The Technological Trajectory of Semantic Analysis: A Historical-Methodological Review of NLP in Social Sciences

Rodrigo Kataishi, Ph.D.
CONICET Research Fellow
National University of Tierra del Fuego
Ushuaia, Argentina
rkataishi@untdf.edu.ar

Abstract

This paper examines the evolution and application of quantitative semantic analysis tools in social sciences, tracking their development from early statistical methods to contemporary large language models. The analysis demonstrates how computational advances have transformed qualitative research capabilities, enabling the systematic analysis of vast textual datasets while maintaining interpretative depth. The study presents a comprehensive review of key methodological approaches, including statistical analysis, topic modeling, semantic networks, and dimensionality reduction techniques, while examining their practical applications in social science research. Special attention is given to recent developments in natural language processing, particularly the emergence of transformer-based models and their impact on research methodologies. The paper provides a detailed typology of cases for applying machine learning strategies in social sciences, covering applications from sentiment analysis to cross-cultural studies. The research concludes by addressing methodological considerations and ethical implications for future research, emphasizing the importance of balancing technological innovation with research integrity and social responsibility.

***K*eywords** Semantic Analysis · Natural Language Processing · Machine Learning · Computational Social Science · Research Methodology

1 Introduction

1.1 Background and Motivation

The digital era has ushered in unprecedented access to a vast array of publicly available documents, research outputs, and governmental records. This surge in accessible information has significantly influenced the social sciences, where qualitative research has become increasingly prominent. Researchers now have the opportunity to explore diverse data sources, including social media content, policy documents, and extensive interview transcripts, to gain deeper insights into complex social phenomena. The integration of computational techniques and specialized software, such as ATLAS.ti, has further enhanced the capacity to analyze qualitative data systematically and efficiently. These tools facilitate the organization, coding, and interpretation of large datasets, enabling researchers to uncover patterns and themes that might remain hidden through manual analysis alone (Halford and Savage, 2017; Woods et al., 2016).

The advent of big data has transformed the landscape of social science research. The sheer volume, speed, and diversity of data generated daily present both opportunities and challenges for researchers. Traditional qualitative methods, while rich in contextual understanding, often struggle to manage and interpret such large-scale datasets. In this context, quantitative semantic analysis has emerged as a pivotal technique, offering systematic and replicable methods to examine text data. By employing computational approaches, scholars can uncover latent patterns and derive insights that may be difficult or time-consuming to achieve through traditional qualitative methods alone. This advancement not only enhances the analytical depth of social science research but also broadens the scope of inquiries into societal behaviors, opinions, and trends (Roberts, 2000; Segev, 2021).

Incorporating methods from fields such as natural language processing (NLP), machine learning, and statistics enables social science researchers to approach qualitative data in ways that complement traditional approaches. This interdisciplinary integration facilitates a richer understanding of textual data, encompassing diverse sources from social media posts and interview transcripts to policy documents and historical texts. Rather than replacing qualitative rigor, computational methods offer a paradigm that supports the examination of data on an unprecedented scale, where patterns, repetitions, and structures within discourse become more apparent. Quantitative semantic analysis bridges the gap between computational power and humanistic inquiry, promoting cross-disciplinary collaborations and introducing innovative methodologies into the social sciences. This paper explores this evolving landscape, tracing the development and applications of these techniques and demonstrating their transformative impact on social science research (Scholz, 2019; Sikstrom and Garcia, 2020).

The primary objective of this work is to highlight the relevance and contributions of quantitative semantic analysis tools in social sciences. By examining the evolution of these methods, from early statistical approaches to modern NLP and large language models, this article aims to provide readers with a comprehensive understanding of how computational analysis has enriched the study of social phenomena. Furthermore, it seeks to establish a typology of cases where different machine learning strategies can be applied within social science research, offering methodological insights that can guide future investigations in this field.

This work is structured to guide readers through the significance, historical development, applications, and considerations of quantitative semantic analysis and NLP in social sciences. Following the introduction, Section 2 discusses the broader relevance of semantic analysis in the social sciences, focusing on its role in enhancing data interpretation, advancing methodological rigor, and enabling interdisciplinary research. Section 3 provides a historical overview of the evolution of semantic analysis tools, from foundational statistical methods to recent advancements in large language models. Section 4 presents a typology of cases where machine learning and NLP techniques can be applied, with examples that illustrate their impact across social science disciplines. Section 5 addresses methodological considerations, covering data collection, model evaluation, and ethical implications. Finally, Section 6 concludes with a summary of findings, implications for future research, and reflections on the ongoing impact of computational methods in social sciences.

1.2 The relevance of Quantitative Semantic Analysis and NLP in Social Sciences

Quantitative semantic analysis has become crucial for processing and interpreting the vast and intricate nature of data within social science research, especially as the field encounters increasingly large and complex textual datasets. The proliferation of public and digital resources—spanning historical archives, government documentation, academic repositories, and cultural collections—has expanded the scope of accessible data. In earlier eras, social science relied heavily on small, curated samples due to limitations in data availability and manual processing capacity. With advancements in digital technology, however, social science researchers now have access to not only traditional forms of qualitative data but also extensive digital repositories containing unstructured textual data, such as online news archives, public forums, and digitized records, as well as original fieldwork results and literature analysis, all critical for understanding historical and contemporary social dynamics (Blei, 2012; DiMaggio et al., 2013; Kataishi and Milia, 2024).

The emergence of quantitative semantic analysis has enabled researchers to approach these large datasets with technical rigor, deploying systematic and scalable methods to analyze text-based data. Historically, such analysis required significant manual effort, including

meticulous coding and thematic categorization, which constrained both the depth and breadth of research. The development of computational techniques—like probabilistic topic modeling—has transformed this process, allowing for automated identification of latent themes and structures across extensive corpora. For instance, topic modeling can reveal recurring discourse structures within policy documents over decades, shedding light on shifts in public priorities or ideological leanings. Additionally, quantitative semantic techniques have allowed researchers to explore longitudinal changes in public and institutional language, providing insight into societal transformations on a broad scale (Roberts et al., 2019).

Another significant advancement in quantitative semantic analysis lies in its ability to extract and interpret hidden patterns within large datasets, enabling researchers to understand complex social narratives at a granular level. Beyond thematic analysis, modern approaches leverage techniques like word embeddings, which represent words in a continuous vector space and capture their nuanced contextual meanings within a text corpus. This technical development has been crucial in fields like cultural sociology and political communication, where understanding the subtle connotations of language—such as shifts in political discourse or evolving representations of social issues—relies on capturing the interplay of terms within their specific contexts. By facilitating pattern detection across extensive and diverse datasets, quantitative semantic analysis complements traditional methods, allowing social scientists to explore historical and emergent themes on a scale previously unattainable (Lazer et al., 2009; Blei, 2012).

Quantitative semantic analysis has significantly bridged methodological gaps between computational sciences and social sciences, revolutionizing collaborative efforts to gain nuanced insights into social phenomena. Traditionally, social sciences relied on qualitative approaches that, while rich in interpretative depth, provided a broad overview of social realities through human-centered coding and thematic interpretation. For decades, this approach remained relatively unchanged, with the sensibility embedded in such analyses largely left to individual researchers' interpretations. This reliance on qualitative methods was partly due to the lack of accessible computational resources and a perceived incompatibility between the nuanced, interpretive nature of social science research and the more rigid, algorithmic structure of early computational methods. However, advancements in quantitative semantic analysis have provided social scientists with scalable, systematic ways to explore these complexities (Kataishi et al., 2023; Scholz, 2019).

The development of computational power and storage capacity over recent decades has paralleled the evolution of algorithmic approaches in semantic analysis. Early methods, such as keyword frequency and simple co-occurrence analyses, were limited by computational constraints but laid the groundwork for more sophisticated models. With the exponential increase in computational power, particularly in the last decade, algorithmic approaches like topic modeling, sentiment analysis, and network analysis have become increasingly

feasible and effective for processing vast textual datasets. Tools such as R, which initially dominated text analysis, provided a robust statistical environment but required significant manual coding for complex models. The more recent transition to Python has catalyzed adoption among social scientists due to its extensive libraries (such as NLTK, spaCy, and Gensim) and an easier learning curve, allowing researchers to integrate advanced NLP techniques more seamlessly (Hovy, 2022).

