# Company Classification: A Machine Learning Approach

This presentation outlines our project to develop a system that automatically classifies companies into relevant insurance taxonomy labels. Using techniques in natural language processing and machine learning, we created a scalable solution that analyzes company descriptions, business tags, and sector information to suggest appropriate insurance categories.

Our approach combines traditional text analysis methods with modern data science practices, offering insights into both the technical implementation and practical applications for the insurance industry. Through this project, we've developed a robust classification system while identifying clear paths for future improvements.

# Project Objectives & Business Context

### Primary Goal

Develop an automated system to classify companies into relevant insurance taxonomy labels using company descriptions and business metadata.

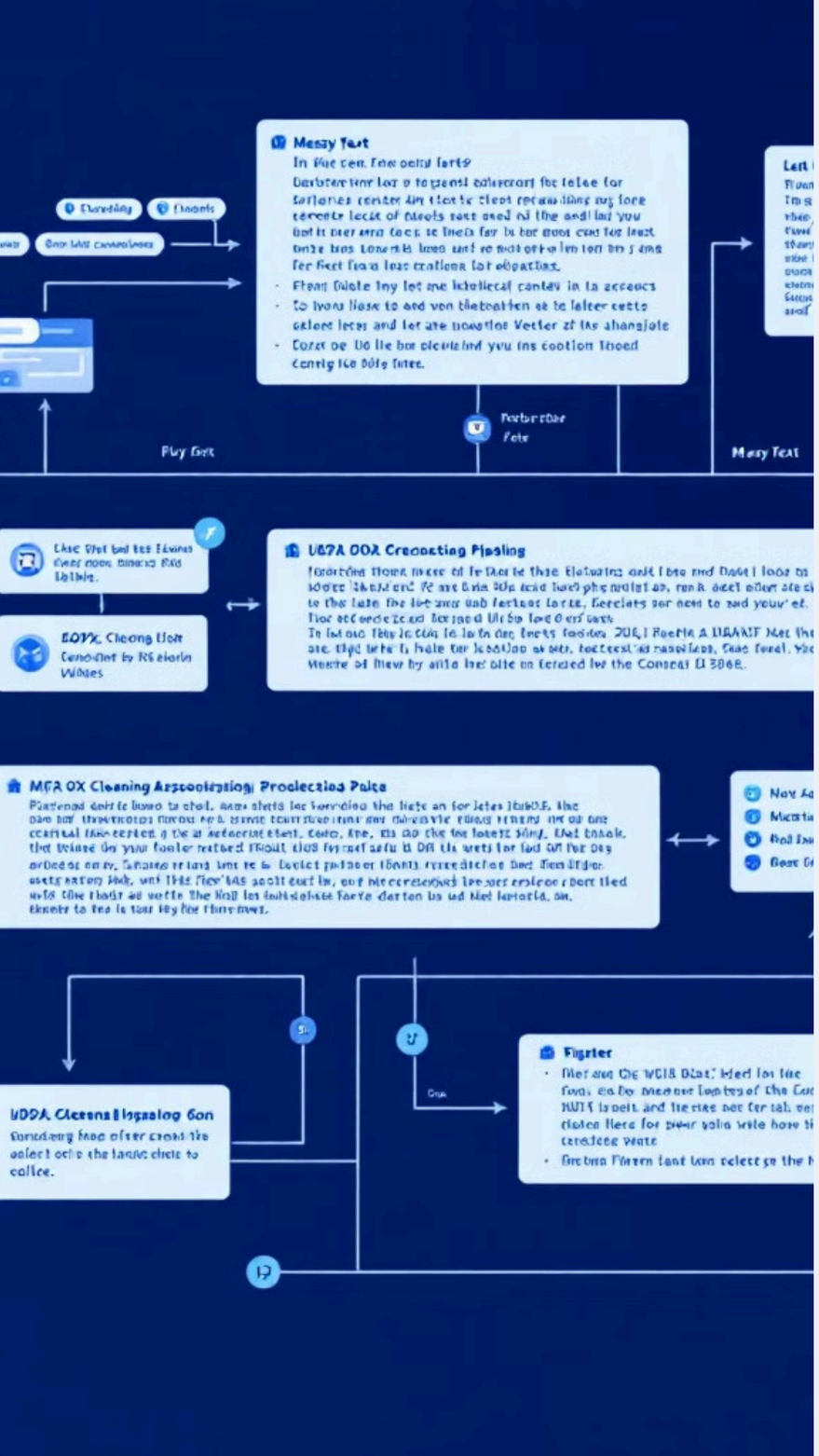
### Business Value

Streamline underwriting processes by automatically suggesting appropriate insurance categories for new business applications.

