

# An Interpretable Classifier based on Large scale Social Network Analysis\*

Monojit Banerjee  
IEEE

With  
In collaboration with Apart Research

## Abstract

Mechanistic model interpretability is essential to understand AI decision making, ensuring safety, aligning with human values, improving model reliability and facilitating research. By revealing internal processes, it promotes transparency, mitigates risks, and fosters trust, ultimately leading to more effective and ethical AI systems in critical areas. In this study, we have explored social network data from BlueSky and built an easy-to-train, interpretable, simple classifier using Sparse Autoencoders features. We have used these posts to build a financial classifier that is easy to understand. Finally, we have visually explained important characteristics.

### Keywords:

*Mechanistic Interpretability Track:* sparse autoencoders, feature visualization, model steering, interpretability tools, neural features, activation patterns, model internals, feature attribution

*Public Education Track Keywords:* AI safety education, learning platforms, interactive learning, curriculum design, educational games, knowledge assessment, mentorship systems, user engagement

*Social Sciences Track Keywords:* machine psychology, behavioral analysis, AI cognition, societal impact, human-AI interaction, institutional transformation, social norms, AI governance

## 1. Introduction

### 1.1. Problem Statement

This research aims to create a simple, interpretable classifier using Sparse Autoencoder (SAE) feature activations for financial sentiment analysis based on social media posts from BlueSky, a rapidly growing social network platform. The model seeks to develop activation-based classifiers that are easy to understand and explain. Additionally, we analyze how frequently users discuss finance—whether positively or negatively—which could aid in building a real-time classifier for integration into agentic workflows, thus facilitating fully automated and safe financial decision-making.

### 1.2. Background and Motivation

Some of the research proposes a supervised sparse coding model that incorporates decision tree classifiers to improve discrimination for nonlinear data (Huang et al., 2019). Similarly, another study highlighted an enhanced sparse autoencoder with Softmax regression to achieve effective

---

\*Research conducted at the Women in AI Safety Hackathon, 2025

feature learning and robust classification in medical diagnosis (Ebiaredoh-Mienye et al., 2020). Interpretable autoencoder utilizing decision trees that effectively manages categorical data without requiring transformation is also an active area of research(Aguilar et al., 2023). Some practical implementation such as increasing classification accuracy and reduce training/testing time for network intrusion detection systems using similar method was also proposed(Al-Qatf et al., 2018). As we can see SAEs are important for feature analysis and visual exploration with interpretable result have tremendous practical use.

### 1.3. Threat Model and Safety Implications

An interpretable classifier using social network data tackles AI safety challenges by enhancing bias detection, ensuring transparency in decision-making, improving robustness against manipulation, fostering user understanding, enabling real-time adaptability, and promoting ethical considerations, ultimately leading to more trustworthy and accountable AI systems in sensitive applications. Our approach combining social network data and visual understanding have addressed these challenges.

## 2. Methods

### 2.1. Approach

The high-level plan for the study is given below.

- Loading and pre-processing a dataset of Bluesky social media posts
- Performing sentiment analysis using a pre-trained model
- Extracting features via Goodfire API
- Training decision tree classifiers with grid search for optimal feature combinations
- Visualizing the best-performing decision tree

In the next section- we have explained the technical implementation.

### 2.2. Implementation

The research was conducted using Marimo notebook for quick visualization. We have used Goodfire ember SDK to create the interpretability pipeline. For variant, 'meta-llama/Meta-Llama-3.1-8B-Instruct' is used as this is relatively smaller model which can help with rapid experiments. For dataset, we have used 'alpindale/two-million-bluesky-posts'. This dataset comprises 2 million public posts collected from Bluesky Social's Firehose API, specifically designed for machine learning research and experimentation with social media data. However- since this data doesn't provide any sentiment analysis details, a fine-tuned model called '/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis' is used. Both the dataset and the sentiment analyzer model are available in hugging-face. The full code is available and attached here as a supplement. The code was executed in an Nvidia 4060 GPU with following library Driver Version: 550.120, CUDA Version: 12.4.

## 3. Results

### 3.1. Analysis and Findings

We have selected 3000 data points randomly from the datasets. After performing financial sentiment analysis steps, we have found that most of the posts are neutral as expected. This is described in Figure 1.

