Documenting Corporate Harm: A Semantic Action Trajectories Approach to the Opioid Industry Document Archive Shared Task

Benjamin Miller
University of Canterbury
Ōtautahi Christchurch, Aotearoa New Zealand
benjamin.miller@canterbury.ac.nz

Abstract

This paper presents a method for modeling change in the possibility space of actors over time as represented in the Opioid Industry Document Archive (OIDA). The approach treats documents as a structured field of actor–action relations and models these relations as *semantic action trajectories* across time. Semantic role labeling (SRL) using the Emory Language and Information Toolkit (ELIT) is applied to extract subject–predicate structures from a corpus of internal industry documents. Subjects are normalized and grouped into actor categories using a combination of rule-based heuristics and constrained language model adjudication. Predicate vocabularies associated with these actors are mapped to psycholinguistic categories using the LIWC lexicon, and random forest feature selection with principal component analysis is used to construct a low-dimensional representation of discourse structure across periods.

The resulting discourse space reveals systematic shifts in how corporate actors, regulators, clinicians, and patients are positioned over time. In particular, corporate entities and the opioid products they produce follow nearly identical semantic trajectories, suggesting that companies and drugs occupy interchangeable roles in the archive's discourse. This method provides a way to analyze changing institutional behavior at scale across heterogeneous litigation and historical archives.

1 Introduction

Large litigation archives provide an unusually detailed record of institutional communication and practice. The Opioid Industry Document Archive (OIDA) contains millions of previously undisclosed corporate documents produced during litigation concerning the opioid crisis. An additional, larger Public Document Repository, mandated by orders issued in the Purdue Pharma bankruptcy proceedings in the United States Bankruptcy Court for the Southern District of New York promises to expand these holdings related to one of the most significant breaches of the public trust in US history (Vadivelu et al., 2018). These documents describe the marketing, sales, research and development, compliance, and regulatory details, along with call notes, procedural descriptions, and trial material that underlie a crisis that nearly tripled the reported drug overdose death rate in the US (Vadivelu et al., 2018). In scope, the court order describes more than 100 million pages of material to be added to a public repository. The OIDA materials and this yet to be released repository provide an important evidentiary record describing an institutional project that contributed to a global health crisis. However, their scale and heterogeneity make systematic analysis difficult.

This paper introduces a computational method for analyzing changes in actor behavior over time as represented by discourse in the OIDA corpus. The central premise is that a heterogeneous corpus like OIDA can be modeled as a set of structured relations linking actors and actions at times. By extracting subject–predicate structures at time from text and aggregating them into subject groups, it becomes possible to model the changing possibility space of actor groups. Following (Mehran et al., 2025), a possibility space refers to the set of allowable actions associated with a subject. In this study, that space is operationalized as the predicates attached to a subject group.

The method builds on prior work applying computational analysis of subject–predicate relations to ideological and institutional discourse (Mehran et al., 2025). More broadly, computational approaches have been used to model narrative and discourse structure across large document collections (Miller et al., 2015). Here we extend those approaches in two directions. First, we model discourse diachronically by associating actor–action relations with historical periods. Second, we intro-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 149–158

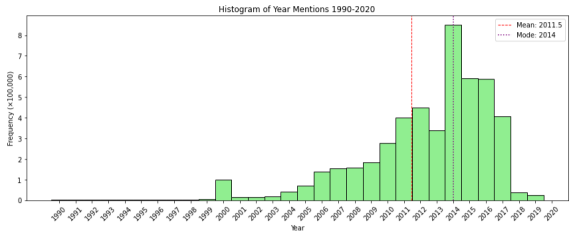

Figure 1: Distribution of year mentions extracted from the corpus. The sample is heavily concentrated in the litigation period of the opioid crisis, with a mean referenced year of 2011.5 and a modal year of 2014.

duce an automated actor grouping procedure combining rule-based classification with constrained large language model adjudication.

The result is a representation of institutional discourse as a set of *semantic action trajectories* that describe how actors move through a conceptual discourse space over time.

2 Task and Data

The NLP+CSS shared task focuses on computational analysis of the Opioid Industry Document Archive. The archive includes internal corporate communications, regulatory materials, litigation documents, and investigative records produced during the development of the opioid crisis.

For this study, a corpus of 10,000 OCR documents was randomly sampled from the public Opioid Industry Documents Archive (OIDA), a repository of corporate documents released through opioid-related litigation (Alexander et al., 2022). This sampled sub-corpus contains approximately 238 million tokens and includes heterogeneous document types such as emails, reports, legal transcripts, and regulatory submissions. The sample was constructed through repeated random traversal of the archive structure, with each traversal yielding one document. The resulting sub-corpus reflects the temporal, topical, and source distributions inherent in OIDA.

Temporal references extracted from documents show a strong concentration in the later years of the archive. Figure 1 shows the distribution of year mentions across the corpus.

Documents were grouped into six temporal periods spanning 1980–2019.

3 Method

After sampling, analysis proceeds in four stages: predicate extraction, actor grouping, semantic feature construction, and discourse space modeling.

Statistic	Value
Predicate instances	246,073
Subject–predicate pairs	184,831
Normalized subject expressions	52,305
Unique predicates	4,548

Table 1: Summary statistics from the SRL predicate extraction stage.

