RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Mentor Name: John Rachlin
Subject/Category : Data Science - Natural Language Processing (NLP)
Class: DS 3500 (Advanced Programming with Data)
Undergraduate

Northeastern University

Natural Language Processing of Appellate Court Insurance Decisions

Oliver Baccay; John Rachlin

Background and Goals

Context

- Insurance litigation has complex linguistic patterns
- Language varies significantly by case type
- Critical for improving risk assessment and settlement strategies

Research Question

How do linguistic patterns vary across different insurance litigation categories?

Approach

- Analyze appellate court insurance decisions using NLP
- Build framework to identify language signatures
- Create tool for quantitative document analysis

Process and Methods

Data Construction

- Eight appellate court insurance decisions from 2020-2025
- Seven jurisdictions: state supreme and federal circuit court
- Major dispute categories: liability, property damage, healthcare fraud, contractual indemnity, regulatory enforcement
- 10-40 pages per decision with detailed legal reasoning

Framework Architecture

- Custom object-oriented Python class
- Structured data storage using nested dictionaries
- Standardized text preprocessing: whitespace, punctuation, stop-word removal
- Extensible parsers and multiple visualization methods for statistical analysis

Case Label	Court Case Name
Progressive	Insurance Professionals, INC. v. Progressive Casualty Insurance Company
Maryland Insurance	Erie Insurance Exchange v. Maryland Insurance Administration
Travelers	Peter D. Protopapas v. Travelers Casualty
American Family	Stephanie Lock v. American Family Insurance Company
Liberty Mutual	Liberty Mutual Insurance v. Digitas, Inc.
Geico	GEICO v. Mayzenberg
State Farm	Schnell v. State Farm
Farmers Insurance	Farmers Ins. v. Minemyer

Figure 1: Insurance Coverage Court Cases Dataset
The court cases listed above were obtained by systematically searching CourtListener's database of recent appellate court decisions. Each case represents a different insurance company and dispute type to ensure linguistic diversity across coverage litigation categories.

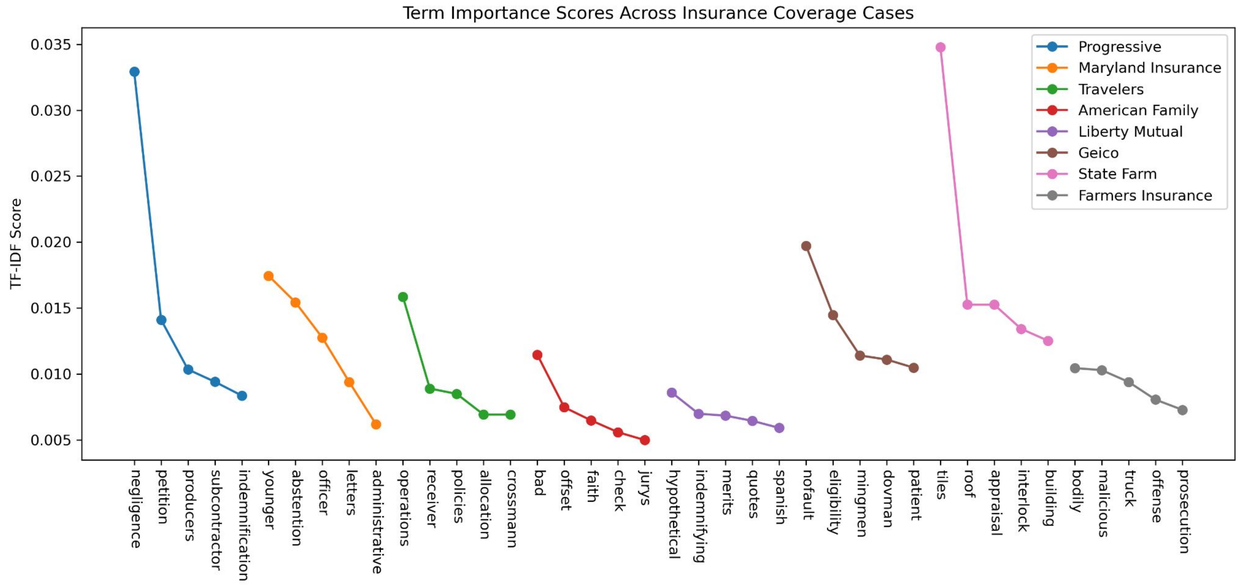

Figure 2: Term Frequency-Inverse Document Frequency Plot (k=5)
TF-IDF importance scores across all analyzed terms, showing how terminology distinctiveness varies by insurance coverage case type. Each line represents a different case, with peaks indicating terms that are uniquely important to that particular dispute category. The Progressive case shows high distinctiveness for certain procedural terms, while other cases peak at different vocabulary clusters, demonstrating the framework's ability to identify case-specific linguistic signatures.