While these advances have enabled large-scale qualitative analysis, they also introduced challenges. Social scientists have expressed concern that algorithmic methods may lack the nuanced sensibility central to qualitative research. Traditional approaches allowed for interpretive flexibility, as researchers could adjust their analyses in real-time based on subjective insights or emerging themes. Algorithmic methods, however, often operate within predefined models that may overlook subtle cultural or contextual variations (Kataishi and Milia, 2024; Musolino et al., 2023). This shift has created a tension between the breadth offered by computational techniques and the depth of interpretative richness that characterizes qualitative analysis. Despite these challenges, improvements in algorithmic sensitivity and adaptability have increasingly aligned quantitative semantic analysis with qualitative goals. By enabling researchers to interact with data on a large scale while maintaining the capacity for detailed interpretation, these methods have expanded the methodological toolkit of social scientists, making it possible to investigate questions that would have been unmanageable just a decade ago (Scholz, 2019).

Quantitative semantic analysis has enabled the development of analytical frameworks that are both replicable and transparent, addressing long-standing issues in qualitative social science research. Traditional qualitative methods, though valuable for their depth and interpretative richness, have often been criticized for lacking replicability due to their subjective nature. The integration of computational methods allows for more systematic tagging, categorization, and interpretation processes that can be reproduced across studies, thus enhancing the reliability of findings. This replicability has allowed social scientists to construct transparent analytical frameworks that can be applied consistently, offering a level of methodological rigor previously difficult to achieve in qualitative research (Blei, 2012; DiMaggio et al., 2013).

A challenge arises, however, with the use of tagging approaches in qualitative discretional analysis. Tagging techniques often require researchers to make decisions about which segments of text to categorize or emphasize, a process that can introduce subjectivity even within an ostensibly objective framework. This discrepancy between the research hypothesis and the tagged text or fieldwork results can lead to a partial view of the discourse, as tagging typically captures specific elements while potentially overlooking broader contextual meanings. For example, a tagged set of sentiments or themes within a corpus might miss underlying, nuanced connections that a qualitative researcher would notice. While these

techniques provide a structured approach to large-scale analysis, they may reduce the depth of discourse representation, highlighting only parts of the discourse rather than describing it entirely (Kataishi et al., 2023; Lazer et al., 2009).

Furthermore, automated tagging and analysis methods help to reduce researcher bias, which can influence the interpretative process in qualitative research. By relying on algorithms to perform initial categorizations or to detect patterns within data, researchers can mitigate personal biases that might shape how themes or sentiments are identified. Automated analysis also allows for an unbiased and consistent approach to tagging, ensuring that similar data segments are categorized in the same way across different datasets. This alignment of qualitative and quantitative methods through automation provides a balanced methodology that combines the interpretative depth of qualitative analysis with the objectivity of computational techniques, creating a complementary approach that enhances both validity and reliability (Blei, 2012; Geuna et al., 2015).

2 Historical Evolution of Quantitative Semantic Analysis Tools

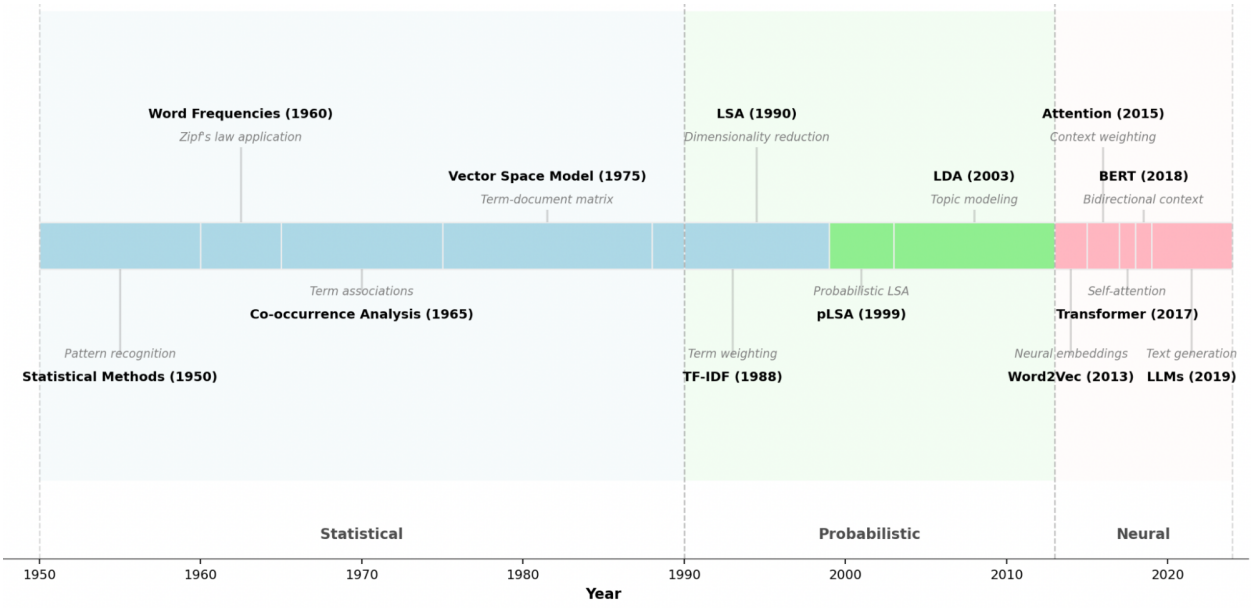
The evolution of quantitative semantic analysis tools in the social sciences can be traced back to foundational statistical methods that laid the groundwork for contemporary computational text analysis. Early statistical techniques in text analysis involved basic word frequency counts and co-occurrence analysis, methods that allowed researchers to detect patterns and commonly occurring terms within text corpora. These techniques provided a structured yet simplistic view of textual data, suitable for analyzing shorter documents with predictable vocabulary. For instance, in early content analysis, researchers manually counted words or themes to analyze public speeches or political texts, an approach that eventually evolved with the advent of computational methods (Kataishi and Milia, 2024).

The development of early NLP methods brought a deeper level of syntactic and semantic understanding to text analysis, shifting from simple word counting to parsing sentence structures and extracting meaning. These involved techniques that segmented sentences into their grammatical components (like nouns, verbs, and adjectives), enabling a more detailed analysis of linguistic structures. Syntactic analysis allowed researchers to analyze sentence-level relationships, providing insights into how language conveyed meaning beyond individual words. Semantic processing followed, focusing on identifying entities, relationships, and the intended meaning of phrases within larger linguistic contexts. This evolution introduced a higher level of interpretative potential in text analysis and began to meet the needs of social sciences research, which demanded context-sensitive insights into language use (Jurafsky and Martin, 2000).

Alongside these advancements, rule-based NLP systems emerged, leveraging pre-defined linguistic rules for applications in information retrieval and extraction. These systems were

designed to recognize certain linguistic patterns and structures within text data, allowing them to categorize information or retrieve relevant data snippets based on keywords or phrases. For example, rule-based approaches were instrumental in developing early search engines, which relied on fixed patterns to match user queries with indexed documents. However, while rule-based systems could handle specific, well-defined tasks, they struggled with ambiguity and lacked flexibility when dealing with unstructured text data, where variations in language use and complex sentence structures are common. This limitation made it challenging for social scientists to apply rule-based NLP systems in nuanced analyses of diverse text sources, such as interviews or open-ended survey responses, where language often deviates from formal structures (Hearst, 1999).

Figure 1: **Technological Trajectories of NLP in a Timeline.**



Source: Own elaboration

The evolution of statistical methods in text analysis, spanning from the 1950s through the 1980s, laid the groundwork for modern quantitative approaches. Early statistical and probabilistic approaches, while limited by the computational power of their time, established essential concepts in text analysis that paved the way for more sophisticated NLP techniques. The primary limitation of these methods was their inability to handle the complexity and vastness of unstructured text data. Nevertheless, their development represents a crucial phase in the evolution of quantitative semantic analysis tools, bridging manual content analysis and the modern, algorithm-driven NLP frameworks used today. These early approaches laid the foundation for the subsequent emergence of machine learning-based NLP systems, which would ultimately expand the social sciences' ability to

analyze large-scale text data with unprecedented depth and accuracy (Geuna et al., 2015; Jurafsky and Martin, 2000).

This era began with basic word frequency analyses and progressed to the introduction of the Vector Space Model in 1975, which conceptualized textual data within a multi-dimensional space, allowing for the quantification of semantic relationships. This advancement was further refined with the development of Term Frequency-Inverse Document Frequency (TF-IDF) in 1988, a technique that assesses term importance by considering both individual document frequency and overall corpus frequency (Salton, 1975; Sparck Jones, 1972; Aggarwal and Zhai, 2012).

The advent of Term Frequency-Inverse Document Frequency (TF-IDF) marked a significant milestone in text analysis by introducing a method for weighting terms in documents based on their importance. TF-IDF is a numerical statistic that reflects the relevance of a word to a document within a larger corpus. The term frequency (TF) component measures how often a word appears in a document, assuming that frequent terms are more significant to the document's content. However, high frequency alone does not equate to importance, as common words like "the" or "and" appear frequently across all documents without providing distinguishing information. To counter this, the inverse document frequency (IDF) component downweights terms that appear in many documents across the corpus, enhancing the weight of words that are unique to particular documents. Together, TF and IDF create a metric that helps researchers identify which words are most informative in distinguishing one document from another, making TF-IDF an essential tool in text analysis, search algorithms, and feature extraction (Sparck Jones, 1972).

