Frontiers in Communication

TYPE Review
PUBLISHED 02 October 2024
DOI 10.3389/fcomm.2024.1457433

Check for updates

OPEN ACCESS

EDITED BY
Koldobika Meso Ayerdi,
University of the Basque Country, Spain

REVIEWED BY
Carlos Lopezosa,
University of Barcelona, Spain
Jon Sedano,
University of Malaga, Spain

*CORRESPONDENCE
Davide Bassi
✉ davide.bassi@usc.es

RECEIVED 30 June 2024
ACCEPTED 13 September 2024
PUBLISHED 02 October 2024

CITATION
Bassi D, Fomsgaard S and Pereira-Fariña M (2024) Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. *Front. Commun.* 9:1457433.
doi: 10.3389/fcomm.2024.1457433

Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion

Davide Bassi[1]*, Søren Fomsgaard[2] and Martín Pereira-Fariña[3]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain, [2]Groupe de recherche en informatique, image et instrumentation de Caen (GREYC), Université de Caen Normandie, Caen, France, [3]Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

The proliferation of digital communication has profoundly transformed the landscape of persuasive discourse. Online platforms have amplified the reach and impact of persuasive techniques. However, they have also enabled the rapid spread of manipulative content, targeted propaganda, and divisive rhetoric. Consequently, a wide range of computational approaches has emerged to address the multifaceted nature of digital persuasion, to detect and mitigate its harmful practices. In light of this, the paper surveys computational methods for detecting persuasive means in digital communication, focusing on how they integrate humanistic knowledge to operationalize this construct. Additionally, special emphasis is placed on models' explainability, a pivotal aspect considering these models are used by institutions to influence societal interactions. For the analysis, two primary perspectives in persuasion are defined: linguistic and argumentative. The linguistic approach analyzes specific textual features, allowing for highly accountable algorithms based on explicit rules. The argumentative approach focuses on broader persuasive mechanisms, offering greater scalability but often resulting in less explainable models due to their complexity. This tension between model sophistication and explainability presents a key challenge in developing effective and transparent persuasion detection systems. The results highlight the spectrum of methodologies for studying persuasion, ranging from analyzing stylistic elements to detecting explicitly propagandist messages. Our findings highlight two key challenges in using these algorithms to tackle societal issues of persuasion misuse: the opacity of deep learning models and the absence of a theoretically grounded distinction between vicious and virtuous persuasion. To address these challenges, we propose integrating social sciences and humanities theories to enhance the effectiveness and ethical robustness of persuasion detection systems. This interdisciplinary approach enables a more nuanced characterization of text, facilitating the differentiation between vicious and virtuous persuasion through analysis of rhetorical, argumentative, and emotional aspects. We emphasize the potential of hybrid approaches that combine rule-based methods with deep learning techniques, as these offer a promising avenue for implementing this interdisciplinary framework. The paper concludes by outlining future challenges, including the importance of multimodal and multilingual analysis, ethical considerations in handling user-generated data and the growing challenge of distinguishing between human and AI-generated persuasive content.

KEYWORDS

persuasion, discourse analysis, natural language processing, machine learning, rhetoric, digital humanities, computational social sciences

1 Introduction

1.1 Defining persuasion

According to Duffy and Thorson (2016), communication as a whole may be understood as a persuasive effort, since speakers interact with each other in a goal-oriented way. This aligns with Mercier and Sperber (2017), who emphasizes that our social interactions and communications commonly aim to justify our beliefs and actions to others, effectively persuading the latter of the value of the former. This phenomenon is particularly evident in democratic contexts, where winning an election hinges primarily upon the quantity of individuals whom the candidate has effectively garnered through discursive means (Partington and Taylor, 2018).

The concept of persuasion emerges as inherently complex and multifaceted, having undergone various conceptualizations and connotations across different historical periods. Initially perceived as an element to foster civic education and engagement, critiques of the deceitful tactics employed by rhetoricians led to the construct of persuasion often being associated with, and sometimes conflated with, negatively connoted constructs such as "manipulation" (Nettel and Roque, 2012; Klemp, 2010) and "propaganda" (Godber and Origgi, 2023; Jowett and O'donnell, 2018). Despite the conceptual distinction between these constructs (Nettel and Roque, 2012; Godber and Origgi, 2023; Jowett and O'donnell, 2018), there is an increasing awareness regarding the detrimental effects of what could be termed "persuasion misuses," which highlight how the potentially harmful applications of persuasion can pose significant threats to the public sphere (we elaborate more on these distinctions in Section 4). This issue is most prominently exemplified by the mass manipulation practiced by totalitarian regimes (Petrova and Yanagizawa-Drott, 2016), often aiming to polarize populations by reinforcing existing beliefs rather than changing minds (Mercier, 2020), but is also evident in more commonplace advertising campaigns (Villarán, 2017).

Particularly in the digital age, the adverse facets of persuasion have garnered significant attention in the scientific community. The rise of internet and data digitization, in fact, has led to an unprecedented surge in data creation, aggregation, and transformation (Haq et al., 2020; Kitchin, 2014). Online platforms not only amass vast amounts of user-generated data, but also facilitate personalized message dissemination to a diverse audience (Zarouali et al., 2022).

One contemporary worry, in a similar vein, stems from the recent advent of large language models, which are capable of writing texts that are both cohesive and coherent enough to come off as being nearly as persuasive human-authored texts to non-expert audiences (Goldstein et al., 2024). As the gap of linguistic differences between human and machine authorship appears to diminish, with no signs of slowing down, we might find ourselves in a future where (illegitimate and unfair) political persuasion can achieve an unprecedented scale in online spaces and thus damage and pollute the informational environments where legitimate democratic discussion takes place.

Various political actors, in fact, have adeptly leveraged these platforms to propagate their ideologies (Zarouali et al., 2022; Haq et al., 2020). While this intensifies the reach and efficacy of efforts to engage the public into political discourse, it also represents a threat to transparent democratic deliberations. This is evidenced by the propagandist campaigns carried out by authoritarian regimes in recent years (Feldstein, 2023), as well as the influence exerted on election outcomes (Goovaerts and Marien, 2020; House, 2019).

