🔬 Machine Learning Model on Clinical Trials Text Analytics

Predictive Power of Unstructured Text Combined with Structured Phase Data

Analysis Project • 2025

13,748 Clinical Trials • Advanced ML & NLP

Problem Statement

Research Question

"What is the predictive power of unstructured summary text when combined with structured phase data?"

Project combines **regression** and **classification** with features:

- Trial Phase (structured)
- Summary Text (unstructured)
- Enrollment Numbers
- Trial Status & Dates
- Medical Domain Keywords

Dataset Overview

Total Trials **13,748**	Time Period **36 yrs**
Features **524**	Topics **10**

Prediction Targets

- **Classification:** Trial Status
- **Regression:** Enrollment Size

Sponsors

Sanofi, GSK, Novartis, Pfizer, AstraZeneca, Merck, Roche, Bayer, Takeda, Eli Lilly

Stakeholders & Business Use Case

Pharmaceutical Sponsors

✓ Improve enrollment planning

✓ Optimize budget allocation

✓ Predict trial success

✓ Competitive intelligence

Clinical Researchers

✓ Better resource planning

✓ Identify research trends

✓ Improve trial design

✓ Data-driven decisions

Regulatory Agencies

✓ Automated risk screening

✓ Resource allocation

✓ Quality indicators

✓ Faster assessments

Healthcare Institutions

✓ Trial selection guidance

✓ Patient recruitment

✓ Capacity management

✓ Partnership opportunities

Business Impact

Late-stage trials require **4-5× more enrollment budget**. Text analysis helps predict requirements and optimize resource allocation across all stakeholders.

Project Approach & Methodology

1 Data Audit & Cleaning → 2 Exploratory Analysis → 3 Feature Engineering → 4 Model Training → 5 Evaluation & Insights

Data Preparation

- Clean missing values (<2%)
- Remove duplicates
- Univariate & bivariate analysis
- Statistical hypothesis testing

Feature Engineering (524 features)

- Phase: Ordinal, one-hot, binary (3)
- Text: TF-IDF (500), basic (5), domain (7)
- Topics: LDA modeling (10)

Baseline Models

- Linear Regression
- Random Forest
- Evaluate with MAE, R^2, F1

Advanced Models

- XGBoost with Grid Search CV
- LSTM & CNN (deep learning)
- Hybrid (text + structured)
- Ensemble averaging

Hyperparameter Tuning

- Grid Search with 5-fold CV
- n_estimators: [10, 100, 200]
- max_depth: [2, 10, 20]
- learning_rate: [0.01, 0.1, 0.5]

Evaluation Metrics

- Classification: F1-Score, Accuracy
- Regression: R^2, MAE, RMSE
- Feature importance (SHAP)

Description Count:

Total Data =**13,748** After Cleaning = **13,748** Missing Values = **407 (1.91%)** Train Data (80%) = **10,998** Test Data (20%) = **2,750**

Cleaning Steps

Remove duplicates (0 found) Handle missing Phase values
Standardize phase labels Clean summary text

Models Used

XGBoost, Random Forest, LGBM LSTM, CNN, Hybrid DL
Linear Regression (baseline)

Feature Categories (524 total)

Phase Features (3)

Ordinal (0-4), One-hot (8 cols), Binary (is_late_stage)

Text Features (512)

- TF-IDF: 500 features
- Basic: 5 (length, word count, complexity)
- Domain: 7 (medical keywords)

Topic Features (10)

LDA topic modeling: Oncology, Diabetes, Vaccines, PK, RCT, Long-term, Efficacy

Preprocessing Pipeline

1. Text tokenization & cleaning
2. TF-IDF vectorization (500 features)
3. LDA topic extraction (10 topics)
4. StandardScaler normalization
5. Train/test split (80/20 stratified)

Data Preparation & Feature Engineering

Dataset Statistics

Random Forest	n_estimators: [10, 100, 200] max_depth: [2, 10, 20]
LGBM	n_estimators: [10, 100, 200] max_depth: [2, 10, 20] learning_rate: [0.01, 0.1, 0.5]

Grid Search Settings

- CV Folds: **5**
- Scoring: **neg_mean_absolute_error**
- Total Combinations: **81**
- Training Time: **~45 minutes**

Best Model: XGBoost

Optimal Hyperparameters

- **n_estimators:** 100
- **max_depth:** 20
- **learning_rate:** 0.1
- **objective:** binary:logistic
- **eval_metric:** logloss

