence-motivated communication, in which investors promote directional views after taking positions, and validation-motivated communication, in which investors engage with others' interpretations to assess whether trading is warranted. Using a novel classification of X.com activity in U.S. grain futures markets, I show that sentiment associated with influence-motivated communication reacts immediately to public information releases, while sentiment associated with validation-motivated communication incorporates information more gradually. These differences extend to market behavior. Validation-oriented communication tends to precede small-speculator trading activity and is associated with narrower bid–ask spreads, whereas influence-oriented communication largely follows trading and is not associated with lower market frictions. Overall, the findings indicate that social networks play a dual role in financial markets. Influence-motivated communication shapes how information is framed after positions are taken, while validation-motivated interaction facilitates belief formation and reduces information integration costs.

*I deeply appreciate my advisors, Brian Bushee (co-chair), Paul Fischer (co-chair), Frank Zhou, and Christina Zhu, for their invaluable mentorship. This paper greatly benefited from discussions with Lihong McPhail, Research Economist and Head of Academic Outreach at the CFTC, and helpful comments and suggestions from Kai Du, Wayne Guay, Luzi Hail, Bob Holthausen, Xu Jiang, Caskey Judson, Eddie Riedl, and Gwen Yu. I also thank seminar participants at the Wharton School of University of Pennsylvania, Duke University, Massachusetts Institute of Technology, Carnegie Mellon University, University of Miami, Pennsylvania State University, and Boston University.

1 Introduction

Investors increasingly use social networks to discuss market developments, interpret public information, and share trading views. A common view in both academic and popular discourse is that such communication is primarily motivated by the desire to *influence* others' beliefs in order to move prices in a favorable direction. High-profile episodes surrounding stocks such as GameStop illustrate how investors with pre-existing positions may use online platforms to promote directional narratives. While influence is clearly an important motive for investor communication, this paper argues that it is not the only one.

I propose that investors also share information on social networks for the purpose of *validation*. Validation-motivated communication reflects an attempt to assess whether trading is warranted at all by engaging with others' interpretations, expressing uncertainty, and soliciting feedback before committing capital. Unlike influence-motivated communication, which typically occurs after positions are taken and seeks to shape others' beliefs, validation-motivated communication arises prior to trading and is inherently interactive. Social networks provide a two-way communication environment in which collaboration can help investors reconcile competing interpretations and reduce uncertainty.[1]

At first glance, information sharing for validation may appear counterintuitive. By sharing information before trading on it, an investor risks accelerating its incorporation into prices and eroding potential profits. This logic, however, relies on a one-way view of information transmission. In a two-way communication environment, interaction itself can be valuable. Through discussion and feedback, investors may reduce ambiguity, assess the credibility of public signals, and determine whether available information is sufficiently reliable to justify trading. When information integration costs are high, such collaborative validation may be an optimal strategy despite the sacrifice

[1] I acknowledge the difference between the two-way exchange of information among investors and that between managers and investors. The feedback effect literature provides evidence that managers also learn from market prices and make better-informed corporate investment decisions (Dye (1983); Dye and Sridhar (2002); Goldstein (2023)). Firms also use social media to disseminate information and help investors process it (Jung, Naughton, et al. (2015); Cade (2018); Blankespoor (2018); Lee and Zhong (2022); T. J. Wong et al. (2024); Caskey, Minnis, and Nagar (2024)). Likewise, prior research studying the conference call setting documents insights from two-way communication between managers and analysts (Jung, M. H. F. Wong, and X. F. Zhang (2018); R. X. Zhang (2022); Bushee, Taylor, and Zhu (2023); Bushee and Y. Huang (2024)). In the two-way exchange between managers and investors, they generally have fixed roles and incentives. Managers (the sender) are incentivized to influence investors' beliefs, while investors (the receiver) aim to validate information. Among investors, these roles and incentives are not fixed but depend on their situations. Additionally, managers do not face the same trade-off to share information that investors do, as they do not decide between information advantages and investment efficiency.

of informational rent.

Despite the intuitive appeal of validation, empirical evidence on this motive remains limited. Existing research on investor communication has largely emphasized influence-oriented behavior and its effects on trading and prices. Whether investors systematically use social networks to validate information, and how such behavior affects sentiment dynamics, trading activity, and market frictions, remains an open question. This paper addresses that gap by distinguishing between influence- and validation-motivated communication and examining how these incentives shape information processing in financial markets.

The empirical setting is U.S. grain futures markets. These markets are well suited for studying investor information processing because speculative traders must form views about fundamentals such as supply, demand, and weather conditions while trading against commercial hedgers who dominate volume. Margin requirements and convergence to cash prices near contract maturity discipline speculative behavior, mitigating concerns about purely non-informational trading.[2] In addition, the Commodity Futures Trading Commission (CFTC) provides detailed data that allow small, unreportable speculator trading to be isolated. At the same time, grain speculators are active on social networks such as X.com, where they discuss public reports and market conditions in real time.

Using a large corpus of X.com posts related to grain futures markets from 2016 to 2023, I classify investor communication into three categories (validation, influence, and unidentified) based on the dominant incentive underlying each post.[3] The classification combines human annotation with supervised machine learning. Posts are labeled by multiple independent annotators using a shared coding protocol and used to fine-tune a RoBERTa language model, which is then applied to the full dataset. Consistent with their interpretation, validation-motivated posts exhibit greater use of uncertainty-related language and interrogative structures, while influence-motivated posts are more assertive. Audience responses further validate the classification: validation posts attract

[2] Prior research has examined commodity futures markets as a setting for studying information dissemination (e.g. Ferracuti (2022)).

[3] The choice of social media platform is driven by both data availability and institutional features relevant to the research question. During the sample period, Reddit contains only a limited number of discussion threads related to grain futures markets, whereas X.com hosts a substantially larger and more active network of speculators engaging in real-time discussion around public information releases. In addition, X.com's reply structure and high-frequency posting allow for a precise measurement of interactive communication and its temporal relationship with trading activity, which are central to the empirical design.

significantly more replies, whereas influence posts receive more passive endorsements.

The paper proceeds in three steps. First, I examine how communication incentives shape the timing of sentiment formation following scheduled public information releases. I introduce a Sentiment Intra-Period Timeliness measure that captures how quickly cumulative abnormal sentiment approaches its post-event saturation level. I find that sentiment associated with validation-motivated communication incorporates public information more slowly than sentiment associated with influence-motivated communication, which reacts sharply and immediately after news releases.

Second, I study whether these differences in communication incentives translate into differences in trading behavior. Using lead–lag regressions relating posting-share measures to small-speculator trading activity, I show that validation-oriented communication tends to precede trading, particularly when expressed through replies. In contrast, influence-oriented communication largely follows trading activity, consistent with post-position belief shaping rather than pre-trade deliberation.

Finally, I examine the implications of validation for market frictions. Focusing on bid–ask spreads as a proxy for information integration costs, I find that days with a higher share of validation-oriented replies are associated with significantly narrower spreads, even after controlling for volatility, trading activity, and macroeconomic conditions. Influence-oriented communication, by contrast, is not associated with lower spreads. These results suggest that collaborative validation improves the information environment faced by liquidity providers by reducing uncertainty and disagreement.

This study contributes to the literature on investor communication and market microstructure by showing that the incentives underlying social network activity shape how information is incorporated into markets. Rather than treating investor communication as homogeneous, the paper distinguishes between influence-motivated communication, which seeks to shape others' beliefs after positions are taken, and validation-motivated communication, which reflects an effort to assess whether trading is warranted through interaction and feedback.

First, the paper provides new evidence on the timing of sentiment formation around public information releases. Using a novel Sentiment Intra-Period Timeliness measure, I show that sentiment associated with validation-motivated communication incorporates public information more slowly than sentiment associated with influence-motivated communication. This finding highlights an important distinction between the speed of directional expression and the process of belief assessment, and demonstrates that slower sentiment dynamics need not reflect inattention or inefficiency.

Second, the paper offers mechanism evidence by examining the temporal ordering of communication and trading activity. Validation-oriented communication tends to precede small-speculator trading, particularly when expressed through replies, whereas influence-oriented communication largely follows trading. These patterns are consistent with validation reflecting pre-trade belief formation and influence reflecting post-position belief shaping.

Third, the paper contributes to the literature on information integration costs.[4] Using bid–ask spreads as a proxy for the costs faced by liquidity providers, I show that periods with a higher share of validation-oriented replies are associated with significantly narrower spreads. This evidence suggests that interactive validation can reduce disagreement about the interpretation of public information, lowering adverse selection risk and improving market liquidity.

More broadly, the paper contributes to the growing literature on investor information sharing by documenting cross-sectional variation in communication incentives among investors. While prior research has emphasized the trading outcomes associated with social networks or strategic information dissemination aimed at influencing prices (Hong, Kubik, and Stein (2005); Cohen, Frazzini, and Malloy (2007); Brown et al. (2007); Ozsoylev et al. (2014); Hvide and Östberg (2015); Ljungqvist and Qian (2016); Farrell et al. (2022)), this study shows that investors also use social networks for collaborative validation. In doing so, the paper helps reconcile mixed evidence on the role of social media in financial markets by examining the importance of communication incentives.

Finally, the paper contributes methodologically by introducing new tools for analyzing social network sentiment.[5] I aggregate sentiment using network-based weights that account for heterogeneity in influence across posts, and introduce a Sentiment Intra-Period Timeliness metric to capture the speed of sentiment incorporation following public information releases.[6] These tools may be useful for future research on investor behavior and information flow.

The remainder of the paper is organized as follows. Section 2 develops the hypotheses. Section

[4]Blankespoor, Dehaan, et al. (2019) suggest that high integration costs may be a key barrier to investors' effective use of information. However, the literature on integration costs remains limited. My paper addresses this gap by analyzing investor social media activities to show how investors with the validation incentive overcome this barrier.

[5]Prior research has shown that simple measures of linguistic tone are associated with future performance, contemporaneous returns, and bias in qualitative disclosures (Davis, Piger, and Sedor (2012); X. Huang, Teoh, and Y. Zhang (2014); Brochet et al. (2019)).

[6]Network-based weights have been applied in the context of boardroom networks in prior research (Larcker, So, and Wang (2013); Akbas, Meschke, and Wintoki (2016); Souther (2018)).

3 describes the institutional setting. Section 4 discusses the data and variable construction. Section 5 presents the empirical analyses, and Section 6 concludes.

2 Hypotheses Development

This section develops testable predictions that link communication incentives to observable patterns in sentiment dynamics and market outcomes. Building on the distinction between validation- and influence-motivated communication introduced in the previous section, the hypotheses focus on how these incentives shape the timing of sentiment incorporation and information integration costs.

2.1 Validation Incentives and the Timing of Sentiment Incorporation

Validation-motivated communication reflects an attempt to assess whether trading is warranted by engaging with others' interpretations and resolving uncertainty. Because this process involves interaction, feedback, and reassessment, it is expected to unfold more gradually than influence-motivated communication, which typically expresses directional views after positions are taken. As a result, sentiment associated with validation-oriented communication should incorporate public information more slowly following scheduled information releases.

Prior research shows that when information is costly to interpret or reconcile with existing beliefs, agents adjust more slowly. Models of limited attention and empirical evidence on investor reactions to complex disclosures document delayed responses when processing demands are high (Peng (2005); Akbas, Markov, et al. (2018)). In a social learning context, exposure to multiple interpretations can further increase the need for deliberate evaluation, reinforcing gradual sentiment formation in validation-oriented communication.

These considerations lead to the following hypothesis:

Hypothesis 1 (H1)**.** *Sentiment associated with validation-motivated communication incorporates public information more slowly than sentiment associated with influence-motivated communication.*

2.2 Validation Incentives and Information Integration Costs

Validation-oriented communication can reduce disagreement among market participants by facilitating interaction and feedback around public information. Through discussion and comparison of interpretations, investors may arrive at more similar assessments of the implications of public signals (Akbas, Meschke, and Wintoki (2016); R. X. Zhang (2022)).

From a market microstructure perspective, disagreement about the interpretation of information increases adverse selection risk faced by liquidity providers (Vashishtha (2014); Stiglitz (2004); Lawrence (2013); Leuz and Wysocki (2008); Roychowdhury, Shroff, and Verdi (2019)). When traders act on heterogeneous interpretations of the same public information, market makers face greater uncertainty about whether incoming orders are information-driven, leading them to widen bid–ask spreads. Conversely, when communication facilitates convergence in how public information is interpreted, adverse selection risk declines and bid–ask spreads narrow.

In the empirical analysis, I use the bid–ask spread as a proxy for information integration costs. To the extent that validation-oriented communication reduces disagreement about the implications of public information, periods with greater validation activity should be associated with lower bid–ask spreads. Because replies reflect interactive discussion rather than unilateral signaling, validation-oriented replies provide a particularly direct measure of collaborative validation.

Hypothesis 2 (H2)**.** *Periods with a higher share of validation-oriented communication (especially replies) are associated with lower bid–ask spreads.*

In addition to testing the hypotheses above, the empirical analysis also examines the temporal ordering of communication and trading activity as complementary evidence on the mechanisms underlying validation and influence. While not framed as a separate hypothesis, differences in whether communication precedes or follows trading help distinguish pre-trade belief formation from post-trade belief shaping and provide additional insight into how communication incentives manifest in observable behavior.

3 The Setting: Grain Futures Markets

3.1 Futures Contract Structure

The setting of this study is the grain futures market, with a specific focus on the three principal commodity futures: corn, wheat, and soybeans. These commodities are traded through the Chicago Board of Trade (CBOT), which is currently a part of the CME Group. Futures contracts are derivatives linked to their underlying assets' market value. Essentially, these contracts forecast the future price of an asset, embodying traders' predictions. For instance, the price of a September corn futures contract reflects the anticipated market value of physical corn in September.

Each grain futures contract has a standardized expiration date—the last business day before the 15th calendar day of the contract's maturity month. By the expiration date, holders of open futures positions can convert these contracts into actual physical grain. The CBOT designates official warehouses to facilitate the physical delivery process, linking these futures contracts closely with the cash grain markets they represent. However, to avoid the complexities of physical delivery, most grain futures contracts are settled before expiration, either rolled over to new contracts or cash-settled. At expiration, the futures prices should align with cash prices at the delivery warehouse, a principle known as convergence. Grain futures contracts start trading years before their actual expiration date, allowing for significant price variance from local cash prices initially. Yet, as a contract nears expiration, its price increasingly aligns with the cash price. Similar to cash markets, grain futures are influenced by global events, including weather conditions, political conflicts, and trade policies, which can significantly impact supply and demand dynamics.