TF-IDF's relevance is further underscored when compared with other term-weighting schemes. Unlike binary term weighting, where a term is merely marked as present or absent in a document, or simple frequency counts, which lack contextual discrimination, TF-IDF offers a balanced approach. It not only considers the frequency of terms but also accounts for their specificity within the corpus. This unique feature has made TF-IDF highly effective for information retrieval tasks, where the goal is to match user queries with relevant documents in a database. In these scenarios, terms that uniquely define documents are prioritized, allowing for more accurate and context-aware retrieval results. In addition to information retrieval, TF-IDF has been widely used in clustering and classification tasks, where it serves as a feature extraction technique to represent textual data in a format that machine learning models can process (Ramos, 2003).

The 1990s marked a pivotal shift in semantic analysis with the introduction of Latent Semantic Analysis (LSA) by Deerwester et al. (1990). LSA represented an early, algebraic approach to extracting hidden, or "latent," structures within text data. By employing a technique called Singular Value Decomposition (SVD), LSA allowed for the mapping

of terms and documents into a lower-dimensional semantic space, with similar themes appearing closer together. This dimensionality reduction enabled researchers to analyze thematic patterns across large collections of text without needing extensive labeling or tagging, offering a systematic view of relationships between words and concepts. However, LSA's linear approach did not capture the probabilistic nature of language, limiting its flexibility and interpretative depth in complex, real-world applications.

Probabilistic Latent Semantic Analysis (pLSA) emerged in 1999 as a significant refinement of LSA (Hofmann, 1999) introducing a probabilistic framework that allowed language variability across different topics, thereby increasing modeling accuracy and flexibility. This approach represents each document as a mixture of underlying topics, with each word assigned a probability within these topics. Despite these advantages, it lacked a mechanism to generalize the approximation for non-pretrained corpora, limiting its application to small, contained datasets.

Building upon pLSA, Blei et al. (2003) introduced Latent Dirichlet Allocation (LDA), establishing a new standard in topic modeling. LDA incorporated a hierarchical Bayesian framework, using Dirichlet distributions, which allowed the creation of a generative model capable of assigning topics to previously unseen documents. In this model, each document is considered a mixture of multiple topics, with each topic defined by a distribution over words. The use of Dirichlet priors added interpretability and robustness, producing coherent and well-defined topics that could be analyzed across both new and existing documents. LDA's generative framework enabled the model to scale to large datasets effectively, making it suitable for a wide range of applications, from academic research to industry use cases, such as content recommendation and information retrieval. LDA's impact has been profound, particularly in social science research, where it facilitates the systematic discovery of hidden patterns in discourse, policy documents, and social communications. Recent studies have demonstrated LDA's efficacy in analyzing large-scale textual data, underscoring its enduring relevance in contemporary research (Jelodar et al., 2019).

While TF-IDF provides a valuable means of feature extraction, it is limited in its capacity to uncover deeper, thematic structures within a text corpus. Latent Dirichlet Allocation (LDA) represents a transformative shift towards probabilistic topic modeling, allowing researchers to identify "hidden" topics within large corpora, with each document potentially associated with multiple topics at varying proportions. For instance, an article about "renewable energy" might encompass topics like "climate change," "policy," and "technology," with each topic contributing differently to the document. LDA enables this layered understanding of text, making it a powerful tool for thematic analysis in vast collections of documents (Blei et al., 2003).

LDA's introduction of probabilistic graphical models marked a shift in the approach to topic modeling. Rather than simply grouping documents by shared words, LDA employs Bayesian inference to generate topic distributions across a corpus, treating each document as a blend of topics rather than a single thematic entity. This probabilistic framework allows for more flexibility in capturing the thematic structure of documents, as it accounts for the variability and overlap of topics within a text. Graphical models in LDA offer a visual and mathematical representation of this structure, enhancing interpretability and providing researchers with insights into the relationships between words, topics, and documents. These graphical models allow researchers to see how topics evolve over time or vary across sources, making LDA suitable for longitudinal studies and comparative analyses (Griffiths and Steyvers, 2004).

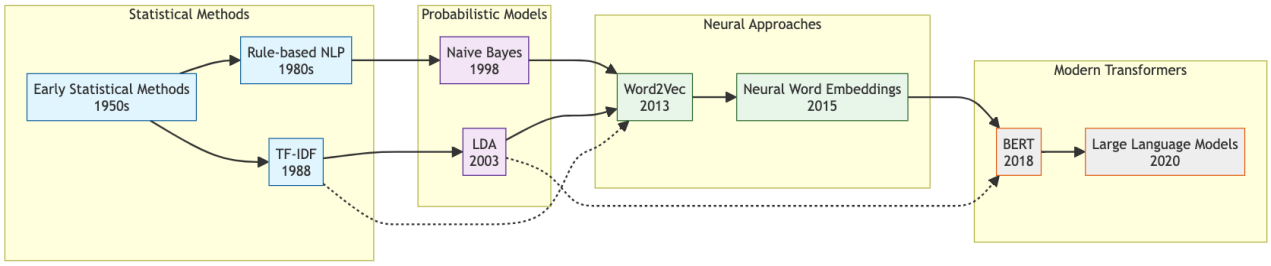
In textual corpora summarization, both TF-IDF and LDA have proven effective, though they approach the task differently. TF-IDF can be used in extractive summarization, where sentences containing high-weighted terms are selected as representative of the document. This approach is straightforward yet effective for summarizing single-topic documents, such as news articles. LDA, however, is more suited for abstractive summarization, generating a concise overview by identifying key topics and phrases representative of the content. This method is particularly useful for multi-topic documents, as it can provide an overview that captures the range of themes within a text. As a result, both TF-IDF and LDA are valuable in quantitative semantic analysis, each bringing unique strengths to textual data analysis (Allahyari et al., 2017).

The progression from algebraic and probabilistic models to fully generative frameworks like LDA marked a pivotal achievement in machine learning for semantic analysis, enabling more accurate and interpretable insights from increasingly vast datasets. However, while these probabilistic methods excelled at identifying thematic structures and latent topics, they still struggled to capture the deeper, context-dependent aspects of language. Recognizing that understanding language requires models capable of representing not only static word meanings but also the subtle shifts that occur depending on context, researchers began developing models focused on dynamic qualities of language.

This shift led to the application of neural networks in semantic analysis, beginning with the introduction of Word2Vec by Mikolov et al. (Mikolov et al., 2013), which marked a fundamental advancement in semantic analysis through embeddings and neural networks. Unlike previous statistical and probabilistic methods, embeddings provide high-dimensional vectorial representations capturing complex relationships based on word context. Each word represents a vector value in a multidimensional plane, enabling the model to process different meanings of a word according to its context. This approach laid the foundation for ELMo, which introduced fully contextualized embeddings that could adjust each word's meaning depending on its sentence-level context (Peters et al., 2018). This new capability

to capture context-aware meaning provided unprecedented depth in analyzing extensive textual data, enabling real-world applications like social media analysis.

Figure 2: **Relations by technology families in NLP**



Source: Own elaboration

The current era represents an enormous difference from previous stages of NLP and ML, and marks the most advanced phase in the evolution of semantic analysis to date. The development of attention mechanisms in 2015 enabled models to focus on relevant parts of input text, leading to the breakthrough Transformer architecture (Vaswani et al., 2017). This architecture's strength lies in its ability to process text in parallel while maintaining attention to relationships between all words in a sequence. The subsequent development of BERT and the GPT family represents a substantial leap in semantic analysis capabilities, enabling unparalleled depth not only in text comprehension but also in text generation.

For social science research, this technological evolution transcends mere methodological innovation. It represents a foundational transformation in how researchers can approach qualitative analysis. The progression from simple word counting to sophisticated neural language models has enabled researchers to analyze discourse on unprecedented scales while preserving analytical depth. This capability is particularly crucial for understanding complex social phenomena, from policy impact analysis to social movement dynamics, where the volume and complexity of textual data previously posed significant methodological challenges.

The integration of these advanced semantic analysis methods into social science research has enabled a more rigorous approach to qualitative analysis. Researchers can now systematically analyze vast corpora of text while maintaining the nuanced understanding traditionally associated with close reading and manual coding (Lazer et al., 2009; Mikolov et al., 2013; Devlin et al., 2019). This blend of computational power and qualitative sensitivity has opened new research avenues, especially in areas where large-scale textual analysis was previously impractical, such as social media discourse analysis, policy document evaluation, and public opinion research.

