Unstructured Data Analysis

Imperial College London

Department of Mathematics

Coursework 2 - Natural Language Processing

Author:
Tom Richardson (CID: 01349943)

Date: January 5, 2026

I have worked independently on this assignment & everything can be found on GitHub - please click here!.

Abstract

1 Problem Statement

Recent rapid growth in generative artificial intelligence (AI) and related products has been accompanied by increased public discourse about whether current market valuations and investment levels constitute an "AI bubble." This project examines public beliefs about the existence and perceived imminence of a potential AI-related market bubble by analysing large-scale social media discourse.
Specifically, the study addresses the following research questions:

1. Do people believe that there is currently an AI bubble (in AI-related stocks or markets)?

2. If so, do people believe that this bubble is about to burst, and what triggers do they cite?

To answer these questions, the project analyses posts from the social media platform X (formerly Twitter) that reference AI, market valuation, bubbles, hype, and related concepts. Each post is treated as a short, noisy, but potentially informative signal of belief, sentiment, and expectation. The analysis is descriptive and focuses on patterns within the sampled discourse rather than population-level inference.

1.1 Approach to Question 1: Belief in the Existence of an AI Bubble

Question 1 is addressed using a simple rule-based stance classification approach. Each post is lowercased and scanned for predefined lexical cues indicating either belief in an AI bubble (e.g. "AI bubble", "bubble burst", "overvalued AI") or rejection of the bubble framing (e.g. "not a bubble", "real value", "long-term growth"). Posts containing only belief cues are classified as expressing belief, posts containing only disbelief cues are classified as rejecting the bubble narrative, posts containing neither are labelled neutral/unclear, and in cases where both are present, disbelief is prioritised to avoid false positives. This transparent keyword-based method was chosen due to the absence of labelled training data and the exploratory nature of the analysis, and it provides an interpretable baseline for estimating the prevalence of explicit bubble-related beliefs in the sampled discourse.
The post-level stance labels are then aggregated to estimate the proportion of posts that express bubble belief among the sampled content and to support further subgroup/temporal breakdowns.

1.2 Approach to Question 2: Perceived Imminence and Causes of a Bubble Burst

Question 2 is addressed in two complementary ways, applied to the subset of posts predicted to support bubble belief:

- **Perceived imminence**: construct time series describing how frequently bubble-believing posts use language associated with bursting or near-term correction (e.g. “burst”, “crash”, “correction”) and measure contemporaneous sentiment. Sentiment is estimated using VADER, a rule-based sentiment model designed for short-form social media text (**?**).
- **Perceived causes**: extract recurring themes in bubble-believing discourse using topic modelling. Latent Dirichlet Allocation (LDA) is used to identify a small set of interpretable topics that summarise major themes (e.g. valuation, profitability, compute constraints, regulation), following the standard probabilistic topic modelling framework introduced by (**?**). Topic prevalence is then tracked over time and contrasted between posts that do vs. do not contain near-term bursting language.

This combination of stance filtering, sentiment trending, and topic modelling is consistent with common text-as-data practice in financial and social media analyses (e.g., social-media sentiment analyses relating to financial markets and influential accounts (**?**)).

2 Data Selection

A key practical challenge in this project was that it was problem-first rather than data-first: the research questions were defined before confirming the feasibility of collecting sufficiently rich, large-scale textual data. In practice, platform API commercialisation and access restrictions significantly shaped the final dataset and consumed a larger share of project time than initially anticipated. For transparency and reproducibility, this section documents the data source exploration and final selection.

2.1 Candidate Data Sources Considered

Several sources were explored:

- **Reddit**: attractive due to longer-form discussion, topic-specific subcommunities, and rich conversational context. However, an application for Reddit API access was submitted and rejected, preventing systematic data collection from this platform.
- **News aggregators (NewsAPI) and Hacker News**: these sources provided structured access to headlines, summaries, and links. In practice, they were

not well aligned with the research questions because the content predominantly reflects journalistic/editorial framing rather than direct expressions of individual belief. Additionally, rate limits and free-tier restrictions reduced the ability to gather a sufficiently rich corpus for robust analysis.