In this context, a significant need for methodologies that leverage human expertise and artificial intelligence to autonomously analyze vast amounts of online data has emerged. Such approaches could equip institutions and citizens with the necessary tools to address the risks associated with online persuasion (Nannini et al., 2024).

1.2 Understanding persuasion

Since the inception of the study of persuasion, scholars have aimed at identifying the most effective ways to craft messages to achieve public endorsement (Demirdöğen, 2016). From the 20^{th} century, the spread of mass media coincided with the rise of other scientific disciplines that pay attention to influence processes. Social psychology, for instance, has aimed at explaining the psychological processes underling social influence and attitude changes (O'Keefe, 2009; Gardikiotis and Crano, 2015). Neuroscience, instead, has focused on elucidating the neurocognitive networks associated with feeling persuaded (Falk et al., 2010).

While these approaches are deeply intertwined and mutually informative, when addressing the automatic detection of persuasion misuses in online environments, language tends to be regarded as the most important (O'Keefe, 2009).

The study of persuasive use of language boasts an ancient and illustrious tradition, that can be organized in two main directions (see also Pauli et al., 2022), although frequently overlapping:

- **Argumentative**: with its roots in Aristotle's "Rhetoric" (especially the *logos* dimension) (Demirdöğen, 2016), this line of research has been devoted to analyze how the argumentative structure of messages can affect persuasiveness. Initially grounded in the assessment of the logical coherence of messages, this line of inquiry progressively has embraced a more pragmatic approach, incorporating for instance informal logic (Wagemans, 2023), as well as the analysis of the strategic use of audience's values and beliefs to generate persuasion effects (Perelman and Olbrechts-Tyteca, 1971).
- **Linguistic/semantic**: contemporary psychologically-oriented methodologies, adopting a more nuanced perspective, have investigated the influence of linguistic units and their semantics on the audience (Mohammad, 2018; Gavenko, 2001). The main assumption is that specific emotional dimensions allows making a certain message more persuasive (Petty and Cacioppo, 1986; Fogg, 2008; Tsinganos et al., 2022). Within this (psycho-)linguistic framework, persuasion is construed as a function of specific linguistic features of a message, such as its concreteness, emotional tone or certainty (Ta et al., 2022).

When it comes to translating these theoretical conceptualizations of persuasion into algorithms, scholars undertake an interdisciplinary effort that has strong repercussions on how the model is devised. Adopting a given perspective-whether argumentative or linguistic/semantic-significantly influences the design and development of algorithms, shaping the ways in which these computational systems analyze persuasive communications.

This choice of how to conceptualize persuasion, not only impacts on the technical application of AI, but also the explainability of these algorithms.[1] Explainability is imperative to ensure that these automated tools do not merely function as black boxes but provide insights that are understandable and actionable for human analysts. When these algorithms are employed by institutions and governments to regulate and monitor the flow of information among citizens, explainability becomes particularly crucial. In such cases, in fact, the ability to audit and justify algorithmic decisions is essential for maintaining public trust and for ensuring that interventions adhere to principles of fairness, transparency, and legality (Nannini et al., 2023).

Bridging theoretical knowledge with practical AI and natural language processing (NLP) application, thus, is a first crucial step toward model explainability (Páez, 2019). Subsequently, this article will provide an overview of AI and NLP techniques employed in the automated analysis of online persuasion (Zarouali et al., 2022). Specifically, the main objective is to elucidate how theoretical knowledge on persuasion is transmuted into measurable indicators and algorithms that can be implemented into computational systems to detect and analyze this phenomenon.

The remainder of this paper is structured as follows: In Section 2, we describe the methodology adopted to collect the analyzed papers, and the theoretical lens we adopted to analyze them. In Section 3 we present the different modalities scholarships adopted to operationalize the construct of persuasion, offering a comprehensive perspective of the state-of-art in this field. Finally, in Section 4 we discuss and critically evaluates the advantages and limitations of these approaches. By doing so, we shed light on the practical implications of these technological methods and their alignment with traditional theoretical frameworks. Finally, in Section 5 we use these elements to pinpoint areas for further research.

2 Methods

2.1 Data collection

Drawing from the above described objective, our methodology aimed at identifying key references underpinning the theoretical foundations applied in studying persuasion through automated methods, namely the ones provided by natural language processing and machine learning. With this focus, to collect the literature we referred to a two-phase approach.

In the first phase, we used "Scopus AI", an AI-driven research tool designed by Elsevier.[2] The tool employs a large language model trained on the Scopus peer-reviewed research repository, more precisely the metadata and abstracts of papers published since 2018.[3] Given a query, in addition to the "Summary" (the AI-generated response to the query) the tool returns multiple outputs, such as: "Foundational Papers" (a list of seminal studies in a given research area), "Topic Experts" (a list of leading experts and their work in a given research area) and "Follow-up questions" (additional prompts to submit to Scopus AI, to expand the initial query) (Aguilera Cora et al., 2024). Considering the aim of the article document, we conducted the first phase of literature search using the following keywords: "Persuasion," "Metrics," "Natural Language Processing," and "Machine Learning." We used the different outputs to understand and navigate the academic content on the topic, as well as to identify the most relevant papers.

In the second phase, we conducted a "backward citation searching" operation. For each article identified in the previous phase, we analyzed the citations to identify related works relevant to the research topic. Following this method we identified 30 documents. We removed all the studies that did not train a model for a total of 15 studies, since we were interested in observing how the theoretical references on persuasion were practically implemented in building the models.

We decided to explore the use of this tool to assess, in a first approach, the potential of this type of technology to boost the literature review for this paper. While we know that a systematic literature review would provide us more exhaustive results, we considered how this field of research is still not very well structured and continuously growing up. Therefore, our approach remains rigorous and pertinent with respect to our objective.