Moreover, this methodological evolution has improved the reproducibility and transparency of qualitative research. The systematic nature of these computational approaches, combined with their ability to process large datasets, enables researchers to validate findings and test hypotheses on scales previously unattainable. This advancement does not reduce the value of traditional qualitative methods but rather complements them by providing additional tools for triangulation and validation.

3 The Relevance of Quantitative Semantic Analysis and NLP in Social Sciences

3.1 From Content to Context: Evolution of Thematic and Topic Analysis

Topic content analysis represents one of the foundational approaches in qualitative research, marking a significant shift from simple content counting to sophisticated contextual understanding. This evolution reflects broader changes in how researchers approach textual data analysis, moving from surface-level examination to deep interpretative frameworks. The method enables researchers to examine underlying meanings, concepts, and relationships embedded in text, transcending basic word analysis to capture the nuanced essence of participants' perspectives. This approach, deeply rooted in social sciences, emerged as a response to the need to organize and interpret complex, unstructured data in a systematic yet flexible manner, making it adaptable to diverse research contexts and questions. Unlike purely quantitative approaches focused on frequency metrics or co-occurrence patterns, thematic analysis prioritizes interpretative depth and contextual understanding, aligning with the exploratory nature of social sciences research (Braun and Clarke, 2006; Guest et al., 2012).

The integration of computational approaches has transformed traditional thematic analysis while preserving its essential interpretative character. Initially conducted through manual examination, where researchers meticulously reviewed texts to identify patterns and themes relevant to research questions, the field has evolved to incorporate sophisticated computational tools. This technological progression became particularly crucial as digital datasets grew in volume and complexity, making exhaustive manual review increasingly impractical. The emergence of specialized software solutions like ATLAS.ti, NVivo, and RQDA has enabled a hybrid approach, combining qualitative sensitivity with computational efficiency. These tools introduce systematic coding capabilities that allow researchers to process large volumes of text while maintaining analytical rigor and ensuring reproducibility in data handling (Guest et al., 2012).

A pivotal development in this evolution has been the sophistication of coding and indexing systems, which have progressed from basic manual frameworks to advanced computer-assisted models supporting both deductive and inductive approaches. Early methodologies

relied heavily on researcher-developed frameworks for manual theme indexing, establishing structured conceptual taxonomies from raw data. The digital revolution has transformed these systems into dynamic, adaptable frameworks capable of evolving with emerging research questions and newly identified themes. This flexibility enables researchers to combine deductive approaches, where predefined categories guide analysis, with inductive techniques that allow themes to emerge organically from the data. Such methodological advancement has proven particularly valuable in fields like cultural research and political discourse analysis, where rigid categorization schemes often prove inadequate for capturing complex social phenomena (Saldaña, 2015).

The role of thematic content analysis in mixed-methods research exemplifies its evolution from a purely qualitative tool to an integrative methodological bridge. In contemporary research designs, it serves as a crucial link between qualitative interpretation and quantitative measurement, enabling researchers to triangulate findings across approaches. This integration extends to the development of quantitative instruments, where thematic analysis informs the creation of survey items and coding schemes grounded in qualitative insights. The advent of machine learning algorithms has further enhanced this integrative capacity, enabling rapid preliminary analysis of extensive datasets while maintaining qualitative depth. This technological augmentation has made it possible to conduct comprehensive analyses of large-scale textual data without sacrificing the nuanced understanding central to qualitative research (Creswell and Plano Clark, 2017; McCusker and Gunaydin, 2015).

The practical applications of evolved thematic analysis span numerous disciplines, demonstrating its versatility as an analytical tool. In media studies, researchers employ this method to uncover patterns in news coverage, social media discourse, and entertainment content, revealing how various narratives construct and disseminate social meanings. For instance, researchers have successfully mapped media representations of social movements, providing crucial insights into public opinion formation and theme resonance across different audiences. In policy analysis, thematic approaches help decision-makers understand public sentiment and societal concerns, complementing quantitative polling data with rich qualitative insights. This application proves particularly valuable in analyzing public response to policy initiatives and government programs, facilitating deeper understanding of community needs and priorities (Boyatzis, 1998; Vaismoradi et al., 2013).

The method's application in cultural research further demonstrates its evolutionary sophistication, particularly in analyzing cultural texts, rituals, and social practices. Contemporary thematic analysis enables researchers to decode complex symbolic meanings embedded within cultural data, moving beyond literal interpretations to uncover deeper societal signifiers. The integration of automated analysis capabilities has expanded this potential, allowing researchers to examine vast cultural datasets, including digital archives and social media repositories, while maintaining interpretative depth. This technological enhancement,

combined with traditional interpretative methods, has created a robust analytical framework capable of processing large-scale textual data while preserving the nuanced understanding essential to cultural analysis. The resulting approximation provides a versatile and adaptive tool for examining textual data across various social science disciplines, representing a significant advancement in qualitative research (Clarke and Braun, 2013; Guest et al., 2012).

3.2 SNA Approaches applied to Semantic Relations

Semantic network analysis emerged as a transformative approach in text analysis, representing a fundamental shift from isolated word analysis to understanding complex relational structures within textual data. This methodology extends beyond basic frequency analysis by examining how words and phrases co-occur and form meaningful semantic associations, making use of social network analysis representation and interpretational background (Geuna et al., 2015). The visualization and quantification of these relationships through network representations, where nodes signify concepts and edges represent their interconnections, enables researchers to uncover the underlying structure of discourse within texts. This network-based perspective proves particularly valuable in revealing complex patterns and relationships that might remain obscured through conventional analytical methods, offering a sophisticated framework for understanding how meaning emerges from textual relationships (Carley, 1997; Doerfel, 1998; Kataishi and Brixner, 2023).

The application domain of semantic network analysis has expanded significantly, particularly in fields requiring deep understanding of semantic structures and discourse evolution. Through systematic examination of textual patterns and relationships, researchers can map the architecture of discourse, track ideological development, and analyze concept linkages across various contexts. In organizational studies, this approach has proven invaluable for analyzing corporate communications, revealing how organizations embed and emphasize specific values within their messaging frameworks. Political discourse analysis has similarly benefited, employing semantic networks to track the evolution of ideological concepts across policy documents and political speeches, illuminating the strategic framing choices of political actors. The method's utility extends to broader social discourse analysis, where it helps researchers understand how complex topics like climate change or social justice are conceptualized and discussed in public spheres (Diesner and Carley, 2005).

The development and implementation of co-occurrence networks marks a crucial advancement in semantic network analysis, providing a sophisticated method for pattern identification based on word proximity relationships. These networks form when words consistently appear together within defined contextual boundaries, such as sentences, paragraphs, or documents. The resulting patterns of co-occurrence reveal implicit semantic relationships, enabling researchers to uncover hidden associations between concepts. When terms like "economy" and "policy" frequently co-occur in political discourse, for instance, this pat-

tern suggests a significant thematic emphasis on economic policy matters. Co-occurrence networks prove particularly valuable in discourse mapping, highlighting central versus peripheral concepts within texts and enabling systematic analysis of thematic structures. This network-centric approach has substantially enhanced researchers' ability to detect and analyze prominent themes within large text corpora (Danowski, 1993; Geuna et al., 2015).

Graph theory provides the mathematical foundation for quantifying and analyzing relationships within semantic networks. Within this framework, various metrics such as centrality, density, and modularity offer precise measures of conceptual relationships and influence within the network structure. Centrality measurements identify key terms that function as conceptual hubs within the discourse, while density metrics evaluate the overall connectedness of the semantic network. Modularity analysis reveals distinct thematic communities within the broader network, showing how different concept clusters form around central themes. These graph-theoretic tools enable researchers to conduct rigorous quantitative analysis of discourse structures, providing insights into the organization and relationships of themes within complex texts (Freeman, 1979; Borgatti and Everett, 1997).

The practical applications of semantic network analysis across various case studies demonstrate its effectiveness in understanding organizational, political, and social discourse dynamics. In organizational contexts, researchers employ these techniques to examine how corporate communications reflect strategic priorities and stakeholder relationships. Political research utilizes semantic networks to investigate issue framing in policy debates, analyzing how different political actors construct and connect concepts around contested topics. In social media analysis, these techniques help track the evolution of public discourse on various issues, identifying emergent themes and opinion shifts over time. These diverse applications highlight the method's versatility in capturing subtle linguistic patterns and meaning structures across different analytical contexts (Knoke and Yang, 2008; Hanneman and Riddle, 2005).

