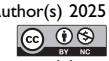
Special Issue

Chinese Journal of Sociology

# Natural language processing for social science research: A comprehensive review

Chinese Journal of Sociology
1–37



DOI: 10.1177/2057150X241306780
journals.sagepub.com/home/chs


**Yuxin Hou**[1,2] **and Junming Huang**[3]

## Abstract

Text data has been a longstanding pivotal source for social science research, providing an informative lens across disciplines including sociology, psychology, and political science. Its salient role in research, combined with the difficulty in numerically digesting unstructured data in natural languages, has been inspiring growing demands for natural language processing techniques to extract meaningful insights from vast text data. Breakthrough advances in natural language processing emerge with the recent expansion in data availability and computational resources, calling for an up-to-date comprehensive review for those methodologies and applications in social science research. This article reviews natural language processing techniques, detailing the procedure from representing unstructured text data to distilling semantic information, with expertise-based algorithms and unsupervised/supervised machine-learning methods. We then introduce their typical applications in producing research outcomes for sociology and political science. Keeping in mind challenges in data representativeness, interpretability, and biases, this review encourages utilizing natural language processing technique responsibly and effectively in social science research to improve quantitative understandings of emerging text data.

## Keywords

Big data/data science, language/linguistics, quantitative methods, natural language processing, text analysis, neural network, topic model

[1]Center for Social Research, Peking University, China
[2]Institution of Education, Tsinghua University, China
[3]Paul and Marcia Wythes Center on Contemporary China, Princeton University, USA

**Corresponding author:**
Junming Huang, Paul and Marcia Wythes Center on Contemporary China, Princeton University, Princeton, New Jersey, USA.
Email: pub@junminghuang.com

## Introduction

Natural language plays an essential role in the field of social science research. Both qualitative and quantitative text analysis techniques are extensively adopted in studies from different social science disciplines. For example, psychologists analyze natural language to grasp individuals' emotions and inner thoughts (Tausczik and Pennebaker, 2010), political scientists extract people's political opinions from online or offline speeches and discussions (Grimmer and Stewart, 2013), and sociologists reveal social mechanisms by manually coding texts in books, diaries, and interviews (Krippendorff, 2019; Schwartz and Ungar, 2015).

We are now entrenched in the epoch of information, or the era of computation (Blei and Smyth, 2017). As of 2024, the monthly active users on X (formerly Twitter) have reached 335 million. These users send short texts to express their feelings or opinions, with X just one of many social media platforms (Statista, 2022). Simultaneously, Gutenberg has amassed over 70,000 electronic books (Gutenberg, 2024) and the volume of the Wikipedia database reaches 4.5 billion words (Wikipedia, 2024). These figures, though substantial, represent only a fraction of the vast amount of digital textual data. Other sources including newspaper archives, online forums, historical datasets, and administrative records further enrich the digital textual resources accessible to social science researchers. This large amount of textual data presents both opportunities and challenges for social science researchers. On the one hand, these data make it possible to study unprecedentedly large populations, but on the other hand, dealing with data on such a large scale is not an easy task.

The good news is that computational techniques offer scientists effective tools to handle large-scale data. The emergence of large-scale data has given rise to a new research field: computational social science. Computational social science has experienced rapid growth in the last two decades, with various techniques such as network analysis, large-scale simulation, and natural language processing (NLP) bringing many new discoveries to various disciplines (Edelmann et al., 2020). However, given the large number of comprehensive reviews on computational social science (Conte et al., 2012; Edelmann et al., 2020; Lazer et al., 2020; Mann, 2016; Theocharis and Jungherr, 2021), we do not intend to provide another comprehensive review of it here, but rather focus on the application of natural language processing in social science research. The current review will introduce NLP techniques that can support social science research and provide insights for more natural language-based studies.

NLP is a subfield of computer science, which involves a range of computational techniques for learning, understanding, and generating natural language (Chowdhury, 2003; Hirschberg and Manning, 2015). This review will discuss NLP techniques and their applications in social science, breaking them down into three layers: an underlying layer representing unstructured text data in a structured form, a middle layer extracting understandable information from that representation, and an upper layer utilizing that information for social science research outcomes. At the end of this review, we will briefly summarize the main challenges in this field.

## Underlying layer: From unstructured text to structured data

Natural language data are inherently unstructured, without a well-defined format. This is very different from survey data—familiar to social scientists—that are usually structured for direct analysis. Therefore, the first challenge for NLP techniques is to transform unstructured text into structured formats that appropriately align with specific research outcomes. This typically involves two steps in practice: preprocessing and representation. Text preprocessing enhances data quality and feature extraction (Naseem et al., 2021), in which researchers reduce the noise in data by eliminating irrelevant information, emojis, spelling errors, etc. This seemingly easy step may have a substantial impact on the final outcomes (Bao et al., 2014).

There is a key difference between English and Chinese in this text preprocessing step: the space token acts as a perfect word divider in English, while no word divider is provided in Chinese. Therefore, word segmentation becomes an important first step when processing Chinese. Relevant techniques range from segmentation standards such as the Penn Chinese Treebank (CTB) (Xue et al., 2005) to deep learning models such as Yang et al. (2018). These techniques break down a Chinese sentence into word pieces.

After preprocessing, the text data are expected to have "informative" content only. Then we utilize representation techniques to convert the processed text into numerical data. Such a representation may directly count raw words as basic units, or infer the intrinsic semantic information embedded within the words. A recent review provides the principles and technical details of over 10 word representation models (Naseem et al., 2021).

### *Raw words: Vector space model, one-hot, bag of words, TF-IDF*

A straightforward representation is representing every unique word as a separate basic unit. For example, in a corpus we define "I" as word #1, "you" as word #2, "utilize" as word #3, etc. That defines a four-dimensional space where each dimension corresponds to a unique word, as follows:

| | |
|---|---|
| I | :[1, 0, 0, 0, 0] |
| you | :[0, 1, 0, 0, 0] |
| utilize | :[0, 0, 1, 0, 0] |
| text | :[0, 0, 0, 1, 0] |
| data | :[0, 0, 0, 0, 1] |

In this space, a sentence "I utilize text data" is simply represented by a vector [1, 0, 1, 1, 1]. This approach leads to the well-known vector space model proposed to numerically represent text data in the early years (Salton, 1989). In the vector space model, words are typically (though not always) considered orthogonal units, and documents are represented as collections of such units, i.e. a bag of words where the sequence of text is ignored for simplicity. The length of every document vector is equal to the number of unique words in the corpus.

It is practically common for a unique word to appear multiple times in a document. One approach to represent this multiplicity, known as one-hot encoding, records only the presence or absence of each word while ignoring the frequency of their occurrence. As a result, a document vector under this approach contains only 1 and 0. This approach trades the information about word frequency for computational simplicity.

A more widely used alternative approach is Term Frequency-Inverse Document Frequency (TF-IDF), which considers word frequency to assess word weight. TF calculates the frequency of a word in a document, while DF represents the overall term frequency in the entire document corpus. The inverse is adopted to mitigate the influence of common words like "I", "the", "a", etc. In TF-IDF representations, a word occurring frequently in a specific document but infrequently in the overall corpus is considered important and will get a high weight. TF-IDF for a particular word $t$ in a document $d$ of a corpus $M$ is calculated as follows:

$$TF - IDF(t, d, M) = TF \times IDF = \frac{N_{d,t}}{\sum_{t'} N_{d,t'}} \log\left(\frac{M}{M_t}\right),$$

where $N_{d,t}$ measures the counts of word $t$ in document $d$, $M$ denotes for the number of documents, and $M_t$ the number of documents containing word $t$. The TF-IDF representation is a matrix of words and documents, marking the importance of each word, supporting subsequent analyses such as classification tasks (Bail, 2016; Egger and Yu, 2022; Pickett and Valdez, 2023; Xiang et al., 2021).

Word representation based on unique words is straightforward to implement. However, its assumption of word orthogonality prevents it from capturing dependence between words, and leads to unnecessary data sparsity and the curse of dimensionality when the number of unique words in a corpus is high. This calls for representation models that capture the intrinsic dependence between words in a lower-dimensional space, enabling a deeper understanding of text.

## *Latent dimensions behind words: Word embedding*

Dropping the assumption of word orthogonality, a recent advancement known as word embedding attempts to construct a latent semantic space with unsupervised machine learning, where words are represented as points. The basic assumption behind word embedding is the distributional hypothesis, which posits that words occurring in similar contexts have similar meanings (Harris, 1954). Generating word embedding requires an unsupervised training process in a large corpus, during which the model counts the natural co-occurrence frequency of words and adjusts the word embedding to maximize the cosine similarity of the words that occur in similar contexts. Therefore, the word embedding can capture the semantic information of words. For example, "cat" in a word-embedding space is closer to "dog" than "car". While training word vectors in large-scale corpora is computationally expensive, researchers can also use pre-trained models to generate word embedding directly.

Compared with raw words-based methods, word embedding captures the semantic information of words, simultaneously avoiding the problem of sparsity and curse of dimensionality, and thus can effectively improve the performance of downstream tasks. Therefore, related algorithms have been widely adopted, with notable examples including Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). Apart from these word embeddings trained on English corpora, there are also those trained on Chinese texts, such as CA8 (Li et al., 2018) and CWE (Chen et al., 2015). These models are trained using Chinese corpora and are optimized for the structure of Chinese text, providing convenience for processing Chinese text. These algorithms are often used by social scientists to explore the meaning of text data. Inspired by commonly used psychological tests such as the implicit association test (IAT) for assessing biases, scholars have developed a word-embedding association test (WEAT) based on word embedding, and measured latent gender and racial biases in historical texts (Caliskan et al., 2017). WEAT has been adopted by many studies. Based on textual data from newspapers, books, websites, and other sources, researchers generate word embedding by GloVe, FastText, and Word2Vec, and then perform WEAT to analyze gender stereotypes in language (Charlesworth et al., 2021; DeFranza et al., 2020). Furthermore, researchers also use word embedding to explore topics such as changes in an artist's reputation following their death (Zhang et al., 2023) and the evolution of collective understandings of social class (Kozlowski et al., 2019).

When extracting semantic information from text, social scientists have a strong interest in the varying semantic meanings of a word across different contexts, such as different time periods, cultural contexts, or political affiliations. However, word-embedding techniques that only provide a global representation for one word struggle with such word ambiguity issues and fail to capture such heterogeneity. For example, the word "bank" can mean either a financial institution or the side of a river, yet it is represented by a single vector in these embedding models. One common approach is to distinguish different corpora and train separate word-embedding models, e.g. Garg et al. (2018). But this approach requires large corpora and substantial computation resources.

Recent models have brought better solutions. Sense2Vec tags words with their context, including the word class and named entity categories. For example, the word "bank" can be marked as "bank_NOUN" and "bank_VERB", which reduces ambiguity to some extent (Trask et al., 2015). However, capturing nuanced contextual changes based solely on these features remains challenging. Other models incorporate contextual information from the text into the embedding process. MUSE, for example, uses embeddings from pre-trained models like Word2Vec and GloVe as base embeddings, and performs cluster analysis to identify different semantic contexts of the same word (Lee and Chen, 2018). However, this approach requires extensive contextual data and involves high computational complexity. A la carte embedding (ALC) also uses pre-trained models to provide global semantic information, and calculates the average vector of other words in the context to generate a new embedding (Khodak et al., 2018). This dynamic adjustment allows the embedding to better fit the current context without retraining the entire model. Researchers have further developed a conText embedding model

based on ALC, applying it to study partisan differences in word usage, UK–US understandings of empire, and sentiment terms in Brexit parliamentary articles (Rodriguez et al., 2023). However, simple averaging may fail to capture deep semantic information and long-distance textual dependencies. Contextual embedding models based on deep learning, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT (Radford et al., 2018), address these issues effectively. They compute a context-based representation for each word, capturing deep meanings within text and considering context from left and right directions. These models achieve state-of-the-art performance across a wide range of natural language processing tasks. For a comprehensive survey on contextual embeddings, please refer to Liu et al. (2020).