2.2 Persuasion modeling framework

To organize the analysis of the gathered literature, we have elaborated a broad taxonomy, drawing from the two main schools of thoughts described in Section 1.
The first category, *Persuasion as a set of Linguistic Style Units*, contains all the research operationalizing the persuasiveness of a text as the result of an interplay of linguistic features. According to these studies, persuasion is realized through a special intertwining of morphosyntactic, psycholinguistic, and rhetorical elements with each other to elicit a specific reaction in the addressee. The studies within this category adopt an approach analogous to that of Gavenko (2001), focusing more on *what is said*, i.e., analyzing the linguistic bricks used to build the persuasive structure.

The second, *Persuasion from an Argumentative Point of View*, instead, revolves around studies characterized by understanding persuasion as the result of specific argumentative structure. The studies falling within this approach, in fact, focus more on issues related to, for example, fallacies and misuses of argumentation. For this reason it can be said that the focus is posed on the "linguistic architecture" of persuasion, *how things are said,* rather than the

[1] Explainability in artificial intelligence (AI) and machine learning (ML) refers to the ability to describe the internal mechanisms and decision-making processes of models in a manner that is understandable to humans. It involves making the functioning of complex models transparent, allowing users to comprehend why a model makes certain predictions or decisions (Adadi and Berrada, 2018).

[2] https://www-elsevier-com.ezbusc.usc.gal/products/scopus/scopus-ai

[3] We signal how the present research has been conducted through the "Beta Version" of the tool, which later has been improved and, nowadays, relies also on papers written from 2003.

materials used to realize it. In this sense they can be connected to the argumentative line of research promoted by, among the others, Aristotle et al. (1909).

Given our focus on explainability, we also assessed the computational models employed in the reviewed studies, particularly in terms of their explainability. We distinguished between "Deep Learning" and "Shallow Learning," both of which are utilized in natural language processing but differ significantly in structure and transparency (Janiesch et al., 2021).

Deep Learning refers to a category of complex neural networks which are structured in multiple layers of algorithms. Each layer processes different aspects of the input data, progressively refining and abstracting the information as it passes from one layer to the next (Lauriola et al., 2022). This hierarchical structure allows deep learning models to perform sophisticated pattern recognition and data inference tasks, making them highly effective for complex language processing tasks in natural language processing (NLP) (Xu, 2023). The depth and complexity of these layers, however, can make it challenging to discern how specific inputs are transformed into outputs, leading to their characterization as "black boxes" (Adnan, 2024).

Shallow Learning, in contrast, involves simpler, more transparent algorithms such as decision trees (DT), logistic regression (LR), and support vector machines (SVM). These models operate with fewer layers of processing and often utilize symbolic representations that make the logic of the algorithm's decision-making process explicit. This feature allows for greater interpretability, as it is easier to see how inputs are directly linked to outputs (Janiesch et al., 2021).

By emphasizing these different approaches, we aimed to elucidate the balance between learning depth required for complex language tasks and the necessity for transparency in how language data is processed and interpreted.

3 Results

3.1 Persuasion as a set of linguistic style units

Linguistic style units play a pivotal role in enhancing persuasive communication through a multifaceted approach. Leveraging human theoretical frameworks on persuasion, it is possible to define the linguistic characteristics underpinning these phenomena, anchoring it in specific syntactic and linguistic features.

Dubremetz and Nivre (2018) adopted this approach for detecting three rhetorical figures based on repetition (Chiasmus, Epanaphora and Epiphora), which proved to be effective in shaping positively the performance of someone (Alkaraan et al., 2023). To assess the use of these linguistic devices, the authors retraced the methodological approach of a previous work (Dubremetz and Nivre, 2015), training three log-linear probability classifiers on a corpus of political debates, obtaining promising results (Chiasmus F1[4] = 0.78; Epanaphora F1 = 0.49; Epiphora F1 = 0.53). The choice of this model was justified by an easier interpretability of the results. Thanks to the "glass box" approach adopted, in fact, the authors were able to carry out an ablation study[5] to keep track of the specific contribution of each feature, using this information to adjust the model and adapt it to the specific figure of speech addressed. Finally, the three algorithms were also applied to datasets belonging to different genres (fiction, science, and quotes) obtaining consistent results. This, in turn, advocate for the cross-domain validity of the methods and, so, for the possibility of applying the classifier for the comparison between different sources.

Another rhetorical figure that has been studied often is "hyperbole," also known as "exaggeration:" a rhetorical figure implemented mainly to create amusement, express emotions and draw attention. Being able to automatically detect hyperbole could allow to evaluate if, and to what extent, political claims constitute a form of puffery or an information disproportion. To tackle this issue, Troiano et al. (2018) created a dataset (HYPO) of 709 hyperboles and trained a pool of shallow learning models. Depending on the particular rhetorical figure, the models exploited different linguistic features, such as: punctuation, sentence size, similarity and lexical structures; combined with different embeddings.[6] The best results in this classification task were obtained using the most explainable of the models adopted (Logistic Regression F1 = 0.76). This result shows how the structured knowledge offered by linguistics can be implemented to build NLP tools able to obtain high performances, without losing in their explainability level.

In the wake of the work inaugurated by Troiano et al. (2018) and Kong et al. (2020) furthered the exploration by developing a Chinese dataset of hyperboles (HYPO-cn), which comprises 4,762 sentences, including 2,680 hyperbolic ones. This focus on Chinese is particularly noteworthy, adding valuable linguistic diversity to the research. On a technical level, similar to Troiano et al. (2018), they initially employed traditional machine learning algorithms. The pivotal aspect of their study, however, was the examination of deep learning methodologies and their effectiveness in enhancing the hyperbole detection task. Specifically, they utilized a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), alongside fine-tuning a pre-trained Chinese BERT model for comparative analysis. The findings of this study were significant, demonstrating a marked superiority of deep learning models over traditional ones in automatically detecting hyperboles, evidenced by an Acc[7] =

[4] The F1 score is a measure of a model's accuracy, considering both the precision (how many selected items are correct) and recall (how many correct items are selected). It is useful for evaluating models, especially when the data has imbalanced classes.