3.3 Mathematical Foundations: Dimensionality and Text Representation

Dimensionality reduction techniques constitute a fundamental mathematical framework in quantitative semantic analysis, addressing the crucial challenge of managing and interpreting high-dimensional textual data. The process begins with vector representation of documents, where each unique word or feature in the vocabulary corresponds to a dimension in a high-dimensional space. This vectorization process, while essential for enabling mathematical analysis of text, creates computational and interpretative challenges due to the "curse of dimensionality"—where data becomes increasingly sparse and difficult to analyze as dimensions increase. Dimensionality reduction techniques provide mathematical solutions to these challenges by transforming high-dimensional data into lower-dimensional

representations while preserving essential structural relationships (Aggarwal and Zhai, 2012).

Principal Component Analysis (PCA) represents one of the foundational mathematical approaches to dimensionality reduction in text analysis. This technique operates by identifying orthogonal directions (principal components) that capture maximum variance in the data. By projecting high-dimensional data onto these principal components, PCA creates a lower-dimensional representation that retains the most significant patterns while discarding less informative variations. In text analysis applications, PCA proves particularly valuable for reducing the dimensionality of term-frequency matrices and other vector-based document representations. However, its linear nature poses limitations when dealing with complex semantic relationships that often exhibit non-linear characteristics in natural language (Jolliffe and Cadima, 2016).

The evolution of dimensionality reduction techniques has led to sophisticated non-linear approaches, notably t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). t-SNE innovates by modeling probabilistic relationships between data points in both high and low-dimensional spaces, attempting to preserve these relationships during dimensionality reduction. This approach excels in visualization tasks, creating clear cluster separations that prove invaluable for identifying thematic groups within text data. However, t-SNE's computational intensity and parameter sensitivity can affect result stability and interpretation. UMAP advances this field by offering enhanced computational efficiency while maintaining the ability to preserve both local and global data structure. Its mathematical framework, based on manifold learning and topological data analysis, provides more stable and interpretable results, particularly beneficial for large-scale text analysis applications (Van der Maaten and Hinton, 2008; McInnes et al., 2018).

Comparative analysis of t-SNE and UMAP reveals their distinct mathematical properties and practical implications for text analysis. While t-SNE excels in revealing detailed local structure and non-linear relationships, its focus on local relationships can sometimes lead to misleading global representations. UMAP addresses these limitations through its mathematical foundation in manifold theory, enabling better preservation of both local and global data relationships. This balanced approach proves particularly valuable in social sciences research, where understanding both detailed thematic relationships and broader conceptual structures is crucial. The mathematical sophistication of these methods allows researchers to explore complex semantic relationships while maintaining interpretability (Becht et al., 2019).

The practical applications of dimensionality reduction in text analysis span various analytical tasks, from clustering and topic modeling to visual data exploration. These techniques

enhance clustering performance by simplifying complex data structures while preserving meaningful relationships. In social media analysis, for example, these methods enable effective clustering of user-generated content, facilitating the identification of emergent themes and patterns. The reduction of high-dimensional text data to visualizable dimensions enables researchers to explore and communicate complex semantic relationships effectively. These mathematical approaches have become essential tools in quantitative semantic analysis, supporting tasks ranging from corpus summarization to thematic analysis while maintaining data structure integrity (Kobak and Berens, 2019).

3.4 The Neural Networks Revolution: Embeddings and Contextual Models

The development of neural embedding technologies, exemplified by Gensim, Word2Vec, and BERT, marks a revolutionary transformation in natural language processing and quantitative semantic analysis. This technological progression represents a fundamental shift from traditional statistical approaches to neural network-based understanding of language, introducing sophisticated methods for capturing semantic meaning and contextual relationships in text. Gensim emerged as a pioneering open-source library specifically engineered for advanced topic modeling and document similarity analysis. Built upon efficient algorithmic foundations, it provides researchers with pre-built models and tools capable of processing large text corpora and extracting latent topics, enabling the discovery of thematic patterns and document relationships without requiring extensive computational resources. This capability has proven particularly valuable in research contexts involving large-scale analysis of academic literature, news archives, and social media content (Řehůřek and Sojka, 2010).

Word2Vec's introduction by Google in 2013 represented a paradigmatic shift in how machines process and understand language, introducing the concept of neural word embeddings that represent words as vectors in a continuous semantic space. This approach marked a decisive break from traditional bag-of-words methods by capturing complex semantic relationships based on word context within sentences. The architecture implements this through two distinct but complementary models: the Continuous Bag of Words (CBOW) and Skip-Gram approaches. CBOW operates by predicting target words from their surrounding context, while Skip-Gram reverses this process, predicting contextual words from a given target word. This dual approach enables Word2Vec to capture nuanced semantic relationships and understand word meanings within their linguistic context, resulting in a vector space where semantically similar words cluster together. This capability has revolutionized applications in sentiment analysis, text classification, and semantic similarity measurement (Mikolov et al., 2013).

The emergence of BERT in 2018 marks another revolutionary advancement through its introduction of contextualized word representations, where word meanings are dynamically

determined by their complete sentence context. BERT's innovation lies in its bidirectional transformer architecture, which processes text by simultaneously considering both preceding and following context. This capability represents a significant advancement over previous models, enabling BERT to capture complex, context-dependent meanings and handle linguistic phenomena like polysemy and ambiguity with unprecedented accuracy. The model's pre-training on extensive text corpora, followed by task-specific fine-tuning, has established new performance standards across numerous NLP benchmarks. BERT's architectural innovations have catalyzed the development of increasingly sophisticated language models, including GPT and RoBERTa, continuously expanding the boundaries of contextual language understanding (Devlin et al., 2019).

The practical impact of Word2Vec's semantic understanding capabilities has been particularly profound in applications requiring sophisticated semantic analysis, such as recommendation systems and information retrieval. The model's ability to embed words in a semantically meaningful vector space enables powerful similarity matching operations, where the semantic proximity of terms can inform content recommendations and information organization. In e-commerce contexts, for example, Word2Vec facilitates product recommendations based on semantic similarities in product descriptions, enhancing user experience through more intuitive and contextually relevant suggestions. The technology's effectiveness in knowledge extraction tasks stems from its ability to recognize semantic relationships and synonymous terms, enabling more accurate entity and relation extraction from unstructured text. This capability has proven particularly valuable in specialized domains like medical research and legal analysis, where precise understanding of technical terminology and relationships is crucial (Le and Mikolov, 2014).

The applications of neural embeddings span diverse analytical tasks, from sentiment analysis to knowledge extraction systems. In sentiment analysis applications, embeddings derived from these models can detect subtle variations in sentiment by considering the contextual usage of words, providing particularly valuable insights for analyzing customer feedback and social media content. Recommendation systems benefit from both Word2Vec and BERT's contextual understanding capabilities, with BERT's superior contextual processing offering enhanced accuracy in suggesting relevant content. In knowledge extraction applications, these embedding technologies enable systems to understand complex relationships between terms in specialized datasets, facilitating automated information extraction and organization. These technological advances have become fundamental tools in contemporary NLP and text analytics, providing researchers with sophisticated, scalable methods for extracting meaningful insights from textual data (Řehůřek and Sojka, 2010; Mikolov et al., 2013; Devlin et al., 2019).

3.5 The Transformative Era of Large Language Models

The emergence of Large Language Models (LLMs) represents a crucial moment in natural language processing, introducing systems capable of understanding and generating human-like text with unprecedented sophistication. These models, trained on massive textual datasets of digitally available information, demonstrate remarkable capabilities in pattern recognition, contextual understanding, and meaning inference that surpass previous technological boundaries. Through the implementation of extensive neural network architectures incorporating billions of parameters, LLMs execute complex language tasks including translation, summarization, and text generation with extraordinary accuracy. Their ability to generate coherent, contextually appropriate language has catalyzed transformative applications across diverse fields, establishing LLMs as essential tools for advanced text analysis in domains ranging from social sciences to medicine and law (Brown et al., 2020; Kaplan et al., 2020).

BERT and its successor transformer-based models, particularly GPT, stand among the most influential developments in the LLM landscape. BERT's introduction of bidirectional transformer architecture in 2018 marked a significant advancement, enabling models to process contextual information from both preceding and following text simultaneously. This innovation substantially improved the handling of linguistic complexities, including polysemy and contextual ambiguity. The transformer architecture underlying BERT established a foundation for subsequent models like GPT, which expanded these capabilities through increased model scale and more extensive pre-training datasets. These developments have continuously pushed the boundaries of language model capabilities, establishing transformers as the cornerstone of modern NLP applications (Devlin et al., 2019; Vaswani et al., 2017).