3.2 Rules of The Markets

Trading commodities involves a high degree of leverage, making it a potentially profitable yet risky activity that requires a significant learning curve for the typical investor. Exchanges enable traders to buy and sell full-size futures contracts without providing the total value of those contracts upfront, effectively using margin—a form of collateral—to control large positions with a relatively small capital outlay. For example, if corn is priced at $5.00 per bushel and deemed unlikely to vary more than $0.50 per week, the exchange might only require a $2,500 margin for a 5,000-bushel corn contract. Should the price fluctuate beyond fifty cents per bushel, the trader faces a margin call, necessitating additional funds to cover potential losses. These margin requirements, aimed at mitigating risk, can change in response to market volatility. The CME Group employs the Standard Portfolio Analysis of Risk system to ensure traders have enough capital to cover 99 percent of potential market outcomes. An increase in margin requirements can occasionally compel some investors to exit their positions, potentially triggering a wave of market liquidation due to the additional financial strain, thereby impacting overall market liquidity and price stability.

The exchange sets regulatory price limits on the daily movements of their commodity futures prices, subject to approval by the CFTC to curb excessive volatility and ensure market integrity. For example, corn futures prices are restricted to a 25-cent move per day, soybeans to 70 cents, and wheat to 30 cents. These limits expand to 150% of their original range the day following a contract being 'locked limit up' or 'locked limit down,' which occurs when prices hit their maximum allowable limit. Unlimited price movement is allowed from the first notice day to the expiration date (during the contract's last two weeks) to facilitate the convergence of futures and cash prices.

Based on demand, the exchange might offer more than a dozen futures contracts for a single commodity (i.e., corn), spanning several years at any given time. The contract nearest its expiration is termed the 'front-month' contract, usually possessing the highest liquidity due to its proximity to delivery. For instance, after the December 2023 contract expires, the March 2024 contract becomes the front-month. Price charts over long periods often use a continuous series of front-month prices, so the data points could skip from one contract to another when the front-month contract expires. This front-month contract typically has the most trading volume and open interest, making it easier for traders to enter or exit positions. However, engaging in a deferred contract can be more cost-effective for investors planning to hold positions for an extended period, avoiding the repeated transaction charges associated with rolling over contracts.

Most commodities exhibit higher deferred prices compared to the nearby (current) prices. This positive futures-spread structure, known in the grain industry as 'carry,' reflects the incremental costs of carrying the grain in storage from one month to the next. For example, the carry between September and December corn futures might typically be 15 cents per bushel, equating to 5 cents per month for three months of storage. Conversely, market conditions can sometimes invert, with near-term futures priced higher than distant futures. We may see July or September corn futures significantly outpacing December futures in price, especially when the market anticipates a substantial incoming harvest by December. Such inversions highlight expectations of tighter near-term supplies or other market dynamics influencing demand and supply perceptions.

3.3 Investors in The Markets

The CFTC is the government agency regulating futures markets. They implement position limits, thereby preventing any single entity or small group from monopolizing a given market. In addition, the CFTC provides a detailed breakdown of market participants and their positions through weekly Disaggregated Commitment of Traders (DCOT) reports. The first group in this report is the 'Producer/Merchant/Processor/User' category, which can be referred to as commercial traders. These traders are hedgers in the commodity markets who make up the vast bulk of grain futures trade. ABCD+ grain trading companies belong to this group.[7] Between 2006 and 2011, commercial entities traded 63% of all corn futures contracts and 76% of all wheat futures contracts cleared by the Chicago Board of Trade (Kub (2012)). The second group is the 'Swap Dealers'. They are typically large banks that write private hedge contracts for their commercial clients and use the futures markets to offset their own subsequent risk. The 'Managed Money' category refers to speculators with capital in managed futures accounts and hedge funds trading commodity futures, and this category can be seen as pure speculators. During the same period, the Managed Money category traded 16% of the corn contracts and 24% of the wheat contracts in the Chicago futures markets (Kub (2012)).

Along with these reportable positions, the CFTC discloses the number of contracts held by nonreportable speculators whose positions do not exceed the threshold and are not obligated to report to the CFTC. The nonreportable speculators are the main interest of the study. I hypothesize the nonreportable speculators are likely to overlap the most with those active on social networks such as X.com to share their information.

[7] Traditionally, the term 'ABCD grain companies' refers to Archer Daniels Midland (ADM), Bunge, Cargill, and Louis Dreyfuss. In recent years, Glencore, COFCO International, and Wilmar have emerged as Asian giants, leading the industry and the media to now refer to the seven as the 'ABCD+' group of grain trading companies. Collectively, ABCD+ handles half of the international trade in grain and oilseed (Fitzgerald (2021)).

4 Sample and Variable Construction

4.1 Social Network Data Collection and Cleaning Process

This section outlines the process of collecting and cleaning X.com posts to construct a sample relevant to grain futures markets. On X.com, users indicate the topics of their posts by including hashtags and cashtags. Since some of these keywords are also associated with other prominent networks, I take additional steps to refine the sample and ensure its relevance to grain futures markets.

Initial sample posts are collected using the hashtags (#wheat, #corn, #soybean, #oatt) and cashtags ($zw, $zc, $zs, $corn, $weat, $wheat, $soyb) that are widely used among speculators trading in grain futures markets. Daily queries are conducted from January 2016 to May 2023. Through exploratory data analysis, I identify three prominent networks using the same hashtags: food/health, research, and producer/farmer networks.[8] The first two categories are easily distinguishable by their unique vocabularies (e.g., 'low carb', 'allergy', 'archaeology', and 'nematode'). However, differentiating producer/farmer networks from investor networks is more challenging due to their shared interests in commodity markets, international trade, weather conditions, and agricultural cycles.

To refine the data, I implement a two-stage filtering process. First, I scrape posts using a combination of keywords and exclusion words that are unique to non-investor networks. This initial filtering results in a more targeted sample of 249,946 posts. In the second stage, I further filter posts algorithmically, using vocabularies specific to non-investor networks that were not included in the initial exclusion due to search option limitations. For instance, food/health networks are identified through terms like 'chowder,' 'bread,' and 'sourdough,' while research networks are filtered using words like 'genetic' and 'phosphorus.' Additionally, producer/farmer-specific hashtags such as #farm365' and terminology like 'drilling' (for planting wheat) are used to further refine the sample.

These non-investor network vocabularies are identified through topic modeling, an unsupervised machine-learning technique that efficiently summarizes large volumes of textual data without requiring manual review. This comprehensive filtering process ensures that each post in the final

[8] Fields such as agronomy, entomology, plant pathology, and environmental science extensively study these crops.

dataset is relevant to grain futures markets, resulting in a curated sample of 148,072 posts. Detailed vocabulary lists for each filtering phase are documented in Appendix A. Appendix B provides additional details on the topic modeling process and highlights the top topics extracted from the final sample, which illustrate speculators' interest in crop conditions, weather forecasts, and the potential impacts of international conflicts. Figure IA.2 in the Internet Appendix further confirms the relevance of the final sample to grain futures markets with a comprehensive list of the topics.

4.2 Classification of Posts

I classify X.com posts into three mutually exclusive categories based on the dominant incentive underlying the post: influence, validation, and unidentified. To do so, I combine manual annotation with supervised machine learning. Posts were independently labeled by three human annotators using a shared coding protocol that defined each category and provided illustrative examples. Annotators completed the initial labeling independently and without access to each other's classifications. When all three annotators agreed, the assigned label was retained. When disagreement occurred, the final label was determined by majority rule (i.e., the category assigned by at least two of the three annotators). This procedure preserves independent judgment while ensuring a systematic and transparent resolution of disagreements. After reconciliation, the final hand-labeled training sample consists of 1,620 posts that are randomly selected, including 713 posts classified as unidentified, 606 as influence, and 301 as validation.

This labeled dataset is used to fine-tune a RoBERTa language model, which is well suited for short, informal text commonly observed on social media. The fine-tuned RoBERTa model is then applied to classify the remaining posts in the full dataset. This procedure allows us to leverage high-quality human judgment while scaling classification to a large corpus. Applying the fine-tuned model to the unlabeled posts yields the following distribution: 69,871 posts classified as unidentified, 44,754 as influence, and 31,827 as validation.

To focus on posts that plausibly generate information sharing and collaboration for validation, I restrict attention to original posts that receive more than one reply. After imposing this restriction, the final sample of original posts consists of 18,983 posts, including 7,673 validating posts, 5,262 influencing posts, and 6,048 unidentified posts. For each original post in the filtered sample, I collect all associated replies. This results in 19,030 replies to influencing posts, 33,009 replies to validating posts, and 28,007 replies to unidentified posts. Combining original posts and replies yields a final sample of 98,342 observations, which constitutes the primary dataset used in the empirical analyses.

To assess the validity of the post classifications, I examine whether the linguistic features of posts align with the underlying incentives implied by each category. Specifically, I analyze the use of linguistic cues associated with uncertainty and information seeking. For each post, I count (i) the number of uncertainty-related words, defined using the Loughran–McDonald Dictionary (Loughran and Mcdonald (2011)) and General Inquirer Dictionary (Stone, Dunphy, and M. S. Smith (1966)), and (ii) the presence of interrogative structures, including question marks ("?") and common question formats (who, what, when, where, why, and how). I then compare these linguistic features across validating and influencing posts. Consistent with the interpretation of validation as information-seeking behavior, validating posts exhibit significantly higher usage of uncertainty-related words and interrogative language.[9] In contrast, influencing posts contain substantially fewer such cues, consistent with their intended role of asserting views or persuading others rather than soliciting feedback. These patterns provide independent linguistic support for the classification scheme.

As an additional validation of the classification method, I examine how audiences respond to posts motivated by different incentives. The underlying premise is that posts driven by distinct communication incentives elicit systematically different types of reactions. Posts that are assertive or confidence-driven are more likely to receive passive forms of endorsement, such as likes and retweets, whereas posts that seek input or express uncertainty are more likely to invite active engagement in the form of replies and discussion. Consistent with this premise, influencing posts receive relatively more passive responses and fewer direct replies, while validating posts attract significantly more active responses in the form of replies (Table 2). These differences in audience behavior are difficult to explain by linguistic variation alone and instead reflect how users interpret and respond to the intent of the original post. Together with the linguistic evidence, these pat-

[9]

	LM Uncertain	*GI Unsure*	*5w1h*	*Inquiry*	*Q Marks*	*Exclam.*	*LM Strong*
Validating (Mean)	0.30	0.52	0.27	0.02	0.35	0.11	0.05
Validating (Std)	0.58	0.78	0.53	0.15	0.67	0.47	0.23
Influencing (Mean)	0.16	0.28	0.07	0.00	0.05	0.10	0.07
Influencing (Std)	0.43	0.57	0.28	0.03	0.33	0.65	0.30

terns provide further support that the classification captures meaningful differences in underlying communication incentives.

4.3 Proxy for Shared Information: Aggregated Sentiment

This study employs aggregated sentiment at the network level as a proxy for shared information, providing an efficient and effective means of summarizing large volumes of textual data. Prior research has established that sentiment correlates with future firm performance, contemporaneous returns, and the bias inherent in qualitative disclosures (Davis, Piger, and Sedor (2012); X. Huang, Teoh, and Y. Zhang (2014); Brochet et al. (2019)), supporting its construct validity. In constructing the aggregate measure, each post's sentiment is weighted by its relative importance within the network.

The sentiment of each post is derived using a RoBERTa-base large language model (LLM) trained on a dataset of 124 million posts from January 2018 to December 2021. This model has been further fine-tuned for sentiment analysis tasks, using the TweetEval benchmark for evaluation (Barbieri et al. (2020); Camacho-Collados et al. (2022); Loureiro et al. (2022)).[10] Using this LLM specialized in X.com sentiment analysis, I assign each post a sentiment label (negative sentiment as -1, neutral as 0, and positive as 1) along with a confidence score, the likelihood that the assigned label is true. The sentiment of each post is then calculated by multiplying the sentiment label by its corresponding confidence score.

Employing LLMs for sentiment analysis offers significant advantages over simply counting negative and positive words. For example, the LLM interprets the phrase "Covid cases are increasing fast!" as negative, assigning it a confidence score of 0.7236, which indicates a strong negative sentiment. Unlike basic word counting, LLMs excel at contextual understanding, preventing potential misinterpretations when the meaning depends on surrounding words and phrases. Moreover, LLMs can handle negations and idioms, as they are trained on large datasets that include such linguistic nuances.[11]

[10] TweetEval is a benchmarking framework designed to assess and compare the performance of LLMs across various tweet-based tasks. See also https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[11] However, one limitation of this approach is that it only uses textual information. People often use memes and emojis on social networks to express their thoughts, mixing visual and textual elements. While LLMs are good at analyzing text, they struggle with multi-modal inputs that combine images and text. Although some models like GPT-4 have beta features for multi-modal analysis, reliably interpreting this complex data is still difficult.

To capture sentiment at the network level, I aggregate individual post sentiment using degree centrality from network theory. This method weights each post by its level of connectedness, reflecting its potential to influence others in the network. As a result, the aggregated sentiment more accurately reflects the overall sentiment circulating within the investor community. This approach moves beyond treating all posts equally and instead recognizes that some voices carry more weight than others. Prior research has shown that connectivity shapes the information environment, particularly in the context of corporate boards (Larcker, So, and Wang (2013); Akbas, Meschke, and Wintoki (2016); Souther (2018)).

As illustrated in Figure IA.1, posts vary in their level of connectivity and, by extension, their influence within the network. The concept of centrality plays a key role in identifying which nodes (in this case, posts) have the highest influence, a principle widely used in social network analysis. To capture this, I use degree centrality, defined as the number of direct connections a node has. In the context of the X.com sample, a post's connections are measured as the sum of likes, reposts, and replies it receives, incremented by one to avoid zero values.

Subsequently, each post's weight within the network is calculated by dividing its number of eangagment metrics by the total number of engagement metrics across all posts in the network. Daily network sentiment is then computed by taking a weighted average: multiplying each post's sentiment score by its corresponding weight and summing these values for all posts on a given day. That is,

$$\text{Weight By Engagement}_i = \frac{|\,\text{Likes, Reposts, and Replies of post}_i\,| + 1}{\sum_{k \in S} |\,\text{Likes, Reposts, and Replies of post}_k\,| + 1},$$

$$\text{Aggregated Sentiment} = \sum_{i \in S} \text{Sentiment Label}_i \times \text{Confidence Score}_i \times \text{Weight}_i,$$

where S is the set of sample posts each day. This measure emphasizes sentiment expressed by users with active real-time engagement at the time of posting.

