

Bags of Words, Bags of Analysts: Two Event Spaces for Industry Classification

Artem Lenskiy

School of Engineering, UNSW Canberra, Australia

a.lenskiy@unsw.edu.au

Abstract

A company’s GICS sector can be inferred from what the firm writes about itself, and it can be inferred from who chooses to write about the firm. We show these are the same inference over two event spaces, and fit one multinomial naïve Bayes to both. In the first a company is a bag of the words of Item 1 of its 10-K; in the second it is a bag of the 814 analysts who covered it, with no use of what they wrote. Sector prototypes are counted log-probabilities in both cases, $\mathbf{w}_s = \log \mathbf{p}_s$, so the two differ only in what is being counted.

Over 426 S&P 500 constituents the word space gives 83.7% top-1 recovery and the analyst space 73.7%, against a 16.2% majority-class baseline. Because the two spaces are conditionally independent given the sector their log-likelihoods simply add — but doing so captures only a quarter of the available gain (+0.75 points). A filing supplies some 2,100 word events against 18 analyst events at the median, and naïve Bayes counts each as independent, so the word log-likelihood varies across the eleven sectors with a median standard deviation of 766 nats against 8.3 for the analyst log-likelihood, a factor of 92: the sum is numerically dominated and analyst evidence rarely moves a decision. Standardising each log-likelihood within a company before adding removes the scale disparity and recovers +2.9 points [0.4, 5.4], reaching 86.5% top-1 and 94.8% top-2.

A dense encoder is not better here, and not combinable on the same terms. Qwen3-Embedding-4B over the same filings reaches 76.9% under the nearest-centroid classifier that dense vectors admit, below naïve Bayes over the same words; and having no event counts it cannot form a third event space, so combining it means leaving the likelihood and choosing a scale.

1 Introduction

Sector labels define fund mandates, group exposures in risk models, and form the peer sets used

throughout empirical finance. Inferring them from data is usually posed as text classification over what a firm says about itself: Item 1 of its annual 10-K, the *Business* section, which underlies text-derived industry classifications (Hoberg and Phillips, 2016) and which recent work encodes with a neural model (Vamvourellis et al., 2023; Ito et al., 2020; Cao et al., 2023).

There is a second, less obvious source. Analysts specialise, so the set of companies an analyst chooses to cover reveals a preference over sectors, and the set of analysts covering a company inverts to evidence about its sector — using only *who* wrote, never what they wrote. Predicting an item’s category from who attends to it has been done for movie genres (Makita and Lenskiy, 2016); here the items are firms and the raters are analysts.

Our point is that these are not two methods but one, over two event spaces. Let a company be a multiset of events and fit a multinomial naïve Bayes: sector prototypes are counted log-probabilities $\mathbf{w}_s = \log \mathbf{p}_s$, and the score is the log-likelihood of the company’s events under sector s . Let the events be *words* and this is the classical text classifier (McCallum and Nigam, 1998); let them be *analysts* and it is a model of coverage behaviour. Because the two spaces are disjoint, evidence combines by adding log-likelihoods rather than by any fitted mixing rule.

That last step is where the model fails, and the size of the failure is the point. Our contributions are:

One multinomial naïve Bayes over two disjoint event spaces — words a firm writes, analysts who write about it — making the textual and behavioural signals instances of a single estimator rather than two systems to be blended (Section 4).

A quantified demonstration that adding their log-likelihoods, though exactly what condi-

tional independence licenses, captures only a quarter of the available gain: unequal event counts leave one space’s log-likelihood spread 92 times the other’s, so the sum is numerically dominated. Per-source normalisation before combining recovers +2.9 points against the sum’s +0.75 (Section 5).

A dense-encoder baseline that neither beats naïve Bayes over the same words (76.9% against 83.7%) nor admits the counts the probabilistic model needs.