3.1 Predicate Extraction

Semantic role labeling (SRL) has long been used to extract predicate–argument structures from text (Gildea and Jurafsky, 2002). For this project, SRL was applied using the Emory Language and Information Toolkit (ELIT) (He et al., 2021). The pipeline performs tokenization, part-of-speech tagging, dependency parsing, and semantic role extraction.

While recent work increasingly relies on transformer-based representations, the present approach intentionally uses a non-neural SRL implementation to produce explicitly structured actor–action relations. This choice prioritizes interpretability and analytical transparency: subject–predicate structures can be directly aggregated into actor-level distributions and inspected without post hoc probing or attribution methods. At the same time, recent work shows that SRL remains a challenging task for large language models, particularly in settings without pre-identified predicates, where performance degrades substantially relative to structured approaches (Li et al., 2025). In such settings, even strong LLM-based methods require retrieval augmentation and task-specific scaffolding to achieve competitive performance.

This motivates the use of a structured SRL pipeline in the present study, where the goal is not maximal benchmark accuracy but a stable and interpretable representation of actor–action relations in noisy, heterogeneous archival data. Table 1 summarizes the resulting predicate extraction statistics.

The corpus contains 5.0 million year mentions, 246,073 SRL predicates, approximately 215k valid predicates, and roughly 185k rows containing subject–predicate pairs. The resulting subject extraction rate is approximately 75%, with a parser error rate of 1.6%. After preprocessing and filtering, the final dataset contains 86,414 subject–predicate–year triples distributed across six temporal periods.

Because many OCR-derived sentences contain

Period	Observations
1980–1994	165
1995–1999	291
2000–2004	3,930
2005–2009	20,074
2010–2014	49,488
2015–2019	12,466

Table 2: Number of cleaned subject–predicate observations by time period.

copular or auxiliary constructions, a subject recovery procedure was applied to identify fallback subjects in otherwise incomplete parses. This process yielded an additional 13.9% subject recoveries, producing a final inventory of 52,305 unique normalized subjects and 4,548 unique predicates across 60,451 sentences.

Year mentions were extracted to associate actor–action relations with the historical periods referenced within documents rather than relying solely on document creation dates(Pustejovsky et al., 2003). Because litigation archives frequently contain retrospective discussion of earlier events, this approach enables temporal indexing of discourse about past regulatory actions, marketing practices, and clinical developments.

Subjects were grouped into a controlled actor ontology consisting of 17 groups that fall broadly into the categories of organizational actors, individual actors, discourse artifacts, and referential placeholders. Table 3 shows the distribution of subject groups after normalization and cleaning. The distribution follows a typical long-tailed pattern, with a small number of high-frequency discourse roles accounting for a large share of predicate instances. Subject groups are defined as follows: clausal_or_artifact_subject (non-agentive linguistic or document artifacts, e.g., clauses, sections), addressee (second-person or recipient roles), individual_actor (named or generic persons), corporate_self (first-person corporate voice, e.g., “we”), information (abstract informational entities, e.g., data, reports), referential (pronouns and discourse placeholders), commercial_products (drug or product names), external_actor (third-party organizations or actors outside the focal firm), patients_consumers (patients or end-users), medical_status (conditions or diagnoses), medical_professionals (clinicians and healthcare providers), corporate_entities (named firms), regulators_government (regulatory or state

Subject Group	Count
clausal_or_artifact_subject	19,226
addressee	15,751
individual_actor	13,250
corporate_self	12,602
information	6,554
referential	5,109
commercial_products	3,359
external_actor	3,303
patients_consumers	1,640
medical_status	1,407
medical_professionals	1,306
corporate_entities	1,194
regulators_government	515
commercial_partners	430
other_actor	334
indefinite_actor	325
interrogative_actor	109

Table 3: Distribution of normalized subject groups after cleaning and actor grouping.

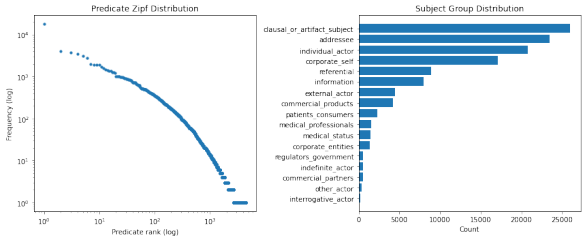

Figure 2: Zipf distribution of predicate frequencies extracted from SRL. The heavy-tailed distribution confirms typical lexical structure and suggests the extraction pipeline preserved natural predicate usage patterns.

actors), commercial_partners (distributors or business partners), other_actor (miscellaneous actors), indefinite_actor (non-specific agents, e.g., “someone”), and interrogative_actor (questioned or unknown agents).

Predicate frequency follows a heavy-tailed Zipf distribution typical of natural language corpora (Zipf, 1949).

Each extracted predicate instance forms a minimal actor–action relation that can be associated with document metadata and temporal references.

3.2 Actor Grouping

Raw subject expressions exhibit substantial lexical variation. Subjects were therefore normalized through a multi-stage procedure consisting of lexical normalization, rule-based classification, and language-model adjudication of ambiguous cases.

