Predicting Market Reaction to Biotech / Pharma News Using Machine Learning and Natural Language Processing

Safiya Zahra Alavi

December 2025

1 Abstract

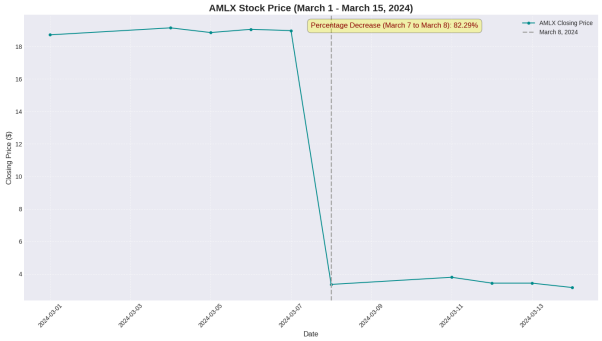
Biotechnology stock prices often exhibit sharp and sudden movements following clinical or regulatory announcements, yet predicting these reactions remains challenging because market responses frequently diverge from the apparent sentiment of the press release. This work develops a two-stage framework that integrates Named Entity Recognition using BioBERT, a domain-adapted NLP model, with an XGBoost classifier trained on financial indicators and NER-derived event features to predict five-day stock-movement classes. The BioBERT NER model achieved an entity-level F1 score of 0.22 with over 60% event-level label accuracy, and the final XGBoost classifier demonstrated strong precision in identifying large drops (0.67) and large increases (0.60), indicating its ability to reasonably detect high-impact events. Feature-importance analysis showed that financial indicators, along with negative regulatory and efficacy events, were the strongest predictors of market movement, demonstrating the value of combining event-specific NLP features with financial data to model biotech market reactions.

2 Introduction

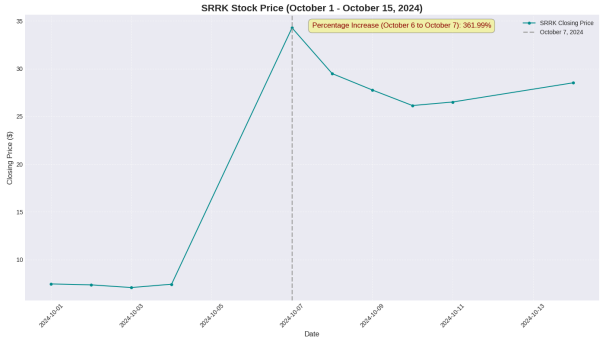
The pharmaceutical and biotechnology industry is characterized by a uniquely large risk reward profile relative to most other sectors. Despite extensive preclinical work designed to identify promising therapeutic targets for diseases with significant unmet needs, there remains considerable uncertainty whether the preclinical findings will translate into meaningful clinical benefit as the programs advance through Phase 1, Phase 2 and Phase 3 clinical trials (Mahalmani et al., 2022). Each stage of development introduces distinct risks, from biological translation and human safety to regulatory uncertainty and an evolving competitive landscape (Thomas et al., 2021).

For many companies, especially those that are smaller or pre-commercial with no revenue, prolonged operating losses are considered the norm (Cleary et al., 2021). The valuation is tied to the perceived strength of the company's clinical programs and the commercial potential of the markets they aim to address. As a result, the outcome of each clinical trial can lead to substantial fluctuations in company value. Market reactions are influenced not only by efficacy and safety results but also by the competitive differentiation and alignment with expectations. In indications with unmet needs or high value markets, companies may command significant value despite having no FDA approved product. However, if negative results are released, this perceived value can drop significantly, as seen with market reaction to the results of Amylyx Pharmaceuticals' Phase 3 trial evaluating the therapy RELYVRIO for people living with amyotrophic lateral sclerosis (Figure 1a). On the other hand, when companies that are significantly undervalued release positive data, the market may react strongly with many investors wanting to come in after additional derisking, driving extreme increases in the share price (Figure 1b). Importantly, even ostensibly positive data may trigger a negative reaction from the market if they fall short of expectations or raise new concerns (Cho et al., 2024).

Given this volatility, it is often challenging to interpret the direction and magnitude of the market's reaction to publicly released clinical trial and company updates. To address this, the goal of this work is to develop a predictive framework that estimates expected market reaction based on the content in press releases related to business activities and clinical-trial-related announcements.