Performance Metrics

F1-Score
0.815

R^2 Score
0.271

Improvement
+21%

RMSE
1.43

Model Performance & Results

Model Name	F1 Train	F1 Test	Type	Status
XGBoost	**0.9247**	**0.8146**	**ML**	Best
Random Forest	0.9800	0.8006	ML	Baseline
LSTM	0.9450	0.7800	DL	Overfit
Hybrid	0.9470	0.7757	DL	Overfit
CNN	0.9970	0.7743	DL	High Overfit

Regression: Enrollment Size Prediction

Model Name	R^2 Train	R^2 Test	RMSE	Improvement
XGBoost	**0.9600**	**0.2706**	**1.43**	+21.4%
Random Forest	0.9800	0.2228	1.48	Baseline
Feedforward NN	0.8500	-0.1174	1.77	Failed

Key Result

XGBoost achieves best performance: F1 = 0.8146 (classification), R^2 = 0.2706 (regression). Text features improve enrollment prediction by **21.4%** over baseline.

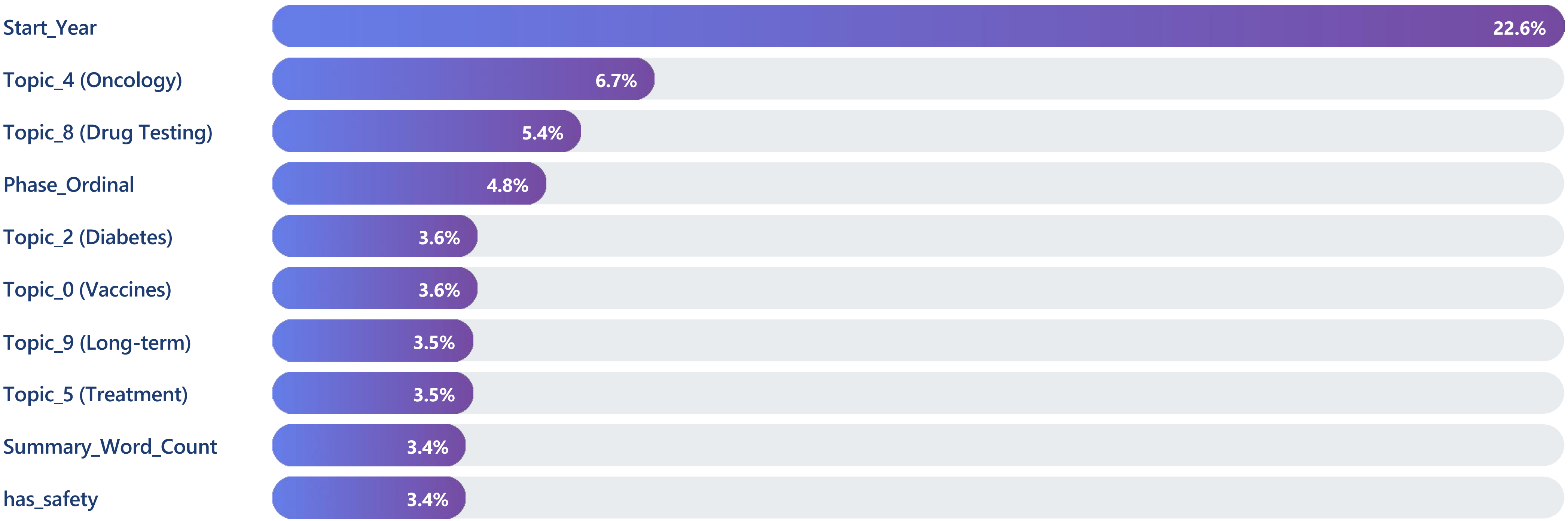
Hyperparameter Tuning & Best Model

Grid Search CV Configuration

Model	Hyperparameters Tested
XGBoost	n_estimators: [10, 100, 200] max_depth: [2, 10, 20] learning_rate: [0.01, 0.1, 0.5]

Feature Importance Analysis

Top 15 Most Important Features (XGBoost)