2 Related Work

The word space. Loughran and McDonald (2011) show general-purpose sentiment lexicons misclassify financial language, an early demonstration that filing vocabulary is idiosyncratic enough to matter. Hoberg and Phillips (2016) build industry classifications from 10-K product descriptions. Vamvourellis et al. (2023) embed 2,590 Russell 3000 business descriptions and predict GICS with logistic regression, reaching 83.6% pre-trained and 90.2% fine-tuned, and report that accuracy rises only marginally from 512 to 1,536 input tokens. Ito et al. (2020) learn company embeddings supervised by sector labels and returns, Cao et al. (2023) tune generative models for sector inference, and Yüksel et al. (2026) add geographic modalities. Such encoders are trained on short inputs (Reimers and Gurevych, 2019), so filings require sparse attention (Beltagy et al., 2020) or chunking with pooling, the route we take with Qwen3-Embedding (Zhang et al., 2025); sparse lexical scoring remains a strong baseline (Robertson and Zaragoza, 2009), and Tang and Yang (2025) find bag-of-words beating dense embeddings on financial semantic similarity.

The analyst space. Predicting an item’s category from who interacts with it rather than from its content has been applied to movie genres (Makita and Lenskiy, 2016) and refined with similarity networks (Ghawi and Pfeffer, 2019); our analyst space applies it to analyst attention. The platform has been studied for the return-predictive content of its articles (Chen et al., 2014); we use coverage alone. What is new is not either space but their unification under one estimator.

3 Task and Data

Given a company c and the eleven sectors of the Global Industry Classification Standard, we predict its sector, and separately rank the ten sectors it is

not assigned to, the top-ranked of which we call its *complement*.

Filings. For each S&P 500 constituent we retrieve the 10-K filed most recently before 15 April 2024 and extract Item 1 by pairing each *Business* heading with the following *Risk Factors* heading and keeping the widest plausible span. Item 1 is recoverable for 426 of 502 constituents; the rest file no 10-K or incorporate Item 1 by reference. The median section is 6,612 words (IQR 3,994–9,549), a median of 8,738 tokens (90th percentile 17,202); 337 filings are from 2024.

Analyst coverage. As a second, non-textual view we use analysts’ coverage decisions: which companies each analyst chooses to write about. From a financial commentary platform we take 16,248 analyst–company pairs from 814 authors over the same period as a binary incidence matrix, 3.98% dense over all 502 constituents. Restricted to the 426 with a filing it is 814×426 with 13,678 pairs: a median of 18 analysts per company, range 3–268. An analyst’s covered firms concentrate in a few sectors, so their presence on a company carries the sector preference learned from the rest of their coverage.

Labels and baselines. Each company carries one of the eleven GICS sectors, and the classes are unbalanced: *Industrials* is largest at 16.2%, *Energy* smallest at 2.1%. Always predicting the largest scores 16.2% top-1 and 30.5% top-2; a uniform guess scores 9.1% and 18.2%.

4 Method

Let a company c be a multiset of events drawn from a finite space $\mathcal{E}$, with $x_{e,c}$ the number of times event e occurs for c . Under a multinomial naïve Bayes the events are independent given the sector, so the log-likelihood of c under sector s is linear in the counts,

$$\begin{aligned} f_s(c) &= \sum_{e \in \mathcal{E}} x_{e,c} \log p_{e,s} + \log \pi_s \\ &= \langle \mathbf{x}_c, \mathbf{w}_s \rangle + \log \pi_s, \end{aligned} \quad (1)$$

with $\mathbf{w}_s = \log \mathbf{p}_s$ and $\hat{s}(c) = \arg \max_s f_s(c)$. The parameters are counted, not fitted: writing $\mathbf{C}$ for the company–event count matrix and $\mathbf{S}$ for the one-hot sector matrix, the sector–event counts are $\mathbf{M} = \mathbf{S}^\top \mathbf{C}$ and

$$p_{e,s} = \frac{\mathbf{M}_{s,e} + \alpha}{\sum_{e'} \mathbf{M}_{s,e'} + \alpha |\mathcal{E}|}. \quad (2)$$

Two matrix products give every posterior; there is no iterative fitting.