The rule-based stage captures high-frequency actors and organizational references. Residual subjects were evaluated using a constrained language model prompt designed for closed-set categoriza-

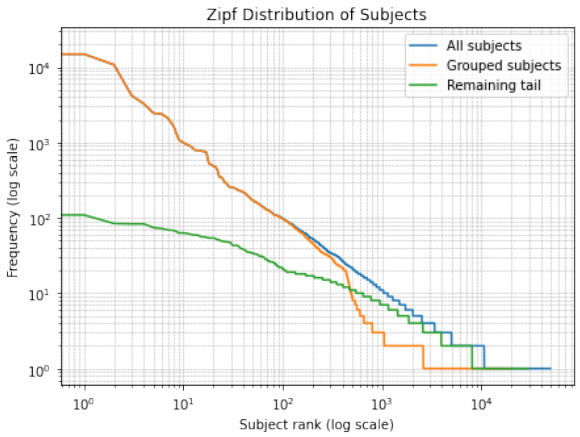

Figure 3: Zipf distribution of normalized subject expressions. The high-frequency head is largely captured by rule-based grouping, while mid-frequency subjects are resolved through LLM-assisted adjudication. The remaining long tail consists primarily of rare expressions and OCR artifacts.

tion. For each of the 267 candidate subjects, the model was shown the subject string, its corpus frequency, and eight example sentence contexts, optionally including the extracted predicate. The model was instructed to assign the subject to the single best label from the predefined actor ontology, to avoid inventing new labels, and to prefer `No Group Found` when evidence was weak or the extraction appeared malformed. Outputs were returned in a structured JSON format including a recommended group, confidence score, optional secondary group, mixed-use flag, and short rationale. A confidence cutoff of 0.9 was used. This stage recovered 267 subject types, including branded opioid products such as Kadian and Eluxadoline, and yielded 10,086 additional predicate–argument–time triples.

Approximately 15% of previously ungrouped subjects were recovered through this procedure.

Figure 3 shows the frequency distribution of subject expressions, which follows a heavy-tailed Zipf distribution typical of natural language. A small number of high-frequency subjects account for a large share of instances, while the majority occur only once or twice.

After removing high-frequency auxiliary predicates and boilerplate discourse markers, 86,414 predicate instances remained.

These distributions provide a semantic representation of actor discourse.

3.3 Semantic Feature Construction

For each actor group and time period, predicate vocabularies were mapped to meaningful psychological and social conceptual categories using the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2015; Tausczik and Pennebaker, 2010).

3.4 Discourse Space Modeling

Random forest classification was used to identify semantic features that distinguish actor groups. A one-versus-rest classification setup was used for each actor group, using LIWC category frequencies as input features. Permutation importance was estimated across 200 bootstrap samples, and features exceeding one standard deviation above the mean feature importance were retained for discourse space analysis. 13 features remained from the initial 118 provided by LIWC.

Principal component analysis (PCA) was applied to the resulting feature matrix. The first two principal components explain 38.0% of the variance in the semantic feature space (PC1: 21.6%, PC2: 16.4%). A third component explains an additional 12.8% of variance but was not included in the analysis in order to preserve a two-dimensional discourse space suitable for visualizing actor trajectories. While 3D visualization is a natural extension, we prioritize a 2D projection for interpretability and leave higher-dimensional visualization as future work.

4 Results

The result of the pipeline is a time-indexed mapping from actor groups to distributions over predicate-linked semantic categories, which can be interpreted as an empirical approximation of each group's "possibility space," or the set of actions attributed to that group within the archive.

Before modeling actor movement in semantic discourse space, we first examine the temporal distribution of subject groups in the corpus. Figure 4 shows period-wise deviations in subject-group frequency relative to each group's overall mean, expressed as z-scores. Positive values shown in red indicate periods in which a subject group is overrepresented relative to its overall distribution, while negative values shown in blue indicate underrepresentation.

The figure suggests three broad phases. First, early periods are characterized by product-,

Feature	PC1	PC2
Cognition	0.472	0.317
cogproc	0.449	0.380
insight	0.391	0.128
perception	0.284	-0.445
allure	0.278	-0.205
motion	0.235	-0.346
attention	0.232	-0.253
focuspresent	0.097	-0.296
cause	0.001	0.358
acquire	-0.079	-0.121
reward	-0.092	-0.242
need	-0.160	-0.016
work	-0.326	0.162

Table 4: Top LIWC feature loadings for the first two principal components of the discourse space. Positive and negative values indicate opposing semantic poles along each component.

medical-, and patient-centered discourse. Second, a middle period shows increasing prominence of corporate entities and external actors, suggestive of branding, distribution, and marketing activity. Third, later periods emphasize corporate, regulatory, and addressee-centered discourse, reflecting both increasing regulatory scrutiny and a greater prevalence of directive communication (e.g., "you will").

Our findings suggest that actor groups occupy distinct regions of the resulting discourse space, and their trajectories across periods reveal systematic changes in their possibility spaces.

Figure 5 plots the positions of selected actor groups from the first period (1980–1994) to the final period (2015–2019). Movement in this space reflects shifts in the semantic framing of actor discourse as captured by LIWC feature distributions.

Two principal semantic dimensions structure this space. Table 4 lists the LIWC features with the highest loadings on the first two principal components.