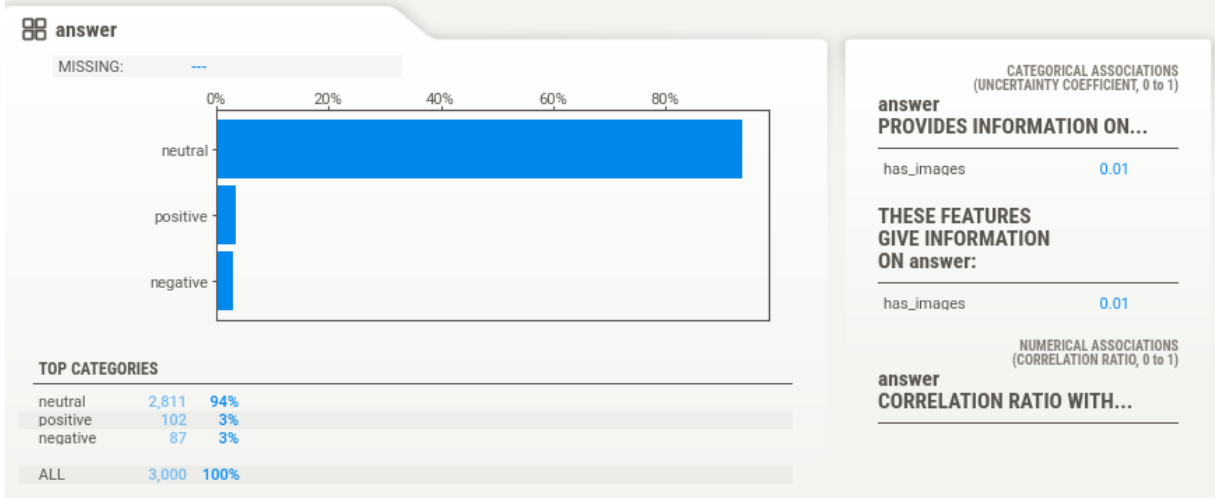


Figure 1: Financial Sentiment Analysis of BlueSky posts

Moreover, we have found following featuregroups are most prevalent and top five of them are FeatureGroup([ 0: "Mentions of Enron Corporation, especially during its controversial period", 1: "Hedging or approximation, particularly before numerical values or extreme statements", 2: "Positive moral or quality evaluation", 3: "Concentration of power and wealth by corporate/wealthy interests", 4: "Start of a new conversation or major topic shift", 5: "Conversation reset marker for context boundaries",) Also, the best combination of top 3 features are given below that our DecisionTreeClassifier is used.

- (Feature("Measurement and quantification tokens in demographic or statistical descriptions"),) 0.7184466019417476 DecisionTreeClassifier(max\_depth=1, min\_samples\_leaf=8, random\_state=42)
- (Feature("Hedging language when analyzing uncertain political outcomes and consequences"), Feature("Formal expressions of moral outrage and condemnation")) 0.7674418604651163 DecisionTreeClassifier(max\_depth=2, min\_samples\_leaf=8, random\_state=42)
- (Feature("Greetings and farewells marking temporal transitions"), Feature("Hedging language when analyzing uncertain political outcomes and consequences"), Feature("Formal expressions of moral outrage and condemnation")) 0.7954545454545454 DecisionTreeClassifier(max\_depth=3, min\_samples\_leaf=8, random\_state=42)

For the best performing classifier- following features are the most important. FeatureGroup([ 0: "Hedging and uncertainty in language patterns", 1: "Narrative flow and transitions in explanatory text", 2: "Past states of employment or utilization", 3: "Statements establishing which groups or entities are included or excluded from negotiations or discussions", 4: "Linguistic markers of hypothetical or unrealized scenarios in formal contexts", 5: "Formal institutional commitments and policy intentions", 6: "Introduction of new methodological frameworks or solutions to problems", 7: "Transitions and unrealized possibilities between states of being", 8: "Declining business conditions and negative economic trends", 9: "Financial analyst predictions and market forecasts" ])

Finally, we have conducted an exhaustive search to identify the optimal combination of SAE features for sentiment classification utilizing decision trees. It systematically evaluates various feature sets by training a regularized decision tree classifier on each combination and monitoring the performance to determine the best set based on the F1 score. We have visualized the max depth tree in Figure 2.