The impact of LLMs on semantic analysis has been particularly revolutionary in social sciences research, enabling the processing of complex textual data at unprecedented scales while maintaining analytical depth. These models facilitate sophisticated analysis of discourse patterns, sentiment variations, and thematic structures across extensive corpora, including social media content, policy documents, and historical texts. LLMs have enhanced researchers' ability to automate qualitative coding processes, identify subtle shifts in public sentiment, and uncover latent topics within large datasets, supporting more comprehensive and nuanced large-scale analyses. Their versatility has enabled innovative applications in analyzing public opinion trends and examining cultural narratives, transcending traditional limitations of manual qualitative research (Bommasani et al., 2021).

The transformer architecture stands as the foundational innovation enabling these advanced language modeling capabilities. Introduced by Vaswani et al. in 2017, transformers revolutionized NLP by enabling parallel sequence processing, departing from the sequential constraints of recurrent neural networks (RNNs). The architecture's attention mechanisms

allow models to process different parts of input text simultaneously, capturing contextual relationships across extended word sequences. This innovation facilitated the development of increasingly larger and more sophisticated models like GPT and BERT, leveraging growing computational resources and data availability. The transformer architecture's scalability and adaptability have made it suitable for diverse NLP tasks, from question answering to complex dialogue generation (Vaswani et al., 2017; Brown et al., 2020).

The deployment of LLMs in sensitive research contexts raises important ethical considerations that require careful attention. While powerful, these models can inadvertently perpetuate or amplify biases present in their training data. Models trained on internet-sourced data may reproduce societal stereotypes or misinformation, raising concerns in applications affecting human decision-making, such as hiring processes or criminal justice assessments. Privacy concerns also emerge, particularly when LLMs process personal or social media data where anonymization measures may prove insufficient. These challenges necessitate careful consideration of data curation practices, model design principles, and usage guidelines to ensure responsible deployment of LLMs in research, especially in domains with direct societal impact (Bender et al., 2021; Mitchell et al., 2019).

The future trajectory of LLM technology suggests numerous promising research directions and applications. As models become increasingly sophisticated and capable of handling multiple tasks simultaneously, researchers are exploring zero-shot learning capabilities—the ability to perform tasks without specific training. This development could lead to more versatile models requiring less domain-specific data, enhancing their applicability across various social science contexts. Ongoing research focuses on improving model interpretability, aiming to make LLM decision-making processes more transparent and understandable. Enhanced transparency could address ethical concerns and increase the trustworthiness of these models in critical applications. The evolution of LLMs continues to emphasize efficiency improvements, bias reduction, and expanded applicability, solidifying their position as crucial tools for advancing knowledge in social sciences and related fields (Bommasani et al., 2021; Raffel et al., 2020).

4 A Typology of NLP Applications in Social Science Research

The application of NLP techniques in social sciences has evolved significantly, addressing various research needs from basic text processing to sophisticated analytical tasks. This section presents a systematic classification of these applications, organizing them by complexity and methodological approach while highlighting their practical implications for social research.

Machine learning approaches to text classification and sentiment analysis have become fundamental for categorizing large volumes of textual data. These techniques enable

Table 1: **NLP Applications and Approaches in Social Science Research**

Application Type	**Key Applications**	**Methodological Approaches**
Text Classification and Sentiment Analysis	Categorizing social media posts, survey responses, and interview transcripts. Public opinion analysis and policy reception. Consumer behavior and market sentiment analysis. Economic sentiment from financial news. Labor market trends through social media and job reviews.	Text classification algorithms (Naive Bayes, SVM). Sentiment analysis models (VADER, BERT). Natural language processing libraries (TextBlob, spaCy). Deep learning approaches for complex classification.
Topic Modeling and Thematic Analysis	Identification of thematic patterns in large text collections. Analysis of policy documents and public discourse. Economic policy discourse across institutions. Monitoring regulatory trends and industry dynamics. Corporate communications and financial report analysis.	Topic modeling techniques (LDA, NMF). Thematic analysis using Gensim. Document clustering algorithms. Hierarchical topic modeling approaches. Temporal topic modeling for trend analysis.
Semantic Network Analysis	Mapping relationships between concepts and terms. Analysis of discourse structures. Knowledge organization patterns. Policy network analysis. Institutional relationship mapping.	Network analysis algorithms. Graph-based text analysis. Semantic similarity measures. Co-occurrence analysis. Social network analysis techniques.
Named Entity Recognition and Relationship Extraction	Entity identification (organizations, individuals, places). Relationship mapping between economic agents. Supply chain and market structure analysis. Regulatory framework mapping. Strategic alliance identification.	Named Entity Recognition (NER) tools (spaCy, NLTK). Relationship extraction algorithms. Dependency parsing techniques. Graph-based network analysis. Entity linking methods.
Inter-temporal Analysis of Large Text Corpora	Analysis of regulatory logs and legislative records. Diplomatic meeting transcriptions. Historical document analysis. Policy evolution studies. Literary and cultural change analysis. Longitudinal institutional studies.	Time-series text analysis. Diachronic word embeddings. Historical text processing. Temporal pattern recognition. Change point detection methods. Sequence analysis techniques.
Predictive Modeling	Economic and social trend forecasting. Policy impact assessment. Market behavior prediction. Demographic trend analysis. Consumer behavior forecasting. Economic cycle prediction. Housing market analysis.	Time-series analysis methods. Machine learning algorithms (ARIMA, LSTM). Regression models for text data. Neural network approaches. Ensemble methods for prediction. Cross-validation techniques.

Source: Own elaboration

automated sorting and labeling of text based on predefined categories, enhancing the speed and accuracy with which researchers can analyze public discourse and social opinions. In public opinion analysis, text classification helps identify the prevalence of specific themes in media discussions, while sentiment analysis assesses the emotional tone of language, allowing for the quantification of attitudes in political impressions and consumer behavior.

In the economic realm, sentiment analysis is increasingly applied to gauge economic sentiment from financial data, corporate reports, and public statements (Kataishi and Milia, 2024. By evaluating the tone and context of published content, researchers can infer broader trends in economic confidence, financial market stability, and investor behavior. This extends into forecasting applications, where sentiment-based predictions of market trends are derived from real-time textual data. The analysis of ICT diffusion (Kataishi and Barletta, 2011) labor market dynamics through job reviews and recruitment platforms provides additional insights into market trends and workforce characteristics.

Topic modeling and thematic analysis play a crucial role in uncovering prevalent themes and recurring patterns in large corpora of qualitative data (Kataishi et al., 2023;Musolino et al., 2023). Algorithms like Latent Dirichlet Allocation (LDA) help discover latent topics within unstructured text, making sense of policy documents, media content, and legislative records. In policy analysis, these techniques monitor evolving regulations and socioeconomic trends, tracking public shifts over time. Economic applications include analyzing discussions on policy across different institutions and examining corporate communications for strategic shifts (Blei et al., 2003).

Semantic network analysis and Named Entity Recognition (NER) enable researchers to map complex relationships between concepts, organizations, and individuals. These techniques are particularly valuable for understanding institutional networks, policy frameworks, and market structures. NER can automatically categorize entities within documents, while relationship extraction determines connections between them, revealing patterns of influence and association in various social and economic contexts (Carley, 1997; Doerfel, 1998).

The addition of inter-temporal analysis capabilities represents a significant advancement, particularly valuable for developing countries and resource-constrained research environments. This approach enables systematic analysis of extensive historical and contemporary documents, including regulatory logs, diplomatic transcripts, and legislative records. For researchers with limited resources for primary data collection, the ability to analyze publicly available documentation provides a cost-effective approach to conducting rigorous research. This is especially important in contexts where extensive fieldwork might be prohibitively expensive or logistically challenging.

Predictive modeling represents the most sophisticated application of these techniques, utilizing machine learning to forecast economic and social trends. By analyzing language

patterns in various data sources, like in-depth interviews (Wilkis et al., 2023) researchers can anticipate changes in economic cycles, consumer behavior, and demographic trends. While powerful, these approaches require careful consideration of methodological limitations and ethical implications, particularly when applied to social phenomena (Kaplan et al., 2020).

The integration of these applications demonstrates the potential of NLP to enhance social science research while acknowledging the importance of combining these tools with established frameworks. The emphasis on processing and analyzing existing documentary sources makes these approaches particularly valuable for researchers working with limited resources, while still maintaining rigorous academic standards. As Kataishi and Milia (2024) suggest, this methodological evolution suggests exciting possibilities for social science research, especially in contexts where traditional data collection may be constrained by resource limitations.