- **X (Twitter)**: selected due to the volume of short, opinionated posts, strong presence of finance/tech discourse, and frequent use of explicitly speculative language (e.g. “bubble”, “hype”, “crash”).

2.2 Choice of X as the Primary Data Source

X was selected as the primary data source because it offers: (i) high-volume short-form text suitable for temporal tracking, (ii) frequent explicit discussion of market narratives, and (iii) a culture of public speculation and prediction. While X users are not representative of the general population, the platform provides a concentrated view of technology and market discourse that is directly relevant to the research questions. Accordingly, the analysis is interpreted as describing beliefs within the sampled X discourse rather than general public opinion.

2.3 Data Acquisition

Data were collected using keyword-based queries targeting AI terms and bubble-related language (e.g. AI, LLM, valuation, hype, bubble, crash, burst). Posts were retained along with associated metadata (e.g. timestamps, engagement metrics) to support temporal analysis and potential weighting/sensitivity checks.

2.4 Filtering and Quality Controls

Several quality controls were applied:

- **Language filtering**: retain only posts marked as English.
- **Basic validity checks**: remove posts with missing or malformed text/timestamps.
- **Text normalisation**: replace URLs and user mentions with placeholders and standardise whitespace to reduce noise while preserving semantics needed for transformer models.
- **Optional content filtering**: metadata such as reply status and engagement measures are retained to enable later sensitivity analysis (e.g. excluding replies that lack standalone content, or comparing high-engagement posts to the full set).

3 Methodology

The methodological approach follows a text-as-data framework in which unstructured social media posts are transformed into structured variables suitable for quantitative analysis. The pipeline consists of four stages: (i) data loading, (ii) preprocessing, (iii) stance classification for Question 1, and (iv) sentiment and topic modelling for Question 2.

3.1 Unit of Analysis

The unit of analysis is an individual post from the platform X (formerly Twitter). Each post is treated as a short textual document associated with a timestamp and minimal metadata. Posts are treated as independent observations for the purposes of aggregation and trend analysis.

3.2 Data Loading and Filtering

Posts were collected externally and stored in JSON Lines (JSONL) format, with one post per line and associated metadata. The dataset was loaded into a pandas DataFrame using a custom JSONL parser. Only posts marked as English-language by the platform were retained. Posts with missing or malformed timestamps or text were removed. Each post was then assigned to a daily time bin based on its timestamp to support temporal aggregation.

3.3 Text Preprocessing

Prior to analysis, raw post text was normalised to reduce noise and sparsity. The preprocessing pipeline consisted of:

- Lowercasing all text.
- Replacing URLs and user mentions with placeholder tokens.
- Removing non-alphanumeric characters and excess whitespace.
- Tokenising text into words.
- Removing standard English stopwords.
- Applying Porter stemming to reduce words to a common morphological form.

This produced a cleaned representation used for vectorisation and modelling.

3.4 Stance Classification for Question 1

3.4.1 Task Definition

Stance classification is framed as a binary classification problem: predicting whether a post supports the belief that an AI bubble exists (`bubble`) or does not support this belief (`not_bubble`). The goal is not to infer the true belief of the author, but to identify linguistic framing consistent with bubble-related discourse.

3.4.2 Weak Supervision

Because no manually labeled data were available, a weak supervision strategy was used to construct an initial training set. A set of heuristic lexical cues was defined:

- Bubble-affirming cues (e.g., bubble, overvalued, mania, burst, crash, correction).
- Bubble-rejecting cues (e.g., not a bubble, no bubble, fundamentals, here to stay).

Posts containing clear instances of these cues were assigned provisional labels accordingly. Posts that did not match either set were excluded from the training data.

3.4.3 Model Training

A logistic regression classifier was trained on TF–IDF representations of the weakly labeled posts. Unigrams and bigrams were used, and extremely rare or extremely frequent terms were filtered to reduce noise and dimensionality. Logistic regression was chosen for its robustness to high-dimensional sparse input and its interpretability.
The weakly labeled data were split into training and validation sets using a stratified 80/20 split. Model performance was evaluated using precision, recall, F1-score, and a confusion matrix. After evaluation, the model was retrained on the full weakly labeled dataset and applied to all posts to generate stance predictions and associated probabilities.