To account for heterogeneity in potential audience reach across posts, I also construct a follower-weighted sentiment measure. This approach assigns greater weight to posts authored by users with larger follower bases, reflecting their higher potential visibility within the network. Specifically, a post's weight is proportional to the number of followers of its author at the time of posting, incremented by one to avoid zero values. Daily network sentiment is then computed as a weighted average of post-level sentiment scores:

$$\text{Weight By Follower}_i = \frac{|\text{Followers of author}_i| + 1}{\sum k \in S |\text{Followers of author}_k| + 1}.$$

This measure emphasizes sentiment expressed by users with broader potential reach, independent of realized engagement.

In contrast to network-based weighting, the simple average aggregation treats all posts as equally influential within the network, regardless of their connectivity or author characteristics. Under this approach, each post contributes identically to the daily sentiment measure. Daily network sentiment is therefore computed as the simple average of post-level sentiment scores across all posts observed on a given day. Formally,

$$\text{Simple Avg.}_i = \frac{1}{|S|}.$$

This measure captures the average sentiment expressed in the network without emphasizing posts that receive greater attention or engagement. Aggregated sentiment measures form the basis of the dependent variables in Table 3, which will be introduced in a later section.

4.4 Proxy for Prevailing Incentive: Posting Share

In addition to aggregating sentiment, I construct *share-based measures* that characterize the *composition* of X.com activity within the network. Whereas aggregated sentiment captures the direction of opinions expressed, share measures capture the relative prevalence of different types of incentives, providing insight into how attention and participation are distributed across the network within a given period.

The simple average share measure treats all posts as equally informative and focuses solely on the relative frequency of posts. For each day (or hour), I compute the fraction of total posts accounted for by a given category. Let N denote the total number of posts observed in a given period, and let N_g denote the number of posts belonging to category g. The simple average share is defined as

$$\text{Share}_g^{\text{Simple Avg.}} = \frac{N_g}{N}.$$

This measure captures the baseline composition of network activity, abstracting from differences in user reach or post-level attention. It reflects how frequently a given type of content appears in the overall information flow, independent of its potential or realized influence.

To account for heterogeneity in potential audience reach across posts, I also construct follower-weighted activity shares. This approach assigns greater weight to posts authored by users with larger follower bases, reflecting their higher potential visibility within the network. Specifically, each post is weighted by the number of followers of its author at the time of posting, incremented by one to avoid zero values. The follower-weighted share for category g is defined as

$$\text{Share}_g^{\text{Follower}} = \frac{\sum_{i \in g}(|\text{Followers}_i| + 1)}{\sum_{k \in S}(|\text{Followers}_k| + 1)},$$

where S denotes the set of all posts observed in the period. This measure captures the relative prevalence of content after accounting for differences in potential audience reach, emphasizing posts authored by users with broader visibility even if their realized engagement is limited.

Finally, I construct engagement-weighted activity shares that reflect the distribution of *realized attention* across posts. Drawing on network theory, I proxy post-level connectedness using degree centrality, measured as the sum of likes, reposts, and replies a post receives, incremented by one to avoid zero values. The engagement-weighted share for category g is defined as

$$\text{Share}_g^{\text{Engagement}} = \frac{\sum_{i \in g}(|\text{Likes}_i + \text{Reposts}_i + \text{Replies}_i| + 1)}{\sum_{k \in S}(|\text{Likes}_k + \text{Reposts}_k + \text{Replies}_k| + 1)}.$$

This measure captures how network attention is allocated across different types of posts, emphasizing content that elicits greater interaction from other users. In contrast to follower-weighted shares, which reflect potential reach, engagement-weighted shares reflect realized influence within the network. Posting share measures used as independent variables in Table 5 and Table 6.

4.5 Descriptive Analysis of Sample Posts

Most of the analysis in this study is conducted at the post level. That is, I do not assume that a speculator's motivation is fixed over time. Instead, the same individual may act as a validator on one day and as an influencer on another, depending on the context and circumstances. However, it is also informative to examine how frequently accounts switch between clusters over time.

In the final sample of 98,342 posts, there are 27,191 unique accounts, of which 17.28% (4,701 accounts) appear in more than one cluster. Figure 1 shows that the correlation between the number of Influencing Posts and Validating Posts at the account level is 71%, suggesting considerable overlap in user participation across these two categories. In contrast, the correlations between Unidentified Posts and Influencing Posts, and between Unidentified Posts and Validating Posts, are lower—at 46% and 56%, respectively. At the account level, 65.8% of the authors of original posts respond to other' posts as well.

It may also be interesting to explore the identity of contributors who do not initiate posts but participate in discussions. As illustrated in Appendix C, many of these individuals appear to be local farmers (who may act as both hedgers and speculators), crop watchers, and market analysts. Based on my observations, I conjecture that these contributors may be driven by a variety of behavioral motives—including quid-pro-quo dynamics and entertainment—in addition to the incentives to influence or validate. In the context of this study, it is important that users of the social network expect to receive meaningful feedback from others. As long as this expectation is, on average, accurate in the setting, disentangling individual motivations of the contributors is for future study.

Another observation from comparing view counts and engagement metrics is the presence of free-riders—users who passively observe others' posts without asking questions or engaging in discussions. This observation has two important implications. First, my empirical tests of the validation incentive rely on observable sharing activities, yet those who actively share on social networks likely represent only a small subset of the broader population of investors with a validation incentive. Second, free-riding is easy but not very beneficial, especially when the validator needs timely feedback. Validators may also build a reputation for helping others and, in turn, become future influencers as shown in Figure 1 or expect reciprocal support when they themselves seek validation.[12]

5 Analyses and Results

5.1 Validation Incentives and the Timing of Sentiment Incorporation (H1)

A central implication of validation-motivated sharing is a difference in the *timing* of information processing. Speculators who engage in influence-motivated communication typically have already taken positions and seek to shape others' beliefs in order to move prices in a favorable direction. In contrast, speculators motivated by validation use social networks to assess whether trading is warranted at all. This objective requires evaluating the credibility of public information, reconciling it with prior beliefs, and refining expectations before committing capital.

As a result, validation-motivated communication is expected to be associated with slower and qualitatively different incorporation of new information into sentiment measures than influence-motivated communication. Rather than expressing immediate directional conviction, validation-oriented posts reflect ongoing assessment and refinement of beliefs, implying delayed or uneven sentiment accumulation following public information releases.

I test this prediction in the context of scheduled public information releases in grain futures markets, which represent the primary source of common information for speculators. These releases include the Acreage Report, the World Agricultural Supply and Demand Estimates (WASDE), the Crop Production Report, and the Crop Progress Report.[13]

To motivate the empirical design, Figure 2 plots the average cumulative abnormal sentiment from 12 hours before to 36 hours after report releases, separately for Influencing and Validating Posts. The figure reveals a clear pattern: sentiment in Influencing Posts reacts sharply immediately following the release, consistent with post-position directional expression, while sentiment in Validating Posts peaks more gradually and continues to adjust beyond the initial response window.

[12] Although I do not formally explore it in this study, reputation building may act as a moderating factor. Some contributors may build a reputation in order to influence more effectively, while others may influence as a way to build their reputation. Validators, too, may be motivated to establish credibility as posters. In the case of news organizations, their involvement amplifies dissemination, but it is difficult to attribute their behavior to either incentive, as they are not trading themselves.

[13] The Acreage Report, released at the end of June, presents planted and harvested acreage by state. WASDE is a monthly report providing global supply–demand forecasts for major grains. The Crop Production Report, also issued monthly, contains U.S. production, yield, and weather information. The Crop Progress Report is released weekly during the growing season and summarizes planting, development, and crop conditions across major producing states.

This visual evidence motivates a formal test of whether validation-motivated sharing incorporates public information more slowly into sentiment measures than influence-motivated sharing.

Sentiment Intra-Period Timeliness. To quantify the speed of sentiment incorporation, I construct an hourly *Sentiment Intra-Period Timeliness* (Sentiment IPT) measure, adapted from the intraperiod timeliness framework used to study price discovery around news events (Butler, Kraft, and Weiss (2007); Bushman, A. J. Smith, and Wittenberg-Moerman (2010); Twedt (2016); Blankespoor, deHaan, and Zhu (2018)). Sentiment IPT captures how quickly cumulative abnormal sentiment approaches its post-event saturation level within a given time window.

Abnormal sentiment is defined at the hourly level as aggregated sentiment within a cluster minus a three-year rolling average baseline computed at the same week-of-year, thereby accounting for strong seasonal patterns in grain markets following planting and harvesting cycles. Let CumAS_T denote cumulative abnormal sentiment from the report release hour (hour 0) through hour T. Sentiment IPT over $[0,T]$ is defined as

$$\text{Sentiment IPT}^{[0,T]} = \frac{1}{2}\sum_{t=0}^{T}\left(\frac{\mathrm{CumAS}_{t-1}+\mathrm{CumAS}_t}{\mathrm{CumAS}_T}\right) = \sum_{t=0}^{T-1}\left(\frac{\mathrm{CumAS}_t}{\mathrm{CumAS}_T}\right)+\frac{1}{2},$$

which corresponds to the area under the normalized cumulative abnormal sentiment curve, approximated by trapezoids. By construction, Sentiment IPT measures the *timing* of sentiment incorporation relative to a cluster's own saturation level, rather than the magnitude of sentiment itself. Because $CumAS_T$ appears in the denominator, Sentiment IPT compares when sentiment arrives rather than how much sentiment arrives.

Empirical design. Rather than comparing separate windows, I adopt a stacked-horizon design that directly isolates delayed incorporation. For each event, cluster, and commodity, I compute Sentiment IPT over two horizons: 12 hours and 36 hours following the release. Most reports, the Acreage Report, WASDE (World Agricultural Supply and Demand Estimates), Monthly Crop Production Reports are scheduled to release at 12:00 PM ET. Thus, the 12-hour window captures immediate same-day responses, while the 36-hour window encompasses overnight digestion and next-day responses. Observations from the two horizons are stacked, and the following regression is estimated:

$$\text{Sentiment IPT}_{eckT} = \alpha + \beta_k \times \text{Cluster}_k + \lambda \cdot 1[T = 36] + \theta_k \cdot 1[T = 36] \times \text{Cluster}_k$$
$$+ \gamma' X_e + \mu_{\text{report type}} + \varepsilon_{eckT}, \quad (1)$$

where e indexes report events, c commodities, k tweet clusters, and $T \in \{12, 36\}$. The omitted cluster is *Unidentified.* Commodity and Report-type fixed effects absorb systematic differences across commodity and report categories, and standard errors are clustered at the report-type level. Additionally, I control for broad asset-market conditions using equity, currency, rates, commodity, energy, metal, and crypto benchmarks.[14]

In this specification, β_k measures differences in early (0–12 hour) sentiment incorporation, while θ_k captures the incremental contribution of sentiment between 12 and 36 hours. H1 predicts that validation sharing incorporates information more slowly, implying a smaller (or even negative) $\theta_{\text{Validation}}$ relative to $\theta_{\text{Influence}}$.

Results. Table 3 reports the main regression results for three sentiment constructions. Across all specifications, Influencing Posts exhibit significantly faster sentiment incorporation within the first 12 hours following public report releases. Validation Posts also respond in the initial window, but to a significantly smaller extent. More importantly, the incremental 12–36 hour response differs sharply across clusters. Sentiment in Influencing Posts continues to accumulate after the initial window, consistent with persistence in expressed directional views, whereas sentiment in Validating Posts exhibits a negative incremental contribution, consistent with ongoing reassessment and refinement of views rather than continued accumulation in a single direction. Note that a negative incremental coefficient does not mean sentiment reverses mechanically, but it means late sentiment contributes less relative to early saturation.

Table 4 formalizes these differences using linear hypothesis tests. Validation sentiment is signif-

[14]Control variables include a stock market index (*SNP500*), a benchmark for movements in the US dollar relative to a basket of world currencies, which can affect international trades (*USD Index Futures*), the risk-free rate (*US 3M Bond*), the long-term interest rate(*US 5Y Bond*), a benchmark that tracks a cryptocurrency market (*Bitcoin Futures*), a benchmark that tracks soft commodities futures markets such as coffee, cocoa, sugar, and cotton (*Dow Jones Softs*), a benchmark for the energy sector, including crude oil, natural gas, heating oil, and gasoline (*Dow Jones Energy*), and a benchmark for the metals sector, including both precious metals and industrial metals (*Dow Jones Metals*). These market variables are log-transformed to mitigate skewness and persistence. *Hedonometer* captures general platform sentiment unrelated to grain markets.

icantly slower than influence sentiment within the first 12 hours, and the incremental 12–36 hour contribution is substantially smaller for validation than for influence across all sentiment measures. Taken together, the results suggest that influence-driven sentiment reacts immediately and continues to adjust, while validation-driven sentiment responds more cautiously and does not contribute to sustained diffusion. These findings are consistent with H1 and highlight a fundamental difference in the role played by influence and validation in shaping information processing dynamics around public information releases.

5.2 Validation Incentives and the Timing of Trading Activity

The preceding analysis shows that validation- and influence-motivated communication differ sharply in how public information is incorporated into expressed sentiment. To better understand how these differences operate in practice, this section examines the temporal relationship between social network communication and *trading behavior*. While not formulated as a separate ex ante hypothesis, differences in the timing of communication and trading provide informative evidence on whether validation-oriented communication reflects pre-trade belief formation or post-trade interpretation, and how influence-oriented communication relates to investors' positioning decisions.

Specifically, if validation-motivated communication reflects an effort to assess whether trading is warranted, such communication should tend to occur prior to trading activity as beliefs are formed and refined. In contrast, if influence-motivated communication is used to shape others' beliefs after positions are taken, it should occur contemporaneously with or following trading. I examine these implications using lead–lag regressions that relate posting activity shares to small-speculator trading intensity.

To examine whether validation- and influence-motivated sharing activity is associated with different trading behaviors, I analyze the dynamic relationship between posting activity composition and small-speculator trading intensity. Importantly, the explanatory variables in this analysis are not sentiment measures. Instead, I use *share-based posting intensity measures* that capture how X.com activity is distributed across different post types within a given hour. Whereas aggregated sentiment reflects the direction of opinions expressed, these share measures characterize shifts in attention and participation across Influencing and Validating Posts.

Empirical design. To examine whether social network activity tends to occur before or after trading activity, I estimate lead–lag regressions that include both past and future values of posting-share measures. For each commodity i and time t, I estimate the following specification:

$$\log(1+Y_{i,t}) = \alpha_i + \phi\log(1+Y_{i,t-1}) + \sum_{k=1}^{K}\beta_{-k}S_{i,t-k} + \sum_{k=1}^{K}\beta_{+k}S_{i,t+k} + \varepsilon_{i,t}, \tag{2}$$

where $Y_{i,t}$ denotes trading activity for commodity i at time t. The dependent variable is measured as the number of trades executed by *small, unreportable traders.*[15] These trades primarily reflect participation by smaller market participants rather than large, reportable institutions.