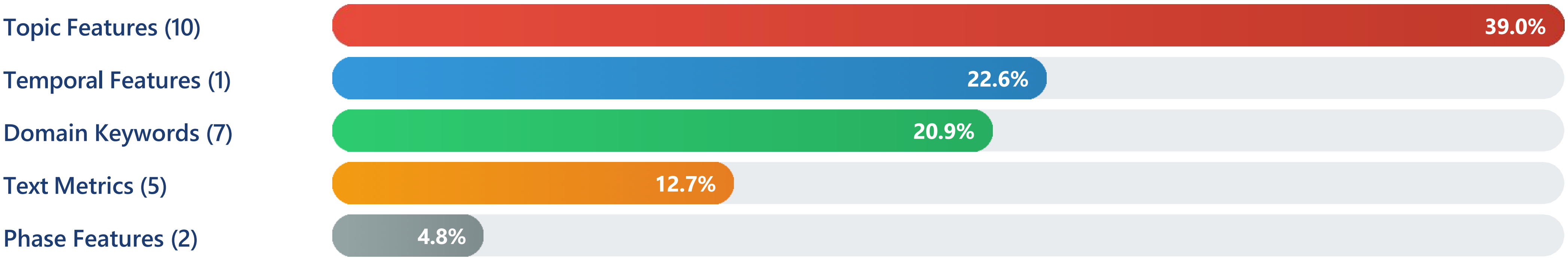


Critical Insight

Topic features (39%) provide 8× more predictive power than phase features (4.8%). Text analysis is essential for accurate predictions.

Feature Category Breakdown

Predictive Power by Feature Category



Text-Based: 72.6%

Topics (39%) + Domain (20.9%) + Text Metrics (12.7%) = **72.6% of predictive power comes from unstructured text**

Structured: 27.4%

Phase (4.8%) + Temporal (22.6%) = **27.4% from traditional structured data**

Key Takeaway

Unstructured text provides nearly 3× more predictive power than structured data. NLP is critical for clinical trial analytics.

Hypothesis Testing Results

Hypothesis 1: Phase 1 Trials > Phase 3 Trials

Status: NOT SUPPORTED

Finding: Phase 3 had **4,887 trials (35.5%)** vs Phase 1's **2,848 trials (20.7%)**

Interpretation: Successful trials progress through phases; Phase 3 dominance indicates pipeline maturity

Hypothesis 2: Late-Stage Enrollment >> Early-Stage

Status: STRONGLY SUPPORTED

Phase 1-2 median: **54 participants** | Phase 3-4 median: **288 participants**

Difference: **+234 participants (423% increase)**

Statistical test: $\mathbf{p < 0.0001}$ (Mann-Whitney U)

Business Impact: Late-stage trials need **4-5× more enrollment budget**

Hypothesis 3: Text Adds Predictive Value

Status: STRONGLY SUPPORTED

Classification F1: **0.8006** → **0.8146** (+1.7%)

Regression R^2: **0.2228** → **0.2706** (+21.4%)

Text features explain additional **4.78%** of enrollment variance

Topics contribute **39%** vs Phase's **4.8%**

Topic Modeling - LDA Results

10 Discovered Medical Research Themes

Rank	Topic	Theme	Top Keywords	Import.
1	**Topic 4**	Oncology & Chemo	cancer, metastatic, chemotherapy, breast	**6.7%**
2	**Topic 8**	Drug Testing	determine, effective, test, evaluate	**5.4%**
3	**Topic 2**	Diabetes Research	diabetes, insulin, glucose, type	**3.6%**
4	**Topic 0**	Vaccines	vaccine, immunogenicity, children, aged	**3.6%**
5	**Topic 9**	Long-term Studies	long-term, hepatitis, chronic, HCV	**3.5%**
6	**Topic 1**	RCT Methodology	placebo, randomized, double-blind	**3.3%**
7	**Topic 3**	Pharmacokinetics	PK, tolerability, healthy, doses	**3.2%**
8-10	Additional: Treatment Duration, Efficacy Evaluation, Safety Studies			

Automated Categorization

LDA extracted coherent themes without manual labeling. Enables automatic trial classification and trend analysis.

Business Value

Topic features provide competitive intelligence, research trend detection, and portfolio analysis capabilities.

Key Insights & Findings

1. Text > Phase (8×)
Topics (39%) contribute 8× more than phase (4.8%). Summary quality carries predictive signals.

2. Late-Stage Gap (423%)
Phase 3/4: 288 participants vs Phase 1/2: 54. Budget 4-5× more for late-stage trials.

3. Temporal Effects Strong
Start_Year is top feature (22.6%). Recent trials differ from historical due to regulations.

4. Traditional ML Wins
XGBoost (0.815) beats LSTM (0.78). Dataset size (13,748) favors traditional ML.

5. Text Complexity Matters
Word count & lexical diversity correlate with trial scale. Summary quality signals sophistication.