(a) The AMLX stock chart between March 1, 2024 and March 13, 2024, following the announcement of the Phase 3 clinical trial results of RELYVRIO evaluated in ALS patients.



(b) The SRRK stock chart between October 1, 2024 and October 13, 2024, following the announcement of the Phase 3 clinical trial results of apitegromab evaluated in spinal muscular atrophy patients

Figure 1: AMLX and SRRK Stock Charts

The proposed framework consists of a sequential two-stage modeling approach that takes the input of free text from a company's press release relating to updates in the clinical program and outputs the predicted magnitude and direction of associated stock price movement. The novelty of this approach stems from the type of integration of natural language processing in the machine learning pipeline. In the first stage, a Bidirectional encoder representations from transformers (BERT) model is employed to extract and classify the information in the press release into predefined categories related to clinical activities and the event sentiment. These features are combined with a number of quantitative financial indicators and company specific fundamentals to predict the degree of stock movement following the announcement utilizing a classification model.

3 Background

Substantial work in financial modeling has focused on predicting stock market behavior, yet the biopharmaceutical space presents a unique challenge due to the volatility and sensitivity to clinical outcomes and regulatory events. This results in traditional predictive methodologies requiring supplementation to meaningfully capture the nuances of this domain. Budennyy et al. (2023) proposed a machine learning framework using a three part model system to classify the magnitude of stock movement following a clinical-trial-related announcement. The system consisted of a BERT based sentiment classifier to evaluate the tone of press releases, a feature extraction model to select the top most predictive features and a long short-term memory (LSTM) neural network for market reaction prediction. The authors reported strong results in predicting extreme movements, particularly for the Extremely Negative, Moderately Negative and Extremely Positive classes of market movement.

Aparicio et al. (2024) pursued a related but distinct approach of deriving sentiment from biopharma press releases to inform a trading strategy. Their approach emphasized fine tuning the publicly available BioBERT model, trained on biomedical related corpora, and the FinBERT model, trained on financial communication text, to produce a hybrid BioFinBERT model that can effectively predict the sentiment of a biopharma press release. Interestingly, the authors found little difference between the models in regard to the quality of the trading strategy produced, and in some scenarios found that their finetuned BioFinBERT model underperformed in relative to its component models.

The approach developed in the present work is differentiated from these two prior studies in that, although the project also utilizes a transformer based language model, the main NLP task is Named Entity Recognition (NER) rather than sentiment classification. This enables more granular and specific categorization of the clinical events, regulatory milestones, or safety signals that are described in the press releases. The NER derived entities span a broader set of clinically meaningful categories which collectively capture the mechanistic drivers of market reaction rather than relying solely on overall sentiment.

4 Methods

4.1 Data Collection

The dataset used to train and evaluate the proposed pipeline was compiled using information curated from BioPharmCatalyst (www.biopharmcatalyst.com). For each press release published in 2024 by micro-, small-, mid- and large-cap biopharmaceutical companies, the company ticker, associated drug, development stage, indication, and the free-text description of the clinical or regulatory event is collected. The events listed include Phase 1 through 3 trial updates, safety findings, regulatory discussions or designations, meeting outcomes, and other important disclosures related to the clinical development. In support of the downstream predictive modeling of the associated return, the dataset was supplemented with financial metrics obtained from S&P Capital IQ Pro. These included the shares outstanding, market capitalization, daily trading volume, short interest rate, beta, and price data for both the company's stock and the XBI index, a modified equal weighted index that provides exposure to small-, mid-, and large-cap biotechnology stocks.

4.2 Large Language Models (LLMs) Evaluated

The first stage of the proposed machine learning pipeline involves the Named Entity Recognition (NER) module designed to identify and categorize the key phrases within clinical-trial-related press releases. A BERT-based architecture was selected for this task due to its bidirectional contextual encoding that has been shown to outperform other recurrent architectures, such as LSTMs and earlier transformer based models, on various NLP tasks including NER (Devlin et al., 2018). Additionally, several BERT derivatives have been pretrained and finetuned on biomedical related corpora, making these models a suitable fit for the NER task relevant to this research problem (Li et al., 2022). To evaluate the impact of the pretrained domain adaptations on the NER performance, this analysis evaluated three additional models: FinBERT, which incorporates financial language patterns and has shown strong performance on tasks related to market relevant text (Huang et al., 2023); BioBERT, a biomedical domain-specific model pretrained on PubMed abstracts and PMC articles; and PubMedBERT, which is also a biomedical specific model but trained specifically on PubMed text, offering strong performance on scientific abstract interpretation.