$S_{i,t}$ denotes a posting-share measure for either validation or influence. The dependent variable is log-transformed to mitigate skewness and persistence. The specification includes commodity fixed effects α_i and a lag of the dependent variable to account for serial correlation in trading activity. Lagged terms $S_{i,t-k}$ capture X.com activity occurring prior to trading at time t, while lead terms $S_{i,t+k}$ capture X.com activity occurring after trading. The regression is estimated separately for validation and influence measures to preserve parsimony and to avoid mechanical correlations induced by share-based measures. Standard errors are heteroskedasticity- and autocorrelation-robust.

Lead–Lag Asymmetry Statistic. To summarize temporal ordering, I construct an asymmetry statistic defined as:

$$\mathcal{A} = \sum_{k=1}^{K}\beta_{-k} - \sum_{k=1}^{K}\beta_{+k}. \tag{3}$$

A positive value of $\mathcal{A}$ indicates that X.com activity tends to precede trading activity, while a negative value indicates that X.com activity tends to follow trading activity. Statistical inference is based on a joint Wald test of the null hypothesis across leads and lags $\sum_{k=1}^{K}\beta_{-k} = \sum_{k=1}^{K}\beta_{+k}$.

Importantly, the lead–lag design does not test predictability. Instead, it provides evidence on the temporal ordering of social network communication and trading activity.

Results. Table 5 reports lead–lag asymmetry results using hourly data. Validation-oriented communication exhibits positive lead–lag asymmetry. For both Validating Posts and Replies to

[15] Defined as trades below the Commodity Futures Trading Commission (CFTC) reporting threshold and classified as unreportable speculator activity in the weekly Disaggregated Commitments of Traders (DCOT) report.

Validating Posts, the sum of lag coefficients exceeds the sum of lead coefficients, indicating that validation activity tends to occur prior to changes in trading activity. The asymmetry is more pronounced for reply-based validation, consistent with collaborative information processing and consensus formation preceding trading decisions.

In contrast, influence-oriented communication exhibits a sharply different pattern. For Influencing Posts, the asymmetry statistic is strongly negative, indicating that trading activity tends to occur prior to influencing Posts. This finding is consistent with post-position signaling, whereby traders act before communicating directional narratives. Replies to Influencing Posts exhibit weaker and less consistent asymmetry, suggesting amplification rather than primary information production. Negative coefficients in Equation (2) indicate lower trading intensity conditional on higher X.com activity. In this setting, reduced trading is economically meaningful and consistent with reduced disagreement or delayed participation. The key result is not the sign of individual coefficients but the relative timing captured by the asymmetry statistic.

Overall, the lead–lag evidence reveals a clear distinction between validation and influence. Validation-oriented communication tends to precede trading, consistent with costly information processing and integration, whereas influence-oriented communication largely follows trading, consistent with post-position signaling or ex post narrative framing. These findings suggest that validation-oriented communication delays directional expression while facilitating earlier resolution of uncertainty prior to trading.

5.3 Validation Incentives and Information Integration Costs (H2)

A central implication of validation sharing is not necessarily that investors uncover directional price signals, but that they reduce the cost of integrating dispersed information into prices through collaboration. When market participants validate one another's interpretations, clarify uncertainty, and converge on shared assessments, liquidity providers face less adverse selection risk. As a result, information integration costs, rather than expected returns, should decline.

To test this mechanism, this section examines whether validation sharing on social networks is associated with lower market frictions. I use the daily bid–ask spread as a proxy for information integration costs faced by liquidity providers. In market microstructure models, wider spreads reflect greater adverse selection risk arising from heterogeneous interpretations of information.

Empirical design. The lead–lag evidence shows that validation-oriented communication is concentrated in replies and that reply-based validation exhibits the strongest temporal precedence relative to trading. Accordingly, the bid–ask spread analysis focuses on the incentive composition of replies rather than original posts. Conceptually, replies are the primary venue for validation because they reflect interaction and feedback rather than unilateral signaling. Methodologically, focusing on reply shares isolates incentive composition while holding total social activity fixed, avoiding mechanical constraints that arise when originals and replies are combined. The spread regressions therefore test whether days with a higher share of validation-oriented replies are associated with lower information integration costs, as reflected in narrower bid–ask spreads.

The key explanatory variables are the daily shares of Replies to Validating Posts and Influencing Posts in total posting activity. Focusing on shares allows the analysis to isolate incentive composition effects while holding overall attention fixed.

$$\begin{aligned}\ln(\text{Bid-Ask Spread}_{k,t}) = {} & \alpha + \beta_1 \,\text{Rep Val.\%}_{k,t} + \beta_2 \,\text{Rep Inf.\%}_{k,t} + \gamma \, \ln(\text{Total Activity}_{k,t}) \\ & + \rho \, \ln(\text{Intraday Price Range}_{k,t-1}) + \delta \, \ln(\text{Daily \# of Trades}_{k,t}) \\ & + \theta' \mathbf{Z}_t + \mu_k + \lambda_{w(t)} + \varepsilon_{k,t},\end{aligned} \tag{4}$$

Formally, I estimate regressions of daily bid–ask spreads on the incentive composition of reply activity, controlling for total activity, lagged daily price range, which proxies for market volatility while avoiding mechanical correlation with contemporaneous bid–ask spreads, and daily number of retail trades, to test whether validation-oriented collaboration reduces information integration costs. All specifications include commodity fixed effects and week fixed effects, ensuring that identification comes from within-commodity variation relative to common weekly market conditions. Standard errors are clustered by week and commodity.

Main results. Table 6 reports the results using three alternative constructions of incentive composition: simple average shares, follower-weighted shares, and engagement-weighted shares. Across all specifications, a higher share of Replies to Validating Posts is associated with significantly lower bid–ask spreads. In contrast, the share of Replies to Influencing Posts is not statistically related to spreads in any specification.

The magnitude of the effect is economically meaningful. A 10 percentage point increase in the share of validation replies is associated with a 1.5–2 percent reduction in daily bid–ask spreads, holding constant volatility, trading activity, macro conditions, and fixed effects. This corresponds to a reduction of approximately 0.2–0.3 basis points in daily spreads in the sample. While modest at the trade level, such reductions represent meaningful improvements in market liquidity when aggregated over time and volume.

Importantly, total posting activity is not significantly related to bid–ask spreads once incentive composition is controlled for. This suggests that the observed relationships are unlikely to be driven by overall attention or noise, but instead reflect differences in communication incentives. In particular, validation-oriented interaction is associated with lower information integration costs, whereas influence-seeking behavior is not.

These findings support the hypothesis that investors use social networks to overcome high information integration costs through collaboration. Validation-motivated replies reflect active engagement with others' interpretations, enabling market participants to refine beliefs and reduce uncertainty. From the perspective of liquidity providers, such collaboration lowers the cost of processing information, resulting in narrower bid–ask spreads. This mechanism differs conceptually from traditional information dissemination. Influence-oriented posts may attract attention or amplify opinions, but they do not systematically reduce market frictions. In contrast, validation-oriented communication appears to improve the information environment by fostering convergence and shared understanding, consistent with the collaborative role of social networks emphasized in this paper.

6 Conclusion

This paper examines how the incentives underlying social network communication shape information processing and market outcomes. While prior research has largely treated investor sentiment and online communication as homogeneous signals, this study shows that why investors communicate is as important as what they communicate. I distinguish between influence-motivated sharing, which promotes directional narratives, and validation-motivated sharing, which reflects collaborative efforts to assess whether trading is justified.

Using detailed data from U.S. grain futures markets and scheduled public information releases, I document systematic differences in how these incentives operate. Influence-oriented sentiment incorporates public information rapidly and continues to accumulate, consistent with persistent directional expression. In contrast, validation-oriented sentiment evolves more slowly and does not contribute to sustained directional diffusion, reflecting ongoing reassessment rather than immediate conviction.

These differences in sentiment dynamics translate into distinct trading patterns. Validation-oriented communication tends to precede small-speculator trading activity, particularly when expressed through replies, suggesting that traders engage in collaborative interpretation before forming positions. Influence-oriented communication, by contrast, tends to follow trading, consistent with post-position signaling or narrative amplification rather than primary information processing.

Most importantly, validation-oriented communication is associated with lower bid–ask spreads, indicating reduced information integration costs. This suggests that collaborative interaction on social networks can reduce disagreement about the interpretation of public information, lowering adverse selection risk faced by liquidity providers.

Taken together, the evidence highlights a dual role for social networks in financial markets. Influence-oriented communication shapes how information is framed and disseminated, while validation-oriented interaction facilitates belief convergence and reduces market frictions. These findings help reconcile mixed evidence in the literature on social media and market efficiency and underscore the importance of accounting for communication incentives when studying investor behavior.

The results have broader implications for understanding information transmission in modern financial markets. As social platforms increasingly serve as venues for interpretation rather than mere dissemination, distinguishing between persuasive and collaborative communication becomes critical for assessing their impact on prices, liquidity, and welfare. Future research may explore how these dynamics vary across assets, market conditions, and platform designs, or how institutional investors interact with validation-oriented communities in shaping market outcomes.

References

Akbas, Ferhat, Stanimir Markov, et al., 2018. Determinants and consequences of information processing delay: Evidence from the Thomson Reuters Institutional Brokers' Estimate System. *Journal of Financial Economics* 127.2, pp. 366–388.

Akbas, Ferhat, Felix Meschke, and M Babajide Wintoki, 2016. Director networks and informed traders. *Journal of Accounting and Economics* 62.1, pp. 1–23.

Barbieri, Francesco et al., 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification.

Blankespoor, Elizabeth, 2018. Firm communication and investor response: A framework and discussion integrating social media. *SSRN Electronic Journal*.

Blankespoor, Elizabeth, E D Dehaan, et al., 2019. Why do individual investors disregard accounting information? The roles of information awareness and acquisition costs: Information awareness and acquisition costs. *Journal of Accounting Research* 57.1, pp. 53–84.

Blankespoor, Elizabeth, Ed deHaan, and Christina Zhu, 2018. Capital market effects of media synthesis and dissemination: evidence from robo-journalism. *Review of Accounting Studies* 23.1, pp. 1–36.

Brochet, Francois et al., 2019. Managers' cultural background and disclosure attributes. *The Accounting Review* 94.3, pp. 57–86.

Brown, Jeffrey R et al., 2007. Neighbors Matter: Causal Community Effects and Stock Market Participation.

Bushee, Brian J and Ying Huang, 2024. Do Analysts and Investors Efficiently Respond to Managerial Linguistic Complexity during Conference Calls? *The Accounting Review* 99.4, pp. 143–168.

Bushee, Brian J, Daniel Taylor, and Christina Zhu, 2023. The dark side of investor conferences: Evidence of managerial opportunism. *SSRN Electronic Journal* 98.4, pp. 33–54.

Bushman, Robert M, Abbie J Smith, and Regina Wittenberg-Moerman, 2010. Price Discovery and Dissemination of Private Information by Loan Syndicate Participants. *Journal of Accounting Research* 48.5, pp. 921–972.

Butler, Marty, Arthur Kraft, and Ira S Weiss, 2007. The effect of reporting frequency on the timeliness of earnings: The cases of voluntary and mandatory interim reports. *Journal of Accounting and Economics* 43.2, pp. 181–217.

Cade, Nicole L, 2018. Corporate social media: How two-way disclosure channels influence investors. *Accounting, Organizations and Society* 68-69, pp. 63–79.

Camacho-Collados, Jose et al., 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media.

Caskey, Judson, Michael Minnis, and Venky Nagar, 2024. A social network transmission model of investor relations. *Journal of Financial Reporting* 9.2, pp. 51–70.

Cohen, Lauren, Andrea Frazzini, and Christopher J Malloy, 2007. The small world of investing: Board connections and mutual fund returns. *The Journal of Political Economy* 116, pp. 951–979.

Davis, Angela K, Jeremy M Piger, and Lisa M Sedor, 2012. Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research* 29.3, pp. 845–868.

Dodds, Peter Sheridan, Eric M Clark, et al., 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences of the United States of America* 112.8, pp. 2389–2394.

Dodds, Peter Sheridan, Kameron Decker Harris, et al., 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PloS one* 6.12, e26752.

Dye, Ronald A, 1983. Communication and Post-Decision Information. *Journal of Accounting Research* 21.2, pp. 514–533.

Dye, Ronald A and S Sridhar, 2002. Resource Allocation Effects of Price Reactions to Disclosures. *Contemporary Accounting Research* 19.3, pp. 385–410.

Dyer, Travis A, 2021. The demand for public information by local and nonlocal investors: Evidence from investor-level data. *Journal of Accounting and Economics* 72.101417, p. 101417.

Farrell, Michael et al., 2022. The democratization of investment research and the informativeness of retail investor trading. *Journal of Financial Economics* 145.2, Part B, pp. 616–641.

Ferracuti, Elia, 2022. Information uncertainty and organizational design. *Journal of Accounting and Economics* 74.1, p. 101493.

Fitzgerald, Timothy, 2021. *out of the shadows: The new merchants of grain*. By Jonathan kingsman, 2019, ISBN 978- 1704267821. *American Journal of Agricultural Economics* 103.5, pp. 1926–1927.

Goldstein, Itay, 2023. Information in financial markets and its real effects. *Review of Finance* 27.1, pp. 1–32.

Grootendorst, Maarten, 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.

Hong, Harrison, Jeffrey D Kubik, and Jeremy C Stein, 2005. Thy neighbor's portfolio: Word-of-mouth effects in the holdings and trades of money managers. *The Journal of Finance* 60.6, pp. 2801–2824.

Huang, Xuan, Siew Hong Teoh, and Yinglei Zhang, 2014. Tone management. *The Accounting Review* 89.3, pp. 1083–1113.

Hvide, Hans K and Per Östberg, 2015. Social interaction at work. *Journal of Financial Economics* 117.3, pp. 628–652.

Jung, Michael J, James P Naughton, et al., 2015. Corporate use of social media. *SSRN Electronic Journal*.

Jung, Michael J, M H Franco Wong, and X Frank Zhang, 2018. Buy-side analysts and earnings conference calls. *Journal of Accounting Research* 56.3, pp. 913–952.

Kub, Elaine, 2012. *Mastering the Grain Markets: How Profits are Really Made*. Kub Asset Advisory, Incorporated.

Larcker, David F, Eric C So, and Charles C Y Wang, 2013. Boardroom centrality and firm performance. *Journal of Accounting and Economics* 55.2-3, pp. 225–250.

Lawrence, Alastair, 2013. Individual investors and financial disclosure. *Journal of Accounting and Economics* 56.1, pp. 130–147.