3.5 Temporal Measures and Sentiment for Question 2

To address the second research question, analysis was restricted to posts predicted to support the existence of an AI bubble.

3.5.1 Burst Imminence Indicator

A lexicon-based “burst imminence” score was computed for each post based on the presence of terms associated with bursting or correction (e.g., burst, pop, crash, collapse) and terms indicating near-term timing (e.g., soon, imminent, about to). Hedging terms (e.g., might, maybe) were used to down-weight speculative language. This produced a continuous indicator of perceived imminence.

3.5.2 Sentiment

Sentiment was estimated using the VADER sentiment analyser, which is designed for short-form social media text and accounts for punctuation, intensifiers, and informal language. The compound sentiment score was computed for each post and aggregated over time.

3.5.3 Temporal Aggregation

Burst imminence and sentiment were aggregated at a daily level. Rolling averages were computed to smooth short-term volatility and highlight broader trends.

3.6 Topic Modelling for Question 2

To identify recurring themes in bubble-related discourse, topic modelling was applied to bubble-believing posts using Latent Dirichlet Allocation (LDA). Text was represented using a count-based bag-of-words model with unigrams and bigrams after standard cleaning and stopword removal. The LDA model produces a distribution over topics for each post and a distribution over words for each topic.

Topic prevalence was aggregated by day to examine how thematic emphasis changed over time. Topic distributions were also compared between posts that contain near-term bursting language and those that do not, in order to identify themes disproportionately associated with imminent-burst narratives.

3.7 Limitations

This methodology is exploratory and descriptive. Weak supervision introduces bias because the classifier is trained on heuristically labeled data and may reproduce the assumptions embedded in the cue lists. Class imbalance further limits the reliability of detecting sceptical posts. In addition, sarcasm, irony, and implicit meaning are difficult to capture with bag-of-words representations.

Finally, the dataset reflects a filtered subset of discourse and is not representative of the general population. Results are therefore interpreted as patterns within this sampled discourse rather than as estimates of public opinion.

Despite these limitations, the pipeline provides a transparent and reproducible framework for transforming social media text into structured signals of belief, sentiment, and thematic emphasis, enabling systematic exploration of narratives surrounding an AI-related market bubble.

4 Results

This section presents the empirical findings derived from applying the methodology described above to the collected X discourse dataset. Results are structured according to the two research questions: (i) whether users express belief in the existence of an AI bubble, and (ii) whether those who believe in a bubble perceive it as imminent and what causes they associate with a potential burst.

4.1 RQ1: Belief in the Existence of an AI Bubble

4.1.1 Stance Distribution

Figure 1 shows the distribution of stance labels across the dataset. The majority of posts are classified as neutral or ambiguous, reflecting that much AI-related discussion does not explicitly express a valuation stance. Among stance-expressive posts, a substantial proportion express belief in the existence of an AI bubble, while a smaller fraction explicitly reject it.

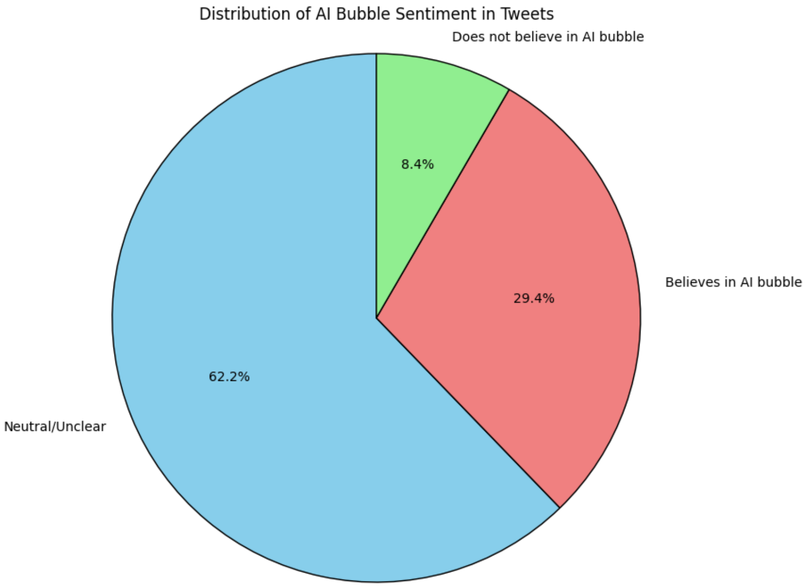

Figure 1: Distribution of stance toward the existence of an AI bubble.