The first component along the x-axis (PC1) contrasts cognitive and perceptual processing language (Cognition, cogproc, insight) with goal-oriented organizational discourse (work). The second component along the y-axis (PC2) contrasts causal analytic reasoning (cogproc, cause) with experiential and perceptual language (Perception, motion, focuspresent). Together these axes differentiate discourse oriented toward explanation and reasoning from discourse oriented toward operational coordination, experiential language, and immediate activity.

No subject group remains semantically stable across the full temporal span of the archive. Several groups—including `patients_consumers`, `regulators_government`, `commercial_partners`, and most dramatically `medical_professionals`—shift substantially within the discourse space. Across these groups, discourse moves away from causal and analytic reasoning toward greater emphasis on experiential and perceptual language.

The `medical_professionals` group exhibits the largest displacement. This suggests a substantial shift in how clinicians are positioned within internal corporate communication across the periods represented in the archive.

Two additional groups display a particularly striking pattern: `corporate_entities` and `commercial_products`. The first includes firms such as Cephalon, McKesson, Endo, Insys, and Teva, while the second includes drug entities such as Exalgo, Opana, and Xartemis XR. The trajectories of these two groups move almost identically through the discourse space. Predicates associated with the companies change in the same direction, and to nearly the same degree, as predicates associated with the drugs themselves.

This parallel movement suggests that corporate actors and the pharmaceutical products they produce are treated almost interchangeably within the predicate structures of the archive. In effect, the companies producing these drugs and the drugs themselves occupy nearly identical semantic roles in the discourse.

Across both groups, discourse shifts away from goal-directed action language toward perceptual processing, while also moving slightly from experiential language toward explanatory reasoning. In practical terms, this corresponds to a reduction in action-oriented framing and a greater emphasis on explanation and interpretation.

This coupling suggests that corporate responsibility and product behavior are linguistically co-constructed within the archive, with actions attributed to drugs mirroring those attributed to the firms that manufacture them.

The only group that moves in the opposite direction is `corporate_self`. Over time this group shifts away from experiential and perceptual language and toward more goal-directed discourse, accompanied by a modest increase in reasoning-oriented language.

To clarify the behavior of the principal institutional actors, Figure 6 shows

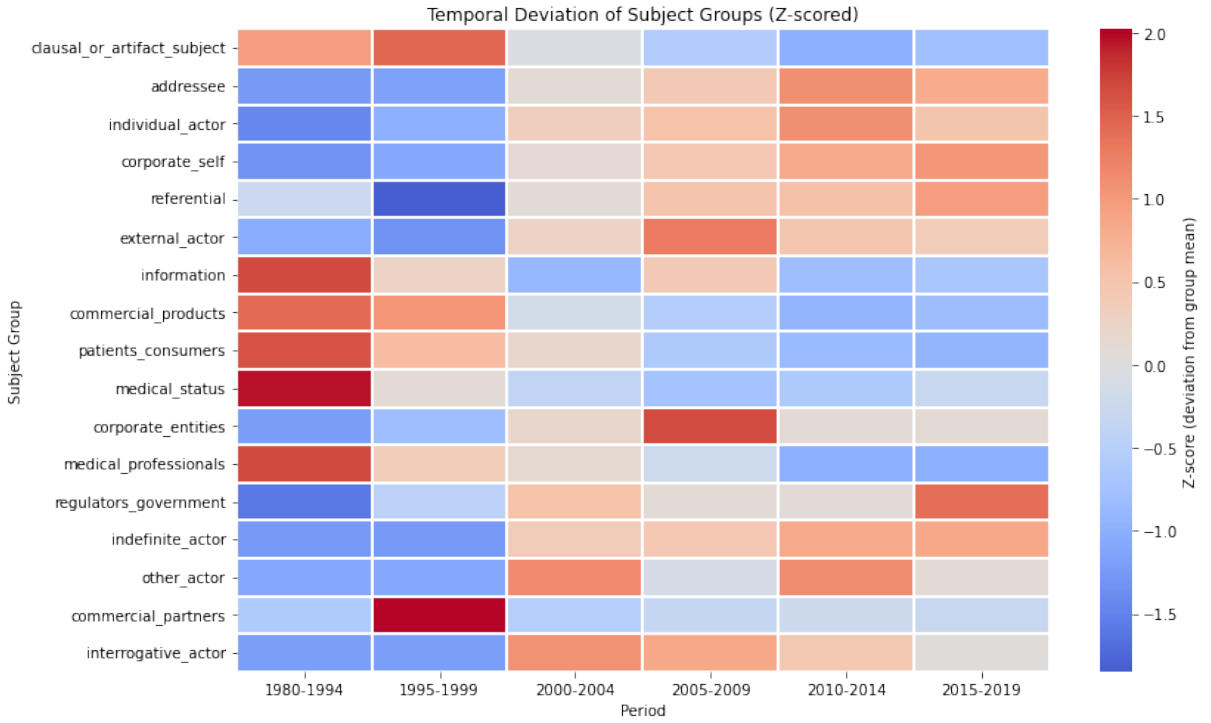

Figure 4: Temporal deviation of subject groups across periods, shown as within-group z-scores relative to each group's mean frequency. Positive values indicate periods in which a subject group is overrepresented relative to its overall distribution; negative values indicate underrepresentation. The figure highlights a temporal shift from product- and patient-centered discourse in earlier periods toward corporate, regulatory, and interactional subject positions in later periods.

the trajectories of five key subject groups across periods: `corporate_self`, `patients_consumers`, `commercial_partners`, `medical_professionals`, and `regulators_government`. Focusing on these actors highlights the most substantial movements in the discourse space and reveals that semantic change across the archive is not uniform across time.