Lee, Charles M C and Qinlin Zhong, 2022. Shall we talk? The role of interactive investor platforms in corporate communication. *Journal of Accounting and Economics* 74.2-3, p. 101524.

Leuz, Christian and Peter D Wysocki, 2008. Economic consequences of financial reporting and disclosure regulation: A review and suggestions for future research. *SSRN Electronic Journal*.

Ljungqvist, Alexander and Wenlan Qian, 2016. How Constraining Are Limits to Arbitrage? *The Review of Financial Studies* 29.8, pp. 1975–2028.

Loughran, Tim and Bill Mcdonald, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 66.1, pp. 35–65.

Loureiro, Daniel et al., 2022. TimeLMs: Diachronic language models from Twitter. *Annual Meeting of the Association for Computational Linguistics*, pp. 251–260.

Ozsoylev, Han N et al., 2014. Investor networks in the stock market. *The Review of Financial Studies* 27.5, pp. 1323–1366.

Peng, Lin, 2005. Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis* 40.2, pp. 307–329.

Roychowdhury, Sugata, Nemit Shroff, and Rodrigo S Verdi, 2019. The effects of financial reporting and disclosure on corporate investment: A review. *Journal of Accounting and Economics* 68.2, p. 101246.

Souther, Matthew E, 2018. The effects of internal board networks: Evidence from closed-end funds. *Journal of Accounting and Economics* 66.1, pp. 266–290.

Stiglitz, Joseph E, 2004. Information and the change in the paradigm in economics, part 2. *The American economist* 48.1, pp. 17–49.

Stone, Philip J, Dexter C Dunphy, and Marshall S Smith, 1966. The general inquirer: A computer approach to content analysis. 651.

Twedt, Brady, 2016. Spreading the Word: Price Discovery and Newswire Dissemination of Management Earnings Guidance. *The Accounting Review* 91.1, pp. 317–346.

Vashishtha, Rahul, 2014. The role of bank monitoring in borrowers discretionary disclosure: Evidence from covenant violations. *Journal of Accounting and Economics* 57.2, pp. 176–195.

Wong, T J et al., 2024. Calling for transparency: Evidence from a field experiment. *Journal of Accounting and Economics* 77.1, p. 101604.

Zhang, Rachel Xi, 2022. Do Managers learn from institutional investors through direct interactions? *Journal of Accounting and Economics* 75.101554, p. 101554.

Appendix A: Data Collection and Cleaning Processes

Category	Word List
Step 1: Querying Sample Posts (Initial Sample: 249,946 Posts)[c]	
Search Hashtags	#wheat, #corn, #soybean, or #oatt (Oils, Agriculture, and Things Traded)
Search Cashtags	$zc, $zs, $zw, $corn, $weat, $wheat, or $soyb
None of these words	grow(year), harvest(year), plant(year), corntownwtf, vodka, beer, whisky, whiskey, recipe, brew, vegan, keto, gmo, genome, genotype, pathology, pathogen, agronomist, phenotype, phenology, agfact, agtwitter, funfact, dyk, pixel, artist, Etsy, decor, sustainable, fanart
Search Periods	Daily intervals From 01.01.2016 to 05.31.2023
Step 2: Algorithmic Filtering With Network Specific Vocabularies (Final Sample: 148,072 Posts)	
Generic	did you know, lyric, poet, artwork, handmade, homemade, vintage, sculpture, clay, Instagram, Tiktok, TBT, throwback, meme, NFT, technology, pioneer, trivia, bitcoin, crypto, film, #national. nike,[a] fashion, sustainab-, asmr, blockchain, selfie, cinema, comic, #art, school, grader, rugby, Blake Shelton,[b] sunset, sunrise, music, photo by, vss365, photography, click here, join us, join me, call us, call me, contact us, contact me, dm us, dm me, student, child, teacher, princess, baby, girl, boy, mother, father, daughter, husband, wife, heaven, peace, born, tribute, movie, family, party, memories, memory, male, female, wild, brain, startup, exhibition, festival, garden, retro, game, golf, ball, holiday, sports, design, shop, collection, history, paint, BBC, #diy, contest, language, generation, color, drawing, craft, membership, register, subscribe, crwd, okta, ddog, tsla, mdb, cybersecurity, panw, twlo, roku
Research Network	genetic, phosphorus, archaeology, geology, nematode, botany, abstract, dissertation, Ph.d., phd, postdoc, thesis, professor, hybrid, sampl, mutant, cultivar, climate change, pollution, innovation, agronomy, science, scientist, research, organic, reject, accept, sponsor, academy, conference, workshop, career, presentation, speaker, survey, award, study

Continued on next page

Category	**Word List**
Food/Health Network	chowder, bread, sourdough, pasta, tortillas, popcorn, fritter, salad, fruit, veggie, grain-free, low carb, calories, cuisine, dish, allergy, fiber, carbs, wholegrain, mustard, kitchen, garlic, spice, butter, cheese, snack, soup, dairy, tasty, baking, BBQ, vitamin, nutritious, delicious, savory, spicy, crunch, kernel, pork, beef, poultry, meat, potato, ingredient, starch, dessert, bake, grill, roast, healthy, sweet, cook, taste, diet, salt, pepper, protein, menu, drink, made with, made of, cake, yum, chips
Producer/Farmer Network	#farm, farm365, pftour, field day, cover crop, money crop, herbicides, grower, weeds, #CargillGrows, seed, sprout, sowing, verification, toleran, bird, insect, beetle, wasp, billbug, caterpillar, cricket, husk, germinat, spray, landscape, drill-, risk management, no till, tillage, tractor, grinder, green thumb, replant, weed control, fungicide, varieties, irrigat-, training, wheat straw, breeding, precision, nitrogen, fixation, planter, chimney, haul, eatwheat, milling, compost, worm, cow, animal, horse, resistan, rotation, bunting, silo, elevator, insurance, rural, ranch, scout, application, productivity, population, feeding, marketing, breeder,agronomy, truck, pest, dryer, agent, summit, safety

This table describes how I collect and clean the X.com sample. I collect posts by searching for a predefined set of keywords daily from January 2016 to May 2023. These keywords are set to optimize search outcomes, as X.com caps search results at about 100 posts regardless of the query. Upon analysis, I identify three prominent networks using the same hashtags: food/health, research, and producer/farmer networks. To refine the data, I implement a two-stage filtering process. First, posts are scraped using a set of exclusion words that are unique and popular to non-investor networks. Second, scraped posts are algorithmically filtered with additional lists of vocabulary that are most likely used among non-investor networks. I iterate the cleaning process until the topic representations of the final sample do not feature subjects that are irrelevant to the grain futures markets.

[a] The brand runs several lines in “wheat” color.
[b] The singer released a song titled “corn” in 2021, and fans were posting about it.

[c] ***Legal Disclaimer****: The author complied with X.com (Twitter)'s code of conduct during the web scraping process. Posts are collected using the log-in information of the author's verified X.com account, respecting the daily user view limit (10,000 posts per day for a premium user.) The paper does not make use of any personally identifiable information or copyright-protected content.*

Appendix B: Top Topics Extracted By LLMs

In topic modeling, I use BERTopic (Grootendorst (2022)), which incorporates LLMs and class-based TF-IDF for its methodology.[16] Similar to traditional TF-IDF, Term Frequency indicates how often a word appears in a post, reflecting its importance to the message. The algorithm performs class-based TF by examining words across many posts about the same topic, identifying key words for that particular topic. It also considers how common or unique these words are across all topics using Inverse Document Frequency, determining if they are widely used or specific to certain discussions. By combining these insights—the TF part about word frequency and the IDF part about word uniqueness—the algorithm calculates a class-based TF-IDF score for each word. This score highlights words that are not just common but especially meaningful to the topics being discussed.

The first step of topic modeling using BERTopic involves converting posts into numerical embeddings, which assign unique numerical values to each word, capturing their semantic meaning. Following this, posts are clustered based on these embeddings, allowing for the identification of distinct topics within these clusters. Then, the algorithm selects the top 50 words for each topic to generate a concise topic representation, ranked by their class-based TF-IDF scores. The higher a word's class-based TF-IDF score, the more it is considered representative of its topic, acting as a measure of information density.

The table reports the top five topics resulting from topic modeling, along with a generic category ('Grain Market News') for posts that do not fit into any specific topic. The Count column indicates the number of posts that fall into each topic, providing a quantitative measure of each topic's prevalence. The Llama2 column lists labels generated by the Llama-2 model, while the KeyBERT and MMR columns display the top ten words for each topic as identified by these respective models. Additionally, the 'Representative Posts' column includes posts selected by the algorithm based on their class-based TF-IDF scores, which closely match the topic's class-based TF-IDF profile.

[16] See also https://maartengr.github.io/BERTopic/index.html

Count	Llama2	KeyBERT	MMR	Representative Posts
20451	Grain Market News[a]	['soybeans', 'soybean', 'commodities', 'grains', 'grain', 'news', 'corn', 'market', 'trade', 'agriculture']	['corn', 'wheat', 'soybeans', 'soybean', 'at', 'http', 'all', 'market', 'from', 'crop']	The grain markets are relatively quiet to start the day, and part of it may be reluctant for traders to do too much with USDA scheduled to release its December supply and demand report this week on Thursday Listen to the morning comments #oatt'
7356	International Grain Trade	['soybeans', 'brazil', 'soybean', 'corn', 'exports', 'brazilian', 'china', 'trade', 'soy', 'tonnes']	['brazil', 'china', 'soybean', 'argentina', 'the', 'corn', 'mmt', 'exports', 'soybeans', 'trade']	[Rainfall over the past week was plentiful for central & southern areas of Brazil's #soybean belt. First-crop #corn planting is at 37%, up from 12% planted at this point last year. Brazil is expected to have favorable growing conditions this year. https://ifcs.co/Grains-Market-Intelligence
5958	Crop Conditions in the US	['wheat', 'kansaswheat', 'crops', 'plains', 'drought', 'crop', 'kansas', 'ks', 'farmers', 'barley']	['wheat', 'winter', 'kansas', 'crop', 'poor', 'spring', 'our', 'bit', 'week', 'at']	Winter #Wheat Crop Conditions This chart shows we went into winter dormancy with wheat at 46% good to excellent. The initial spring rating came in at 53% good to excellent. At this time the record cold temps don't seem to have affected the winter wheat crop #oatt
2165	Wheat Market Analysis	['bullish', 'trading', 'wheat', 'commodities', 'highs', 'trade', 'chart', 'futures', 'trend', 'from']	['wheat', 'kc', 'weat', 'chart', 'zw', 'market', 'of', 'oatt', 'futures', 'short']	#Wheat prices have been extremely volatile up and down. We are well below last week's highs and well above the lows from early May,in fact we are close to the middle of our long-term trading range. The rain in the plains has certainly put a damper on the wheat market. #oatt
2083	Global Wheat Supply Disruption	['ukraine', 'russia', 'wheat', 'exports', 'turkey', 'russian', 'from', 'grain', 'countries']	['russia', 'ukraine', 'russian', 'exports', 'wheat', 'sizov', 'from', 'war', 'putin', 'global']	CHART: #Russia and #Ukraine account for a quarter of global grains exports. #Commodities #wheat
2006	Agricultural Weather Forecast	['soybeans', 'drought', 'temps', 'forecast', 'precipitation', 'agwx', 'corn', 'weather', 'ag', 'plains']	['agwx', 'belt', 'next', 'bamwx', 'gfs', 'week', 'midwest', 'forecast', 'weather', 'plains']	[As we hinted at yesterday, the data over the next 7-10 days is quite soaked across the main growing areas from the Plains to the eastern #corn belt. Below is both the Euro and GFS Ensemble probabilities of 1"+ over the next 10 days #AGwx #OATT #soybeans #wheat

This table presents the top 5 topics extracted from the dataset to ensure the quality of the sample data. Topic modeling, a form of unsupervised machine learning, automatically identifies patterns within the data, thereby summarizing vast amounts of textual information without the need for manual review. If the sample included irrelevant posts to the degree that impacted the analysis, the extracted topics would feature random subjects not pertinent to the grain futures markets. The 'Count' column indicates the number of posts that fall into each topic, providing a quantitative measure of each topic's prevalence. The 'Llama2' column lists labels generated by the Llama-2 model, while the 'KeyBERT' and 'MMR' columns display the top ten words for each topic as identified by these respective models. Additionally, the 'Representative Posts' column includes posts selected by the algorithm based on their c-TF-IDF scores, which closely match the topic's c-TF-IDF profile.

[a] "Agricultural Market Update" is a generic topic created by the BERTopic model for posts that cannot be assigned to a specific subtopic. Posts in this category are depicted as gray dots in Figure IA.2 in the Internet Appendix.

Appendix C: Examples of Influencing and Validating Posts

Figure C.1: Examples of Influencing Posts

[redacted]
@[redacted]

Acreage will be in focus but don't forget about stocks on Friday. Trade sees June 1 #corn stocks up 0.5% from a year ago but #soybeans up 53% to record levels. Corn would be 3rd largest after '87 & '88. #Wheat stocks (effectively 18/19 carryout) are pegged identical to last year.