Two event spaces. We instantiate Equation (1) twice. In the *word* space $\mathcal{E}_W$ is the terms of Item 1 surviving a document-frequency filter fitted per fold on training documents (12,768 on average) and $x_{w,c}$ is the term frequency, which is the classical multinomial text classifier (McCallum and Nigam, 1998). In the *analyst* space $\mathcal{E}_A$ is the 814 authors and $x_{a,c} \in \{0, 1\}$ records whether analyst a covered c ; the sentiment of the article is discarded, so a firm is a bag of analysts exactly as it is a bag of words. With binary indicators the multinomial is Bernoulli naïve Bayes with the absence terms dropped (McCallum and Nigam, 1998); sweeping an absence weight confirms dropping them is right here: on a held-out split the indicator form scores 75.2% against 61.6% when absences are weighted in. Smoothing α is selected for *both* spaces from one grid on one inner split, so neither is tuned at the other’s expense (median 0.03, 0.1).

Combining them. Treat the two spaces as conditionally independent given the sector, so the joint likelihood factorises and the log-likelihoods add,

$$f_s^{WA}(c) = f_s^W(c) + f_s^A(c) - \log \pi_s, \quad (3)$$

subtracting the prior counted twice; no mixing weight appears. This is a product of two multinomials, not a single multinomial over $\mathcal{E}_W \cup \mathcal{E}_A$, which would additionally allocate mass between the spaces and contribute a sector-dependent term Equation (3) omits; the factorised form is the standard treatment of two feature blocks. Section 5 shows what this leaves on the table, and that replacing f by its within-company standardisation $z(f) = (f - \bar{f})/\text{sd}(f)$ before adding recovers it.

A dense comparator. Equation (1) needs counts, which an embedding does not provide, so we use the classifier dense vectors do admit: a nearest centroid, scored by cosine, applied to Qwen3-Embedding-4B in its strongest configuration and, as a like-for-like reference, to TF-IDF over the same vocabulary.

Protocol. Repeated stratified 10-fold cross-validation, 20 repeats, so folds partition rather than overlap. Counts, smoothing, vocabulary and centroids are all estimated on the fold’s training companies. Intervals are a bootstrap over companies, the company being the unit of inference.

Model	Top-1	Top-2
<i>naïve Bayes, Equation (1)</i>		
word space	83.7	94.2
analyst space	73.7	87.0
both spaces, summed	84.4	94.3
both, standardised	86.5	94.8
<i>nearest centroid (no counts required)</i>		
TF-IDF	77.7	86.2
Qwen3-Emb 4B	76.9	90.6
majority class	16.2	30.5

Table 1: Sector recovery (%) under repeated stratified 10-fold cross-validation, 20 repeats, except the Qwen row, which is the 50-split protocol under which its pooling variants were compared; under cross-validation it scores 75.1/90.0. Qwen is mean-pooled over token positions on the whole filing, its best configuration.

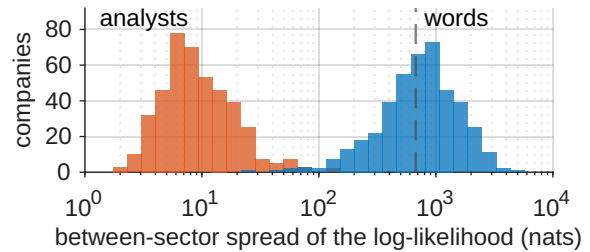


Figure 1: Standard deviation of a company’s eleven sector log-likelihoods in each event space; one observation per company, matched smoothing. The median is 766 nats for words against 8.3 for analysts, a factor of 92 (unchanged under largest-minus-smallest: 2,673 against 29). The axis is logarithmic because the two do not otherwise share a panel. The dashed line is the word model’s median top-1/top-2 margin, 672 nats: an analyst spread must reach it to reorder a word decision, and 2.1% do.

5 Results

Both event spaces carry sector information.

Table 1 gives the arms. Words give 83.7% top-1, 95% CI [80.3, 86.9]; analysts give 73.7% [70.0, 77.3], from an 814×426 matrix that is 3.9% dense and contains no text. Both far exceed the 16.2% majority-class baseline, and the analyst space comes within ten points of the word space while observing nothing a firm says.

Adding log-likelihoods under-delivers, and why.