The trajectories support the earlier interpretation that the corpus reflects three broad discursive phases: an early period oriented toward medical discussion, a middle period emphasizing distribution and commercial coordination, and a later period dominated by regulatory scrutiny and investigative discourse.

For example, the `corporate_self` group begins in a region of the discourse space associated with experiential and cognitive language. During the second period (1995–1999) it shifts further toward experiential framing, before moving sharply away from this modality in later periods. This later movement corresponds to the increasing prevalence of corporate email communication and the transition toward distribution and regulatory investigation.

By contrast, the `patients_consumers` grouping initially moves toward more experiential action language during the 1995–1999 period, before shifting toward more cognitive and perceptual discourse. This pattern is consistent with the later emphasis on retrospective patient testimony and the collection of investigative evidence.

The `commercial_partners` group exhibits the greatest overall movement, though not the greatest displacement from its starting position. This group undergoes the most pronounced shifts in discourse framing as its institutional role changes across the periods represented in the archive.

Taken together, these trajectories illustrate how actor positions within the discourse space evolve and drift across time, revealing which institutional actors occupy the most unstable or rapidly changing semantic roles. In this sense, trajectory instability provides a quantitative indicator of shifting institutional roles within the evolving discourse of the opioid crisis.

4.1 Feature Trajectories

Finally, temporal feature trajectories across periods by group highlight which semantic dimensions

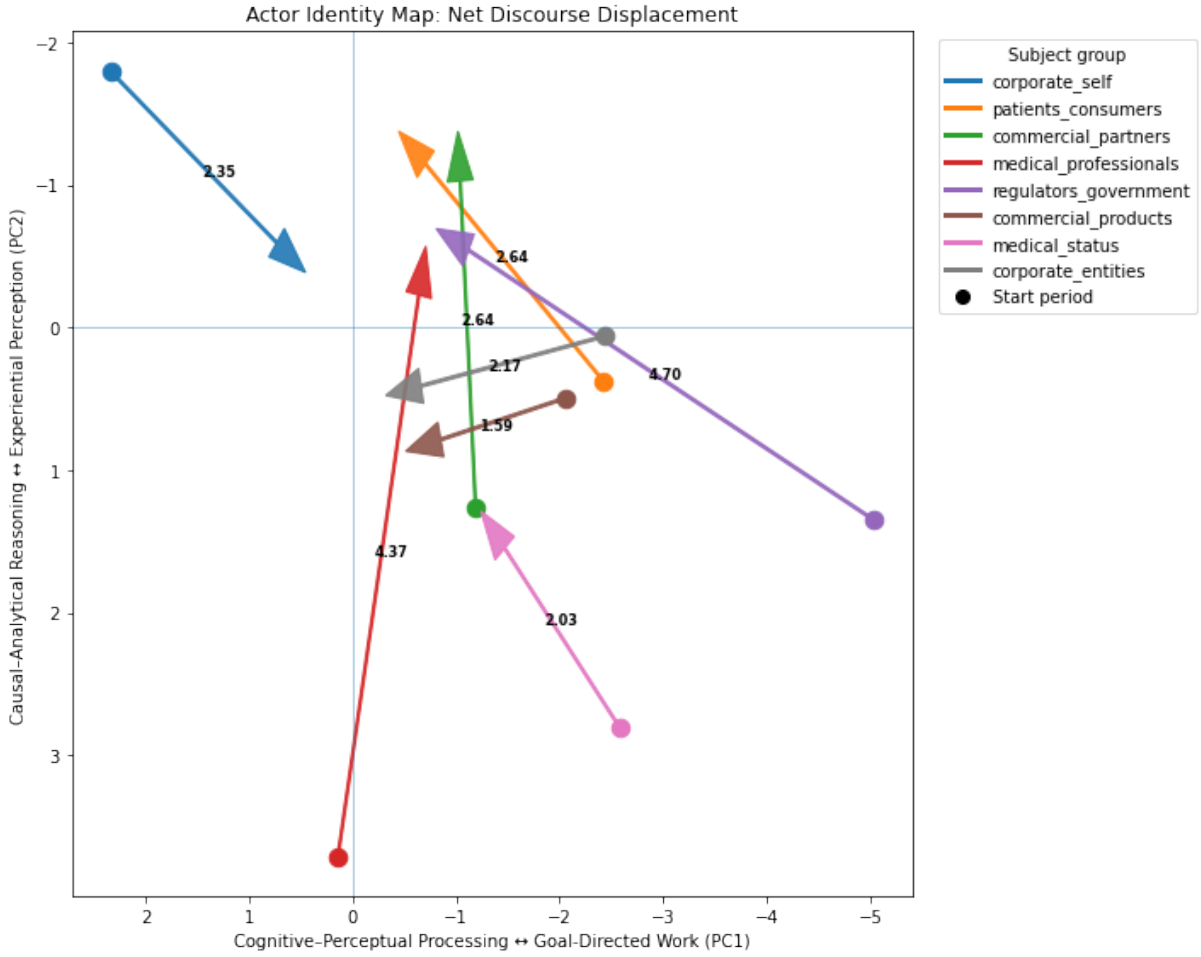

Figure 5: Actor trajectories in discourse space across historical periods.