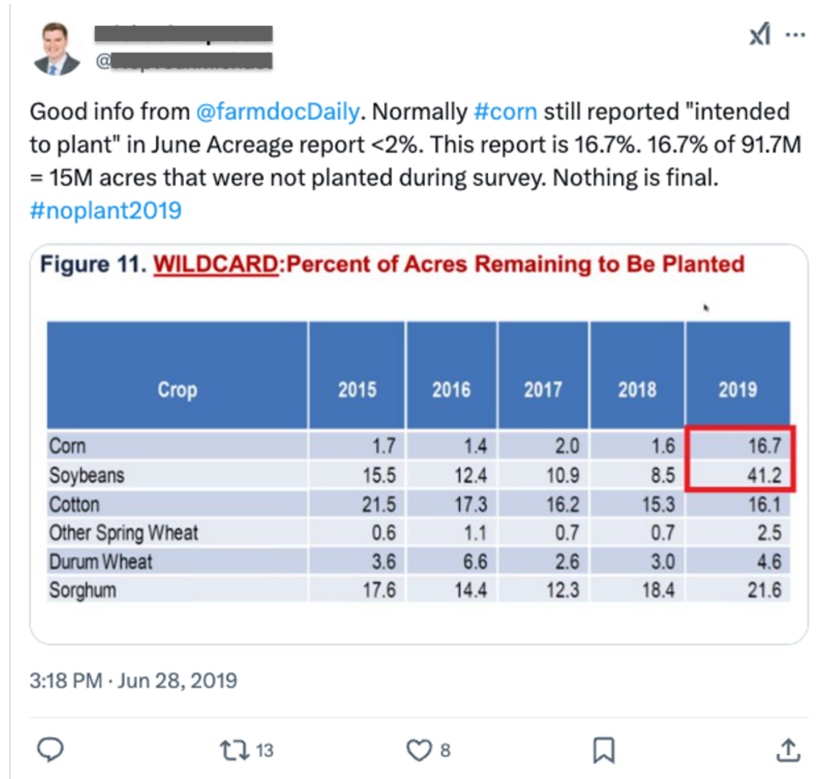
U.S. Quarterly Grain Stocks: Trade Predictions
(billions of bushels)

	CORN	SOY	WHEAT
June 1, 2019 – Average	5.332	1.861	1.100
June 1, 2019 – Range	5.174 – 5.497	1.700 – 1.962	1.077 – 1.160
June 1, 2018 Stocks	5.305	1.219	1.099
Record June 1 Stocks	6.332 ('87)	1.219 ('18)	1.905 ('86)

Data sources: USDA; Reuters
@kannbwx

2:56 PM · Jun 25, 2019

2 9 17 1



[redacted]
@[redacted]

Good info from @farmdocDaily. Normally #corn still reported "intended to plant" in June Acreage report <2%. This report is 16.7%. 16.7% of 91.7M = 15M acres that were not planted during survey. Nothing is final. #noplant2019

Figure 11. WILDCARD:Percent of Acres Remaining to Be Planted

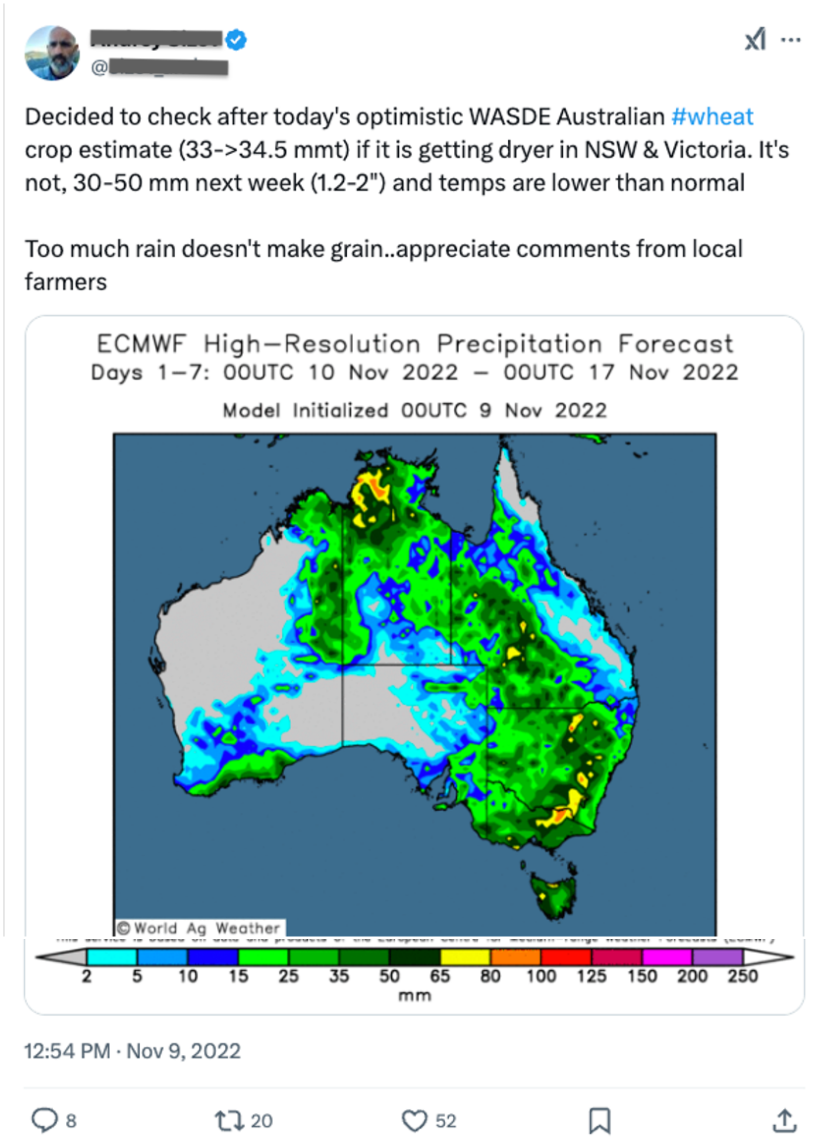
Crop	2015	2016	2017	2018	2019
Corn	1.7	1.4	2.0	1.6	16.7
Soybeans	15.5	12.4	10.9	8.5	41.2
Cotton	21.5	17.3	16.2	15.3	16.1
Other Spring Wheat	0.6	1.1	0.7	0.7	2.5
Durum Wheat	3.6	6.6	2.6	3.0	4.6
Sorghum	17.6	14.4	12.3	18.4	21.6

3:18 PM · Jun 28, 2019

13 8

Note: These are examples of Influencing Posts from the First Notice Day period, when the incentive to influence is likely heightened. The posts highlight overlooked details in public reports and share the poster's own interpretation or insight. Based on engagement metrics, each post received several indirect responses (e.g., likes), but few or no direct responses (e.g., replies).

Figure C.2: An Example of a Validating Post and Replies



Note: The left-hand side presents an example of a Validating Post during the Russo-Ukrainian War period. The right-hand side shows replies to this post (continued on the next page).

Figure C.3: An Example of a Validating Post and Replies

Note: These are replies to the Validating Post example shown on the previous page. The original post poses a question about interpreting a public crop estimate report (WASDE), highlighting a contradiction between optimistic planting-based projections and unfavorable weather forecasts. In response, local farmers and crop watchers contribute their perspectives, discussing the effects of prolonged rain on harvest timing and, ultimately, crop yields. This exchange illustrates how investors use social networks to validate and refine their interpretations of public information. It also aligns with prior findings that local investors often possess superior information due to their proximity and firsthand knowledge (Dyer (2021)).

Appendix D: Classification Protocol

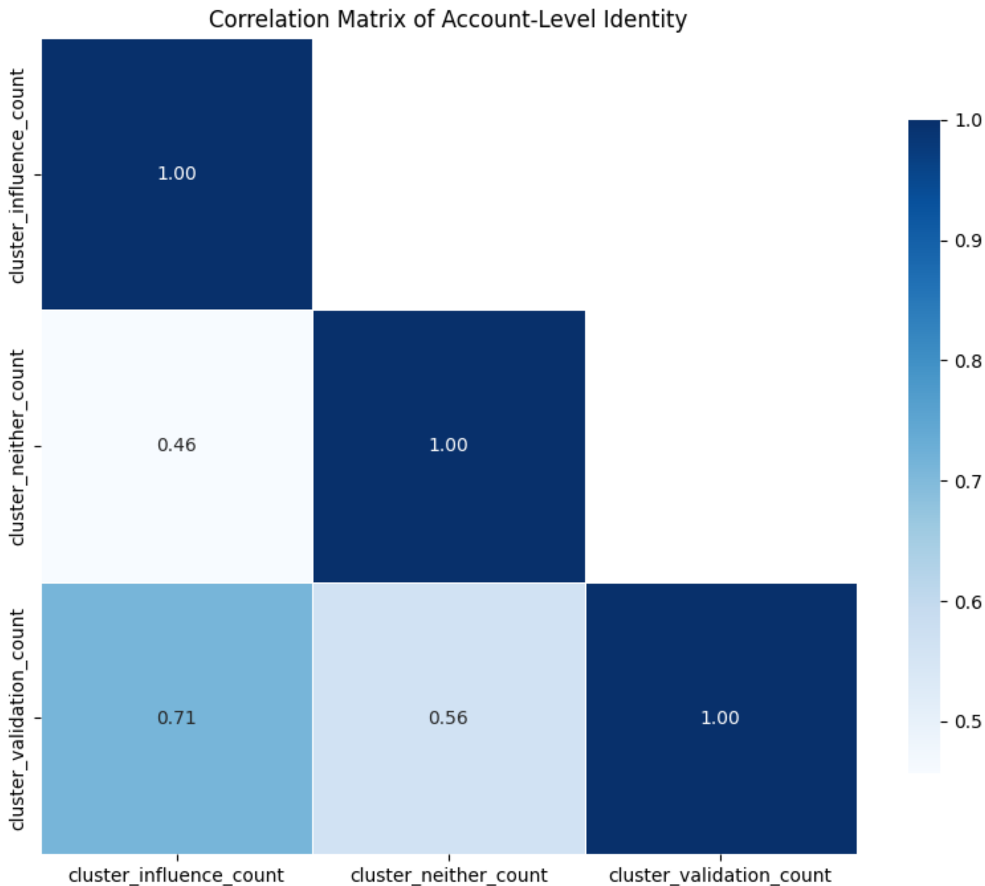
I classify X.com posts into three mutually exclusive categories (validation, influence, and unidentified) based on the dominant communication incentive expressed in the post. Validation posts seek input, express uncertainty, or invite collaborative interpretation of public information (e.g., questions or requests for clarification). Influence posts assert a directional interpretation or aim to persuade others (e.g., confident claims or trading recommendations). Posts that do not clearly reflect either incentive are classified as unidentified.

Posts were independently labeled by three human annotators, including the researcher, using a shared coding protocol that emphasized observable linguistic cues rather than inferred motives. Annotators worked independently and without access to each other's labels. When all three annotators agreed, the assigned label was retained; when disagreement occurred, the final label was determined by majority rule. The resulting labeled sample was used to fine-tune a RoBERTa language model, which was then applied to the full dataset. To assess validity, I examine linguistic features and audience responses across categories and show that validation posts exhibit greater use of uncertainty-related language and receive more replies, while influence posts are more assertive and receive more likes.

A detailed coding protocol (IA-1) and calibration examples (IA-2) shared with annotators are provided in the Internet Appendix.

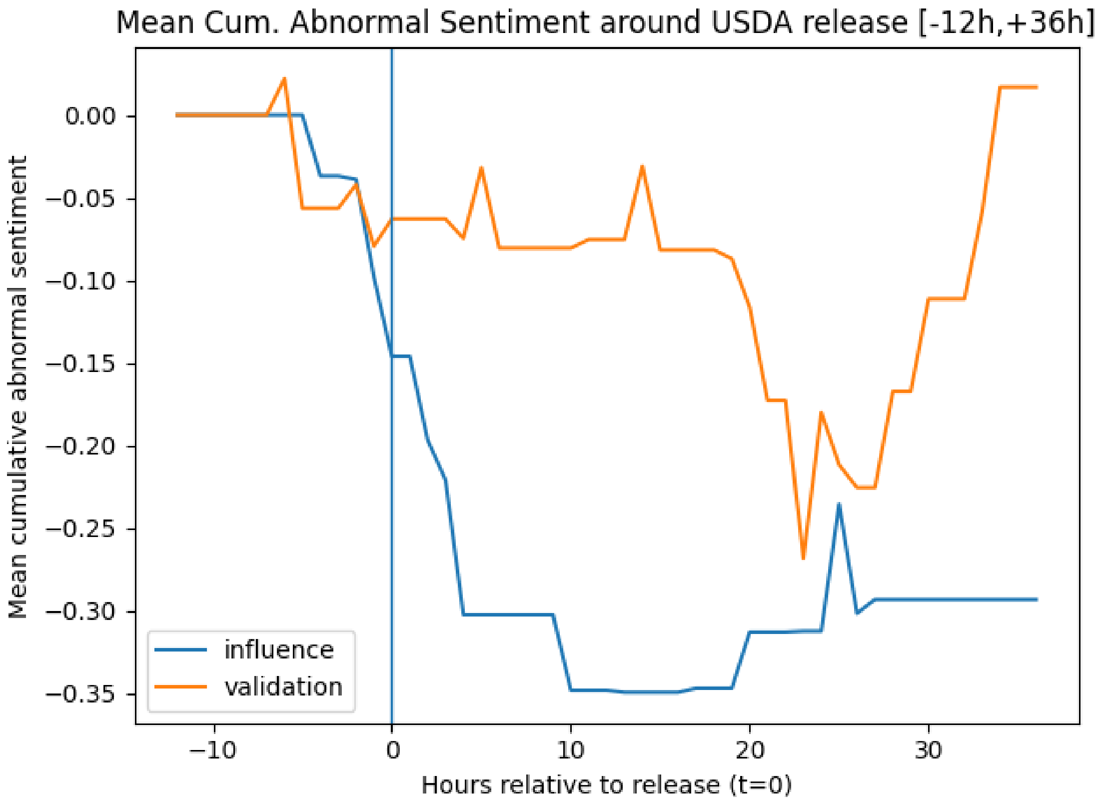
Figures

Figure 1: Correlation Matrix of Acccount Level Identity



Note: This figure shows the correlations between user activity across post types (Influencing, Validating, and Unidentified) at the account level.

Figure 2: Abnormal Cummulative Sentiment Around Report Dates



Note: The figure plots the average cumulative abnormal sentiment from 12 hours before to 36 hours after report releases, separately for Influencing and Validating Posts. The figure reveals a clear pattern: sentiment in Influencing Posts reacts sharply immediately following the release, consistent with post-position directional expression, while sentiment in Validating Posts peaks more gradually and continues to adjust beyond the initial response window. This motivates the stacked-horizon design used in the subsequent regression analysis.