## Intermediate layer: From structured data to semantic information

Numeric representation of text data provides a foundation for semantically understanding text. Semantic extraction methods vary from dictionary-based methods that are straightforward to implement and compute, to clustering and topic modeling that captures the intrinsic relationship between words and documents, to more computationally intensive deep learning models achieving the most advanced understanding on text.

### *Flag of expertise: Dictionary*

Initially, researchers mainly adopted dictionary-based approaches, i.e. a set of predefined "flag" keywords that characterize the text. Such a set is usually encoded in a dictionary—a data structure containing pairs of values and keys, typically mapping keywords to corresponding categories (Stoltz et al., 2024). For example, a simple sentiment lexicon can be defined as {"abandon": "negative", "benefit": "positive", "report": "neutral"}, where each word is associated with the indicated sentiment. When using this lexicon to analyze the sentence "She was abandoned on a winter night", researchers can count the frequency of each kind of sentiment and conclude that the sentiment of this sentence is negative.

Dictionaries have been extensively used in social science research over a long period. This can be traced back to the General Inquirer, in which researchers developed a lexicon to compare the tone of political speeches (Stone et al., 1966). Today, dictionaries remain popular and are widely used across various research topics. Researchers use dictionaries to extract features like linguistic abstractness (Snefjella and Kuperman, 2015) or cultural embeddedness (Goldberg et al., 2016) in a large corpus. The most prevalent application of dictionary-based methods is sentiment analysis. Researchers use sentiment lexicons to measure the sentiment of social media posts, enabling them to investigate how public sentiment changes in response to particular policies and events (Ahmed et al., 2017; Flores, 2017; Havey, 2020; Wei et al., 2023; Yu et al., 2022).

The wide usage of dictionary-based techniques is primarily due to their simplicity and low computational cost. Researchers easily get desired information by simply matching words and phrases in text, categorizing the text, or counting the word frequencies (Kroon

et al., 2022). However, dictionary-based techniques have limitations. Firstly, constructing a dictionary requires extensive time and effort. Additionally, dictionaries also ignore contextual information and are thus susceptible to ambiguity. For instance, the word "bright" is often considered "positive" in sentiment lexicons, and researchers might incorrectly categorize the sentence "This room was bright" as "positive". As a result, users of the General Inquirer often need to augment it to address ambiguity (Young and Soroka, 2012). Besides, many researchers also doubt the representativeness, effectiveness, and accuracy of dictionary-based methods (Guo et al., 2016; van Atteveldt et al., 2021).

Nowadays, the intensive effort required for preparing dictionaries has been partly mitigated by emerging advancements in crowdsourced and automatic dictionary construction. Crowdsourced dictionaries rely on web platforms like Amazon's Mechanical Turk, enabling researchers to recruit hundreds of annotators to expedite the labeling process, thereby reducing time cost and avoiding biases introduced by a small group of experts (Schwartz and Ungar, 2015). Deriving dictionaries from text starts with a substantial volume of text with specific labels. Initially, the text is segmented into words, and then the correlation between a word and the outcomes is identified through techniques such as pointwise mutual information. Finally, a dictionary is generated based on these findings (Schwartz and Ungar, 2015). For instance, scholars have generated a dictionary of new terms based on scientific publications and utilized it to investigate the idea diffusion process in science (Cheng et al., 2023). Furthermore, researchers may consider using open-source dictionaries. For example, the General Inquirer can be employed for political speech analysis, while the Linguistic Inquiry and Word Count (Pennebaker et al., 2015) can be used to quantify words reflecting emotions, thinking styles, and social concerns. There are many open-source dictionaries for measuring text sentiment, as well as online platforms that integrate various dictionaries (van Atteveldt et al., 2021; Zhao and Wong, 2023).

### *Co-occurrence: Clustering and topic modeling*

The intensive human labor for creating dictionaries make dictionary-based methods less favored when large-scale corpora emerge with new terms. That motivates the adoption of machine-learning methods to automatically discover the hidden semantic information behind text. Such methods could be unsupervised, utilizing word co-occurrence relationships to generate categories and distill topics, or supervised, targeting at mapping from the word space to a predefined label space.

Unsupervised learning methods solely rely on the data to categorize text into categories without requirements of additional textual knowledge. Typical unsupervised learning methods include clustering techniques such as K-Means, and topic modeling methods like latent Dirichlet allocation (LDA) (Blei et al., 2003). These methods are helpful for social scientists, as they can efficiently uncover hidden patterns in large-scale textual data, and categorize these data into topics or clusters for further analysis.

Clustering algorithms group data into clusters based on feature similarity, maximizing intra-cluster similarity and inter-cluster distance (Ezugwu et al., 2022; Xu and Tian, 2015). Traditional clustering algorithms are categorized by different strategies, among

which partition-based clustering, hierarchy-based clustering, and model-based clustering are commonly used by social scientists. Partition-based clustering algorithms, such as K-Means (Macqueen, 1967) and K-Medoids (Park and Jun, 2009), iteratively assign data points to clusters, updating centers until stabilization, but require a predefined number of clusters, which can be challenging in many contexts. Hierarchy-based clustering creates a tree-like structure by gradually splitting or merging data based on distance (Johnson, 1967), without the need of a predefined number of clusters. However, both partition-based and hierarchical clustering struggle with overlapping clusters and non-spherical shapes. Model-based clustering, such as Gaussian mixture models (GMM) (Rasmussen, 1999), can partially address these issues. These algorithms assume data as a mixture of probability distributions and can accommodate more complex data shapes and sizes, although the model distribution assumptions may also affect the clustering results. Clustering algorithms can effectively handle the classification of short texts, and many researchers apply clustering algorithms for event detection in large-scale social media texts (Mukherjee and Bala, 2017; Vijayakumar and Rajam, 2024). For a more systematic review of different types of clustering algorithms, please refer to the survey by Xu and Tian (2015).

While clustering groups data into similar clusters, topic modeling focuses on identifying the latent themes within a collection of documents and revealing their semantic structure. LDA is one of the most common topic modeling techniques used by social scientists. It is proved to be highly effective for analyzing large-scale text data. Utilizing word embedding containing semantic information, this method measures the relationships between texts through matrix operations to detect latent topics present in the text (Wilkerson and Casas, 2017). In LDA, each document is assumed to contain multiple topics, and each topic is generated by a group of words. LDA generates two distributions, namely a topic distribution for each document and a word distribution for each topic. The distributions are determined by the co-occurrence of words, and a document's topics are inferred based on the distributions. LDA is frequently employed to study discussion topics in online communities. For example, some researchers used LDA to investigate different themes of vaccine-related misinformation on X (Valdez et al., 2023) and others analyzed anti-vaccination sentiments on Facebook (Smith and Graham, 2019). Similarly, scholars used LDA to summarize online discussions and sentiments of Weibo users during the COVID-19 pandemic (Shi et al., 2022; Xie et al., 2021) and to analyze the development of public opinion (Han et al., 2020; Zhu et al., 2020). LDA has also been adopted in political science to study databases such as leader speeches to analyze important dynamics of political agendas (Catalinac, 2016; Quinn et al., 2010).

However, LDA also presents several limitations. First, LDA relies on word co-occurrence to extract topics, which leads to poor performance when handling short texts from social media (Hong and Davison, 2010). To address this issue, scholars have proposed the biterm topic model, which was specifically designed for short texts such as tweets (Cheng et al., 2014). Second, LDA determines topics based on the distribution of words, leading to the ignorance of contextual information. The structural topic model (STM) addresses this limitation by involving document-level metadata as covariates (Roberts et al., 2013, 2014). STM-based text analysis allows researchers to include

features such as publication time, location, and demographic information of the authors. These features are crucial for social group analysis, making STM a popular choice among social scientists. STM is frequently adopted to analyze open-ended survey questions, as demonstrated in studies by Enria et al. (2021), Rothschild et al. (2019), Tvinnereim and Fløttum (2015), and Yan et al. (2024). Third, LDA utilizes the bag-of-words model to represent documents, which ignores the order of words and cannot capture deep semantic information. Recent topic models involve more comprehensive word-embedding techniques, such as Top2Vec (Angelov, 2020), which uses Word2Vec, and BERTopic (Grootendorst, 2022), which uses BERT. These models effectively address this issue by capturing more subtle semantic differences in the text. Egger and Yu (2022) conducted a survey over four topic modeling techniques, namely LDA, non-negative matrix factorization (NMF), Top2Vec, and BERTopic. Based on the results from 50,000 English X (Twitter) posts related to travel and COVID-19, it reported that while LDA revealed more topics related to geography and borders, it also generated more meaningless topics. Top2Vec was more policy-oriented, while BERTopic's topics were more related to aviation issues. However, unsupervised learning techniques do not have standardized evaluation criteria, and the interpretation of model results depends on the specific research context and domain knowledge (Hannigan et al., 2019).

## *Mapping: Naïve Bayes, SVM, tree*

Supervised learning seeks for a mapping from input text to desired research outcomes. Researchers need to provide both input text and human-labeled results as ground truth to the model simultaneously. During the training process, the model utilizes this data for "learning", namely continuously adjusting parameters to minimize the gap (often called "loss") between the model outputs and the ground truth. After this training process converges, the fitted model is a function mapping input text to the desired output, thus accomplishing the task intended by the researchers (Grimmer and Stewart, 2013). Depending on the features of the desired output, supervised learning models can perform tasks including classification and regression. Specifically, all supervised learning involves three steps:

1. Data preparation: Researchers need to specify coding schemes and obtain a dataset with input text and desired output through human labeling. To ensure the reliability of annotations, it is generally necessary to have at least two annotators and report the kappa statistic (McHugh, 2012).
2. Model training: Splitting the dataset into training and test sets in certain proportions (e.g. 0.8, 0.2), we use the training set to train the model. During training, the model continuously adjusts parameters to minimize the loss, thereby continuously optimizing the model.
3. Model validation: The model is trained to label a larger-scale text data instead of human annotators. Therefore, measuring the performance of the model is crucial to ensure the reliability of the results generated by the model. Poor performance may lead to bias or even errors in the results. A common practice is to use the fitted

model to make predictions on a test set with known human labels and compare the results and the labels. To avoid the influence of randomness, cross-validation can be used to validate the model and select the best model (Arlot and Celisse, 2010).

Various supervised learning algorithms have been developed and widely used, each with its own strengths and weaknesses. No single machine-learning method is universally superior to any other, which is known as the "no free lunch" theory (Wolpert and Macready, 1997). Too simple models may fail to capture the features of the data, while overly complex models are more prone to overfitting and require more computational resources. The task is not to choose the best method but to select the most appropriate method based on the current research context. Here, we introduce several commonly used supervised learning algorithms and their applications.