Equation (3) is what conditional independence licenses, and it does help: the sum scores 84.4%, +0.75 points over words alone, 95% CI [0.27, 1.35]. But standardising first is worth nearly four times as much. The reason is not numerical — single rather than double precision changes no prediction — but that the two spaces supply very unequal

evidence. A filing contributes a median of 2,078 word events over the retained vocabulary against 18 analyst events, and naïve Bayes counts each as independent, so the word log-likelihood varies across the eleven sectors with a median standard deviation of 766 nats against 8.3 for the analyst log-likelihood, a factor of 92 (Figure 1). The word space’s median margin between its best and second-best sector is 672 nats, which the analyst spread exceeds for 2.1% of companies.

This is the independence assumption failing asymmetrically — a known failure in a less familiar guise. Domingos and Pazzani (1997) show that naïve Bayes can classify well while its probability estimates are badly wrong, and Rennie et al. (2003) correct those estimates for text. Both concern a single classifier, where mis-calibration is harmless because only the argmax is read. It stops being harmless once two such models are added. Words in a filing are strongly dependent, so counting two thousand of them as independent inflates the word likelihood far beyond its evidential content; analysts are not independent either — they follow one another onto a story — but two dozen events admit far less over-counting. The rule weights the two by their inflation rather than their information.

The remedy is standard: normalise each source before combining (Rennie et al., 2003). Standardising each space’s eleven log-likelihoods within a company preserves its ordering while removing its scale, and recovers most of what the raw sum leaves: 86.5% top-1, +2.9 points over words alone [0.4, 5.4], $p = 0.023$. The gain is on exact recovery only; on top-2 the two rules are indistinguishable (+0.6, $p = 0.51$). No mixing weight is fitted, though standardising is itself a per-company rescaling by $1/\text{sd}(f)$ rather than a weight of one.

The dense encoder is not better, and not combinable on the same terms. Under the nearest centroid that dense vectors admit, Qwen3-Embedding-4B reaches 76.9% top-1 (75.1% under cross-validation), below naïve Bayes on the same filings either way, though it ranks the neighbouring sector better (90.6% top-2 against 86.2% for TF-IDF under that classifier). Replacing the centroid by naïve Bayes on an identical per-fold vocabulary is worth +5.9 points [2.6, 9.3], $p = 0.0006$, so the classifier matters more here than the lexical-or-neural choice. And the embedding cannot form a third event space: no counts, so no multinomial. It can still be combined — a density over embeddings,

or the standardised blend above — but only by leaving Equation (3) and choosing a scale, which counted events avoid.

The spaces fail on different sectors. *Energy* recovers at 66.7% in the word space but 98.9% in the analyst space — nine firms analysts cover almost interchangeably — while *Materials* reverses, 66.1% against 33.0%. Macro-F1 rises from 83.2 and 74.3 to 87.4, further than accuracy does. The two also propose different *second* sectors, agreeing for 21–23% of companies against 7–10% by chance ($\kappa = 0.15$), where two text encodings agree with each other for 54% ($\kappa = 0.50$): the disagreement is between event spaces, not encodings of one.

The posterior below the argmax. Equation (1) returns a distribution, not a label, and its runner-up is informative. Taking each company’s highest-ranked *non-assigned* sector as a proposed *completion*, we test it on data the model never sees: over 571 trading days we residualise daily returns on the equal-weighted market and ask whether that sector is the one it co-moves with most among the ten it is not assigned to. It is, for 17.0% of companies against a marginal-preserving permutation null of 11.1% ($z = 4.1$). This survives projecting out three and five principal components but not ten, by which point the within-minus-cross sector correlation gap GICS itself carries has fallen 62% — that control removes the signal being measured. It is robust to moderate factor controls and untested beyond them.

6 Conclusion

Classification from filings and from coverage behaviour is one estimator over two event spaces, and stating it that way makes the shortfall of the exact combination diagnosable: the sum is not wrong, it is dominated. Naïve Bayes’ over-confidence, harmless when one argmax is read, becomes decisive once two models are added, and normalising each source first recovers what the sum leaves. A dense encoder is outperformed here and, having no counts, cannot join as a third space.