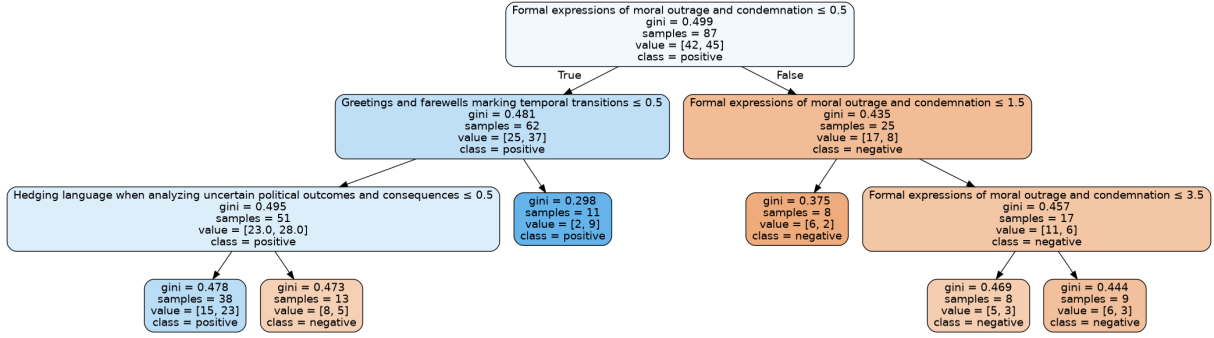


Figure 2: Representation of the best individual features

### 3.2. Impact Assessment

The interpretable classifier developed for financial sentiment analysis using Sparse Autoencoders (SAEs) effectively addresses several AI safety challenges. By enhancing transparency in decision-making, the solution promotes trustworthy AI systems in sensitive financial applications. The classifier’s ability to provide clear insights into its decision-making process through feature importance and decision tree visualizations fosters user understanding and accountability.

## 4. Discussion and Conclusion

In terms of AI safety, the solution demonstrates robustness against manipulation by leveraging interpretable features, making it harder for adversaries to exploit the model. The real-time adaptability of the classifier aligns with the goal of creating safe and responsive AI systems in dynamic financial environments. However, limitations exist. The reliance on social media data may not fully capture the complexity of financial sentiment, potentially introducing biases. Additionally, the model’s performance is constrained by the quality of the pre-trained sentiment analysis model used. Potential improvements include integrating more diverse data sources to enhance the model’s generalizability and refining the feature extraction process to improve accuracy and interpretability. Expanding the dataset beyond BlueSky posts and incorporating fairness-aware algorithms could further mitigate biases. Practically, this work contributes to safer automated financial decision-making systems, aligning with AI safety goals by promoting transparency, reducing biases, and enhancing user trust in AI-driven financial analysis.

## 5. Appendix

This project is heavily influenced by the Goodfire Decision-trees discussion. We thank the Goodfire team for creating the easy-to-follow tutorial and the SDK. We also extend our gratitude to the Lambda lab for offering credit and to the Apart research team for this opportunity.

## References

- Aguilar, D. L., Medina-Pérez, M. A., Loyola-González, O., Choo, K.-K. R., & Bucheli-Susarrey, E. (2023). Towards an interpretable autoencoder: A decision-tree-based autoencoder and its application in anomaly detection. *IEEE Transactions on Dependable and Secure Computing*, 20, 1048–1059. <https://api.semanticscholar.org/CorpusID:246571971>
- Al-Qatf, M., Yu, L., Al-habib, M., & Al-Sabahi, K. (2018). Deep learning approach combining sparse autoencoder with svm for network intrusion detection. *IEEE Access*, 6, 52843–52856. <https://api.semanticscholar.org/CorpusID:52964800>

- Ebiaredoh-Mienye, S. A., Esenogho, E., & Swart, T. G. (2020). Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis. *Electronics*. <https://api.semanticscholar.org/CorpusID:228986425>
- Huang, Y., Quan, Y., & Liu, T. (2019). Supervised sparse coding with decision forest. *IEEE Signal Processing Letters*, 26, 327–331. <https://api.semanticscholar.org/CorpusID:58672495>