4.3 Named Entity Recognition Analysis

In the NER task developed for this work, spans were classified into categories relevant to clinical and regulatory events: positive or negative clinical efficacy outcomes, safety findings, regulatory actions, trial enrollment updates, clinical trial operations, and conference related disclosures. Training and evaluating in an NER task requires a dataset annotated using a Beginning, Inside, and Outside (BIO) tagging schema. The initial iteration of the BIO-tagged data was generated using regular expression (regex) based pattern matching. Certain phrases were identified using a comprehensive list of predefined words in particular orders and matched to their corresponding category. However, regex alone had limited coverage and struggled to capture the variability and complexity of the biopharmaceutical language and the magnitude of effects. To improve the annotation quality, OpenAI's GPT-4.1 was incorporated into the tagging pipeline. A structured prompt was written to extract the beginning and end of key phrases and the corresponding category the phrase belonged to. The implementation of GPT-4.1 significantly increased the number of correctly captured entities and reduced the need for manual tagging. This dual-method approach produced a substantially richer

supervised dataset, yielding nearly 10,000 tags and enabling more effective learning during model training (Figure 2).

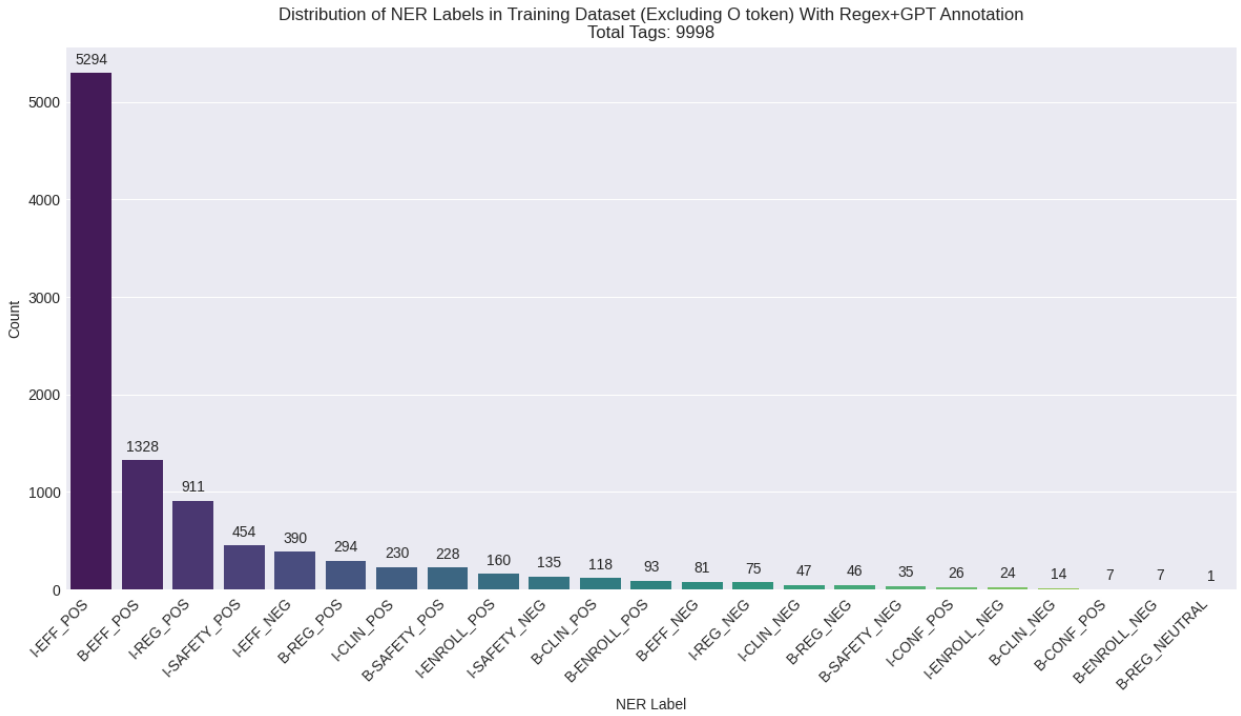


Figure 2: The distribution of the different tags in the training dataset after the implementation of the regex and GPT annotation methodology for the BIO tagging.