[5] Ablation Study is a research method used to assess the impact of specific components on a model's performance by systematically removing them and observing the outcome. This process helps identify critical elements and optimize the model, especially in complex systems like machine learning and natural language processing.

[6] Embeddings are numerical representations of words or phrases in a continuous vector space, where semantically similar words are closer together, allowing algorithms to process text more effectively.

[7] Accuracy (Acc) is a measure of how often the model correctly predicts the outcome, calculated as the ratio of correct predictions to the total number of predictions.

0.85 (≈+0.1). While acknowledging the improved performance of these models, the study also underscores their "black box" nature, which obscures the understanding of how specific features influence the model's predictions. This highlights a key comparison between the two techniques: while deep learning models offer enhanced performance, they lack the interpretability that traditional models provide.

Al Khatib et al. (2020) developed a system for the automatic analysis of syntactic-based persuasive devices, including "pysma," "epizeuxis," and "polysyndeton." The authors, using a finite set of elements (e.g., "Cl" for Clause, "N" for Noun) devised a formalized definitions for each rhetorical figure. "Pysma," for instance, a rhetorical figure characterized for asking multiple questions successively, was formalized as <"Cl?"> <"Cl?">. To perform such a challenging annotation, they employed Apache Ruta (Kluegl et al., 2016), a rule-based script language, to facilitate the annotation process within their algorithm. The implementation of the model was then performed on the outputs of the Stanford Parser, an extensible pipeline that provides core natural language analysis. The model succeeded in identifying the rhetorical devices with a substantial score (F1 = 0.70 on average), indicating a high effectiveness of the approach in providing a quantitative analysis of the rhetorical figures in a text, maintaining a high level of explainability thanks to the symbolic representation at the base of the algorithm. Subsequently, the authors implemented the model to analyze the use of these persuasive devices by different political actors (with a special focus on Trump and Clinton). This analysis allowed the researcher both to provide a detailed comparative analysis of the rhetorical style between the two political actors, and to assess the different use of rhetorical figures between, for example, monologs (newspapers articles) and dialogues (political debates).

To identify persuasive texts, Tan et al. (2016) created a dataset starting from `/r/ChangeMyView`. `/r/ChangeMyView` is a subreddit where a user publishes a post regarding a certain issue, and other users discuss it in order to try to change the perspective of the publisher. The convincing arguments are tagged with a △. Drawing from this interactive environment, thus, the researcher had at their disposal a dataset of persuasive (tagged) and one of non-persuasive (non-tagged) posts. The research hypothesis, thus, was to observe a higher frequency of persuasive linguistic features in the texts of tagged (i.e., persuasive) posts. More precisely, they retrieved from the psycholinguistic literature different features associated with the persuasiveness of an argument. The level of arousal (the intensity of an emotion), concreteness (denoting something perceptible), dominance (expressions of control), and valence (words' pleasantness), for instance, were assessed through the LIWC dictionary (Chung and Pennebaker, 2012), assuming they had an impact on the persuasiveness. The authors, thus, trained a logistic regression model using the linguistic features retrieved by the psycholinguistic literature and integrating them with different text representation models (BOW,[8] POS,[9] Number of Words and a combination of all of them). The study confirmed the relation between linguistic features and level of persuasiveness. Moreover, thanks to the theory-driven and symbolic-based (the dictionaries) approach adopted, the researcher managed to track and explain the impact of each feature.

A similar study, which used a wide taxonomy of persuasion linguistic features, is Addawood et al. (2019). Starting from the Interpersonal Deception Theory (IDT) (Buller and Burgoon, 1996), the authors identified 49 linguistic cues indicators of persuasive language. To do so, they recurred to different already available dictionaries: MPQA (Wilson et al., 2017) and LIWC (Chung and Pennebaker, 2012). In NLP, dictionaries are human-knowledge-based lexicons that categorize and analyze words in text to extract nuanced information from language. LIWC, for example, is a software designed to connect the extracted linguistic features of a text to 80 different psychological categories (e.g., anger, anxiety etc.). In this way, thus, the authors aimed at detecting the psychological (LIWC) and sentimental (MPQA) indicators, expressed by the linguistic style, which, according to the IDT, should be characteristics of persuasion. Using these tools, thus, the utilization of persuasive language cues was quantified by analyzing tweets from suspected political trolls (assumed to be more users posting harmful persuasive content) and contrasting them with those from a control group of non-troll users. Finally, they tested the effectiveness of the taxonomy in detecting trolls, assessing, at the same time, which features were most important in distinguishing between trolls and non-trolls. To do so, they recurred to two machine learning algorithms: Random Forest (RF) and Gradient Boosting Classifier (GBC). The model was able to identify trolls with high accuracy (RF F1 = 0.8; GBC F1 = 0.82), showing how theory-driving approaches, by linking social phenomena to specific linguistic features, can provide useful insights to help tackle real-world critical issues.

Ahmad and Laroche (2015) translated a psychological theory in computational terms as well. They followed the "Cognitive Appraisal Theory" (Smith and Ellsworth, 1985), according to which emotions are induced by the person's evaluation of the situation she is interacting in. Starting from this, their hypothesis was that persuasive text are characterized for being particularly certain. Consequently, they worked to detect the level of emotions linked to certainty expressed in a text, and test if they correlated with the effectiveness of the text. On a computational level, Ahmad and Laroche (2015) recurred to a quantitative content analysis, namely Latent Semantic Analysis (LSA): a NLP information retrieval technique used for uncovering the hidden semantic structure within a collection of text documents. This approach, although it has some limitations connected to polysemy and context-grounded meanings, offers a highly interpretative approach. The results proved the research hypothesis, showing how language increases its persuasive power when wording refers to concrete objects contextualized with perceptibility (concreteness), as opposed to being abstract and alluding to intangible qualities (abstraction). Regarding this study, we highlight how it is a good example of the virtuous interplay that can rise between social and

[8] A text representation method that treats a document as a collection of individual words, ignoring grammar and word order, and focuses on the frequency of each word.