drive actor movement in the discourse space. For example, reward language drove early variation in the 1995–1999 period for the `corporate_self` grouping, while need did the same in the 2000–2004 period for the `medical_status` group. Some features reveal relatively little change over time despite notable early spikes, such as attention, while others reveal interesting colinear pairings, such as `medical_professionals` and `patients_consumers` in the 2010–2014 and 2015–2019 periods relative to actions LIWC labeled as *acquire*. In effect, this figure provides a quantitative perspective on the actions ascribed to and undertaken by broad classes of subjects, people, and institutions as they navigated a developing catastrophic public health crisis.

5 Analysis

The results suggest a gradual reorientation of discourse across the archive. Earlier documents tend to situate discussion around products and patients, while later communication increasingly centers on corporate entities and regulatory actors.

This shift corresponds partly to changes in document type, particularly the emergence of internal corporate email around 2000. However, the semantic structure of discourse also changes, suggesting a broader transition in institutional communication as the crisis developed.

Within the discourse space, corporate actors exhibit comparatively stable positioning, while medical and regulatory actors show larger movement across periods. The near-identical trajectories of `corporate_entities` and `commercial_products` suggest that firms and the drugs they produce occupy closely aligned semantic roles within the archive. In practice, similar types of actions are attributed to both companies and their products across time. One interpretation is that product behavior and corporate behavior are linguistically co-constructed in the documents, such that actions described in relation to drugs (e.g., efficacy, risk, usage) mirror those attributed to the firms themselves (e.g., development, marketing,

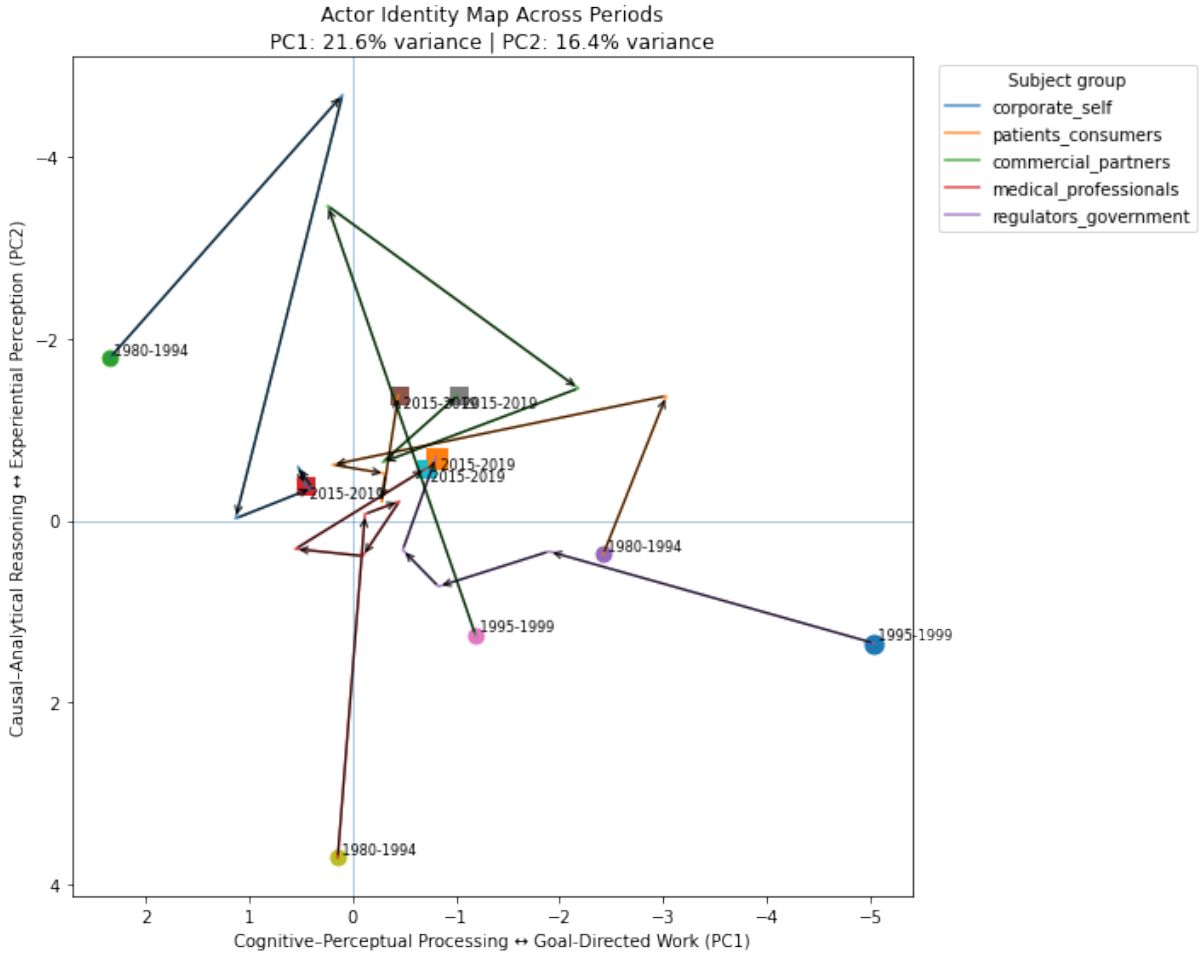

Figure 6: Trajectories of selected actor groups across periods in the PCA discourse space. Each line traces the movement of an actor group's semantic position across six temporal periods.