4.1.2 Interpretability via Model Coefficients

Inspection of learned coefficients reveals that terms such as *bubble*, *burst*, *pop*, *crash*, and *hype* strongly push predictions toward the bubble class, while terms such as *fundamentals*, *revenue*, *enterprise*, and *capex* push toward the not-bubble class. This aligns with intuitive distinctions between speculative versus fundamentals-based framing.

4.2 RQ2: Perceived Imminence and Causes of a Bubble Burst

All subsequent analyses condition on posts classified as bubble-believing.

4.2.1 Temporal Trends in Imminence and Sentiment

Figure 2 shows the 7-day rolling average of the Burst Imminence Index and the share of posts mentioning bursting or correction. Both measures exhibit temporal variation, indicating that perceived urgency is not constant but fluctuates in response to events or narratives.

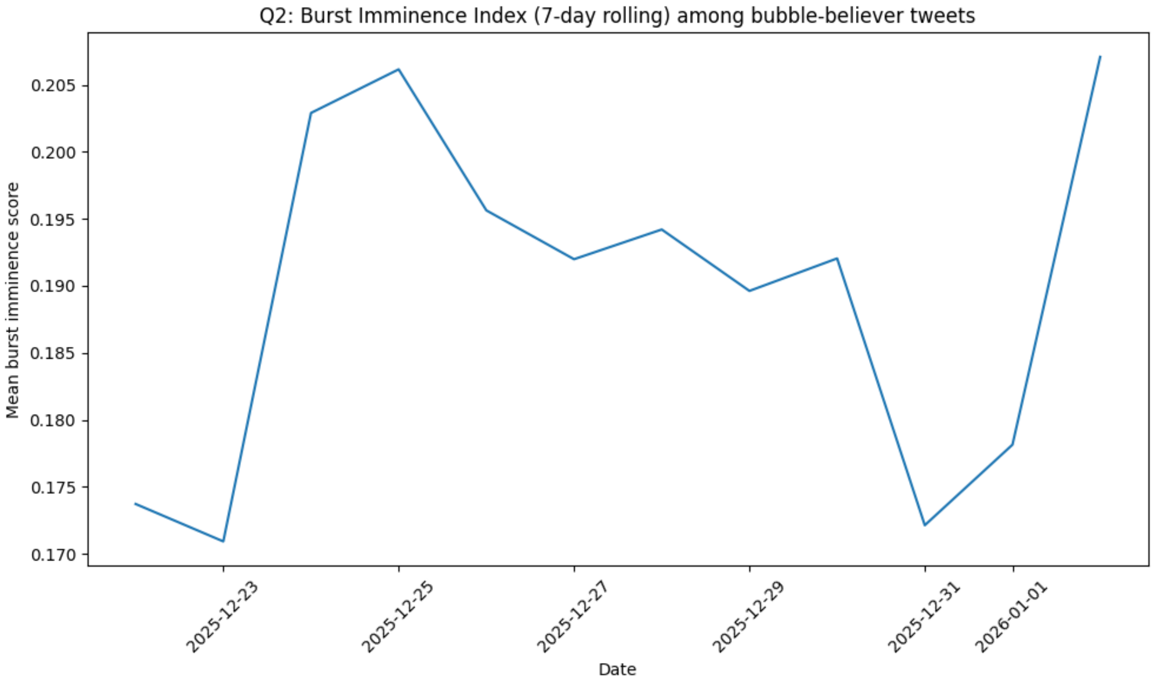

Figure 2: Temporal trends in bursting language among bubble-believing posts.

Figure 3 shows the rolling mean sentiment. Sentiment becomes more positive over the period, suggesting that increased imminence does not necessarily coincide with pessimistic emotional tone.