[9] A linguistic categorization that assigns words to specific grammatical roles (such as nouns, verbs, adjectives) within a sentence, aiding in syntactic and semantic analysis.

computational sciences: the first providing theoretical references to build explainable systems and, the latter, providing technical tools that can be used to test theory-driven hypothesis on a large quantity of data, thus improving the generalizability of the conclusions.

3.1.1 Lexicon inducted persuasive features

We have described how theories can inform the building and the use of computational tools. Nevertheless, this relation can also be structured the other way around. This approach is called *lexicon induction* (Hamilton et al., 2016; Pryzant et al., 2018), precisely for the direction it imparts to the knowledge process: from the specifics of each message's occurrence to a broad comprehension of the phenomenon under investigation, such as persuasion. The methodological praxis of this approach can be described as follows. Firstly, texts considered persuasive are collected (i.e., texts successful in realizing what they were created for). Secondly, the most distinctive features of these persuasive messages are extrapolated. Finally, the extracted features are connoted as persuasive by virtue of their efficacy in the real-world situation in which they were employed.

An example of this inductive approach is Khazaei et al. (2017). As Tan et al. (2016), they worked with \r\Changemyview subreddit to collect two groups of texts: persuasive and non-persuasive. What distinguishes this study from Tan et al. (2016), is that the authors did not refer to any theory to choose the specific features to observe in the text. In fact, they analyzed the dataset employing all the 80 LIWC categories, i.e., the different "psychological values" that can be attributed to the text using specific linguistic features. After that, they ran a t-test and found that 34 linguistic categories were statistically more frequent in one of the two groups of texts. This study showed how surface-based linguistic attributes can enhance text persuasiveness. Moreover, it shows how lexicon induction study can be conducted also with traditional algorithm and, thus, how human-knowledge can effectively be implemented to increase the interpretability of the algorithm.

Pryzant et al. (2018) conducted a study following this approach. They collected texts that proved to be effective in different domains, such as selling a product and directing a university choice. Then, they used two deep learning algorithms to extract the words that are, at the same time, predictive of their target and decorrelated from confounding variables. Subsequently, they compared the performance of the proposed algorithm in detecting words correlated with successful outcomes, with other shallow learning methods. The results showed a general trend: "deep learning approaches" outperform the "shallow learning ones." On this regard, we remark how, despite relying on "deep learning algorithms," the inductive form of this experiment allow the researcher to make its system more interpretable: both by linking the results to the analyzed outcome and by making explicit the function of the different modules of the learning algorithms employed. At the same time, we highlight how, given its nature, this approach is strongly dependent on the chosen dataset: with critical pitfalls being the generalizability of the results and the emergence of possible biases characteristic of the dataset.

This inductive approach has certain advantages. It grounds the produced knowledge in empirical outcomes, providing data that are directly connected to the everyday experience of people. Moreover, identifying features indicative of a certain outcome and decoupling them with confounds, promotes a better understanding and interpretability of machine learning models in NLP. In this regard, the inductive perspective could provide useful insights in the field of *causal inference using texts* (Egami et al., 2022; Sridhar and Blei, 2022), a research branch aimed at using large quantities of text data to inductively discover measures that are useful for testing social science theories. Many studies in this field are mostly unconcerned with the underlying features and algorithmic interpretability. Athey (2017) and Pryzant et al. (2018) showed how the lexicon inducted approach could be applied to increase the explainability of the algorithms. Considering this, with respect to the theme of persuasion, using this approach could help in isolating the "active ingredient" of persuasive narratives: rooting it in a pragmatic and empiric foundation.

Finally, we highlight some critical points that it is possible to anticipate. This approach, strongly relying on the dataset features to define what persuasion is, can be subject to certain biases. The persuasive linguistic features for a topic or a certain group of people, could ineffective when applied in a different context or theme. Another problem is related to the platform used for the dataset building. The community of \r\Changemyview, for instance, is composed by a set of people who start premising an openness to changing one's point of view, which vitiates the generalizability of the results. Considering this, it is possible to anticipate a proliferation of studies using different samples and, in turn, the generation of contrasting or, even contradictory results regarding persuasion (as discussed in Section 1, see also Druckman, 2022).

3.2 Persuasive from an argumentative point of view

This section will describe a group of studies aimed at capturing the argumentative essence of persuasion; i.e., how the different contents are ringed and combined between each other to build convincing texts.

A seminal work in this area is the one of Da San Martino et al. (2019), who elaborated an algorithm to perform a fine-grained analysis of propaganda in texts. Previous methodologies, in fact, operated on a "full-text level," i.e., by labeling the entire article as propagandist or not. This raises different criticalities, both by creating a noisy golden label (affecting in turn the quality of the learning of the system) and by exacerbating the lack of explainability. To tackle these issues, they proposed a new task: detecting all the text fragments of an article containing propaganda techniques, and then identifying their type. In this work they recurred to a taxonomy of 18 persuasion techniques, combining the ones identified by Miller (1939) and Weston (2018), choosing them in relation to the type of content available on newspapers. We highlight how, according to our definition, some of these techniques fall under the "linguistic" POV, however we included the study in this section for its strong focus on the argumentative ones. After annotating the corpus, they fine-tuned a BERT with a novel multi-granularity neural network and showed how it outperforms several strong BERT-based baselines. The aforementioned task has then been used to create a SemEval Task in 2020 (Da San Martino et al., 2020). Finally, a software has been created (Prta – Propaganda Persuasion Techniques Analyzer) (Da San Martino et al., 2020) allowing users to explore the articles crawled, discover the persuasion techniques used in them and have a statistical report about the use of the techniques overall and over the time. This seminal work inaugurated a prominent branch of research, resulting in different shared tasks (Piskorski et al., 2023; Alam et al., 2022).