- **Support vector machine** (Cortes and Vapnik, 1995): Support vector machine (SVM) is a commonly used classification algorithm. All inputs are considered as sample points in an n-dimensional space, and the goal of SVM is to find a hyperplane such that the distance from all sample points to the hyperplane is maximized, thus enabling classification of the points. As a margin maximizing classifier, SVM often outperforms probability classifiers like naïve Bayes and can effectively handle high-dimensional and sparse data. However, SVM has a high time complexity, meaning its training efficiency is lower when dealing with large volumes of text data. Therefore, SVM is more commonly used for small-scale text classification tasks. Political scientists have applied SVM to tasks such as classifying Militarized Interstate Dispute 4 (MID4) data (D'Orazio et al., 2014) and analyzing climate change-related news articles (Boussalis et al., 2018).
- **Naïve Bayes**: Unlike SVM and many other classification methods, naïve Bayes is a probability-based classifier. The basic idea of naïve Bayes is to directly learn the joint distribution $P(X, Y)$ between output features $Y$ and input features $X$, and generate the conditional distribution $P(Y|X)$ using Bayes' theorem $P(Y|X) = \frac{P(X, Y)}{P(X)}$. The classification result is given by maximizing the conditional distribution $P(Y|X)$. Its name "naïve" comes from the assumption that all features in $X$ are mutually independent. Naïve Bayes is simple, with a high training efficiency and strong interpretability. In existing research, naïve Bayes has been used for tasks such as predicting student performance (Pujianto et al., 2017), event encoding (Hillard et al., 2008), and text classification of user reviews (Lam and Chan, 2024).
- **Decision trees and random forests**: Decision trees (Quinlan, 1986) simulate a series of decision-making processes to achieve the desired results. The decision tree algorithm constructs a tree with $n$ nodes, with each node splitting the dataset according to specific features and the decision conditions of the node, thereby generating a model that can classify the data. Decision trees also have good interpretability and can handle multi-classification problems, but a single decision tree is susceptible to noise and prone to overfitting. Therefore, more useful algorithms based on decision trees have emerged, such as random forests (Breiman, 2001) and gradient-boosted

decision trees (GBDT) (Friedman, 2001). These algorithms have better robustness and algorithm performance when faced with large amounts of data and noise, and have been widely used in social science research (Markowitz, 2022; Matalon et al., 2021).

- **Multilayer perceptron**: Multilayer perceptron (MLP), also known as artificial neural networks (ANN), is a deep learning model based on feedforward neural networks. All neural networks consist of input layers, hidden layers, and output layers. The input layer receives features, the output layer provides the final prediction, and the hidden layer, also known as neurons, extracts features and performs nonlinear transformations on the output of the previous layer. In MLP, each neuron is fully connected to the previous layer. The layer structure of MLP provides advantages in fitting high-dimensional data and handling nonlinear problems. In a study on social media, researchers used a three-layer perceptron to classify the political orientation of Twitter (now X) bots, achieving precision and recall rates both exceeding 90% and obtaining good classification results (Stukal et al., 2019). However, MLP requires high consumption of computational resources and lacks interpretability.

## *Sequence: Deep learning models*

Most of the aforementioned models use vectorized text, namely word embeddings, as input and treat these vectors as independent data points. Consequently, they rely solely on individual words or local contextual information, ignoring the crucial sequence information. This limitation results in poor performance in complex language processing tasks such as sentiment analysis, question answering, and language translation, which demand a deep understanding of complex language patterns like metaphors and rhetoric, as well as the ability to capture long-distance dependencies within the text.

Deep learning models have produced promising results in these tasks (LeCun et al., 2015). Like MLPs, the structure of a deep learning model is a stack of multiple layers of neural networks, but they typically have a larger number of layers (so-called "deep"). The nonlinear activation functions within each neural network enable the model to perform nonlinear transformations, and the multilayer structure allows it to represent deep features in a sequence of data. The most typical application of deep learning models is also in supervised learning, which involves data preparation, model training, and model prediction, as mentioned above. During training, the model updates the parameters of each layer based on the loss between the predicted and ground-truth labels. Although the large number of parameters makes the update steps seem computationally intensive, the back-propagation algorithm based on chain rules makes the training process quite manageable (Rumelhart et al., 1986). After training, trained models are expected to reach a good performance on the designed tasks. Many deep learning models have been used to tackle NLP tasks in various contexts. Here, we mainly introduce three classic models: RNN, LSTM, and GRU.

- **Recurrent neural network**: Recurrent neural networks (RNN) are deep learning models specifically designed to handle sequential data, such as speech and text.

At a time step $t$, an RNN focuses on one element in the input sequence, but its recurrent input structure allows it to incorporate information from the previous time step $t-1$ to update the current hidden state. The state of hidden layer is updated by:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$$

where $\sigma$ is the activation function, $W_h$ is the weight matrix for the previous hidden state, $W_x$ is the weight matrix for the current input, and $b$ is the bias. $W_h$, $W_x$, and $b$ are model parameters that remain constant while processing the same sequence. $h_{t-1}$ is the hidden state at time step $t-1$, and $h_t$ is the current hidden state. This recurrent input structure equips RNNs with memory capabilities to better capture context information and long-distance dependencies in text. RNN is a classic application of the backpropagation algorithm, but it requires backpropagation through time (BPTT) to update parameters, which involves multiplying gradients at each time step. This leads to issues such as gradient explosion or vanishing when processing long sequences (Bengio et al., 1994).

- **Long short-term memory**: Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) is a special type of RNN designed to address the issues of gradient explosion and vanishing when processing long sequences. LSTM introduces an independent explicit memory cell responsible for storing input information over long periods. Besides, LSTM incorporates gating mechanisms: the input gate determines which information from the current input and the hidden state of last time step should be used to update the memory cell; the forget gate decides which part of information should be discarded; the output gate determines which part of information will be output to the next step. With these gating mechanisms, LSTM can dynamically adjust the content of the memory cell and effectively improve performance on long sequences. Based on LSTM, researchers further proposed the bi-LSTM structure, which trains two LSTMs simultaneously—one from front to back and the other from back to front on the input sequence—thus utilizing context information from both sides (Schuster and Paliwal, 1997).
- **Gated recurrent unit**: Gated recurrent unit (GRU) (Cho et al., 2014) is a simplified version of RNN designed to address the issues of gradient explosion and vanishing in RNNs. Unlike LSTM, GRU does not use explicit memory cells, but combines the memory cell within the hidden state instead. GRU involves two gating units, namely the reset gate and the update gate. The reset gate determines how much information from the hidden state of the previous time step should be forgotten when generating the current candidate hidden state. The update gate decides how much of the hidden state in the current time step should be retained from the past hidden state and from the new candidate hidden state. Compared to LSTM, GRU is simpler and requires less computational resource.

These deep learning models produce promising results in complex text processing tasks. Sentiment analysis is one of the most classic applications. During the COVID-19 pandemic, many researchers adopted these models to conduct sentiment classification of COVID-19-related text on social media platforms (Arbane et al., 2023; Nemes and Kiss, 2021; Xu et al., 2019). Besides, deep learning models are also commonly used for detecting online extremism (Gaikwad et al., 2021), cyberbullying (Fang et al., 2021; Murshed et al., 2022), and fake news (Ajao et al., 2018). Additionally, the excellent sequence data processing capabilities of deep learning models is also helpful in time series data. Some researchers use LSTM to train on the time series data of hot topics on Weibo, enabling them to predict changes in public opinion trends related to these topics (Mu et al., 2023). LSTM is also used to predict the COVID-19 pandemic trends in different countries (Wang et al., 2020). Some researchers also use RNNs to conduct counterfactual inference based on their performance in sequential prediction tasks (Poulos and Zeng, 2021).

### *Integration: Large language models*

Recently, transformer-based large language models (LLM) like BERT (Devlin et al., 2019) and generative pre-trained transformer (GPT) (Radford et al., 2018) have attracted significant attention and become focal points of discussion. LLMs are pre-trained on large-scale datasets to acquire a foundational understanding of common language patterns. These models demonstrate impressive ability in various complex tasks, including natural language understanding, translation, text summarization, question answering, and even reasoning tasks, even reaching a higher score than humans in some benchmarks (Bang et al., 2023; Street et al., 2024; Qin et al., 2023). The impressive performance of LLMs can be attributed primarily to three features: the transformer-based architecture, pre-training, and the large scale of parameters.

- **Transformer-based architecture**: In 2017, Google introduced the transformer model, a revolutionary architecture in natural language processing. Unlike previous models that heavily relied on sequential processing, the transformer utilized attention mechanisms, enabling it to capture global dependencies within the input sequence. The transformer achieved significant advancements in language understanding and performance across various tasks (Vaswani et al., 2017).
- **Pre-training**: The concept of pre-training is derived from transfer learning in computer vision (Zhuang et al., 2020). Pre-training refers to training a model on a large text dataset to learn general language patterns. This process enables the model with a broad understanding of language, which can be transferred to various downstream tasks. Typical pre-training tasks involve masked language modeling or next sentence prediction, enabling the model to capture syntactic and semantic information in text data (Devlin et al., 2019; Radford et al., 2019).
- **Large scale of parameters**: The large parameter scale of LLMs is a key factor in their impressive performance. Early models like BERT contained about 110 million parameters for the base version, and the parameter scale of recent models has dramatically

increased, reaching up to hundreds of billions (e.g. the recent 405-billion-parameter LLaMA 3.1). Researchers have demonstrated that model performance follows the scaling law, exhibiting a power-law relationship with the number of model parameters, dataset size, and the amount of compute used for training (Kaplan et al., 2020).

BERT is one of the earliest and most extensively applied LLMs. BERT is a pre-trained deep bidirectional encoder representations model based on transformers. It conducts self-supervised pre-training by simultaneously considering the context of the text. BERT was pre-trained using BooksCorpus of 800 million words and English Wikipedia of 2.5 billion words. The pre-training phase involves two unsupervised tasks: masked language model and next sentence prediction. Pre-trained BERT models offer substantial language understanding, serving as valuable initialization for new tasks. The open sourcing of BERT models allows researchers to skip the expensive pre-training phase. Pre-trained BERT models are capable of achieving state-of-the-art performance across different tasks with just one output layer (Devlin et al., 2019). RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) further enhance BERT by incorporating dynamic masking, parameter sharing, and modifying pre-training tasks. Pre-trained models' ability relies on the pre-training corpus. The aforementioned BERT series models are mainly trained on English corpora, and thus unable to handle non-English texts. Researchers from China have released the Chinese-BERT-wwm model which is pre-trained on Chinese Wikipedia corpus and extended Chinese datasets (Cui et al., 2021). This model has proved to perform well in many Chinese-based language processing tasks.

BERT-based models have been applied in social sciences to various complex language understanding tasks, including sentiment classification (Field et al., 2022; Huang et al., 2021; Sivakumar and Rajalakshmi, 2022; Xie et al., 2024), measuring abstract concepts in political texts such as populism, nationalism, and authoritarianism (Bonikowski et al., 2022), and detecting polarization and ideology in texts and videos (Han, 2022; Lai et al., 2022).

Since the release of the GPT series by OpenAI (Brown et al., 2020; Radford et al., 2018, 2019), there has been an increasing emergence of models with billions of parameters. For example, GPT-3, with its massive scale of 175 billion parameters and 570 gigabytes of training data, has garnered significant attention and demonstrated effectiveness in various few-shot (even zero-shot) NLP tasks (Brown et al., 2020). The powerful text generation capabilities of GPT-3 enable its application in diverse domains, including question answering, summarization, conversation, basic arithmetic computation, and generating various types of text. However, the GPT series models after GPT-3 are no longer open-source, which means users must access these models through OpenAI's API. This also makes it more challenging to fine-tune the models for specific tasks. Consequently, other open-source large models like the Llama series (Dubey et al., 2024; Touvron et al., 2023a, 2023b) and Falcon (Almazrouei et al., 2023) have gained widespread use. Additionally, researchers from China have developed many large models optimized for Chinese, such as the GLM series (Zeng et al., 2024), Qwen (Bai et al., 2023), and Pangu-$\alpha$ (Zeng et al., 2021). More details can be found in the comprehensive survey on large language models by Zhao et al. (2023) and Chang et al. (2023).