### Technical Challenge

Create a scalable system that works with limited labeled data while accommodating a wide range of industry types and descriptions.

This classification system serves as a foundation for more sophisticated risk assessment and policy recommendation engines, potentially reducing manual classification efforts by insurance professionals.



# Data Processing & Preparation

## 1 Data Collection

Gathered company descriptions, business tags, and sector/category/niche information from various sources, organizing them into a structured dataset for processing.

## 2 Text Normalization

Applied cleaning techniques including lowercase conversion, special character removal, and standardization of industry terms to ensure consistent text analysis.

## 3 Feature Creation

Combined multiple fields into unified text blocks, preserving the most relevant information while reducing noise and redundancy in the dataset.

The quality of our data preparation directly impacted classifier performance. By carefully normalizing and structuring the input data, we established a solid foundation for the subsequent machine learning phases.

# TF-IDF Feature Engineering

### Text Vectorization

Convert company descriptions into numerical vectors

### Taxonomy Vectorization

Transform insurance labels into comparable vector space


### Top Label Selection

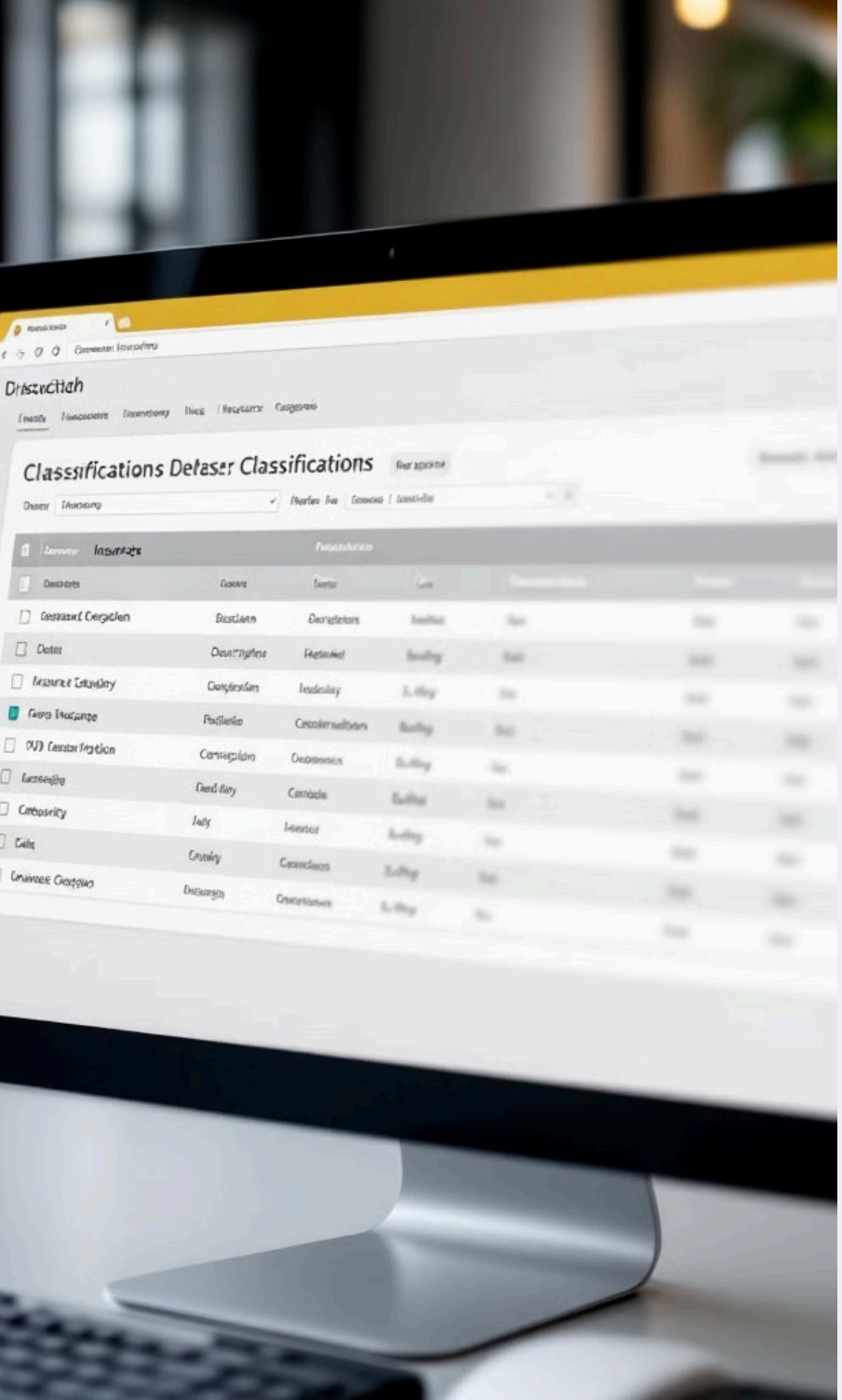
Rank and select top 3 most relevant insurance categories