distribution). This pattern suggests that responsibility, agency, and outcomes are distributed across both entities and products, rather than sharply distinguished between them. Further work could disaggregate these subject positions into individual actors, clarifying which actors are foregrounded or backgrounded in the attribution of responsibility.

A key interpretive caveat is the shift in corporate communication associated with the adoption of email. To assess whether the observed discourse shifts could be driven primarily by communication medium, we approximate the prevalence of email-style documents using synonym-based detection of header markers. Table 5 summarizes opioid-term frequency and the estimated prevalence of email markers. The results show a sharp increase in email-style communication after 2000, indicating that part of the observed shift reflects changes in archival composition. However, the persistence of structural changes in actor–action relations suggests that the trajectories capture more than a simple medium effect.

Period	Opioid Mentions	Share of Documents	Email Markers
1980–1994	72	14–19%	0.0%
1995–1999	196	∼24%	3.3%
2000–2004	796	∼9%	9.7%
2005–2009	7,041	∼17%	11.5%
2010–2014	9,632	∼10%	18.8%
2015–2019	2,159	∼8%	15.3%

Table 5: Temporal distribution of opioid-term mentions and email markers across periods.

Email markers include header elements such as *From*, *To*, *Subject*, and timestamp fields.

The sharp shift after 2000 partly reflects the increasing presence of internal communication such as email, but the change in subject roles cannot be explained solely by communication medium. Instead, the corpus reveals a broader transition from product- and patient-centered discourse toward organizational coordination, corporate entities, and

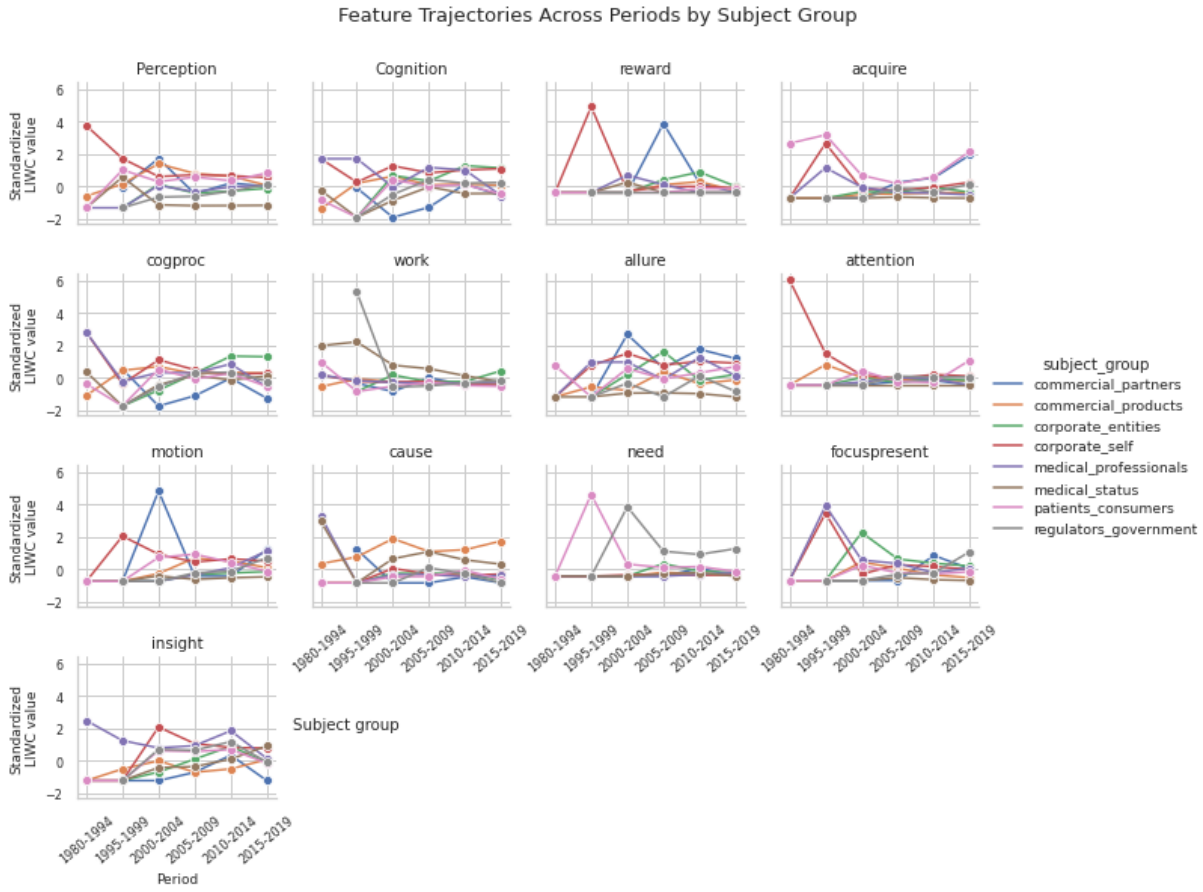

Figure 7: Temporal trajectories of selected LIWC features across actor groups.

regulatory actors.