Starting from the work of Da San Martino et al. (2019) and Vorakitphan et al. (2021) (see also Vorakitphan, 2021) tried to enhance the performance and the explainability of the algorithm for the detection of the same persuasion techniques. To do so, they selected a set of semantic, sentimental and argumentative features assumed to play a persuasive role in texts. They run an ablation test to select the most performing features and, finally, used them to fine tune a BERT-based model. They compared the performance of this model with the state-of-art models for persuasion detection (retrieved from Da San Martino et al., 2019; Yoosuf and Yang, 2019; Jurkiewicz et al., 2020) and observed that the implementation of the features generated an improvement of $F1 = +0.10$. This study exemplifies how, to tackle the complex task of detecting persuasion techniques in texts, the argumentative approach can be combined with the linguistic one to improve the performance of the algorithm as well as it explainability.

Given the promising results obtained, this study has been followed by another one focusing specifically on political debates. Goffredo et al. (2022) retrieved 31 political debates from the US presidential campaigns and annotated them with six categories of fallacious arguments. In addition to the logical fallacies, they made use also of argumentative contextual information, namely "premise," "claim," "attack," "support," and "equivalence." To accommodate these features, the researcher used two Pre-Trained Language Models: Longformer and Transformers-XL, which have longer maximum sentence lengths than BERT. They compared the performance of these models with the ones of BERT models which didn't employ argumentative information. Interestingly, compared to the results obtained by Vorakitphan et al. (2021) (see above), these contextual information helped to substantially improve the performance of the model, which reached an average $F1 = 0.84$ ($\approx +0.2$). We highlight how, as the database consists of debates collected from numerous politicians in an extensive historical period, the work allows the researcher to compare both the different use of persuasive techniques by the different politicians, and how this use varies along the time.

A similar work to Da San Martino et al. (2019), is Jin et al. (2022). Starting from the assumption for which persuasion can be conveyed through the structure and the form of the argument, Jin et al. (2022) worked to create a model particularly focused on the argumentative structure of the text. To do so, they took the cue from the architecture of natural language inference systems and designed a "structure distillation method." This method involves concealing key content words in the premise, thus generating a logical form with placeholders: this to prioritize the structural aspects over specific content. For instance, the specific contents of the statement "Jack is a good boy. Jack comes from Canada. Therefore, all Canadians are good boys" were masked, returning the string "[MSK1] is a [MSK2]. [MSK1] comes from [MSK3]. Therefore, all [MSK3] are [MSK2]." On a computational level, firstly, they used the CoreNLP package (Manning et al., 2014) for the coreference resolution. Subsequently, they identified word spans that represent paraphrased content, considering solely non-stop words, lemmatizing them via the Stanza package (Qi et al., 2020), and representing each word using context-based embeddings generated by Sentence-BERT (Reimers and Gurevych, 2019). Finally, they calculated the similarity between these pairs of words. If the similarity surpassed a predetermined threshold, the words were classified as similar. This "masked data" were then used to train a deep learning model aimed at detecting 13 different persuasive technique. Compared to the language models fine-tuned in the "standard way," the proposed one showed an increased performance: $F1 = 58.77(+0.05)$, and $Accuracy = 0.48(+0.12)$. The outcomes of the study, therefore, indicate a promising future regarding the implementation of the logical structure within persuasion detection tasks. At the same time, provide an example of how human-based knowledge can be embedded into deep learning models, improving their potentials and increasing their explainability.

Sheng et al. (2021), aimed at investigating "ad hominem" attacks in social media interactions. They worked to understand how "ad hominem" Twitter responses vary according to the different topics analyzed, which, in turn, covered political and non-political topics. To this end, they extracted English post responses pairs on different topics from Twitter, such as: working from home, black-lives-matters, or the #metoo movement. Thanks to this training data, the authors managed to fine-tune also a chatbot (DialoGPT) to generate automatic answers to the different posts on Tweet (this way they worked both with "naturally-generated-answers" and "synthetic-answers"). Subsequently, they annotated all the gathered texts (user and machine generated) tagging the posts containing "ad hominem" attacks. The dataset was used to fine-tuned a deep learning model (BERT based) for the detection of "ad hominem" attacks, with encouraging results ($F1 = 0.8$). The results of the study allowed the researcher to notice how responses from both humans and DialoGPT contain more "ad hominem" attacks for discussions about marginalized communities. Moreover, they observed that different quantities of "ad hominem" in the training data can influence the likelihood of generating "ad hominem" in the chatbot algorithm. On the face of this, the authors used a list of "ad hominem" phrases as a soft constraint to avoid generating responses that contained these phrases. The authors found that their constrained decoding technique was effective in reducing the number of "ad hominem" generated by the DialoGPT model: showing one of the possible practical application deriving from the computational study of persuasion. Moreover, this study exemplifies how the analysis power provided by the application of AI and ML in the NLP field can contribute to uncover social phenomena that would, otherwise, be overlooked, such as the correlation between ad hominem attacks and marginalized communities.

Finally, we conclude with Pauli et al. (2022). Starting from the problem of persuasion theoretical fragmentation, proposed a novel way to group the persuasion techniques. More precisely, referring to the classic Aristotelian tripartition of the elements of rhetoric (Ethos, Logos, and Pathos) (Aristotle et al., 1909), each persuasion technique is understood as a misuse of one of those elements. This way, in turn, the researcher is provided with a theoretical framework able to group the techniques and, thus, reduce their numbers. The authors used this taxonomy to train three RoBERTa models, one for each rhetorical category. Subsequently, they applied the models on five different misinformation datasets to test whether the misuse of persuasive techniques was more frequent in false claims. Their hypothesis proved to be right, therefore this study, in addition to a broader and more transversal theoretical structure for the study of persuasion techniques, constitutes an interesting example of how persuasion knowledge and methodologies can be effectively applied in different domains.