Zero-shot or few-shot LLMs' striking ability presents new possibilities for computational social science. Studies have shown that LLMs can be applied to analyzing psychological constructs across different languages (Rathje et al., 2024) and analyzing political stances and ideologies (Wu et al., 2023). Ziems et al. (2024) conducted a systematic analysis about the performances of different LLMs on various types of research tasks. The results indicate that zero-shot LLMs perform well in classification tasks on stance, emotion, figurative language, and utterance-level ideology. Although they do not perform better than carefully fine-tuned RoBERTa models, they offer an approach that can avoid the high expense of human labeling. However, researchers should also be aware that LLMS do not perform well in complex classification tasks like event argument extraction, semantic change, empathy or toxicity detection, and stereotype detection, which require expert opinions. Ziems et al. (2024) also pointed out that LLMs produce better results than human crowdsourcing in generation tasks including emotion-specific summarization, misinformation explanation, language reframing, etc. Beyond classification and generation tasks, researchers can also develop agents based on LLMs for social simulation experiments. For example, some researchers use LLMs to simulate human subjects (Argyle et al., 2023b) and social media dynamics (Gao et al., 2023a), or to conduct experiments related to personality (Jiang et al., 2024) or persuasion (Karinshak et al., 2023).

However, along with the impressive performance of LLMs come various threats, and social bias in LLMs is a typical example. LLMs are mainly trained on raw internet-based content, which contains various biases, stereotypes, misrepresentations, and other patterns that may affect marginalized groups. Numerous studies have demonstrated that the content generated by LLMs inherits and even exacerbates these social biases (Abid et al., 2021; Bender et al., 2021; Gallegos et al., 2024; Kotek et al., 2023; OpenAI et al., 2024). Investigating these biases and intervening through various alignment techniques has become a new topic in the field of computational social science (Wang et al., 2023).

## Upper layer: From semantic information to social science research outcome

After text representation and understanding, semantic information is now readily accessible for social science research. This distilled semantic information provides a new lens to advance research on traditional social science topics, such as bias, elections, cultures, and the science of science. Moreover, social media emerge with millions of users engaged in communication, interaction, and expression of emotions and ideas through short textual messages. This online world presents dynamics distinct from the physical society. The recent availability of large-scale textual data on social media, along with advanced natural language processing techniques offering refined granularity of text understanding, have together inspired an emerging community studying online engagement and communications on social media with an unprecedented level of detail. In this section, we demonstrate the application of semantic information to social science research, using

representative examples from sociology and political science. These two fields have been at the forefront of adopting natural language processing techniques, primarily due to their rich availability of textual data sources and the complexity of the relationships and interactions they address. It is also worth noting that disciplines such as cognitive science and psychology are increasingly utilizing NLP techniques. Although they are not covered extensively in this review, their growing adoption is significant and should not be overlooked.

## *Sociology: Social bias, culture, science of science*

*Social bias.* **Social bias and stereotypes** were traditionally studied with self-reported survey data. While psychologists have developed relevant tests, it is impossible to acquire large-scale experimental data due to the high cost of human labor. The advent of NLP has enabled researchers to extract attitudes and viewpoints directly from text, providing novel insights into studying bias on gender, age, and ethnicity.

One of the most prevalent areas of study involves the measurement of gender bias implicit within language. Cognitive psychologists have utilized word-embedding techniques across diverse text datasets, ranging from books, dictionaries, and web pages, to lyrics and textbooks, to measure the presence of gender biases in language. This body of work has consistently identified pervasive implicit biases against women (Bailey et al., 2022; Betti et al., 2023; Charlesworth et al., 2021; DeFranza et al., 2020; Jiao and Luo, 2021; Lucy et al., 2020; Napp, 2023). However, some research indicates a gradual reduction in gender stereotypes over time (Jones et al., 2020). Beyond the analyses based on word embedding, scholars have employed various NLP techniques to explore potential gender disparities. For instance, Markowitz (2022) analyzed patient–physician records using the Linguistic Inquiry and Word Count tool to uncover gender differences in patient–physician relationships. Research findings indicate that physicians tend to pay more attention to the emotions of female patients compared to male patients. Similarly, Czymara et al. (2021) employed topic modeling to examine the experiences during the COVID-19 lockdown in Germany, revealing a greater negative impact for women on both physical and cognitive levels of work, which may exacerbate gender inequalities. Additionally, Parthasarathy et al. (2019) examined the relationship between gender and political influence using transcripts from constitutionally mandated village assemblies, revealing women's relative disadvantage and silence compared to men.

Scholars have conducted research on ethnicity bias based on text data. Similar to gender bias, some scholars employ word embedding to measure the association between ethnicity-related vocabulary and other terms. For instance, a study based on Texas history textbooks revealed that the most commonly mentioned individuals are predominantly White, while Black individuals are often depicted as having limited agency and power (Lucy et al., 2020). Another study utilizing the Contemporary American English Corpus found that the United States' racial framework is deeply ingrained in American English, with racial/ethnic groups being differently associated with notions of superiority and Americanness (Lee et al., 2024). Moreover, van Loon et al. (2022)

discovered that merely assessing the frequency of Black names could predict anti-Black bias across various regions. Markowitz (2022) conducted research based on patient–physician records, revealing that physicians also attended to fewer emotions expressed by Black/African and Asian patients compared to White patients. Additionally, Kennedy et al. (2021) utilized topic modeling to analyze the text of rental advertisements in Seattle, investigating how neighborhoods' racial composition was described. The findings indicated that while White neighborhoods emphasized trust and connections to neighborhood history and culture, listings from non-White neighborhoods tended to offer more incentives and focused on transportation and development features, thus demonstrating the existence of racialized neighborhood discourse.

Researchers have also conducted text analyses on discrimination and biases concerning older adults. Analyzing Twitter (now X) data related to older adults during the COVID-19 pandemic, studies have identified instances of discrimination and negative emotions directed towards the elderly or other vulnerable groups. Interestingly, despite this negativity, it has been observed that such negative sentiments are gradually decreasing over time (Ng et al., 2022; Xiang et al., 2021).

*Cultural sociology.* Cultural sociology examines culture from a sociological perspective, exploring its formation, transformation, and influence. Text serves as one of the vital carriers of culture and is central to the study of cultural sociology (Bail, 2014), hence NLP has introduced new possibilities for cultural sociology. Michel et al. (2011) utilized a database containing 5,195,769 digitized books and employed frequency-based calculations to measure vocabulary, syntactic changes, fluctuations in the fame of prominent figures, and shifts in collective memory. Kozlowski et al. (2019) also utilized a dataset comprising millions of books to train word-embedding models, measuring the cultural significance of social classes in text and finding that the markers of class changed continuously during the 20th-century economic transformation while their cultural dimensions remained stable. Zhang et al. (2023) constructed a vast historical corpus from digitized texts of 20 newspapers spanning 1795 to 2020 and researched collective memory about artists. They built word-embedding models for different periods and measured the association between artist names and vocabulary related to reputation, thus gauging artists' reputations. The results indicated that most artists attained peak reputations before death, followed by a decline, losing nearly one standard deviation per century. These studies exemplify excellent use of large-scale data for researching cultural transformations.

In addition to the aforementioned studies based on large-scale historical data, NLP also supports more micro-level research. For instance, Light and Odden (2017) employed text data from music review websites to conduct topic modeling on factors related to consumer reviews of music, thereby understanding how contemporary consumers assess the value of cultural products. McCumber and Davis (2024) investigated changes in the standards of "elite environmental aesthetics" based on articles from the *New York Times* travel section from 2000 to 2019 and explored the impact of climate change on these standards.

*Science of science.* Science of science is an emerging interdisciplinary research field that investigates the mechanisms behind scientific research (Fortunato et al., 2018). Leveraging extensive bibliographic data, researchers can construct intricate citation networks, collaboration networks, and study topics such as idea diffusion and teamwork in science (Edelmann et al., 2020). However, besides these network-based computational techniques, text-based analysis also provides many new insights. For instance, McMahan and Evans (2018) developed an information-theoretic statistical model to compute the ambiguity of given words in scientific texts. Their model revealed that humanities, law, and environmental and earth sciences exhibit the highest ambiguity. Vilhena (Vilhena et al., 2014) measured the communicative efficiency of specialized knowledge and language across different fields based on citation structures and phrase frequencies in articles, finding that communicative efficiency decays with citation distance in a field-specific manner.

Some scholars utilize NLP to study the dynamics of scientific development in specific disciplinary domains. For example, a study explored the divide in sociological methodology. Analyzing word frequencies in 8737 articles from 1995 to 2017, they demonstrated the existence of methodological divergence but found a slight increase in quantitative research published in comprehensive journals over time (Schwemmer and Wieczorek, 2020). Another study addressed the decline discourse in organizational sociology by automatically classifying articles published in comprehensive sociological journals since the 1950s using SVM. They applied topic modeling to organize the themes in the relevant articles, finding that while the overall publication level of organizational sociology has not significantly decreased compared to 20 years ago, there has been a decline in theoretical and methodological diversity (Grothe-Hammer and Kohl, 2020).

NLP has also been applied to study gender inequality in science. Key and Sumner (2019) examined systematic difference in topic selection between male and female researchers. They used topic modeling to analyze abstracts of 2055 political science papers, inferred the gender of researchers from their names, and summarized gender differences in research topics. The results revealed that women are more interested in topics such as race, healthcare, narrative and discourse, and branches of government, while big topics in political science, such as voting, campaigns, and congress, are predominantly dominated by men. This finding partly explains the lower publication rates and citation counts among women. Larregue and Nielsen (2024) explored gender differences in research funding. Combining interview data with content analysis of funding proposals, they found that gender differences in funding might be related to gender differences in disciplinary focus, thematic specializations, and methodologies.

## *Political science: Election, engagement, polarization*

*Campaigns.* Natural language processing techniques have offered new tools to analyze campaigns like election for political science research. In early years, topic modeling was commonly used for text analysis. For instance, Catalinac (2016) analyzed 7497 election manifestos from the 1994 Japanese elections using topic models to understand different strategies of candidates under various electoral systems. Similarly, DiMaggio et al.

(2013) adopted LDA to analyze how governments assist artists and arts organizations. Recently scholars started to utilize pre-trained models for more complex text classifications. Bonikowski et al. (2022) analyzed speech records of Democratic and Republican presidential candidates from 1952 to 2020 to investigate whether frames of radical-right campaigns have gradually spread to centrist parties. They employed a pre-trained RoBERTa model, fine-tuning it based on human-labeled data to accurately identify political frames in 71,808 segments of election speeches, achieving good accuracy. The study revealed trends in the presence of different political frames and the strategies employed by candidates in their speeches.

Scholars have also focused on election campaigns based on social media. Barack Obama's successful strategy on social media helped him set records of donations and grassroots mobilization (Tumasjan et al., 2011), making social media platforms pervasive tools in election campaigns. Numerous studies have discussed the use of Twitter/X in elections, exploring the tendency of individuals with different demographic features to use Twitter/X and its connection with electoral opportunities, with more emphasis on how political parties and candidates use Twitter/X (Jungherr, 2016). In these studies, NLP techniques are applied to text analysis of party and candidate social media posts to understand the sentiments and political inclinations of their content. Tumasjan et al. (2011) conducted sentiment analysis on 100,000 tweets containing party or candidate mentions during the 2009 German federal election, finding that the mere number of party mentions can reflect the election result, and tweet sentiment is highly correlated with voters' political preferences. Other scholars studied the relevance of topics discussed by electoral members online to the public. They used naïve Bayes to identify political topics in the text and observed the relationship between candidates and the public on different platforms. The research indicated that politicians and their audiences discuss different topics on social media, and politicians use Facebook and Twitter/X for different purposes, related to the distinct target groups candidates encounter (Stier et al., 2018). Livne et al. (2011) studied the use of Twitter by House, Senate, and gubernatorial candidates during the 2010 US elections, finding significant differences in social media usage patterns among people of different political affiliations, suggesting that conservative candidates more effectively use social media platforms.