Tables

Table 1: Variable Definitions

Name	**Description**	**Data Source**
OG/RP Val./Inf.% (Sim./Fol./Eng.)	Fraction of original/reply posts classified as validating/influencing within an hour or a day, computed using unweighted post counts (Sim.) or weighted by the author's follower count at the time of posting (Fol.) or weighted by post-level engagement (sum of replies, retweets, and likes, (Eng.))	X.com/custom
ln(Total Posts)	Natural logarithm of one plus the total number of posts (original posts and replies) within the aggregation window.	X.com/custom
ln(Total Followers)	Natural logarithm of one plus the total number of followers aggregated across all authors posting within the window.	X.com/custom
ln(Total Engagement)	Natural logarithm of one plus total engagement (sum of replies, retweets, and likes) across all posts within the window.	X.com/custom
Sentiment IPT Val./Inf./NA (12h/36h)	Sentiment IPT is constructed as the trapezoidal approximation to the area under the normalized cumulative abnormal sentiment curve. The measure captures the average timing of sentiment incorporation. Sentiment IPT Val. (12h/36h) are aggregated over validating/influencing/unidentified posts (both original and replies) within a 12-hour (36-hour) window following a public report event, which typically released at noon (ET).	X.com/custom
ln(Daily Bid-Ask Spread)	Natural logarithm of the daily average level-1 bid–ask spread, measured using futures market quote data.	CME Historical Market Dept Data
ln(Intraday Price Range)	Natural logarithm of the daily high–low price range computed from intraday transaction prices.	CME Historical Market Dept Data

Continued on next page

Table 1 – continued from previous page

Name	**Description**	**Data Source**
ln(Hourly # of Trades)	Natural logarithm of total retail (nonreportable speculators) number of trades aggregated at the hour level.	CME Historical Market Dept Data
ln(Daily # of Trades)	Natural logarithm of total retail (nonreportable speculators) number of trades aggregated at the day level.	CME Historical Market Dept Data
Hedonometer	The daily X.com sentiment index that is measured based on 10% random sampling of the roughly 500 million Posts every day.	Dodds, Harris, et al. (2011); Dodds, Clark, et al. (2015)
ln(SNP500)	Natural logarithm of the S&P 500 index level.	S&P Global
ln(USD Index Futures)	The ICE US Dollar Index futures contract is a benchmark for the international value of the USD and the USD index relative to a basket of world currencies.	Intercontinental Exchange (ICE)
ln(US 3M Bond)	The United States 3-Month Bond Yield is the interest rate paid by the U.S. government on its 3-month Treasury bills.	NYSE
ln(US 5Y Bond)	The United States 5-Year Bond Yield is the interest rate paid by the U.S. government on its 5-Year Bonds.	NYSE
ln(Bitcoin Futures)	Natural logarithm of Bitcoin futures prices.	S&P Global
ln(Dow Jones Softs)	The Dow Jones Softs Index is a benchmark that tracks the performance of futures contracts in the soft commodities sector. This sector includes agricultural products such as coffee, cocoa, sugar, and cotton.	S&P Global
ln(Dow Jones Energy)	The Dow Jones Energy Index tracks the performance of futures contracts in the energy sector, including crude oil, natural gas, heating oil, and gasoline.	S&P Global
ln(Dow Jones Metals)	The Dow Jones Metals Index measures the performance of futures contracts in the metals sector, including both precious metals (such as gold and silver) and industrial metals (such as copper and aluminum).	S&P Global

Table 2: Summary statistics

	N	Mean	Std	Min	P25	Median	P75	Max
Panel A: Validating Posts								
Followers Count	7354.00	37386.00	183879.63	4.00	6199.00	18530.00	39201.00	8133713.00
Following Count	7354.00	2256.38	4610.78	0.00	553.00	1361.00	2659.00	279047.00
Retweet Count	7360.00	7.23	26.48	0.00	0.00	2.00	7.00	1300.00
Reply Count	7360.00	4.10	5.43	2.00	2.00	3.00	4.00	210.00
Like Count	7360.00	25.80	67.09	0.00	7.00	15.00	29.00	4300.00
Photo/Video Attached	7360.00	0.50	0.50	0.00	0.00	1.00	1.00	1.00
View Count	673.00	11923.22	13472.85	184.00	5039.00	9096.00	14968.00	245435.00
Fog Index	7360.00	7.98	3.86	0.80	5.33	7.87	9.94	26.95
Direct/Indirect Ratio	7360.00	0.27	0.32	0.01	0.09	0.17	0.33	8.00
Panel B: Influencing Posts								
Followers Count	5118.00	115032.98	604732.67	84.00	9171.00	38153.00	87640.00	13962724.00
Following Count	5118.00	1838.02	2758.52	0.00	154.00	1096.50	2554.00	23596.00
Retweet Count	5134.00	13.67	35.37	0.00	1.00	6.00	16.00	1200.00
Reply Count	5134.00	3.70	6.05	2.00	2.00	2.00	4.00	214.00
Like Count	5134.00	36.50	114.11	0.00	8.00	19.00	39.00	6200.00
Photo/Video Attached	5134.00	0.54	0.50	0.00	0.00	1.00	1.00	1.00
View Count	451.00	11127.12	16375.21	202.00	3195.00	7649.00	14679.00	263172.00
Fog Index	5134.00	8.58	4.29	0.80	5.67	8.04	11.00	30.13
Direct/Indirect Ratio	5134.00	0.18	0.21	0.01	0.06	0.11	0.21	3.00
Panel C: Unidentified Posts								
Followers Count	5766.00	87807.37	562229.99	11.00	4218.00	12682.00	39202.00	16467445.00
Following Count	5766.00	2166.05	4796.46	0.00	490.00	1482.00	2659.00	215554.00
Retweet Count	5802.00	12.81	69.44	0.00	1.00	3.00	9.00	2700.00
Reply Count	5802.00	5.96	42.75	2.00	2.00	3.00	4.00	2000.00
Like Count	5802.00	45.37	239.30	0.00	7.00	16.00	36.00	13000.00
Photo/Video Attached	5802.00	0.63	0.48	0.00	0.00	1.00	1.00	1.00
View Count	465.00	10047.83	34091.47	126.00	1989.00	3964.00	9099.00	524124.00
Fog Index	5802.00	8.64	5.24	0.40	4.40	8.14	11.67	31.60
Direct/Indirect Ratio	5802.00	0.23	0.32	0.00	0.07	0.14	0.28	6.64

Continued on next page

Table 2 continued

	N	Mean	Std	Min	P25	Median	P75	Max
Panel D: Replies to Validating Posts								
Followers Count	33009	9637.18	25414.65	0.00	434.00	1699.00	7759.00	1277984.00
Following Count	33009	1549.01	2329.85	0.00	383.00	891.00	1983.00	74609.00
Retweet Count	33009	0.15	1.67	0.00	0.00	0.00	0.00	110.00
Reply Count	33009	0.51	0.78	0.00	0.00	0.00	1.00	26.00
Like Count	33009	1.81	8.74	0.00	0.00	1.00	2.00	638.00
Photo/Video Attached	33009	0.09	0.29	0.00	0.00	0.00	0.00	1.00
View Count	3476	587.24	944.60	1.00	156.00	346.50	680.25	17521.00
Fog Index	33009	9.56	6.95	0.40	5.55	8.28	11.67	41.60
Direct/Indirect Ratio	33009	0.26	0.43	0.00	0.00	0.00	0.50	6.00
Panel E: Replies to Influencing Posts								
Followers Count	19030	15929.08	130681.60	0.00	325.25	1508.00	8686.75	13963318.00
Following Count	19030	1383.73	2268.94	0.00	244.00	705.00	1749.00	62448.00
Retweet Count	19030	0.36	3.32	0.00	0.00	0.00	0.00	230.00
Reply Count	19030	0.48	0.84	0.00	0.00	0.00	1.00	36.00
Like Count	19030	2.21	9.56	0.00	0.00	1.00	2.00	503.00
Photo/Video Attached	19030	0.11	0.31	0.00	0.00	0.00	0.00	1.00
View Count	1808	540.86	1259.73	1.00	70.00	179.00	431.25	14584.00
Fog Index	19030	10.09	7.12	0.40	5.70	8.51	12.49	42.40
Direct/Indirect Ratio	19030	0.24	0.42	0.00	0.00	0.00	0.33	5.00
Panel F: Posting Intensity Measures								
OG Val.% (Sim.)	47579	0.07	0.20	0.00	0.00	0.00	0.00	1.00
RP Val.% (Sim.)	47579	0.33	0.42	0.00	0.00	0.00	0.75	1.00
OG Val.% (Fol.)	47509	0.08	0.21	0.00	0.00	0.00	0.00	1.00
RP Val.% (Fol.)	47509	0.33	0.42	0.00	0.00	0.00	0.74	1.00
OG Val.% (Eng.)	47579	0.10	0.24	0.00	0.00	0.00	0.00	1.00
RP Val.% (Eng.)	47579	0.31	0.41	0.00	0.00	0.00	0.66	1.00
OG Inf.% (Sim.)	47579	0.06	0.18	0.00	0.00	0.00	0.00	1.00
RP Inf.% (Sim.)	47579	0.20	0.36	0.00	0.00	0.00	0.33	1.00
OG Inf.% (Fol.)	47509	0.06	0.19	0.00	0.00	0.00	0.00	1.00
RP Inf.% (Fol.)	47509	0.20	0.36	0.00	0.00	0.00	0.28	1.00

Continued on next page

Table 2 continued

	N	Mean	Std	Min	P25	Median	P75	Max
OG Inf.% (Eng.)	47579	0.07	0.21	0.00	0.00	0.00	0.00	1.00
RP Inf.% (Eng.)	47579	0.19	0.35	0.00	0.00	0.00	0.21	1.00
ln(Total Posts)	194976	0.27	0.54	0.00	0.00	0.00	0.00	4.67
ln(Total Followers)	194976	0.67	1.23	0.00	0.00	0.00	0.00	6.48
ln(Total Engagement)	194976	0.34	0.68	0.00	0.00	0.00	0.00	4.92
Panel G: Sentiment Measures								
Sentiment IPT Val. 12h	563	9.28	2.75	−1.65	7.50	9.50	11.50	14.61
Sentiment IPT Val. 36h	738	19.88	8.86	−5.63	13.50	18.01	26.33	47.78
Sentiment IPT Inf. 12h	656	10.48	2.72	0.50	9.21	11.50	12.50	15.41
Sentiment IPT Inf. 36h	767	24.53	10.34	−3.68	16.33	24.58	35.50	41.37
Sentiment IPT NA 12h	551	8.71	3.49	−2.26	6.54	9.48	11.51	13.48
Sentiment IPT NA 36h	839	20.63	10.44	−4.64	12.50	18.50	30.52	47.46
Panel H: Market Outcome Measures								
ln(Daily Bid-Ask Spread)	5587	−6.86	1.29	−8.47	−7.51	−7.22	−6.97	−1.54
ln(Intraday Price Range)	4851	−2.88	1.32	−5.51	−3.67	−3.24	−2.80	0.86
ln(Hourly # of Trades)	81296	7.08	1.38	0.69	5.99	7.03	8.30	11.28
ln(Daily # of Trades)	5587	11.39	0.57	0.69	10.98	11.39	11.79	13.11
Panel I: Macroeconomic Measures								
Hedonometer	2709	5.99	0.06	5.63	5.97	6.00	6.03	6.30
ln(SNP500)	2709	8.03	0.25	7.51	7.84	7.98	8.27	8.48
ln(USD Index Futures)	2709	4.57	0.05	4.48	4.54	4.57	4.60	4.74
ln(US 3M Bond)	2709	−0.56	1.73	−27.63	−1.70	0.01	0.72	1.69
ln(US 5Y Bond)	2709	0.42	0.71	−1.65	0.13	0.58	0.98	1.49
ln(Bitcoin Futures)	2289	9.31	1.12	6.79	8.78	9.26	10.24	11.12
ln(Dow Jones Softs)	2709	5.06	0.21	4.74	4.89	4.97	5.22	5.49
ln(Dow Jones Energy)	2709	4.54	0.32	3.40	4.33	4.51	4.74	5.35
ln(Dow Jones Metals)	2709	5.16	0.20	4.70	5.01	5.10	5.35	5.65

Notes: Panel A through E report summary statistics for post-level characteristics across different post categories. Followers and Following denote the author's follower and following counts at the time of posting. Retweet, Reply, Like, Photo/Video Attached (an indicator variable), and View (only available after late 2022) capture post-level engagement measures. Fog Index is the Gunning fog index computed from post text. Direct/Indirect measures the ratio of direct replies to indirect responses (likes and retweets) to a post. All variables are defined in Table 1.

Table 3: Same-Day and Next-Day Sentiment IPT Test (H1)

	(1) Simple Avg.	(2) By Follower	(3) By Engagement
Dependent variable: Hourly Sentiment IPT (Same-day [0,+12h], Next-day [13,+36h])			
$InfluencingPosts$	1.751***	1.792**	1.799**
	(0.170)	(0.196)	(0.202)
$ValidatingPosts$	0.678**	0.554*	0.793**
	(0.095)	(0.149)	(0.119)
$NextDay(ND)$	12.118***	12.163***	12.155***
	(0.363)	(0.483)	(0.519)
$InfluencingPosts \times ND$	2.318	1.723	2.235*
	(0.801)	(0.741)	(0.631)
$ValidatingPosts \times ND$	−1.559**	−1.478***	−1.454***
	(0.193)	(0.127)	(0.081)
$Hedonometer$	25.167	23.054	23.405**
	(10.221)	(11.160)	(4.901)
$ln(SNP500)$	0.164	−0.531	1.137
	(7.308)	(8.264)	(5.388)
$ln(USDIndexFutures)$	−10.966	−6.805	−10.373
	(4.019)	(5.050)	(6.085)
$ln(US3MBond)$	1.178	1.146	1.195
	(0.429)	(0.410)	(0.425)
$ln(US5YBond)$	−1.177	−1.260	−1.294
	(1.334)	(1.480)	(1.615)
$ln(BitcoinFutures)$	0.473	0.943	0.360
	(0.588)	(0.908)	(1.000)
$ln(DowJonesSofts)$	−10.042*	−10.690**	−10.743**
	(2.499)	(2.164)	(1.999)
$ln(DowJonesEnergy)$	−1.433	−0.629	−1.001
	(2.582)	(3.384)	(2.567)
$ln(DowJonesMetals)$	10.838	10.156	11.207
	(3.982)	(6.457)	(4.610)
Commodity FE	Yes	Yes	Yes
Report Type FE	Yes	Yes	Yes
Observations	1,373	1,361	1,380
R^2	0.425	0.408	0.421

Notes: Report type and commodity fixed effects included. Controls included. Standard errors clustered at event type and reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels. The dependent variable is the Sentiment Intra-Period Timeliness (IPT) measure, constructed as the normalized area under the cumulative abnormal sentiment curve. The analysis employs a stacked-horizon design with 12- and 36-hour windows following public information releases (typical release time: 12 pm ET). The omitted group is Unidentified Posts. Coefficients on Influence and Validation capture differences in sentiment incorporation within the first 12 hours, while interaction terms with Next Day measure the delayed response between 12 and 36 hours relative to the baseline.

Table 4: Validation vs Influence Coefficient Comparisons (H1)

	(1) Simple Avg.	(2) By Follower	(3) By Engagement
Dependent variable: Hourly Sentiment IPT (Same-day [0,+12h], Next-day [13,+36h])			
$ValidatingPosts - InfluencingPosts$	-1.074^{***}	-1.237^{***}	-1.006^{**}
	(0.088)	(0.062)	(0.133)
$Validating \times ND - Influencing \times ND$	-3.877^{**}	-3.201^{*}	-3.689^{**}
	(0.670)	(0.845)	(0.712)
Commodity FE	Yes	Yes	Yes
Report Type FE	Yes	Yes	Yes
Observations	1,373	1,361	1,380
R^2	0.425	0.408	0.421

Notes: Each entry reports the estimated contrast from a linear hypothesis test based on the corresponding main regression in Table 4. Standard errors are clustered at report type. ***, **, and * denote significance at the 1%, 5%, and 10% levels. Each row reports a linear contrast between the coefficients for Validating and Influence Posts from the stacked-horizon specification. The first row measures the difference in Sentiment IPT within the initial 12-hour window following the public information release. The second row measures the difference in incremental sentiment incorporation between 12 and 36 hours. Negative values indicate slower incorporation for Validating relative to Influencing Posts.