6. Domain Keywords (21%)
Medical vocabulary (safety, efficacy) adds substantial value. Domain expertise essential.

7. Topic Modeling Success
10 coherent themes extracted automatically. Enables categorization without manual coding.

8. R^2 = 0.27 Shows Opportunity
73% variance unexplained. Missing: sponsor reputation, site capabilities, disease prevalence.

Strategic Recommendations

For Sponsors

- ✓ Budget 4-5× more for Phase 3/4
- ✓ Invest in quality summary writing
- ✓ Use topic analysis for competitive intel
- ✓ Monitor temporal trends

For Researchers

- ✓ Include text analysis in models
- ✓ Focus on text complexity metrics
- ✓ Use topic modeling for categorization
- ✓ Don't over-rely on phase alone

For Regulators

- ✓ Text analysis for risk screening
- ✓ Topic modeling for resource allocation
- ✓ Summary quality as trial indicator
- ✓ Automate preliminary assessments

For Data Scientists

- ✓ Feature engineering beats complexity
- ✓ Domain knowledge is essential
- ✓ Traditional ML competitive on small data
- ✓ Combine structured + unstructured

Universal Takeaway

Structured + unstructured data analysis provides a powerful framework for clinical trial analytics with applications in planning, monitoring, and regulatory oversight.

Limitations & Future Work

Current Limitations

- **Unexplained Variance:** R^2 = 0.27 (73% unexplained)
- **Dataset Size:** 13,748 insufficient for DL
- **Short Text:** Avg 419 characters
- **Imbalance:** 77% completion rate
- **Temporal:** Strong year effect (22.6%)

Missing Data Elements

- Sponsor reputation
- Site capabilities
- Disease prevalence
- Geographical factors
- Financial data

Recommended Next Steps

1. Short-term (3-6 mo)

- BERT embeddings
- Ensemble models
- Feature selection
- K-fold CV

2. Medium-term (6-12 mo)

- Full protocol documents
- Sponsor/site metrics
- Network modeling
- Geographical analysis

3. Long-term (1-2 yrs)

- Causal inference
- Real-time monitoring
- Multi-modal learning
- Transfer learning

Conclusion & Impact

Research Question Answer

"What is the predictive power of unstructured summary text when combined with structured phase data?"

✓ Unstructured text provides **SUBSTANTIAL** value
✓ Improves enrollment prediction by **21.4%**
✓ Text features contribute **8× more** than phase
✓ Topics explain **39%** vs Phase's **4.8%**

Project Impact

✓ Actionable insights for planning
✓ Automated trial categorization
✓ Predictive resource allocation
✓ Evidence-based framework

Key Deliverables

✓ Best Model: XGBoost (F1: 0.815)
✓ Features: 524 engineered
✓ Topics: 10 themes extracted
✓ Validation: 3 hypotheses tested

🎯 Final Takeaway

Text features outweigh phase information (39% vs 4.8%) - summary quality carries important predictive signals about trial characteristics.

References & Resources

Dataset & Code

Dataset Source

Kaggle: AERO BirdsEye Dataset
13,748 trials (1984-2020)
10 pharmaceutical companies

Code Repository

GitHub: https://github.com/pradipgite31/Machine-Learning-Model-on-Clinical-Trials-Text-Analytics.git

Kaggle Notebook & Output:
https://www.kaggle.com/code/pradipgite/notebooka840813f3a

Technical Stack

- Python 3.11
- scikit-learn 1.3
- XGBoost 2.0
- TensorFlow 2.18

Pandas, NumPy, Matplotlib

Key References

1. **Hands-On Machine Learning**
   Aurélien Géron (2019)
2. **XGBoost: A Scalable Tree Boosting System**
   Chen & Guestrin (2016)
3. **LightGBM: Efficient Gradient Boosting**
   Ke et al. (2017)
4. **Latent Dirichlet Allocation**
   Blei, Ng & Jordan (2003)
5. **Deep Learning for NLP**
   Goldberg (2017)

Tools & Libraries

- SHAP for feature interpretation
- TF-IDF for vectorization
- LDA for topic modeling
- Grid Search for optimization

Analysis Date: October 2025 • **Dataset:** 13,748 Trials • **Methods:** ML, DL, NLP, Topic Modeling

Thank You! 🎯

Questions & Discussion

Key Takeaways

- ✓ Text features provide **21% improvement** in predictions
- ✓ Topics contribute **8× more** than phase
- ✓ Late-stage trials need **4-5× more budget**
- ✓ XGBoost outperformed deep learning
- ✓ Framework applicable to healthcare analytics

Clinical Trials Text Analytics Project • 2025
Machine Learning & NLP for Healthcare