To evaluate the quality of entity detection and classification, the entity-level (span-level) F1 score was used. Unlike token level accuracy, which can be artificially high due to a high proportion of other ("O") tokens or partially recognized entities, the entity level F1 score directly measures whether the model correctly identifies both the category and the exact start and end boundaries of each span. This metric is critical in the biomedical and regulatory contexts, where multi-token phrases such as *"did not meet the primary endpoint"*, *"received a Complete Response Letter"*, or *"statistically significant improvement versus placebo"* must be captured as coherent units to preserve their relevance for downstream modeling. Misidentification of boundaries can distort the inferred sentiment, ultimately degrading feature reliability in predicting the market reaction. The baseline model for the NER task was the original BERT-Base model introduced by Devlin et al. (2018) published in the Hugging Face library. All subsequent domain-adapted models were evaluated relative to this baseline.

For the downstream predictive modeling task, the primary features derived from the NER system are the counts of each event category identified within a press release. The underlying assumption is that the frequency of positive or negative event-related phrases, such as efficacy signals, safety findings, regulatory updates, or operational milestones, serves as a proxy for the overall direction and intensity of the news. For each catalyst in the dataset, a global net-effect score is computed by summing the total number of positive classified spans and subtracting the total number of negatively classified spans. This score is calculated separately as well for major event types, including the clinical efficacy, regulatory actions, safety, trial operations and conference communications. Catalysts were assigned a coarse sentiment class utilizing the global net-effect score (positive for total counts over zero, negative for less than zero, or neutral for zero). These event-specific net-effect metrics and the global net-effect metric constitute the key structured features utilized in the final modeling dataset. The NER-derived features essentially transforms unstructured press

releases into quantitative indicators of news polarity and content, which enables the downstream classification model to learn relationships between the nature of the disclosed events and the magnitude of subsequent stock price movements.

4.4 Predictive Modeling Approach

For the predictive modeling component, an XGBoost classification model was developed to estimate the magnitude of stock-price movement following each catalyst event. The feature set included: the key financial indicators (market capitalization, volatility, trading volume, beta, short interest), the clinical stage associated with the event, and the net-effect scores identified from the NER system. The target variable was defined as the five day return following the press release, discretized into five ordinal classes that represent the degrees of market reaction: LARGE DROP (-35% or more), DROP (-5% to -35%), NEUTRAL (5% to -5%), INCREASE (between 5% to 35%), LARGE INCREASE (35% or more).

The XGBoost classifier was trained using a learning rate of 0.01, a tree depth of 6, and the number of estimators were 500, parameters selected through hyperparameter tuning. Feature importance was conducted to analyze both the most influential predictive features and interactions among clinical, regulatory, and financial variables.

5 Results

5.1 Performance of BERT Models on the NER Task

Multiple BERT models were evaluated using the entity-level F1 score, an appropriate metric for span-based NER. Performance varied across the label types; however, BioBERT achieved the highest overall F1 score, closely followed by the BERT-Base model. This similarity is likely attributed to the distribution of the data: the most common categories (EFF_POS, REG_POS, and SAFETY_POS) were sufficiently represented which enabled BERT-Base to learn their nuances without requiring domain-specific knowledge (Table 1).

Unexpectedly, FinBERT outperformed PubMedBERT in the overall score, which is driven by stronger results on regulatory categories given that the positive and negative labels in this category represent a large portion of the data. PubMedBERT under performance compared to BioBERT is likely attributable to the differences in the pretraining corpora, with BioBERT having a more diverse set than PubMedBERT.

Table 1: The F1 scores across the various NER labels utilizing different BERT models.