4 Discussions

4.1 Technical issues in computational persuasion analysis

In the previous sections, we discussed some insights derived from the current state of automated persuasion analysis. Despite not being a systematic analysis, the info-graphic in Figure 1 shows some trends worth to be discussed. Table 1 allows the reader to trace back the studies depicted in the figure. Key insights include:

- Over the years, there has been a decrease in studies conducted with reference to "shallow" learning models, i.e., models characterized by easily explainable and interpretable learning processes, in favor of deep learning models. As outlined in Section 1, these latest methods, can handle a higher level of complexity and possesses significant scalability potential, thereby facilitating generalization. However, its reliance on non-linear and non-intuitive feature interactions reduces its explainability in how inputs impact predictions. Stakeholders using these models to address societal issues involving manipulative content may encounter critical challenges in terms of:
    - *Accountability*: drawing from Floridi (2013), an agent is deemed accountable for an action when it serves as the causal origin of that action. Hence, accountability is particularly vital in assessing artificial intelligence systems, since it establishes the fundamental requirements for transparency and explainability necessary to attribute responsibility-praise or blame-to the appropriate entities in morally sensitive situations involving AI. In essence, without a clear causal link between the inputs and outputs of a model, such as in black box models, assigning blame becomes infeasible when the outcomes are ethically adverse (Novelli et al., 2023).
    - *Fairness*: fairness in AI involves identifying, measuring, and improving algorithmic fairness to prevent discriminatory outcomes. In the context of evaluating AI systems throughout their lifecycle (design, data collection, training, deployment, and regulation), some relevant symmetries of this kind are resources (access to compute, data collection and ownership of data, and models) and outcomes (identifying benefits, harms, beneficiaries, and potential victims of harms) (Pessach and Shmueli, 2022). A common example of unfairness in AI models that cuts across these symmetries is the issue of bias. Models may inadvertently learn biases present in the training data, and perpetuate them or even amplify them, generating unfair or discriminatory outcomes (Mutlu et al., 2022).
    - *Trust*: as a consequence of the two previous points, stakeholders might be hesitant to trust those tools, reducing the virtuous impact they can play on society (Sethumadhavan, 2018).
- Another trend is the decrease in the number of studies conducted following a linguistic point of view and, at the same time, an increase in the number of studies adopting an argumentative one. All the research in the second section, in fact, have been conducted recurring to argumentative approaches. Given the complexity of understating the argumentative processes underlying persuasion, this change can be definitely linked to the increased availability of deep learning systems. This, on one side, constitutes an important step forwards for the computational study of persuasion since it allows researchers to employ all the theories elaborated in the fields of argumentation or rhetoric. At the same time, it exposes to the issues outlined above related to deep learning systems.
- Finally, we highlight how the 90% of the studies were conducted on English datasets. Hindering the generalization of these models to non-English languages is a concern. Additionally, relying solely on one language in a multi-cultural and multi-lingual online environment can reduce the impact of the computational study of persuasive devices on the community.

As discussed above, the multi-faceted nature of persuasion requires a constant interplay between the theoretical knowledge grounding this construct and the methodological possibilities generated by the technological development. On this regard, it is possible to observe how, despite the argumentative theories on persuasion existed since the age of Aristotle, only the most recent advances in ML and AI allowed their computational study. Considering then the criticalities connected to these new technologies, it emerges the necessity to adopt more interpretable systems. With respect to this last aspect, we stress how, very few of the analyzed studies made use of hybrid methodologies (Panchendrarajan and Zubiaga, 2024). This is particularly critical since, by leveraging the complementary aspects of rule-based and deep learning approaches, hybrid NLP models in the study of persuasion could enhance explainability, transparency, and ethical considerations: contributing to more responsible and effective computational persuasion systems (we elaborate more on this in Section 4.2.2).

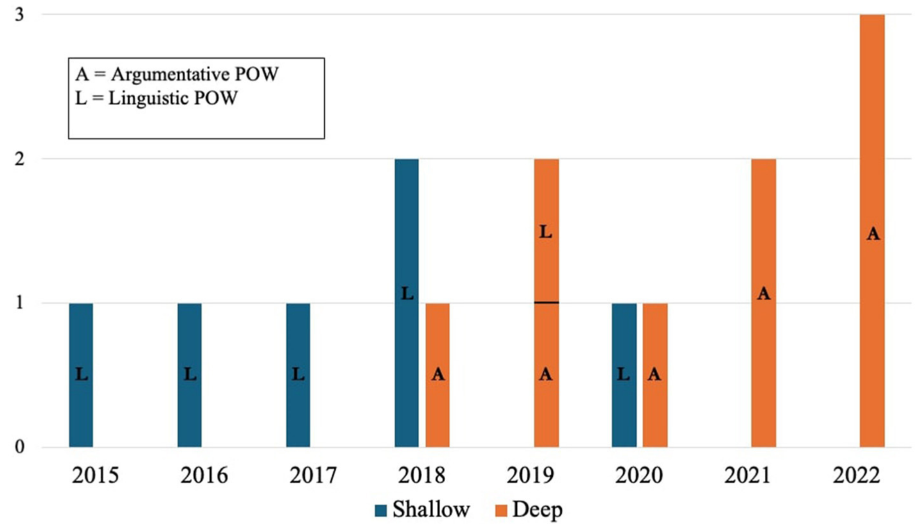

FIGURE 1
Studies compared for year and approach.

4.2 The moral status of persuasive communication

Persuasive strategies are distinguished by specific structures depending on the context and the objective for which they are used (Tindale, 2007). Consequently, as demonstrated in Section 3, a spectrum of methodologies for studying persuasion has emerged. These methodologies range from analyzing purely stylistic elements-such as the identification of rhetorical figures like Chiasmus or Hyperbole-to detecting explicitly propagandist methods designed to manipulate audience opinions.