Table 5: Lead–Lag Tests of Social Network Activity and Trading (H1)

	Sum of lags	Sum of leads	**Lags − leads**	***p*-value**
Dependent variable: ln(1 + Hourly # of Trades)				
OG Val.% (Eng.)	7.819	3.062	4.757	0.145
RP Val.% (Eng.)	2.480	−1.474	3.954	0.072
OG Inf.% (Eng.)	−13.564	−0.811	−12.753	0.001
OG Inf.% (Eng.)	5.308	3.207	2.101	0.061
Number of lags/leads = 12 ; SE = HAC(maxlags=12); Commodity FE				

Notes: This table reports lead–lag asymmetry tests from hourly regressions of $\log(1 + \text{number of trades})$ on K lags and K leads of the indicated post-share measures, including a lag of the dependent variable and commodity fixed effects. "Sum of lags" is $\sum_{k=1}^{K} \hat{\beta}_{-k}$ and "Sum of leads" is $\sum_{k=1}^{K} \hat{\beta}_{+k}$. Positive (lags − leads) indicates post activity tends to precede trading; negative indicates post activity tends to follow trading. p-values are based on a joint test of (sum of lags = sum of leads).

Table 6: Dominating Incentives and Information Integration Cost (H2)

	(1) Simple Avg.	(2) By Follower	(3) By Engagement
Dependent variable: ln(Daily Bid-Ask Spread)			
RP Val.% (Sim.)	-0.156^{***}		
	(0.056)		
RP Inf.% (Sim.)	-0.090		
	(0.068)		
RP Val.% (Fol.)		-0.146^{***}	
		(0.057)	
RP Inf.% (Fol.)		-0.073	
		(0.069)	
RP Val.% (Eng.)			-0.147^{**}
			(0.063)
RP Inf.% (Eng.)			-0.037
			(0.081)
ln(Total posts)	0.007		
	(0.022)		
ln(Total followers)		-0.003	
		(0.018)	
ln(Total engagement)			-0.001
			(0.019)
$ln(IntradayPriceRange)_{-1}$	0.606^{***}	0.606^{***}	0.606^{***}
	(0.032)	(0.032)	(0.032)
ln(Daily # of Trades)	0.217^{***}	0.220^{***}	0.217^{***}
	(0.060)	(0.060)	(0.060)
ln(SNP500)	0.407	0.434	0.437
	(1.292)	(1.295)	(1.296)
ln(UDSIndexFutures)	6.520^{*}	6.630^{*}	6.648^{*}
	(3.777)	(3.778)	(3.768)
ln(US3MBond)	0.002	0.002	0.002
	(0.008)	(0.008)	(0.008)
ln(US5YBond)	-0.376	-0.383	-0.387
	(0.492)	(0.492)	(0.491)
ln(BitcoinFutures)	-0.002	-0.004	0.006
	(0.125)	(0.125)	(0.124)
ln(DowJonesSofts)	0.427	0.425	0.413
	(1.174)	(1.173)	(1.172)
ln(DowJonesEnergy)	0.665	0.688	0.705
	(0.633)	(0.634)	(0.633)
ln(DowJonesMetals)	1.929	1.934	1.920
	(1.837)	(1.839)	(1.839)
Commodity FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Observations	3,790	3,790	3,790
R^2	0.641	0.641	0.641

Notes: Standard errors are two-way clustered by week and commodity. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The table examines whether collaborative validation on social networks reduces information integration costs in grain futures markets. The dependent variable is the logarithm of the daily relative bid–ask spread, which proxies for the cost faced by market participants in processing and incorporating dispersed information. The key explanatory variables capture the incentive composition of social media activity, measured as the shares of validation- and influence-motivated replies in total posting activity under alternative weighting schemes. Across all specifications, a higher share of validation-motivated communication is associated with significantly lower information integration costs, while influence-motivated communication has no such effect.

Internet Appendix

IA-1. Coding Protocol for Communication Incentives

Purpose: Each X.com post is classified into one of three mutually exclusive categories (*Validation*, *Influence*, or *Unidentified*) based on the *dominant communication incentive expressed in the text.* The classification reflects what the post communicates, not the author's true underlying motive.

Unit of Annotation: The unit of annotation is a single X.com post. Annotators base their labels solely on the text of the post (and any quoted text embedded within it). Profile information, prior posts, and engagement metrics are not used in labeling.

Category Definitions:

- **Validation (Initiating Collaboration).** A post is classified as *Validation* if its primary purpose is to seek input, express uncertainty, or invite collaborative interpretation of information relevant to grain futures markets. Typical signals include questions, requests for explanation, or language expressing uncertainty or confusion.
- **Influence (Asserting a View).** A post is classified as *Influence* if its primary purpose is to assert a directional interpretation or persuade others. Typical signals include confident or assertive language, explicit or implicit trading recommendations, or narrative framing intended to shape beliefs.
- **Unidentified.** A post is classified as *Unidentified* if it does not clearly reflect validation or influence. Common examples include neutral information sharing, price quotes without interpretation, administrative or personal comments, humor without market meaning, or posts whose intent is ambiguous.

Decision Rules: Annotators apply the following rules:

1. *Dominant signal rule:* The label reflects the primary purpose of the post.
2. *Question rule:* A question is not automatically classified as Validation; rhetorical questions used to persuade are classified as Influence.
3. *Recommendation rule:* Explicit trade recommendations are classified as Influence unless clearly framed as uncertainty.
4. *Ambiguity rule:* If the post cannot be confidently classified after rereading once, it is labeled Unidentified.

Annotation Process: Posts are independently labeled by three human annotators using this shared protocol. Annotators work without access to each other's labels. When all annotators agree, the assigned label is retained. When disagreement occurs, the final label is determined by majority rule.

IA-2. Annotation Calibration Examples

Instructions: Below are example X.com posts used to calibrate annotators prior to full labeling. For each post, assign one label: *Validation*, *Influence*, or *Unidentified*. Labels should reflect the dominant communication incentive expressed in the text.

Calibration Posts:

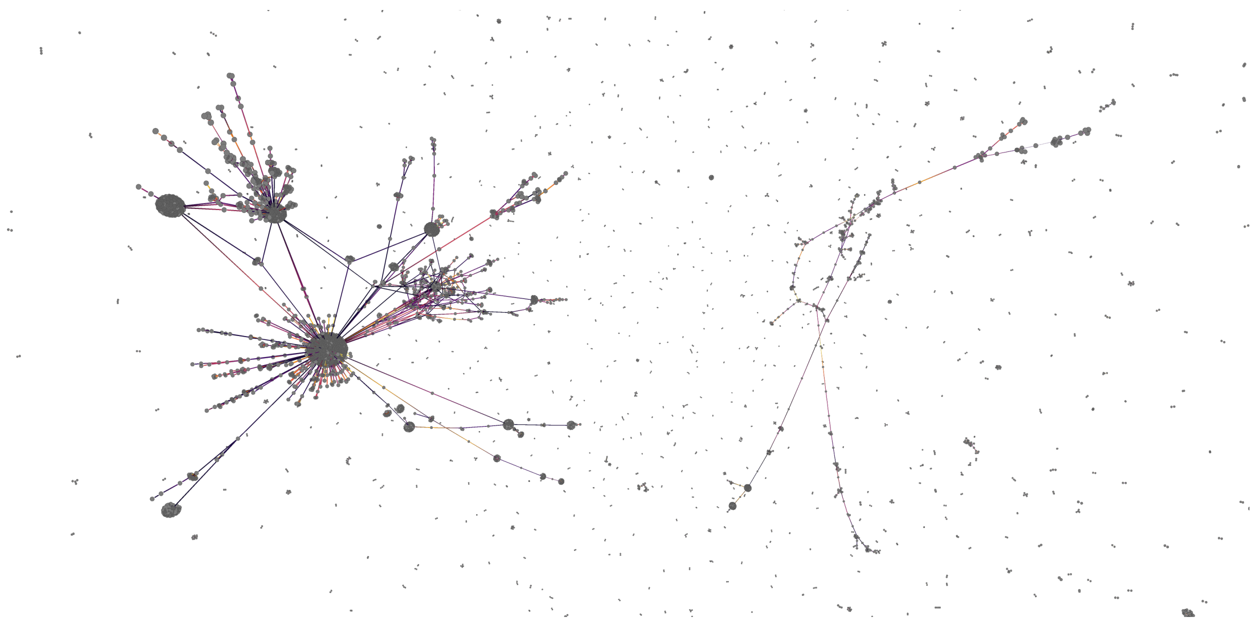
1. "WASDE out — honestly not sure how to read the stocks revision. Anyone have thoughts?"
2. "Corn selling off makes zero sense. This is clearly bullish long term."
3. "ZS 1365.25"
4. "Trying to understand why wheat didn't react more to the acreage number."
5. "Stop panicking. This report changes nothing — buy the dip."
6. "WASDE released at noon ET today."
7. "Am I missing something here, or is this move just noise?"
8. "Market is wrong again. Funds will chase this higher."
9. "Interesting spread behavior after the report."
10. "People don't get it — this setup is obviously bearish."

Answer Key and Explanations:

Post	Label	Explanation
1	Validation	Explicit uncertainty and request for input.
2	Influence	Confident directional assertion.
3	Unidentified	Pure price quote without interpretation.
4	Validation	Seeks interpretation of market response.
5	Influence	Persuasive language and trading recommendation.
6	Unidentified	Neutral informational announcement.
7	Validation	Question expressing uncertainty.
8	Influence	Assertive narrative framing.
9	Unidentified	Ambiguous observation without clear intent.
10	Influence	Persuasive and dismissive tone asserting direction.

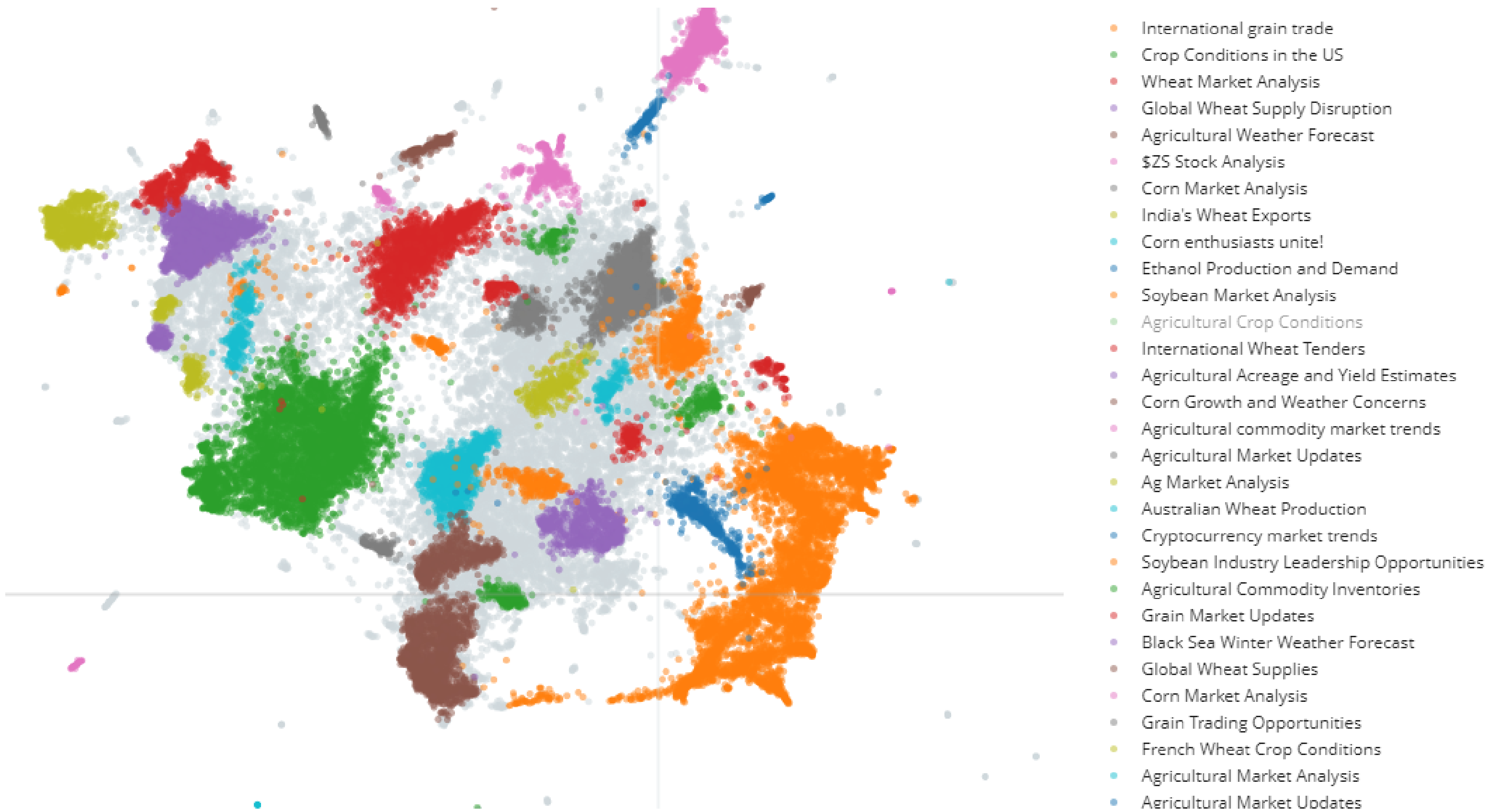
Notes: Posts 9 and 10 are included to highlight common boundary cases between Influence and Unidentified. Annotators reviewed disagreements during training to ensure consistent interpretation of the protocol.

Figure IA.1: Degree Centrality of X.com Sample (07.13.22 - 07.19.22)



Note: The figure illustrates how different X.com accounts have varying levels of connectivity (two accounts are connected if a post is passed between those accounts by likes, reposts, and replies) and, thus, can exert different influences on the network. To aggregate sentiment within a network, I use the notion of degree centrality, defined as the number of connections each post has, which captures this differential weight in the network.

Figure IA.2: Extracted Topics



Note: The figure depicts the universe of extracted topics. For visualization purposes, the dimensionality of the data is reduced to two dimensions. Therefore, the exact position of each topic cluster does not convey much meaning beyond illustrating the clustering of posts.