4.2.2 Topic Modelling: Perceived Causes and Narratives

Latent Dirichlet Allocation with six topics was applied to the bubble-believing subset. Table 1 summarises representative topic keywords.

Topic	Top terms
0	bubble, hype, market, crypto, year
1	build, infrastructure, agent, project
2	burst, pop, stock, bubble burst
3	verify, proof, trust, lab
4	investor, money, valuation, company
5	token, trade, launch, narrative

Table 1: Summary of LDA topics with representative terms.

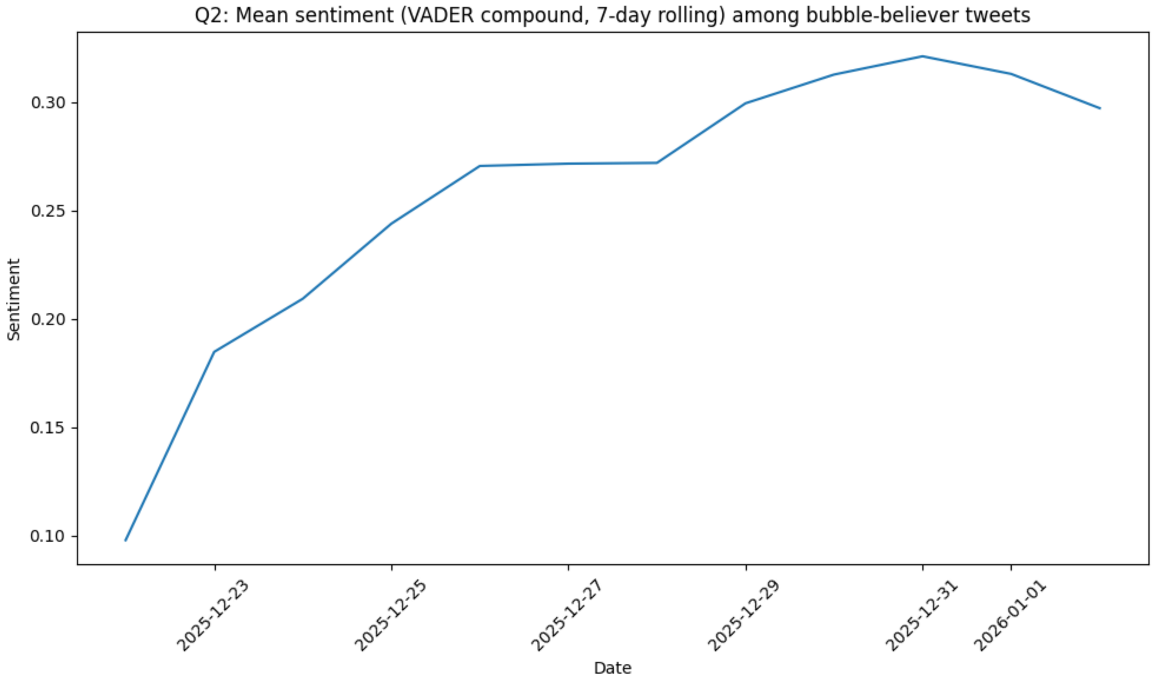

Figure 3: Mean sentiment (VADER compound) among bubble-believing posts over time.

Finally, topics were compared between posts mentioning imminent bursting and those that did not. Topic 2 (explicit bursting language) shows the strongest positive association with imminence, while infrastructure and long-term build narratives are negatively associated, as shown in Table 2.

Topic	Mean Difference (Imminent – Not)
Topic 2 (burst/pop)	+0.32
Topic 0 (bubble/hype)	+0.07
Topic 1 (infrastructure)	-0.28

Table 2: Topic association with imminent bursting language.

5 Conclusion

This project began from a problem-first research goal and encountered substantial constraints in data access due to API restrictions and commercialisation. Despite these constraints, the final pipeline demonstrates how large-scale X discourse can be transformed into structured signals of stance, sentiment, and themes to study speculative market narratives. Future work would prioritise richer manual labelling, improved handling of sarcasm/irony, and broader triangulation across platforms (e.g. Reddit, news, forums) if access permits.