*Polarization.* Another line of research focuses on the polarization of viewpoints in traditional media such as newspapers and explores its impact on individuals. Using a dictionary-based approach, Hart et al. (2020) examined the politicization and polarization of COVID-19-related content in US newspapers and online news, finding higher levels of politicization in newspapers compared to online news and suggesting that such politicization and polarization may contribute to the polarization of attitudes towards COVID-19 in the United States. Similarly, Chinn et al. (2020) measured climate change-related news content and found increasing politicization, a rise in political actors, a decline in scientific actors, and increasing polarization. Huang et al. (2021) investigated how opinions in news reports about China influence public opinion. They used BERT to label opinions in China-related reports from the *New York Times* and combined this with survey data

to demonstrate how events in international relations shape social media opinions and, consequently, public opinion.

Within discussions of social media dynamics, polarization has recently attracted significant attention. Polarization refers to the tendency of more extreme opinions and sentiments on controversial topics. Data suggests that political polarization has intensified over the course of modern US history, and this trend may also exist in other countries (Geiger, 2014). Consequently, an increasing number of studies are focusing on the causes and potential consequences of polarization (Ferguson, 2021).

Some studies have explored the contribution of social media use to polarization. Researchers have extracted and analyzed the sentiment, polarity, and topics of the textual social media data, and provided a detailed investigations of polarization in social media. For instance, Zollo et al. (2015) utilized SVM to analyze over one million Facebook comments, examining the sentiment dynamics within and between communities discussing science and conspiracy news. They found that regardless of content, the longer the discussion continues, the more negative the emotions become. Similarly, Quattrociocchi et al. (2016) studied Italian and US Facebook users, conducting sentiment analysis on comments in online debates. Combining this with social network analysis, they discovered that the more active a polarized user is, the more they tend towards negative sentiments on both science and conspiracy posts.

Various hypotheses exist to explain social media polarization, and among them the most famous one is the "echo chamber effect" (Cinelli et al., 2021). This effect describes how social interactions on social media are driven by homophily, and thus users with similar ideologies tend to congregate, leading individuals to be surrounded by homophilic information. Colleoni et al. (2014) employed NLP techniques to analyze political orientation in Twitter texts, thereby measuring political homophily. Their findings showed that Democrats demonstrated higher political homophily, whereas Republicans following official Republican accounts exhibited even greater levels of homophily. Similarly, Jiang et al. (2021) used BERT to analyze COVID-19-related discourse on Twitter, confirming the existence of echo chambers. Gao et al. (2023b) focused on short video platforms, utilizing BERT to study comments on Chinese short video platforms. They found that echo chamber members tend to showcase themselves to attract peer attention, and cultural differences can impede the development of echo chambers. In addition, a separate study discovered that emotionally charged Twitter messages tend to be retweeted more frequently and rapidly compared to neutral ones (Stieglitz and Dang-Xuan, 2013).

While the aforementioned studies primarily adopted sentiment analysis to analyze text, some scholars have explored polarization through topic modeling methods. For example, Farrell (2016) utilized structural topic modeling to study texts related to climate change counter-movements, identifying key themes and measuring their prevalence. They found that organizations sponsored by businesses were more likely to write and disseminate texts aimed at polarizing climate change issues.

*Engagement.* Social media plays an increasingly crucial role in the political arena. Social media platforms provide individuals with a platform for large-scale, open, real-time

political discussion, enabling people to share information without geographic limits (Spaiser et al., 2017). They also facilitate the dissemination of political movement-related information, as seen in events like the "Twitter revolution" during the Arab Spring (Cottle, 2011). Individual expressions of political attitudes on social media, coupled with authorities' dissemination of information, have introduced new data and research topics into political science.

One of the widely discussed topics is how social media affects people's political engagement. Social media provides individuals with a space for political expression across distances, and numerous studies have explored individual online political engagement. In these studies, sentiment analysis based on text plays a crucial role. For instance, Shugars and Beauchamp (2019) investigated the reason and motivation for engagement in online political debates. Based on the results of sentiment analysis and topic modeling, they constructed a model to predict one's engagement in online debates, with an accuracy exceeding 98%. Field et al. (2022) utilized RoBERTa to study the emotions expressed in tweets related to participating in "Black Lives Matter" protests, finding that posts expressed a high degree of anger and disgust, but positive emotions such as friendship and pride resulting from the protests may outweigh other expressed emotions. Scholars have also inferred demographic information such as gender and ethnicity from Twitter/X data and analyzed the online political behavior and participation of different groups (Brandt et al., 2020).

However, social media is not merely a platform for individual expression. Similar to traditional media, various forces in social media potentially control the discourse. A study analyzing Twitter messages related to the 2011–2012 Russian protests found that pro-government users employed a variety of communication strategies to shift political discourse and marginalize opposite voices on the platform (Spaiser et al., 2017).

Social media is also seen as revolutionary and capable of sparking offline social movements (Harlow, 2012). Online connections reduce organizational costs, making activities more likely to erupt unexpectedly (Enikolopov et al., 2020). Related studies have investigated the dynamics from online discussions to offline movements. By conducting sentiment analysis on 65,613 tweets related to the Indian Nirbhaya protests, researchers found a significant similarity between the emotional patterns of online discussions and offline protests, suggesting resonance between online discussions and offline activities (Ahmed et al., 2017). Reda et al. (2024) conducted sentiment analysis and topic modeling on millions of tweets, constructed a score to quantitatively assess social movement tendencies, and found that this score can accurately predict resource mobilization within the same time frame. Gallacher et al. (2021) analyzed online conversations between members of protest groups from opposite sides of the political spectrum and violence occurrence during these protests and rallies. By examining 25 events, including protests, marches, or gatherings, they found that increased engagement between groups online is associated with increased violence when these groups meet in the real world.

*Bots.* The prevalence of automated accounts, such as social bots, is one of the distinguishing features of the online world. Social bots are programmed to fulfill specific tasks,

including disseminating messages or engaging in particular social behaviors, thereby influencing the online society (Ferrara et al., 2016). Detecting and analyzing content generated by these bots has become a new object of research attention.

Scholars have developed algorithms to detect the presence of social bots and measure their political inclinations (Stukal et al., 2019; Sanovich et al., 2018). Others have focused on the impact of social bots' activities on human behavior. For example, studies have revealed that social bots tend to propagate negative and inflammatory content, contributing to the polarization of online discussions (Stella et al., 2018). In another study, researchers analyzed the diffusion structure and content of political events based on sentiment analysis and network analysis, revealing that verified accounts were more visible than unverified bots in event coverage, but social bots attracted more attention than human accounts (González-Bailón and De Domenico, 2021). Researchers have also explored Coronavirus Conspiracy Talk, finding that both social bots and humans contribute to related discourse. In these scenarios, social bots are designed to create moral panic, while humans exploit conspiracy talk to gain attention (Greve et al., 2022).

Bots can also serve very positive purposes. Argyle et al. (2023a) developed a social bot to act as an at-scale, real-time moderator in divisive political conversations. This social bot can provide suggestions on language use during live discussions. Evidence suggests that these intervention measures enhance conversation quality and democratic reciprocity.

## Challenges and directions for future research

### *Challenges*

*Representativeness.* Although natural language processing provides researchers with a range of tools for processing and studying text data, it also raises some concerns. One of the common criticisms researchers face is the question of whether the samples of digitized text represent the overall population of interest. Although decreasing in recent years, a longstanding digital divide exists (DiMaggio and Bonikowski, 2008). On platforms like X, the users are compositionally different from and do not perfectly represent the entire population of the United States (Adams-Cohen, 2020; Diaz et al., 2016). Survey evidence indicates that around 42% of young people aged 18–29 use X, while this figure is only 6% among those aged 65 and above. Additionally, 25% of urban residents use X compared to only 13% of rural residents (Pew Research Center, 2024). These biases in user distribution pose potential threats to research results, and researchers must be aware of these threats.

*Inherent bias.* Another potential threat of NLP technology is the bias inherent in the models. Biases inherent in language can be absorbed during the machine-learning process and may alter NLP models' predictions (Bail, 2024). Researchers summarized four types of biases that NLP models may exhibit: label bias, where the output labels in the training data diverge substantially from the real world; selection bias, which refers to non-representative observations; over-amplification, where the model tends to

amplify small differences in predicted outcomes; and semantic bias, where embedding contains societal stereotypes (Shah et al., 2020). Almost all language models, from word embedding to large language models like BERT and GPT, cannot perfectly avoid semantic bias. Rozado (2023) conducted a multilingual political bias test on ChatGPT and found that it tends to hold left-leaning views, which may be related to the fact that the training materials for large language models are mainly sourced from the Internet, which is dominated by influential institutions in Western society. In addition to political bias, it is evident that NLP models may also exhibit ethnicity bias, gender stereotypes, and other biases (Gross, 2023; Rozado, 2020).

*Interpretability.* NLP techniques also face challenges in interpretability. Although some statistically based supervised machine-learning methods have good interpretability, deep learning models that handle more complex tasks are often considered black boxes. Apart from results and performance, it is difficult for us to understand what happens within the model. This to some extent limits the application scenarios of NLP models in social science research.

## Directions for future research

In the previous section, we primarily focused on how NLP technologies can assist researchers in extracting semantic information. However, with the rapid developments in LLMs, these models have demonstrated impressive capabilities in text understanding and generation. These advancements open up new possibilities for computational social science research, allowing us to explore how models can be adopted in various stages of research and in tasks beyond text analysis. Here, we will discuss three possible directions.

*Assist in research design and data labeling.* LLMs have the potential to serve as research assistants. For instance, researchers can use LLMs to help generate necessary research materials, such as images or texts needed in psychological experiments to evoke different emotions in subjects, or political texts conveying different ideologies (Bail, 2024). Besides, LLMs can also be applied in data annotation and thus replace the expensive and time-consuming human labeling. Many studies have shown that zero-shot GPT models can reach similar or even better performances than crowdsourcing platforms like Amazon Mechanical Turk in annotation tasks (Mellon et al., 2024; Ziems et al., 2024). Ziems conducted a systematic review about using LLMs for annotation tasks. They examined the performance of 13 different language models across 20 classification tasks at the word, sentence, and document levels, as well as five generation tasks. The results indicated that while zero-shot LLMs may not perform as well as fine-tuned RoBERTa in classification tasks, they can achieve or even exceed human annotators. Moreover, in generation tasks, LLMs generally outperform trained human annotators. Therefore, they suggest researchers to use zero-shot LLMs to assist data annotation tasks, but to be careful when conducting research in sensitive topics. Additionally, researchers can guide LLMs to complete more challenging generation tasks required in research.

In this part, three future topics are worth attention: (1) the potential biases that models may introduce into the labeled data, and how upcoming models will address this issue; (2) the possibility of multimodal data labeling (such as audio, images, etc.) brought by multimodal LLMs; (3) the performance of LLMs in labeling tasks across different languages and cultures, e.g. annotation tasks based on Chinese text.

*Simulate social behavior.* LLMs have been proved to have the ability to simulate individual personalities or possible behaviors (Argyle et al., 2023b; Bail, 2024), which makes it possible for LLM agents to simulate real social processes. This social simulation has two possible directions. The first is using LLM agents to replace certain participants in research. Argyle et al. (2023b) indicated that algorithmic biases in LLMs are related to demographics, and LLM-based samples are able to simulate real-world samples in various aspects. Similarly, Jiang et al. (2024) used prompts to assign different personas to LLMs and then conducted Big Five personality tests and story-writing tasks on these agents. The results showed that the agents behaved consistently with their assigned personas, both in the Big Five test and in the story-writing task. Larger models and more prompting data are expected to further enhance the performances of simulation. Systematic research is still necessary on how different models and different input would produce different simulation results.