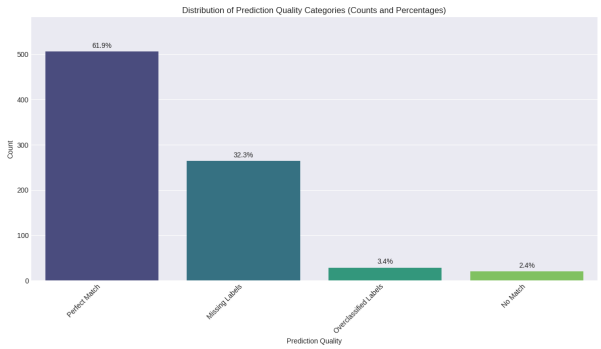
	Baseline (All classified as "0")	BERT-Base	Modern BERT	BioBERT	PubMedBERT	FinBERT
CLIN_NEG	0	0	0	0	0	0
CLIN_POS	0	0.3125	0.2857	0.3509	0.3448	0.3175
CONF_POS	0	0	0	0	0	0
EFF_NEG	0	0.2	0.1900	0.1667	0.1790	0.1550
EFF_POS	0	0.1642	0.1292	0.1615	0.1132	0.1113
ENROLL_NEG	0	0	0	0	0	0
ENROLL_POS	0	0.3684	0.3158	0.5625	0.2857	0.2710
REG_NEG	0	0.3571	0.3871	0.2441	0.2727	0.3704
REG_POS	0	0.4804	0.3706	0.4884	0.1633	0.4667
SAFETY_NEG	0	0	0	0	0	0
SAFETY_POS	0	0.3543	0.3155	0.3945	0.3552	0.3830
OVERALL	0	0.2300	0.1882	0.2383	0.1570	0.1961

5.2 Span-Level Evaluation

While entity recognition was the primary task, evaluating classification consistency at the event level was essential. The entity-level F1 score was critical for assessing whether the model correctly identified relevant spans and assigned them to the appropriate category. Nevertheless, for the downstream predictive modeling, the key information used is the count of each type of label. In the held-out testing set, approximately 62% of the samples achieved a perfect match of the labels, meaning the spans were correctly classified. About 32% of the samples were missing one or more labels, suggesting under identification or misidentification of some multi token spans. Only about 3% exhibited over classification and 2% had no matches at all. A key limitation is that the initial annotations were generated using GPT prompting rather than manually curated labels, thus improvements to the BIO tagging procedure, particularly including hybrid human annotation, may yield gains in the NER performance. (Figure 3)

Predicted Labels	True Labels	Catalyst	Prediction_Quality
[]	[CLIN_NEG, SAFETY_NEG]	Phase 2 trial voluntarily halted all dosing du...	Missing Labels
[EFF_POS]	[SAFETY_POS]	Phase 2 trial met primary safety endpoint, not...	No Match
[]	[CLIN_NEG]	Development of ulliedlimab is being paused, no...	Missing Labels
[ENROLL_POS]	[ENROLL_POS]	Phase 2b enrollment completed, noted January 6...	Perfect Match
[REG_POS]	[REG_POS]	Orphan Drug Designation granted by the FDA, no...	Perfect Match

(a) Example of the prediction quality for the predicted labels from the NER model versus the true labels on rows in the testing dataset. Some rows were completely missing the correct label and some rows had the perfect match.



(b) Distribution of the quality of the classifications on the testing set.

Figure 3

5.3 Exploratory Analysis of Returns by Sentiment Class

To characterize the general behavior of the percent returns, the XBI-adjusted day-5 returns (calculated as the company's 5-day return subtracted by the XBI's return over the same period) were stratified by the catalyst's overall sentiment, as determined by the net-effect scores (positive for total counts over zero, negative for less than zero, or neutral for zero). The positive and neutral catalysts displayed distributions centered near zero with positive skewness, driven by rare but large upside outliers. In contrast, the negative catalysts had an average return of -14% with a noticeable left skew, indicating that negative disclosures are punished more severely than positive catalysts are rewarded. This asymmetry is intuitive from a clinical development and commercial standpoint: negative results often introduce or confirm seemingly irreversible program risk leaving limited routes for recovery; whereas positive catalysts generally advance a program forward but do not remove all uncertainties regarding ultimate approval, competition, or commercial feasibility (Figure 4).