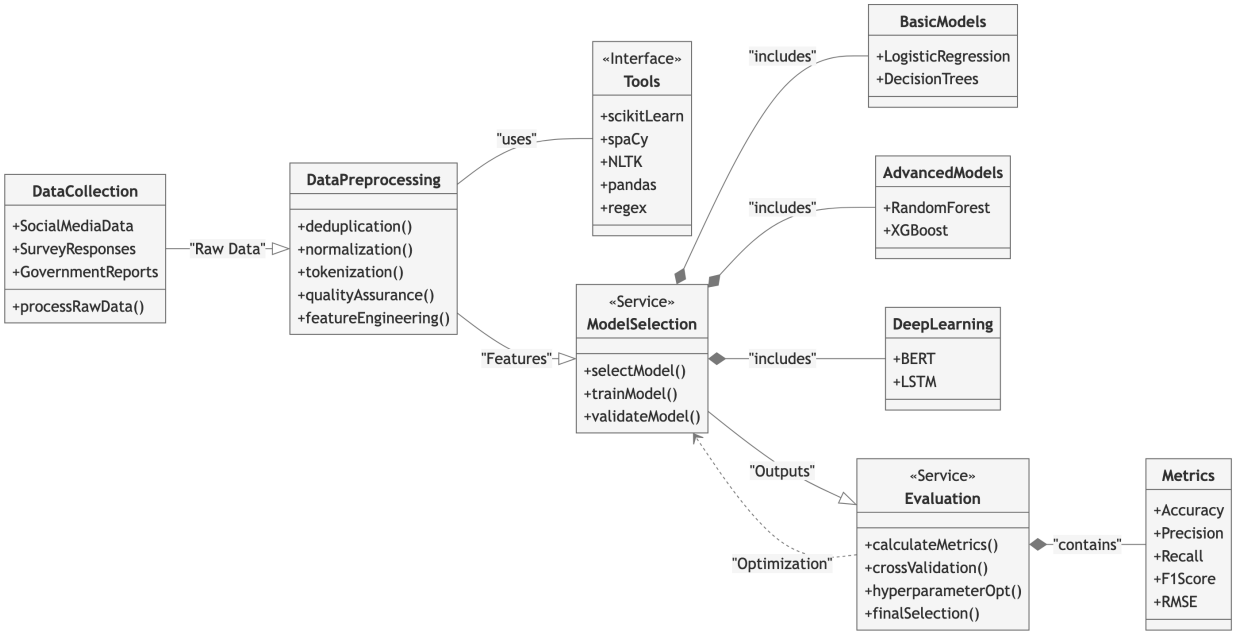
5 Methodological Considerations

The application of machine learning in social sciences requires careful attention to methodological issues that ensure the validity, reliability, and ethical integrity of research outcomes. This section outlines critical considerations in data collection, model selection, evaluation, and key practices that guide rigorous and responsible use of NLP and machine learning in social sciences.

Data collection and preprocessing are foundational steps in any machine learning workflow, this is particularly so where data often comes from diverse sources with varying structures. Obtaining large datasets presents several challenges, including restricted access to proprietary data, the high cost of data acquisition, the challenges of deploy an original fieldwork, and many other limitations regarding primary or secondary data collection. Furthermore, data gathered from sources like social media, survey responses, and government reports are often unstructured and require the key process of data setup, which imply significant preprocessing to standardize formats, filter irrelevant information, and address missing or incomplete entries. The structuration of unestructured data, then, becomes a key step in validation of the entire process. Effective data preprocessing is essential to ensure that texts are accurately represented in the dataset, and that the target phenomena and the models applied yield meaningful results.

The implementation of NLP techniques in social sciences requires careful attention to data preprocessing and quality assurance. These preparatory steps, while often overlooked in methodological discussions, are crucial for ensuring robust and replicable research outcomes. This section details the technical approaches and tools commonly employed in social science research preprocessing workflows.

Figure 3: **NLP Methods and Workflows for Social Sciences**



Source: Own elaboration

Data deduplication and normalization processes typically leverage *scikit-learn*'s `TfidfVectorizer` for computing text similarities, while *pandas* provides `DataFrame` operations for efficient duplicate removal. Text normalization combines multiple tools: *spaCy* or *NLTK* for basic text processing, the `re` library for pattern matching and cleaning, and specialized libraries like *textblob* or *transformers* for advanced normalization tasks. These tools are particularly valuable when dealing with social media data or informal text sources that require extensive cleaning.

Tokenization approaches vary based on research requirements and data characteristics. Researchers typically employ either *spaCy*'s comprehensive language models or *NLTK*'s specialized tokenizers (`word_tokenize`, `sent_tokenize`). *spaCy*'s pipeline approach offers integrated linguistic analysis, while *NLTK*'s modular toolkit provides fine-grained control over the tokenization process. For languages with specific tokenization challenges, such as Chinese or Japanese, specialized tokenizers like *jieba* or *MeCab* are often integrated into the preprocessing workflow.

Data quality assurance methods encompass several complementary approaches. Outlier detection typically employs statistical methods through *numpy* and *scipy*, or more sophisticated approaches using *sklearn*'s outlier detection algorithms such as `IsolationForest` or `LocalOutlierFactor`. Consistency checks leverage *pandas*' string methods and regular expressions, while sample balancing often utilizes *imbalanced-learn* (`imblearn`) for techniques like SMOTE or random under/over-sampling.

Feature engineering in NLP contexts combines multiple specialized tools. Linguistic features are commonly extracted using *spaCy*'s annotation system, while sentiment and subjectivity scores typically come from *textblob* or `vaderSentiment`. More sophisticated feature engineering often employs word embeddings through *gensim*'s implementation of `Word2Vec` or `fastText`, or contextual embeddings from transformer-based models accessed through the *transformers* library.

For large-scale text collections, researchers often implement out-of-memory processing using *dask* or *vaex*. These tools enable efficient handling of datasets that exceed available RAM, a common challenge when working with social media data or extensive document collections. The entire preprocessing workflow can be standardized and made reproducible using *scikit-learn*'s `Pipeline` class, ensuring consistent application of preprocessing steps across different research contexts.

Model selection represents a crucial methodological decision point in machine learning applications for social science research. The *scikit-learn* ecosystem offers a spectrum of models, each with distinct trade-offs between interpretability and predictive power. For structured data analysis, researchers often begin with interpretable models like `LogisticRegression` or `DecisionTreeClassifier`, which provide clear insights into feature importance and decision boundaries. These models, implemented through *sklearn.linear_model* and *sklearn.tree* respectively, allow for straightforward examination of coefficients and decision paths.

When dealing with text data, the transition from basic models to more sophisticated approaches often follows a clear progression. Simple but interpretable text classification might employ *sklearn*'s `MultinomialNB` or `LinearSVC`, particularly effective with `TF-IDF` features. For more complex textual patterns, researchers might utilize *XGBoost* or *LightGBM* implementations, which offer a balance between performance and partial interpretability through feature importance rankings.

The consideration of model complexity becomes particularly nuanced in social science applications where explaining societal phenomena is as crucial as predicting them. While deep learning frameworks like *PyTorch* and *TensorFlow* offer powerful `LSTM` and `BERT` implementations, their application requires careful justification. Tools like *SHAP* (`shap.TreeExplainer`) and *LIME* provide post-hoc interpretability for complex models, helping bridge the gap between predictive power and explanatory needs.

For model validation and selection, researchers typically employ *sklearn*'s `GridSearchCV` or `RandomizedSearchCV` with custom scoring metrics relevant to social science objectives. Cross-validation strategies, implemented through `sklearn.model_selection`, often require modification to account for temporal dependencies in social data. The *optuna* framework offers more sophisticated hyperparameter optimization when dealing with complex model architectures.

Ensemble methods, available through *sklearn.ensemble*, provide a pragmatic compromise between complexity and interpretability. `RandomForestClassifier` and `GradientBoostingClassifier` deliver robust performance while maintaining some level of feature importance interpretation. For cases requiring explicit uncertainty quantification, Bayesian approaches implemented in *PyMC3* or *Stan* offer probabilistic interpretations particularly valuable in social science contexts.

Evaluation metrics play a crucial role in assessing model performance and determining its suitability for a given application. Standard metrics, such as accuracy, precision, recall, and F1-score, are essential for evaluating classification models, while metrics like mean absolute error (MAE) and root mean square error (RMSE) are common in regression tasks. In addition to these, researchers often use cross-validation to assess model robustness by splitting the data into training and testing sets multiple times, which helps gauge a model's generalizability to new data. Model tuning techniques, such as hyperparameter optimization, are also vital to enhance model performance by fine-tuning key parameters. Given the interdisciplinary nature of social sciences, it is often necessary to balance model complexity with interpretability and to use evaluation metrics that align with both technical performance and the practical relevance of findings.

6 Conclusions

The evolution of quantitative semantic analysis in social sciences represents a unique moment in research history, marked by an unprecedented convergence of knowledge and technology. This convergence is not merely a linear progression of computational capabilities, but rather a complex intersection of theoretical understanding, methodological innovation, and technological advancement that is fundamentally changing how we conduct research. The emergence of new methodological approaches reflects this singular process, where traditional research practices meet modern computational power in ways that were unimaginable just decades ago.