The second direction is at the system level, where researchers can use LLM agents playing different roles to observe social processes within specific social contexts. Park et al. (2023) presented one of the most classic LLM-based social simulation researches, in which they constructed a small town with 25 agents, each with different personalities. Similar to human society, these agents exhibited emergent social behaviors in the town. This makes it possible to simulate social processes using LLM agents, and such simulation allows social science scholars to observe the interaction between individual micro-behaviors and macro-phenomena within specific social structures. Some researchers have constructed social media simulation systems, in which LLM agents trained with social media data interacted with each other (Gao et al., 2023a). The agents successfully predicted the diffusion of information and emotions. However, more discussion is needed on whether the simulation results of LLM agents are reliable.

*Facilitate causal inference.* Recently, the interdisciplinary field of LLMs and causal science has gained significant attention. In fact, there exists a mutually supportive relationship between LLMs and causality: LLMs can enhance causal estimation, while causality can also increase the robustness of LLMs and reduce issues such as bias and hallucination (Feder et al., 2022; Liu et al., 2024). Here, we will briefly discuss three examples of how LLMs enhance causality. Firstly, the collaboration between LLMs and causality makes more complex causal estimation possible, such as where textual data serve as confounders, treatments, or outcomes. For a comprehensive review, please read Feder et al. (2022). Multimodal models further enable causal estimation involving modalities such as images and sounds. Secondly, the language understanding capabilities of LLMs allow them to better extract commonsense causality from text. Compared to past methods based on keywords or linguistic patterns, LLMs can handle complex causal

relationships in text more effectively, enabling researchers to extract causal relationships from knowledge bases or scientific literature (Cui et al., 2024). Thirdly, researchers can utilize the generation ability of LLMs for data augmentation, generating counterfactual data that cannot be obtained in reality to achieve better causal estimation (Li et al., 2024).

With the release of more causal benchmarks and the enhancement of models' causal abilities, LLMs may be able to handle higher-dimensional data, discover more complex causal structures, and perform more robust causal estimation. However, the integration of LLM-based causal inference into social science research is still insufficient.

NLP is advancing rapidly, with new breakthroughs emerging every few months, or even monthly. We believe that social scientists need to proactively engage in this interdisciplinary dialogue, exploring more data, applying innovative techniques, and drawing more conclusions. This is not only because NLP makes research more convenient but also because in the era of data explosion, embracing new technologies to conduct research based on large-scale data is the only way to keep up with understanding and interpreting the world.

## Conclusion

Natural language processing techniques have long been accelerating social science research. Early attempts like dictionary-based methods are accurate and compact but require expert knowledge, and therefore pose scalability challenges particularly in large-scale complicated scenarios. An alternative solution extracts semantic information by representing documents in a word space, which may or may not be further projected onto an Euclidean space for geometric interpretation. Machine-learning techniques have empowered us to explore high-dimensional text data. Unsupervised learning methods like topic modeling are straightforward to implement but may produce results that are difficult to interpret if not appropriately configured. Conversely, supervised learning methods such as support vector machine and decision trees rely on human inputs to target well-defined research outcomes, at a significantly higher expense in hiring human helpers. To address the challenges of sparseness and high cost of human inputs, and the increasing need of processing complex text, pre-trained large language models are proposed to offer a promising head start on traditional natural language processing tasks. Instead of training from scratch, such tasks are training upon a pre-trained neural network which has encoded rich linguistic and semantic information extracted from vast corpora. Such methods significantly enhance text understanding capabilities and expand the boundary of text analysis with explosively new scenarios that were not possible before. However, along with their powerful ability of text understanding come high extensive computational expenses. The above natural language processing techniques have transformatively impacted social science research such as analyzing online engagement and exploring the science of science. Combined with unprecedented availability of digital text data, those techniques offer a powerful toolkit for revealing insights from vast and diverse data sources.

Natural language processing techniques do not come without challenges. Concerns in data representativeness, amplified social biases, and limited interpretability together call

for cautious engagement of social scientists. It is imperative for social science researchers to not only address methodological issues but also proactively engage with the natural language processing community to (re)design models in a manner more aligned with ethical regulations.

## Author contributions

Yuxin Hou and Junming Huang wrote and revised the manuscript.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by High-performance Computing Platform of Peking University.

## ORCID iDs

Yuxin Hou https://orcid.org/0000-0003-1368-7326
Junming Huang https://orcid.org/0000-0002-2532-4090

## References

Abid A, Farooqi M and Zou J (2021) Persistent anti-Muslim bias in large language models. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306.

Adams-Cohen NJ (2020) Policy change and public opinion: Measuring shifting political sentiment with social media data. *American Politics Research* 48(5): 612–621.

Ahmed S, Jaidka K and Cho J (2017) Tweeting India's Nirbhaya protest: A study of emotional dynamics in an online social movement. *Social Movement Studies* 16(4): 447–465.

Ajao O, Bhowmik D and Zargari S (2018) Fake news identification on Twitter with hybrid CNN and RNN models. In: *Proceedings of the 9th International Conference on Social Media and Society*, pp. 226–230.

Almazrouei E, Alobeidli H, Alshamsi A, et al. (2023) The Falcon series of open language models. arXiv: 2311.16867. https://arxiv.org/abs/2311.16867.

Angelov D (2020) Top2Vec: Distributed representations of topics. arXiv: 2008.09470. https://arxiv.org/abs/2008.09470.

Arbane M, Benlamri R, Brik Y, et al. (2023) Social media-based COVID-19 sentiment classification model using bi-LSTM. *Expert Systems with Applications* 212: 118710.

Argyle LP, Bail CA, Busby EC, et al. (2023a) Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120(41): e2311627120.

Argyle LP, Busby EC, Fulda N, et al. (2023b) Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3): 337–351.

Arlot S and Celisse A (2010) A survey of cross-validation procedures for model selection. *Statistics Surveys* 4: 40–79.

Bai J, Bai S, Chu Y, et al. (2023) Qwen technical report. arXiv: 2309.16609. https://arxiv.org/abs/2309.16609.

Bail CA (2014) The cultural environment: Measuring culture with big data. *Theory and Society* 43(3): 465–482.

Bail CA (2016) Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences* 113(42): 11823–11828.

Bail CA (2024) Can generative AI improve social science? *Proceedings of the National Academy of Sciences* 121(21): e2314021121.

Bailey AH, Williams A and Cimpian A (2022) Based on billions of words on the internet, people = men. *Science Advances* 8(13): eabm2463.

Bang Y, Cahyawijaya S, Lee N, et al. (2023) A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 675–718. DOI: 10.18653/v1/2023.ijcnlp-main.45.

Bao Y, Quan C, Wang L, et al. (2014) The role of pre-processing in Twitter sentiment analysis. In: Huang D-S, Jo K-H and Wang L (eds) *Intelligent Computing Methodologies*. Cham: Springer International Publishing, pp. 615–624.

Bender EM, Gebru T, McMillan-Major A, et al. (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*, pp. 610–623.

Bengio Y, Simard P and Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on Neural Networks* 5(2): 157–166.

Betti L, Abrate C and Kaltenbrunner A (2023) Large-scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science* 12(1): 1–22.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Blei DM and Smyth P (2017) Science and data science. *Proceedings of the National Academy of Sciences* 114(33): 8689–8692.

Bojanowski P, Grave E, Joulin A, et al. (2017) Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135–146.

Bonikowski B, Luo Y and Stuhler O (2022) Politics as usual? Measuring populism, nationalism, and authoritarianism in U.S. Presidential campaigns (1952–2020) with neural language models. *Sociological Methods & Research* 51(4): 1721–1787.

Boussalis C, Coan TG and Holman MR (2018) Climate change communication from cities in the USA. *Climatic Change* 149(2): 173–187.

Brandt J, Buckingham K, Buntain C, et al. (2020) Identifying social media user demographics and topic diversity with computational social science: A case study of a major international policy forum. *Journal of Computational Social Science* 3(1): 167–188.

Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.

Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877–1901.

Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Catalinac A (2016) From pork to policy: The rise of programmatic campaigning in Japanese elections. *The Journal of Politics* 78(1): 1–18.

Chang Y, Wang X, Wang J, et al. (2023) A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15(3): 1–45.

Charlesworth TES, Yang V, Mann TC, et al. (2021) Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science* 32(2): 218–240.

Chen X, Xu L, Liu Z, et al. (2015) Joint learning of character and word embeddings. In: *Twenty-fourth International Joint Conference on Artificial Intelligence*, pp. 1236–1242.

Cheng M, Smith DS, Ren X, et al. (2023) How new ideas diffuse in science. *American Sociological Review* 88(3): 522–561.

Cheng X, Yan X, Lan Y, et al. (2014) BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26(12): 2928–2941.

Chinn S, Sol Hart P and Soroka S (2020) Politicization and polarization in climate change news content, 1985–2017. *Science Communication* 42(1): 112–129.

Cho K, van Merrienboer B, Gulcehre C, et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 1, pp. 1724–1734. DOI: 10.3115/v1/D14-1179.

Chowdhury GG (2003) Natural language processing. *Annual Review of Information Science and Technology* 37(1): 51–89.

Cinelli M, De Francisci Morales G, Galeazzi A, et al. (2021) The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118(9): e2023301118.

Colleoni E, Rozza A and Arvidsson A (2014) Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication* 64(2): 317–332.

Conte R, Gilbert N, Bonelli G, et al. (2012) Manifesto of computational social science. *The European Physical Journal Special Topics* 214(1): 325–346.

Cortes C and Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3): 273–297.

Cottle S (2011) Media and the Arab uprisings of 2011: Research notes. *Journalism* 12(5): 647–659.

Cui S, Jin Z, Schölkopf B, et al. (2024) The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, pp. 16722–16762. DOI: 10.18653/v1/2024.emnlp-main.932.

Cui Y, Che W, Liu T, et al. (2021) Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 3504–3514.

Czymara CS, Langenkamp A and Cano T (2021) Cause for concerns: Gender inequality in experiencing the COVID-19 lockdown in Germany. *European Societies* 23(sup1): S68–S81.

DeFranza D, Mishra H and Mishra A (2020) How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology* 119(1): 7–22.

Devlin J, Chang M-W, Lee K, et al. (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*, Vol. 1, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Diaz F, Gamon M, Hofman JM, et al. (2016) Online and social media data as an imperfect continuous panel survey. *PLOS One* 11(1): e0145406.

DiMaggio P and Bonikowski B (2008) Make money surfing the web? The impact of internet use on the earnings of U.S. Workers. *American Sociological Review* 73(2): 227–250.

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.

D'Orazio V, Landis ST, Palmer G, et al. (2014) Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis* 22(2): 224–242.

Dubey A, Jauhri A, Pandey A, et al. (2024) The Llama 3 herd of models. arXiv: 2407.21783. https://arxiv.org/abs/2407.21783.

Edelmann A, Wolff T, Montagne D, et al. (2020) Computational social science and sociology. *Annual Review of Sociology* 46(1): 61–81.

Egger R and Yu J (2022) A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology* 7: 886498.

Enikolopov R, Makarin A and Petrova M (2020) Social media and protest participation: Evidence from Russia. *Econometrica* 88(4): 1479–1514.

Enria L, Waterlow N, Rogers NT, et al. (2021) Trust and transparency in times of crisis: Results from an online survey during the first wave (April 2020) of the COVID-19 epidemic in the UK. *PLOS One* 16(2): e0239247.

Ezugwu AE, Ikotun AM, Oyelade OO, et al. (2022) A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence* 110: 104743.

Fang Y, Yang S, Zhao B, et al. (2021) Cyberbullying detection in social networks using bi-GRU with self-attention mechanism. *Information* 12(4): 171.

Farrell J (2016) Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences* 113(1): 92–97.

Feder A, Keith KA, Manzoor E, et al. (2022) Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics* 10: 1138–1158.