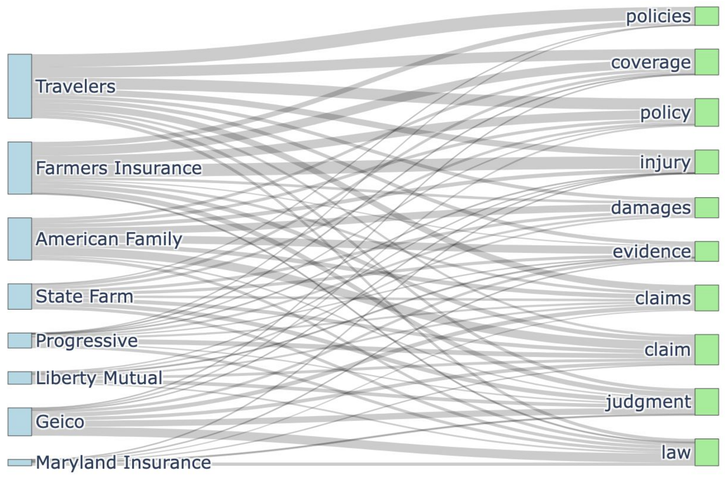

Figure 3: Sankey Diagram (k=10)
Sankey diagram illustrating the relationship between court cases (left) and their most frequent words (right). The flow thickness represents the strength of association between each case and key terms. This visualization reveals how specific legal concepts cluster around different insurance companies and dispute types. Coverage-related terms flow primarily to certain cases while procedural terms like 'judgment' and 'evidence' connect to most cases, highlighting the systematic linguistic differences in legal reasoning patterns.

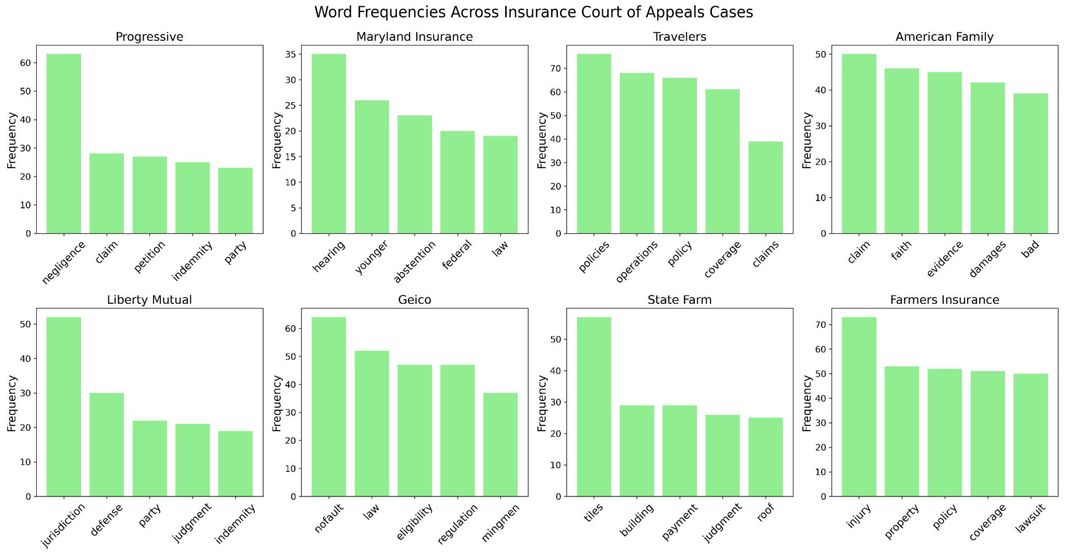

Figure 4: Bar Chart (k=5)
Word frequency analysis showing the most common terms used in each insurance case. Each bar chart displays the top 5 most frequently occurring words after preprocessing for different dispute categories. The variation in terminology across case types demonstrates distinct linguistic patterns. For example, Progressive emphasizes 'coverage' and 'claim' language, while Geico focuses more on 'policy' and 'law' terminology, indicating different legal reasoning approaches across insurance litigation categories.

Conclusion and Next Steps

Key Findings & Impact

- Distinct linguistic signatures identified across insurance coverage dispute types, enabling systematic analysis
- Framework successfully distinguishes between different insurance companies' legal language patterns
- Applications: Enhance risk assessment models, predict litigation costs, and inform settlement strategies

Future Steps

- Expand to 25+ cases across more jurisdictions and coverage types
- Develop predictive capabilities using linguistic signatures to forecast case outcomes and settlement amounts
- Apply to real-time applications in insurance underwriting and claims processing