The historical development of quantitative semantic analysis tools has demonstrated how computational advancements have transformed text analysis in social sciences. From early statistical methods to sophisticated language models, these tools have enhanced our ability to examine societal trends, economic sentiment, and cultural narratives. However, the major breakthrough lies in understanding how to effectively combine these new capabilities with established research practices.

A critical insight from this work concerns the essential integration of classical qualitative methods with robust NLP techniques, particularly those from the pre-probabilistic approach. Traditional methods like tagging and citation analysis, when thoughtfully combined with NLP, allow researchers to establish a solid descriptive foundation before addressing specific

research questions. This crucial preliminary step - systematically describing fieldwork and results in an analytically agnostic way - is often overlooked in contemporary research, largely due to the entrenchment of classical techniques. While these pure qualitative approaches have undoubtedly produced valuable insights and evolved over time, they remain fundamentally rooted in mid-20th century methodological perspectives ans frameworks. Their limitations become increasingly apparent when confronting modern research challenges, particularly with large-scale fieldwork, extensive interview series, or the vast textual landscapes emerging from social media and the internet era.

The role of NLP and Semantic Analysis in social science presents an intriguing paradox. Despite being an undeniably groundbreaking contribution to social studies, its methodological adoption remains surprisingly limited, as it can be considered as an emerging niche in the broader landscape of social research. This limited penetration stands in contrast to NLP's profound influence on the development of machine learning applied to data processing, large language models and text generation systems, technologies that are rapidly reshaping how we approach investigative tasks. An in-depth understanding of these technologies offers more than just technical advantages; it provides opportunities to generate more robust research contributions by advancing in this technological trajectory, fostering methodological innovation in social sciences.

Perhaps most fascinating is the unexpected renaissance of traditional techniques in the context of emerging LLM challenges. Classical approaches to text analysis are finding new relevance and application in modern contexts, particularly in retrieval-augmented generation (RAG) strategies and the development of sophisticated chunking and embedding methods. The application of topic modeling to enhance context awareness in these systems demonstrates how established methodological principles can be reimagined and repurposed. Social sciences have several aspects to contribute in this matter, as the revival of rooted techniques is not merely a return to old practices, but rather a creative reinterpretation of foundational methods in light of contemporary technological and interpretative challenges of neural networks based systems.

Looking forward, the described methodological evolution suggests exciting possibilities for social science research. The development of models that combine interpretability with analytical power, alongside cross-disciplinary approaches merging social science theory with machine learning algorithms, offers potential for more nuanced understanding of social phenomena. The emergence of multimodal analysis capabilities presents opportunities to capture the richness of human communication and cultural expression in previously impossible ways. However, this evolution brings important ethical considerations, that may be developed in a subsequent work. As machine learning tools become more integrated into social science research, vigilance regarding potential biases, data privacy concerns, and

transparency remains crucial. These challenges require ongoing attention to ensure our methods contribute positively to both academic understanding and broader society.

The convergence of qualitative expertise with computational capabilities represents more than just a change in tools - it marks a fundamental transformation in how we approach social science research in the digital age. While classical qualitative methods have provided valuable insights, their integration with modern computational approaches opens new possibilities for addressing complex social questions. By acknowledging both the potential and limitations of current approaches while working to improve our methods, we can contribute to developing more effective and responsible research practices. As these methods continue to evolve, their thoughtful application promises to enrich our understanding of social phenomena while maintaining the rigorous standards essential to scientific inquiry.

References

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Springer.

Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., von Arx, S., and Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Borgatti, S. P. and Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks*, 19(3):243–269.

Boyatzis, R. E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development*. Sage Publications.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, 18(S1):533–558.

Clarke, V. and Braun, V. (2013). Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, 26(2):120–123.

Creswell, J. W. and Plano Clark, V. L. (2017). *Designing and Conducting Mixed Methods Research*. Sage Publications, 3rd edition.

Danowski, J. A. (1993). Network analysis of message content. *Progress in Communication Sciences*, 12:197–222.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Diesner, J. and Carley, K. M. (2005). Exploration of communication networks from the enron email corpus. In *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*. Retrieved from https://www.casos.cs.cmu.edu/publications/papers/2005ExplorationOfCommunicationNetworks.pdf.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606.

Doerfel, M. L. (1998). What constitutes semantic network analysis? *Connections*, 21(2):16–26.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239.

Geuna, A., Kataishi, R., Toselli, M., Guzmán, E., Lawson, C., Fernandez-Zubieta, A., and Barros, B. (2015). Sisob data extraction and codification: A tool to analyze scientific careers. *Research Policy*, 44(9):1645–1658.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235.

Guest, G., MacQueen, K. M., and Namey, E. E. (2012). *Applied Thematic Analysis*. Sage Publications.

Halford, S. and Savage, M. (2017). Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology*, 51(6):1132–1148.

Hanneman, R. A. and Riddle, M. (2005). *Introduction to Social Network Methods*. University of California, Riverside.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296.

Hovy, D. (2022). *Text Analysis in Python for Social Scientists: Classification and Prediction*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kataishi, R. and Barletta, F. (2011). Diffusion of ict in the argentine productive fabric: A review of available evidence / difusión de las tic en el tejido productivo argentino: una revisión de la evidencia disponible. Technical report.

Kataishi, R. and Brixner, C. (2023). La política industrial en el entramado normativo: una aproximación sistémica del desarrollo del subrégimen de promoción industrial de tierra del fuego. In *XXVIII Reunión Anual de la Red PYMES MERCOSUR*, Córdoba, Argentina. CONICET-UNTDF. Globalización, desarrollo y desigualdad productiva: las pymes ante el desafío de la digitalización.

Kataishi, R., Brixner, C., Calá, C. D., and Niembro, A. (2023). Crisis, resilience, and innovation in strategic sectors: Reconfigurations in the tourism complex of tierra del fuego / crisis, resiliencia e innovación en sectores estratégicos: reconfiguraciones en el complejo turístico de tierra del fuego. *Tiempo de Gestión*, 2(33):7–30.

Kataishi, R. and Milia, M. (2024). Statistical-semantic analysis as a methodological approach: Reflections on its relevance in the study of latin american issues - chapter 7. / capítulo 7. el análisis semántico-estadístico como estrategia de abordaje metodológico: reflexiones sobre su pertinencia en el estudio de problemáticas latinoamericanas. In Natera, J. M. and Suárez, D., editors, *Methods for Analyzing Science, Technology, and Innovation Processes: Tools for Studying Development in Latin America / Métodos para el análisis de los procesos de ciencia, tecnología e innovación: herramientas para el estudio del desarrollo de América Latina*, page 265. Universidad Nacional de General Sarmiento and Universidad Autónoma Metropolitana, Los Polvorines, Argentina and Mexico City, Mexico, 1st edition.

Knoke, D. and Yang, S. (2008). *Social Network Analysis*. Sage Publications, 2nd edition.

Kobak, D. and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature Communications*, 10(1):5416.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., and Van Alstyne, M. (2009). Computational social science. *Science*, 323(5915):721–723.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32, pages 1188–1196.

McCusker, K. and Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion*, 30(7):537–542.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.

Musolino, D., Bruognolo, D., and Kataishi, R. (2023). The development of highly peripheral areas on a continental scale: The case of tierra del fuego, in argentina, and calabria, in italy. *Rivista economica del Mezzogiorno, Trimestrale della Svimez*, 3-4:623–668.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, pages 133–142.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality & Quantity*, 34(3):259–274.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). stm: An r package for structural topic models. *Journal of Statistical Software*, 91(2):1–40.

Saldaña, J. (2015). *The Coding Manual for Qualitative Researchers*. Sage Publications, 3rd edition.

Salton, G. (1975). A vector space model for information retrieval. *Communications of the ACM*, 18(11):613–620.

Scholz, R., editor (2019). *Quantifying Approaches to Discourse for Social Scientists*. Springer.

Segev, E. (2021). *Semantic Network Analysis in Social Sciences*. Routledge.

Sikstrom, S. and Garcia, D. (2020). Statistical semantics: Methods and applications. Springer.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Vaismoradi, M., Turunen, H., and Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences*, 15(3):398–405.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wilkis, A., Kataishi, R. E., and Robert, V. (2023). Indebtedness during the pandemic: Replicas of a financialized society / los endeudamientos en la pandemia: Réplicas de una sociedad financiarizada. Technical report, Ministerio de Ciencia, Tecnología e Innovación Productiva de la República Argentina, Agencia I+D+i.

Woods, M., Macklin, R., and Lewis, G. K. (2016). Researcher reflexivity: Exploring the impacts of caqdas use. *International Journal of Social Research Methodology*, 19(4):385–403.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.