Ferguson CJ (2021) Does the internet make the world worse? Depression, aggression and polarization in the social media age. *Bulletin of Science, Technology & Society* 41(4): 116–135.

Ferrara E, Varol O, Davis C, et al. (2016) The rise of social bots. *Communications of the ACM* 59(7): 96–104.

Field A, Park CY, Theophilo A, et al. (2022) An analysis of emotions and the prominence of positivity in #BlackLivesMatter tweets. *Proceedings of the National Academy of Sciences* 119(35): e2205767119.

Flores RD (2017) Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology* 123(2): 333–384.

Fortunato S, Bergstrom CT, Börner K, et al. (2018) Science of science. *Science* 359(6379): eaao0185.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5): 1189–1232.

Gaikwad M, Ahirrao S, Phansalkar S, et al. (2021) Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access* 9: 48364–48404.

Gallacher JD, Heerdink MW and Hewstone M (2021) Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society* 7(1): 2056305120984445.

Gallegos IO, Rossi RA, Barrow J, et al. (2024) Bias and fairness in large language models: A survey. *Computational Linguistics* 50(3): 1097–1179.

Gao C, Lan X, Lu Z, et al. (2023a) S3: Social-network simulation system with large language model-empowered agents. arXiv: 2307.14984. https://arxiv.org/abs/2307.14984.

Gao Y, Liu F and Gao L (2023b) Echo chamber effects on short video platforms. *Scientific Reports* 13(1): 6282.

Garg N, Schiebinger L, Jurafsky D, et al. (2018) Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16): E3635–E3644.

Geiger A (2014) Political polarization in the American public. Available at: https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/.

Goldberg A, Srivastava SB, Govind Manian V, et al. (2016) Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *American Sociological Review* 81(6): 1190–1222.

González-Bailón S and De Domenico M (2021) Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences* 118(11): e2013443118.

Greve HR, Rao H, Vicinanza P, et al. (2022) Online conspiracy groups: Micro-bloggers, bots, and Coronavirus Conspiracy Talk on Twitter. *American Sociological Review* 87(6): 919–949.

Grimmer J and Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv: 2203.05794. https://arxiv.org/abs/2203.05794.

Gross N (2023) What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI. *Social Sciences* 12(8): 35.

Grothe-Hammer M and Kohl S (2020) The decline of organizational sociology? An empirical analysis of research trends in leading journals across half a century. *Current Sociology* 68(4): 419–442.

Guo L, Vargo CJ, Pan Z, et al. (2016) Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly* 93(2): 332–359.

Gutenberg (2024) Project Gutenberg. Available at: https://www.gutenberg.org/.

Han S (2022) Elite polarization in South Korea: Evidence from a natural language processing model. *Journal of East Asian Studies* 22(1): 45–75.

Han X, Wang J, Zhang M, et al. (2020) Using social media to mine and analyze public opinion related to COVID-19 in China. *International Journal of Environmental Research and Public Health* 17(8): 2788.

Hannigan TR, Haans RFJ, Vakili K, et al. (2019) Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals* 13(2): 586–632.

Harlow S (2012) Social media and social movements: Facebook and an online Guatemalan justice movement that moved offline. *New Media & Society* 14(2): 225–243.

Harris ZS (1954) Distributional structure. *Word* 10(2-3): 146–162.

Hart PS, Chinn S and Soroka S (2020) Politicization and polarization in COVID-19 news coverage. *Science Communication* 42(5): 679–697.

Havey NF (2020) Partisan public health: How does political ideology influence support for COVID-19-related misinformation? *Journal of Computational Social Science* 3(2): 319–342.

Hillard D, Purpura S and Wilkerson J (2008) Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics* 4(4): 31–46.

Hirschberg J and Manning CD (2015) Advances in natural language processing. *Science* 349(6245): 261–266.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Hong L and Davison BD (2010) Empirical study of topic modeling in Twitter. In: *Proceedings of the First Workshop on Social Media Analytics*, *SOMA'10*. New York, NY: Association for Computing Machinery, pp. 80–88.

Huang J, Cook GG and Xie Y (2021) Large-scale quantitative evidence of media impact on public opinion toward China. *Humanities & Social Sciences Communications* 8: 181.

Jiang H, Zhang X, Cao X, et al. (2024) PersonaLLM: Investigating the ability of large language models to express personality traits. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. arXiv: 21305.02547: 3605–3627. DOI: 10.18653/v1/2024.findings-naacl.229.

Jiang J, Ren X and Ferrara E (2021) Social media polarization and echo chambers in the context of COVID-19: Case study. *JMIRX Med* 2(3): e29570.

Jiao M and Luo Z (2021) Gender bias hidden behind Chinese word embeddings: The case of Chinese adjectives. In: *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, Vol. 1, pp. 8–15.

Johnson SC (1967) Hierarchical clustering schemes. *Psychometrika* 32(3): 241–254.

Jones JJ, Amin MR, Kim J, et al. (2020) Stereotypical gender associations in language have decreased over time. *Sociological Science* 7: 1–35.

Jungherr A (2016) Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics* 13(1): 72–91.

Kaplan J, McCandlish S, Henighan T, et al. (2020) Scaling laws for neural language models. arXiv: 2001.08361. https://arxiv.org/abs/2001.08361.

Karinshak E, Liu SX, Park JS, et al. (2023) Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1): 1–29.

Kennedy I, Hess C, Paullada A, et al. (2021) Racialized discourse in Seattle rental ad texts. *Social Forces* 99(4): 1432–1456.

Key EM and Sumner JL (2019) You research like a girl: Gendered research agendas and their implications. *PS: Political Science & Politics* 52(4): 663–668.

Khodak M, Saunshi N, Liang Y, et al. (2018) A la carte embedding: Cheap but effective induction of semantic feature vectors. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 12–22. DOI: 10.18653/v1/P18-1002.

Kotek H, Dockum R and Sun D (2023) Gender bias and stereotypes in large language models. In Proceedings of the ACM Collective Intelligence Conference, 12–24.

Kozlowski AC, Taddy M and Evans JA (2019) The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84(5): 905–949.

Krippendorff K (2019) *Content Analysis: An Introduction to Its Methodology*, 4th Edition. Thousand Oaks, CA: SAGE Publications, Inc.

Kroon AC, Van der Meer Toni GLA and Vliegenthart R (2022) Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research* 4(2): 528–570.

Lai A, Brown MA, Bisbee J, et al. (2022) Estimating the ideology of political YouTube videos. *Political Analysis* 32(3): 345–360.

Lam C and Chan CS (2024) A text mining and machine learning study on the trends of and dynamics between collective action and mental health in politically polarized online environments. *Journal of Computational Social Science* 7: 1379–1401.

Lan Z, Chen M, Goodman S, et al. (2020) ALBERT: A lite BERT for self-supervised learning of language representations. arXiv: 1909.11942. https://arxiv.org/abs/1909.11942.

Larregue J and Nielsen MW (2024) Knowledge hierarchies and gender disparities in social science funding. *Sociology* 58(1): 45–65.

Lazer DMJ, Pentland A, Watts DJ, et al. (2020) Computational social science: Obstacles and opportunities. *Science* 369(6507): 1060–1062.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Lee G-H and Chen Y-N (2018) Muse: Modularizing unsupervised sense embeddings. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, pp. 327–337. DOI: 10.18653/v1/D17-1034.

Lee MHJ, Montgomery JM and Lai CK (2024) America's racial framework of superiority and Americanness embedded in natural language. *PNAS Nexus* 3(1): pgad485.

Li S, Zhao Z, Hu R, et al. (2018) Analogical reasoning on Chinese morphological and semantic relations. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, pp. 138–143. DOI: 10.18653/v1/P18-2023.

Li Y, Xu M, Miao X, et al. (2024) Prompting large language models for counterfactual generation: An empirical study. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Vol. 1, pp. 13201–13221.

Light R and Odden C (2017) Managing the boundaries of taste: Culture, valuation, and computational social science. *Social Forces* 96(2): 877–908.

Liu Q, Kusner MJ and Blunsom P (2020) A survey on contextual embeddings. arXiv:2003.07278. https://arxiv.org/abs/2003.07278.

Liu X, Xu P, Wu J, et al. (2024) Large language models and causal inference in collaboration: A comprehensive survey. arXiv:2403.09606. https://arxiv.org/abs/2403.09606.

Liu Y, Ott M, Goyal N, et al. (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692. https://arxiv.org/abs/1907.11692.

Livne A, Simmons M, Adar E, et al. (2011) The party is over here: Structure and content in the 2010 election. *Proceedings of the International AAAI Conference on Web and Social Media* 5(1): 201–208.

Lucy L, Demszky D, Bromley P, et al. (2020) Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *AERA Open* 6(3): 2332858420940312.

Macqueen J (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Oakland, CA: University of California Press, pp. 281–297.

Mann A (2016) Computational social science. *Proceedings of the National Academy of Sciences* 113(3): 468–470.

Markowitz DM (2022) Gender and ethnicity bias in medicine: A text analysis of 1.8 million critical care records. *PNAS Nexus* 1(4): pgac157.

Matalon Y, Magdaci O, Almozlino A, et al. (2021) Using sentiment analysis to predict opinion inversion in tweets of political communication. *Scientific Reports* 11(1): 7250.

McCumber A and Davis A (2024) Elite environmental aesthetics: Placing nature in a changing climate. *American Journal of Cultural Sociology* 12(1): 53–84.

McHugh ML (2012) Interrater reliability: The kappa statistic. *Biochemia Medica* 22(3): 276–282.

McMahan P and Evans J (2018) Ambiguity and engagement. *American Journal of Sociology* 124(3): 860–912.

Mellon J, Bailey J, Scott R, et al. (2024) Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics* 11(1): 20531680241231468.

Michel J-B, Shen YK, Aiden AP, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182.

Mikolov T, Chen K, Corrado G, et al. (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781. https://arxiv.org/abs/1301.3781.

Mu G, Liao Z, Li J, et al. (2023) IPSO-LSTM hybrid model for predicting online public opinion trends in emergencies. *PLOS One* 18(10): e0292677.

Mukherjee S and Bala PK (2017) Sarcasm detection in microblogs using naïve Bayes and fuzzy clustering. *Technology in Society* 48: 19–27.

Murshed BAH, Abawajy J, Mallappa S, et al. (2022) DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access* 10: 25857–25871.

Napp C (2023) Gender stereotypes embedded in natural language are stronger in more economically developed and individualistic countries. *PNAS Nexus* 2(11): pgad355.

Naseem U, Razzak I, Khan SK, et al. (2021) A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Transactions on Asian and Low-Resource Language Information Processing* 20(5): 74:1–74:35.

Nemes L and Kiss A (2021) Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication* 5(1): 1–15.

Ng R, Indran N and Liu L (2022) Ageism on Twitter during the COVID-19 pandemic. *Journal of Social Issues* 78(4): 842–859.

OpenAI, Achiam J, Adler S, et al. (2024) Gpt-4 technical report. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774.

Park H-S and Jun C-H (2009) A simple and fast algorithm for K-Medoids clustering. *Expert Systems with Applications* 36(2): 3336–3341.

Park JS, O'Brien JC, Cai CJ, et al. (2023) Generative agents: Interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Vol. 2, pp. 1–22. DOI: 10.1145/3586183.3606763.

Parthasarathy R, Rao V and Palaniswamy N (2019) Deliberative democracy in an unequal world: A text-as-data study of South India's village assemblies. *American Political Science Review* 113(3): 623–640.

Pennebaker JW, Boyd RL, Jordan K, et al. (2015) The development and psychometric properties of LIWC2015. Available at: http://hdl.handle.net/2152/31333.

Pennington J, Socher R and Manning C (2014) GloVe: Global vectors for word representation. In: Moschitti A, Pang B and Daelemans W (eds) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Doha: Association for Computational Linguistics.