Limitations

The evaluation covers 426 large-capitalisation US firms at one point in time; neither smaller firms nor other markets nor other periods are tested, and Item 1 extraction fails for 76 constituents in a way

that is unlikely to be random, since filings that incorporate Item 1 by reference are plausibly the more structurally complex firms. A manual audit of 30 sampled extractions found 25 beginning at the true *Business* heading and 5 beginning instead at an internal cross-reference to it, and so covering only the tail of the section. Detecting that pattern automatically flags 46 of 426 filings (10.8%); removing them leaves the comparison unchanged (TF-IDF 78.7% against Qwen 76.7%, +2.0 points [−1.6, +5.7]), so extraction noise is real but does not favour either representation.

Qwen3-Embedding is mean-pooled over token positions; attention-weighted or late-interaction pooling might order the encoders differently. It is also one general-purpose family under one shared classifier, and a finance-tuned encoder or a tuned supervised head would score higher (Vamvourellis et al., 2023), so the dense arm is a reference point rather than a ceiling.

The co-movement check is contemporaneous with the filings and establishes association rather than prediction, and residualising on the equal-weighted market leaves size and style exposure in the residuals. The 200 folds reuse the same 426 companies, so fold-level intervals understate uncertainty; every comparison is therefore tested on paired company-level outcomes. The analyst data are licensed and cannot be redistributed, which limits exact reproduction of the coverage arm; the construction itself is not restricted, since it needs only a writer–company incidence matrix, which I/B/E/S supplies directly and public bylines approximate. The platform’s contributor base is self-selected and skews towards liquid, widely held names. Finally, the ground truth is itself a provider’s taxonomy whose schemes disagree materially with one another (Bhojraj et al., 2003), so recovery measures agreement with GICS rather than correctness.

References

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *CoRR*, abs/2004.05150.
- Sanjeev Bhojraj, Charles M. C. Lee, and Derek K. Oler. 2003. [What’s my line? a comparison of industry classification schemes for capital market research](#). *Journal of Accounting Research*, 41(5):745–774.
- Lele Cao, Vilhelm von Ehrenheim, Astrid Berghult, Cecilia Henje, Richard Anselmo Stahl, Joar Wandborg, Sebastian Stan, Armin Catovic, Erik Ferm, and Hannes Ingelhart. 2023. A scalable and adaptive system to infer the industry sectors of companies: Prompt + model tuning of generative language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing*, pages 55–62.
- Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2014. [Wisdom of crowds: The value of stock opinions transmitted through social media](#). *The Review of Financial Studies*, 27(5):1367–1403.
- Pedro Domingos and Michael Pazzani. 1997. [On the optimality of the simple Bayesian classifier under zero-one loss](#). *Machine Learning*, 29(2–3):103–130.
- Raji Ghawi and Jürgen Pfeffer. 2019. [Movie genres classification using collaborative filtering](#). In *Proc. 21st Int. Conf. on Information Integration and Web-based Applications & Services (iiWAS ’19)*, pages 1–10.
- Gerard Hoberg and Gordon Phillips. 2016. [Text-based network industries and endogenous product differentiation](#). *Journal of Political Economy*, 124(5):1423–1465.
- Tomoki Ito, Jose Camacho-Collados, Hiroki Sakaji, and Steven Schockaert. 2020. Learning company embeddings from annual reports for fine-grained industry characterization. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 27–33.
- Tim Loughran and Bill McDonald. 2011. [When is a liability not a liability? textual analysis, dictionaries, and 10-ks](#). *The Journal of Finance*, 66(1):35–65.
- Eric Makita and Artem Lenskiy. 2016. [A multinomial probabilistic model for movie genre predictions](#). *CoRR*, abs/1603.07849.
- Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.
- Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive Bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 616–623.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.

- Yixuan Tang and Yi Yang. 2025. [FinMTEB: Finance massive text embedding benchmark](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3620–3638.
- Dimitrios Vamvourellis, Máté Tóth, Snigdha Bhagat, Dhruv Desai, Dhagash Mehta, and Stefano Pasquali. 2023. [Company similarity using large language models](#). *arXiv preprint arXiv:2308.08031*.
- Arda Yüksel, Gabriel Thiem, Susanne Walter, Patrick Felka, Gabriela Alves Werb, and Ivan Habernal. 2026. [MONETA: Multimodal industry classification through geographic information with multi agent systems](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14790–14814.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.