5.4 Predictive Modeling Performance

The XGBoost classifier was trained using the financial indicators, clinical stage variables, and NER derived features. Feature importance analysis highlighted that while the financial features (volatility, trading volume, market cap) were among the most influential, several NER-based features also shaped the predictive performance. In particular, REG_NEG, EFF_NEG, total_net, and regulatory_net were ranked prominently, indicating that negative regulatory signals and clinical efficacy markers materially affect the market movement predictions. Intuitively, this is consistent with the expectation that a program failing or an adverse regulatory outcome can result in a biotechnology company significantly losing value (Figure 5). After fine

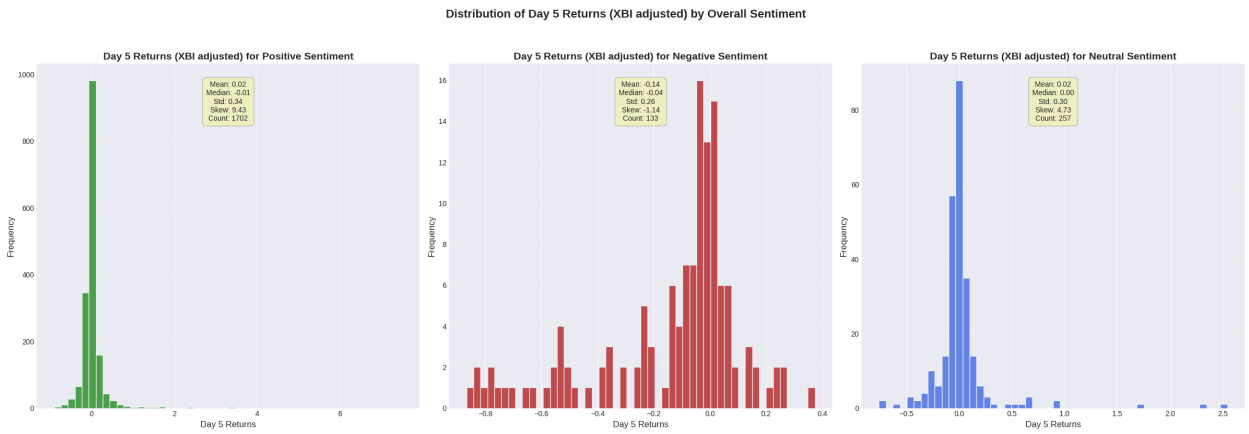


Figure 4: Distribution of the 5-day XBI-adjusted returns stratified by the catalyst sentiment, which was determined by the calculated net_effect from the NER labels.

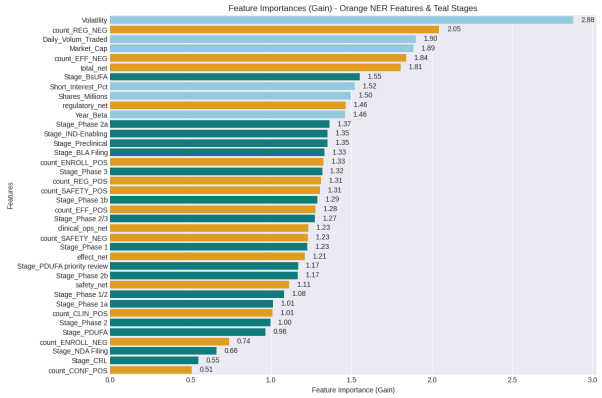
tuning, the final model achieved an accuracy of 50.12% across the five stock movement classes. For comparison, a baseline classifier predicting only the majority NEUTRAL class would achieve 41.68% accuracy, indicating that the features are providing approximately 9% improvement over baseline. From a trading and investment perspective, precision is a critical metric because false positives can be costly. For example, if the model incorrectly predicts a large drop and a position is exited prematurely, the investor forfeits potential gains; conversely, an incorrect prediction of a large increase may lead to buying additional shares that fail to appreciate, resulting in inefficient capital allocation. Notably, the model achieved its highest precision scores in the LARGE DROP class (0.67) and LARGE INCREASE class (0.60). However, sensitivity in the INCREASE class was low due to frequent misclassifications of the events as neutral (Table 2).

Table 2: Precision, recall, and F1 scores on the 5 classes of market reaction.