Peters ME, Neumann M, Iyyer M, et al. (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*, Vol. 1, pp. 2227–2237. DOI: 10.18653/v1/N18-1202.

Pew Research Center (2024) Social Media Fact Sheet. Available at: https://www.pewresearch.org/internet/fact-sheet/social-media/.

Pickett AC and Valdez D (2023) Mining online discourse related to transgender exclusive policies in interscholastic sport: An exploratory natural language processing study. *Sexuality Research and Social Policy* 20(3): 936–949.

Poulos J and Zeng S (2021) RNN-based counterfactual prediction, with an application to homestead policy and public schooling. *Journal of the Royal Statistical Society Series C: Applied Statistics* 70(4): 1124–1139.

Pujianto U, Azizah EN and Damayanti AS (2017) Naive Bayes using to predict students' academic performance at faculty of literature. In: *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pp. 163–169. DOI: 10.1109/ICEEIE.2017.8328782.

Qin C, Zhang A, Zhang Z, et al. (2023) Is ChatGPT a general-purpose natural language processing task solver? In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Vol. 1, pp. 1339–1384. DOI: 10.18653/v1/2023.emnlp-main.85.

Quattrociocchi W, Scala A and Sunstein CR (2016) Echo Chambers on Facebook. Available at: https://ssrn.com/abstract=2795110 or http://dx.doi.org/10.2139/ssrn.2795110.

Quinlan JR (1986) Induction of decision trees. *Machine Learning* 1(1): 81–106.

Quinn KM, Monroe BL, Colaresi M, et al. (2010) How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1): 209–228.

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving language understanding by generative pre-training. *OpenAI Blog*. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford A, Wu J, Child R, et al. (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1(8): 9. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rasmussen C (1999) The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems* 12: 554–560.

Rathje S, Mirea D-M, Sucholutsky I, et al. (2024) GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* 121(34): e2308950121.

Reda AA, Sinanoglu S and Abdalla M (2024) Mobilizing the masses: Measuring resource mobilization on Twitter. *Sociological Methods & Research* 53(1): 153–192.

Roberts ME, Stewart BM, Tingley D, et al. (2013) The structural topic model and applied social science. In: *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, Harrahs and Harveys, Lake Tahoe, USA, Vol. 4, pp. 1–20.

Roberts ME, Stewart BM, Tingley D, et al. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Rodriguez PL, Spirling A and Stewart BM (2023) Embedding regression: Models for context-specific description and inference. *American Political Science Review* 117(4): 1255–1274.

Rothschild JE, Howat AJ, Shafranek RM, et al. (2019) Pigeonholing partisans: Stereotypes of party supporters and partisan polarization. *Political Behavior* 41: 423–443.

Rozado D (2020) Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLOS One* 15(4): e0231189.

Rozado D (2023) The political biases of ChatGPT. *Social Sciences* 12(3): 148.

Rumelhart DE, Hinton GE and Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088): 533–536.

Salton G (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc.

Sanovich S, Stukal D and Tucker JA (2018) Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia. *Comparative Politics* 50(3): 435–482.

Schuster M and Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11): 2673–2681.

Schwartz AH and Ungar LH (2015) Data-driven content analysis of social media: A systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science* 659(1): 78–94.

Schwemmer C and Wieczorek O (2020) The methodological divide of sociology: Evidence from two decades of journal publications. *Sociology* 54(1): 3–21.

Shah DS, Andrew Schwartz H and Hovy D (2020) Predictive biases in natural language processing models: A conceptual framework and overview. In: Jurafsky D, Chai J, Schluter N, et al. (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5248–5264. Association for Computational Linguistics. DOI: 10.18653/v1/ 2020.acl-main.468.

Shi W-Z, Zeng F, Zhang A, et al. (2022) Online public opinion during the first epidemic wave of COVID-19 in China based on Weibo data. *Humanities and Social Sciences Communications* 9(1): 1–10.

Shugars S and Beauchamp N (2019) Why keep arguing? Predicting engagement in political conversations online. *Sage Open* 9(1): 2158244019828850.

Sivakumar S and Rajalakshmi R (2022) Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers. *Social Network Analysis and Mining* 12(1): 04.

Smith N and Graham T (2019) Mapping the anti-vaccination movement on Facebook. *Information Communication & Society* 22(9): 1310–1327.

Snefjella B and Kuperman V (2015) Concreteness and psychological distance in natural language use. *Psychological Science* 26(9): 1449–1460.

Spaiser V, Chadefaux T, Donnay K, et al. (2017) Communication power struggles on social media: A case study of the 2011–12 Russian protests. *Journal of Information Technology & Politics* 14(2): 132–153.

Statista (2022) X/Twitter: Number of users worldwide 2024. Available at: https://www.statista.com/statistics/303681/twitter-users-worldwide/.

Stella M, Ferrara E and De Domenico M (2018) Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115(49): 12435–12440.

Stieglitz S and Dang-Xuan L (2013) Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems* 29(4): 217–248.

Stier S, Bleier A, Lietz H, et al. (2018) Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication* 35(1): 50–74.

Stoltz DS, Taylor MA and Dudley JSK (2024) A tool kit for relation induction in text analysis. *Sociological Methods & Research.* Epub ahead of print. DOI: 10.1177/00491241241233242.

Stone PJ, Dunphy DC and Smith MS (1966) *The General Inquirer: A Computer Approach to Content Analysis.* Oxford, England: M.I.T. Press.

Street W, Siy JO, Keeling G, et al. (2024) LLMs achieve adult human performance on higher-order theory of mind tasks. arXiv:2405.18870. https://arxiv.org/abs/2405.18870.

Stukal D, Sanovich S, Tucker JA, et al. (2019) For whom the bot tolls: A neural networks approach to measuring political orientation of Twitter bots in Russia. *Sage Open* 9(2): 2158244019827715.

Tausczik YR and Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.

Theocharis Y and Jungherr A (2021) Computational social science and the study of political communication. *Political Communication* 38(1-2): 1–22.

Touvron H, Lavril T, Izacard G, et al. (2023a) Llama: Open and efficient foundation language models. arXiv:2302.13971. https://arxiv.org/abs/2302.13971.

Touvron H, Martin L, Stone K, et al. (2023b) Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. https://arxiv.org/abs/2307.09288.

Trask A, Michalak P and Liu J (2015) Sense2Vec – a fast and accurate method for word sense disambiguation in neural word embeddings. arXiv:1511.06388. https://arxiv.org/abs/1511.06388.

Tumasjan A, Sprenger TO, Sandner PG, et al. (2011) Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review* 29(4): 402–418.

Tvinnereim E and Fløttum K (2015) Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change* 5(8): 744–747.

Valdez D, Soto-Vásquez AD and Montenegro MS (2023) Geospatial vaccine misinformation risk on social media: Online insights from an English/Spanish natural language processing (NLP) analysis of vaccine-related tweets. *Social Science & Medicine* 339: 116365.

van Atteveldt W, van der Velden MACG and Boukes M (2021) The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures* 15(2): 121–140.

van Loon A, Giorgi S, Willer R, et al. (2022) Negative associations in word embeddings predict anti-Black bias across regions—but only via name frequency. *Proceedings of the International AAAI Conference on Web and Social Media* 16: 1419–1424.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: *Advances in Neural Information Processing System*s, Vol. 30. Curran Associates, Inc.

Vijayakumar AP and Rajam VMA (2024) Detection and context reconstruction of sub-events that influence the course of a news event from microblog discussions. *Journal of Computational Social Science* 7: 1483–1517.

Vilhena DA, Foster JG, Rosvall M, et al. (2014) Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science* 1: 221–223.

Wang P, Zheng X, Ai G, et al. (2020) Time series prediction for the epidemic trends of COVID-19 using the improved lstm deep learning method: Case studies in Russia, Peru and Iran. *Chaos, Solitons & Fractals* 140: 110214.

Wang Y, Zhong W, Li L, et al. (2023) Aligning large language models with human: A survey. arXiv:2307.12966. https://arxiv.org/abs/2307.12966.

Wei L, Yao E and Zhang H (2023) Authoritarian responsiveness and political attitudes during COVID-19: Evidence from Weibo and a survey experiment. *Chinese Sociological Review* 55(1): 1–37.

Wikipedia (2024) Wikipedia: Size of Wikipedia. Wikipedia. Available at: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

Wilkerson J and Casas A (2017) Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science* 20: 529–544.

Wolpert DH and Macready WG (1997) No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* 1(1): 67–82.

Wu PY, Nagler J, Tucker JA, et al. (2023) Large language models can be used to estimate the latent positions of politicians. arXiv:2303.12057. https://arxiv.org/abs/2303.12057.

Xiang X, Lu X, Halavanau A, et al. (2021) Modern senicide in the face of a pandemic: An examination of public discourse and sentiment about older adults and COVID-19 using machine learning. *The Journals of Gerontology: Series B* 76(4): e190–e200.

Xie R, Chu SKW, Chiu DKW, et al. (2021) Exploring public response to COVID-19 on Weibo with LDA topic modeling and sentiment analysis. *Data and Information Management* 5(1): 86–99.

Xie Y, Yang F, Huang J, et al. (2024) Declining Chinese attitudes toward the United States amid COVID-19. *Proceedings of the National Academy of Sciences* 121(21): e2322920121.

Xu D and Tian Y (2015) A comprehensive survey of clustering algorithms. *Annals of Data Science* 2: 165–193.

Xu G, Meng Y, Qiu X, et al. (2019) Sentiment analysis of comment texts based on biLSTM. *IEEE Access* 7: 51522–51532.

Xue N, Xia F, Chiou Fd, et al. (2005) Building a large annotated Chinese corpus: The Penn Chinese treebank. *Journal of Natural Language Engineering* 11(2): 207–238.

Yan Y, Li Z and Meng T (2024) Conceptions of democracy in China: New evidence from the structural topic model. *International Political Science Review*. DOI: 10.1177/01925121241266228.

Yang J, Zhang Y and Liang S (2018) Subword encoding in lattice LSTM for Chinese word segmentation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1: 2720–2725. DOI: 10.18653/v1/N19-1278.

Young L and Soroka S (2012) Affective news: The automated coding of sentiment in political texts. *Political Communication* 29(2): 205–231.

Yu S, He S, Cai Z, et al. (2022) Exploring public sentiment during COVID-19: A cross country analysis. *IEEE Transactions on Computational Social Systems* 10(3): 1083–1094.

Zeng A, Xu B, Wang B, et al. (2024) ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools. arXiv:2406.12793. https://arxiv.org/abs/2406.12793.

Zeng W, Ren X, Su T, et al. (2021) Pangu: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. arXiv:2104.12369. https://arxiv.org/abs/2104.12369.

Zhang L, Banerjee M, Wang S, et al. (2023) The fragility of artists' reputations from 1795 to 2020. *Proceedings of the National Academy of Sciences* 120(35): e2302269120.

Zhao WX, Zhou K, Li J, et al. (2023) A survey of large language models. arXiv: 2303.18223. https://arxiv.org/abs/2303.18223.

Zhao X and Wong C-W (2023) Automated measures of sentiment via transformer- and lexicon-based sentiment analysis (TLSA). *Journal of Computational Social Science* 7: 145–170.

Zhu B, Zheng X, Liu H, et al. (2020) Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos, Solitons & Fractals* 140: 110123.

Zhuang F, Qi Z, Duan K, et al. (2020) A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1): 43–76.

Ziems C, Held W, Shaikh O, et al. (2024) Can large language models transform computational social science? *Computational Linguistics* 50(1): 237–291.

Zollo F, Novak PK, Del Vicario M, et al. (2015) Emotional dynamics in the age of misinformation. *PLOS One* 10(9): e0138740.