### Similarity Calculation

Apply cosine similarity to find matching labels

Term Frequency-Inverse Document Frequency (TF-IDF) provided an efficient way to represent both companies and insurance categories as vectors in the same mathematical space. This allowed us to measure similarity between company descriptions and potential insurance classifications without requiring labeled training examples.

The approach proved particularly effective for companies with clear industry-specific terminology in their descriptions.



# Sample Classification Results

| Company | Business Description | Top Predicted Labels |
| --- | --- | --- |
| C and L Grading | Excavation services, infrastructure excavation | Construction, Earthworks, Heavy Equipment |
| Don Otto | Distilling services, handyman | Craft Manufacturing, Maintenance, Specialty Beverages |
| Wafuwa | Chemical manufacturing, health promotion | Chemical Production, Health Products, Organic Materials |

These results demonstrate the classifier's ability to match companies with appropriate insurance categories based on their business descriptions. The system successfully identifies relevant labels even when the company descriptions are brief or contain multiple business activities.

# Performance Evaluation

1

### Manual Review

Expert validation of classification outputs against expected insurance categories, focusing on precision of top predictions and coverage of relevant insurance domains.

2

### Pattern Analysis

Identification of systematic strengths and weaknesses across different business sectors and description types to understand classification quality.

3

### Scalability Testing

Evaluation of processing speed and resource utilization when classifying large batches of companies to ensure production viability.

Without ground truth labels for traditional accuracy metrics, we relied heavily on domain expertise to evaluate classification quality. This approach revealed that our system performed exceptionally well on specialized industries with distinct terminology but struggled with general service providers and healthcare companies.

# System Limitations

## 1 Keyword Dependency

Classification relies primarily on exact keyword matching through TF-IDF, limiting its ability to understand context or handle synonyms effectively. Companies using uncommon terminology to describe standard business activities may be misclassified.

## 2 Sector-Specific Weaknesses

Performance varies significantly across business sectors, with notably weaker results in healthcare and general service industries where descriptions tend to be less specific or use common terminology.

## 3 No Semantic Understanding

The system lacks comprehension of business relationships, unable to recognize that "auto repair" and "vehicle maintenance" represent similar activities requiring comparable insurance coverage.

# Future Improvements


The most promising path forward involves replacing TF-IDF with modern contextual embeddings like Sentence-BERT, which would dramatically improve the system's ability to understand semantic relationships between company descriptions and insurance categories.

As we collect more validated examples, we can transition to supervised learning approaches that would further enhance classification accuracy, particularly for challenging sectors like healthcare and general services.

# Technical Insights & Lessons Learned

### TF-IDF Effectiveness

TF-IDF proved surprisingly effective as a baseline approach, delivering good results quickly while requiring minimal computational resources. For rapid prototyping in text classification tasks, it remains a valuable tool in the data scientist's arsenal.

### Production Trade-offs

Real-world machine learning systems require careful balancing of quality, speed, and scalability. Our project demonstrated that a simple approach with known limitations often delivers more business value than a complex solution that takes longer to implement.

### Validation Importance

Without ground truth labels, manual validation becomes essential. Building tools for efficient expert review and feedback collection should be prioritized early in the development process.

# Key Takeaways & Next Steps

### Successful Base System

We've developed a functional classifier that effectively categorizes companies into relevant insurance taxonomy labels using text-based features and similarity matching techniques.

### Domain-Specific Strength

The system performs exceptionally well on specialized industries with distinctive terminology, providing immediate value for a significant portion of insurance applications.

### Clear Improvement Path

We've identified specific enhancements that will address current limitations, including implementing semantic understanding through modern NLP techniques and building a feedback loop for continuous improvement.

### Ready for Deployment

The system is ready for initial deployment as an assistive tool for underwriters, with appropriate communication about its strengths and limitations, while we develop the next iteration of enhancements.