As a second validation test, the frequency of mentions of opioids as a category was also measured. Early periods contain a higher proportional share of direct references to opioid products and medical conditions, but later periods increasingly center on corporate actors, regulatory institutions, and directive communication among organizational members. This suggests a shift in the archival record from product- and patient-focused discourse toward internal coordination and regulatory engagement.

Taken together, these results and tests indicate that the observed semantic trajectories reflect changes in what actions are attributed to actor groups, rather than shifts in latent topic or document similarity alone.

6 Conclusion

This paper introduced a computational pipeline for modeling actor discourse trajectories in the Opioid Industry Document Archive. By extracting predicate structures, grouping actors, and constructing a semantic discourse space, the method produces interpretable representations of institutional communication and semantic change of subject groupings over time.

The resulting actor trajectories reveal systematic differences in how corporate actors, regulators, clinicians, and patients are positioned within internal communication and how these positions evolve across historical periods.

These results demonstrate how structured actor–action representations can reveal shifts in institutional discourse that are not captured by topic-based or document-level analyses alone, and provide a foundation for more granular analysis of responsibility, agency, and role attribution within corporate structures.

Limitations

A key limitation is the uneven temporal distribution of documents in the archive. Early periods (pre-2000) contain substantially fewer documents than later periods, which are dominated by internal corporate email. The archive also contains hetero-

geneous document types, including reports, emails, and legal transcripts, as would be normal for any comprehensive corporate archive. As a result, the observed diachronic shifts cannot be interpreted purely as changes in that underlying corporate behavior; they also reflect changes in document production, preservation, and legal disclosure. The analysis and signal therefore also captures changes in the archival representation of institutional discourse, rather than a fully controlled sample of communication across time.

To mitigate the effects of temporal imbalance, actor–action relations are indexed using within-period normalization, and analysis focuses on relative deviations (z-scores) rather than raw counts. However, comparisons between early and late periods should be interpreted cautiously, with greater confidence placed on within-period structure and post-2000 trends where document density is higher.

Explicit modeling of semantic action trajectories in pre- and post-email corporate regimes could help disentangle the effects of communication medium from underlying institutional change. Additionally, early periods remain relatively sparse despite targeted sampling. A supplemental sampling strategy for pre-1990 documents was explored, but the combination of OCR noise, data sparsity, and deviations from the random sampling design led to its exclusion. Future work could focus on improving SRL robustness for noisy OCR data.

Acknowledgements

I thank the organizers of the NLP+CSS OIDA Shared Task; the United States Bankruptcy Court for the Southern District of New York, whose orders in the Purdue Pharma bankruptcy proceedings require the creation of a Public Document Repository on the subject; and the University of California, San Francisco and Johns Hopkins University for maintaining and providing access to the Opioid Industry Documents Archive. I am grateful to Lawrence Fogelman for discussions of the legal context and interpretive considerations relevant to this work. Portions of the analysis pipeline code were developed with the assistance of a large language model (GPT-5.3). The model was used for programming assistance, debugging, and LaTeX formatting; all methodological decisions, analyses, and interpretations were conducted by the author.

References

G. Caleb Alexander, Lisa A. Mix, Sayeed Choudhury, Rachel Taketa, Cecilia Tomori, Mehdi Mooghali, Andrew Fan, Sarah Mars, Daniel Ciccarone, Michael Patton, Dorie E. Apollonio, Laura Schmidt, Michael Steinman, Jeremy Greene, Pamela Ling, Andrew K. Seymour, and Stanton Glantz. 2022. The opioid industry documents archive: A living digital repository. *American Journal of Public Health*, 112(8):1126–1129.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Han He, Liyan Xu, and Jinho D Choi. 2021. Elit: Emory language and information toolkit. *arXiv preprint arXiv:2109.03903*.

Xinxin Li, Huiyao Chen, Chengjun Liu, Jing Li, Meishan Zhang, Jun Yu, and Min Zhang. 2025. Llms can also do well! breaking barriers in semantic role labeling via large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23162–23180.

Weeda Mehran, Ben Miller, and Stephen Herron. 2025. Nothing in common? analysis of moral, psychological, and social factors in the identity construction of far-right and violent jihadi extremists. *Studies in Conflict & Terrorism*, pages 1–24.

Ben Miller, Jennifer Olive, Shakthidhar Gopavaram, Yanjun Zhao, Ayush Shrestha, and Cynthia Berger. 2015. A method for cross-document narrative alignment of a two-hundred-sixty-million word corpus. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1673–1677. IEEE.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Nalini Vadivelu, Alice M Kai, Vijay Kodumudi, Julie Sramcik, and Alan D Kaye. 2018. The opioid crisis: a comprehensive overview. *Current pain and headache reports*, 22(3):16.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*.