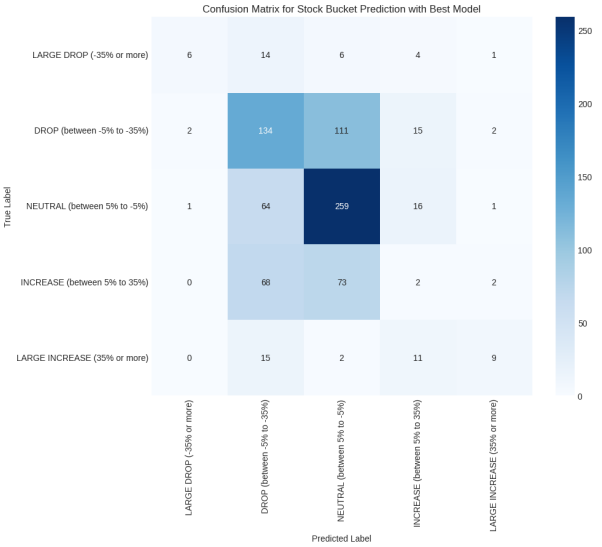
	Precision	Recall	F1-score	Support
LARGE DROP (-35% or more)	0.67	0.19	0.30	31
DROP (between -5% to -35%)	0.45	0.51	0.48	264
NEUTRAL (between 5% to -5%)	0.57	0.76	0.65	341
INCREASE (between 5% to 35%)	0.04	0.01	0.02	145
LARGE INCREASE (35% or more)	0.60	0.24	0.35	37
Accuracy			0.50	818
Macro avg	0.47	0.34	0.36	818
Weighted avg	0.45	0.50	0.46	818

6 Conclusion

The work presents an end-to-end pipeline focused on predicting market reaction to biopharmaceutical press releases, which can often vary based on terminology and verbal cues mentioned in these announcements. A BioBERT NER system performed with an F1 score of 0.22, with 62% of the events in the catalysts tagged correctly, which was sufficient for downstream feature engineering. The NER outputs were aggregated into event-specific net-effect scores and combined with the financial indicators to train an XGBoost classifier, which performed with an accuracy of 50% on a five-class stock-movement task. Notably, the precision was maximized the the LARGE INCREASE and LARGE DROP classes, indicating that the model performs sufficiently well for investor interest. Key areas for improvement include enhancing the BIO tag-



(a) Feature importance from the XGBoost Classifier final model.



(b) Confusion Matrix of the classifications from XGBoost Classifier final model.

Figure 5

ging methodology, potentially through expanded GPT-assisted annotation or human guided correction, and incorporating additional financial or market structure features.

References

Aparicio, V., Gordon, D., Huayamares, S. G., & Luo, Y. (2024). *Biofinbert: Finetuning large language models (llms) to analyze sentiment of press releases and financial text around inflection points of biotech stocks* [Unpublished manuscript], Quantitative & Computational Finance Program, Georgia Institute of Technology, GA, 30332, USA.

Budennyy, S., Kazakov, A., Kovtun, E., & Shalyto, A. (2023). New drugs and stock market: A machine learning framework for predicting pharma market reaction to clinical trial announcements. *Scientific Reports*, *13*, 12817. https://doi.org/10.1038/s41598-023-39301-4

Cho, J., Singh, M., & Lo, A. W. (2024). How does news affect biopharma stock prices?: An event study. *PLOS ONE*, *19*(1), e0296927. https://doi.org/10.1371/journal.pone.0296927

Cleary, E. G., Jackson, M. J., Folch, B., Bell, J., Jackson, M. L., Avorn, J., & Kesselheim, A. S. (2021). Comparing long-term value creation after biotech and non-biotech ipos, 1997–2016. *PLOS ONE*, *16*(1), e0243813. https://doi.org/10.1371/journal.pone.0243813

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Huang, A. H., Wang, H., & Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/https://doi.org/10.1111/1911-3846.12832

Li, J., Wei, Q., Ghiasvand, O., Weng, C., & Zhang, Y. (2022). A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Medical Informatics and Decision Making*, *22*(Suppl 3), 235. https://doi.org/10.1186/s12911-022-01967-7

Mahalmani, V., Kapoor, R., Jain, S., Mahapatra, A., Arul, J., Kishor, K., Thakur, P., & Kumar, S. (2022). Translational research: Bridging the gap between preclinical and clinical research. *Indian Journal of Pharmacology*, *54*(6), 393–396. https://doi.org/10.4103/ijp.ijp_860_22

Thomas, J., Fink, L., Shah, J., & Moustakas, V. (2021). *Clinical development success rates and contributing factors 2011–2020* (tech. rep.). Biotechnology Innovation Organization (BIO).