However, employing persuasion detection algorithms to tackle societal challenges, necessitates a strict distinction between what we could call "virtuous persuasion"—integral to numerous interactions and democratic processes—and "vicious persuasion"—such as propaganda and manipulation. Lexicon inducted studies such as Tan et al. (2016) and Khazaei et al. (2017), for instance, empirically determined the persuasive elements of text through a "bottom-up" approach. However, this method lacks criteria to discern if the language used promotes fair discussion or deceives to circumvent critical thinking. Conversely, studies like Troiano et al. (2018) and Ahmad and Laroche (2015), theoretically delineated linguistic devices in effective persuasive messages. Yet, these theories overlook the social and interactive effects of rhetoric, limiting qualitative and moral assessments needed to identify harmful persuasion.

To tackle this issue, below we delineate and discuss criteria specifically aimed at distinguishing "virtuous" from "vicious" uses of persuasive means, thus promoting the development of effective, precise, and explainable algorithms to address such challenges.

4.2.1 Telling apart fair and unfair persuasion

Evaluating the moral status of communication can be done using classic normative ethics. Adopting a deontological view, the cognitive and epistemic autonomy of the recipient should be respected, in alignment with the Humanity Formulation of the Kantian Categorical Imperative (Allison, 2011). This means that, when engaging in a dialogical or monological argumentative setting, persuasion should be reached honoring the recipient's rational capacity, i.e., without overshadowing or distorting the informational content of the message.

Additionally, a satisfactory account of vicious persuasion needs to pay attention to both the rational and rhetorical dimension, since all argumentative communication involves a trade-off between reasoning and presentation.

To address this issue, Godber and Origgi (2023) proposes distinguishing legitimate forms of persuasion from their misuses based on how the rhetorical devices employed by the speaker impact on audience's intellectual autonomy. *Intellectual autonomy*[10] refers to the capacity of individuals to think critically and independently, forming beliefs based on their own reasoning and evidence (Carter, 2020).

Using these criteria, Godber and Origgi (2023) proposes a taxonomy dividing these forms of persuasion into the following categories:

- *Rational Persuasion (RP)*: this typology encompasses discourses that are based on facts, evidence, and sound logical reasoning. This type of persuasion seeks to present sufficient grounds for its claims, ensuring that the information is balanced and avoids misleading suggestions. The goal is to respect the intellectual autonomy of the audience by clearly presenting and defending the arguments, thereby allowing the

[10] According to the Kantian definition, intellectual autonomy contrasts with "intellectual immaturity," which is characterized by cowardice when one's thinking is willfully guided by external influences. However, in line with Godber and Origgi (2023), we adopt Carter (2020)'s more interactive definition, which involves the possibility to appropriately rely on external sources while maintaining intellectual self-direction (see also Roberts and Wood, 2007).

TABLE 1 Research in linguistic and argumentative AI methods from 2015 to 2022.

		2015	2016	2017	2018	2019	2020	2021	2022
Linguistic	Shallow	Ahmad and Laroche, 2015	Tan et al., 2016	Khazaei et al., 2017	Dubremetz and Nivre, 2018; Troiano et al., 2018	Addawood et al., 2019	Al Khatib et al., 2020		
	Deep				Pryzant et al., 2018				
Argumentative	Shallow								
	Deep					Da San Martino et al., 2019	Kong et al., 2020	Vorakitphan, 2021; Sheng et al., 2021	Goffredo et al., 2022; Jin et al., 2022; Pauli et al., 2022

audience to understand, analyze, and potentially counter the proposed reasons. Given these characteristics, RP represents the clearest example and prototype of "virtuous persuasion."

- *Non-Rational Persuasion (NRP)*: this category of persuasion strategies seeks adherence from the audience not through reasons (logos) but by appealing to emotions (pathos) or the speaker's virtues (ethos). The former seeks to emphasize certain elements of the message by employing what Manson (2012) refers to as lexical spin, which involves the use of inapt terms or linguistic devices such as metaphor and hyperbole to achieve desired effects. The latter, conversely, endeavors to project or transfer qualities associated with one entity, such as an expert or authoritative figure, onto another. As shown in Figure 2, this form of persuasion is mainly characteristic of vicious persuasive efforts. However, as emphasized by Godber and Origgi (2023), when used alongside rational persuasive means (RP+NRP in Figure 2), appeals to emotions or authorities can support argument presentation without compromising argumentative soundness.
- *Rational Manipulative Persuasion (RMP)*: this form of persuasion is characterized for appealing to facts and shareable evidences, however it does so "disingenuously," in a manner similar to Manson (2012)'s "aspect-based spin," according to which facts are selectively included or omitted (e.g., cherry-picking). This misleading presentation contributes to the creation of a biased narrative that subverts rational processes and manipulates the audience. This type of persuasion falls entirely within the macro category of "vicious persuasion," as it operates within the content dimension of the message and compromises the audience's intellectual autonomy by depriving them of the necessary elements to evaluate and counter the speaker's arguments effectively.
- *Irrational Persuasion (IP)*: messages crafted through irrational persuasive means rely on fallacies and outright falsehoods. Fallacies are deliberately employed to exploit the audience's cognitive biases, thereby manipulating their reasoning and decision-making processes. Falsehoods, on the other hand, can serve two distinct purposes: they may be used to promote a specific narrative, or they may be used to undermine the possibility of consensus by saturating media ecosystems with false and conflicting information, thereby creating confusion and distrust. This form of persuasion is inherently unfair. By relying on these strategies, it aims not to promote healthy debate, but rather to polarize the audience and foster an environment of skepticism and uncertainty, thereby eroding the foundation of truth.

Godber and Origgi (2023) devised this taxonomy to rhetorically define "Propaganda"—untangling it from (fair) persuasion. The former, in fact, is built through specific combinations of rhetorical strategies, namely **RMP + NRP**, **IP + NRP**, or **RMP + IP + NRP**.

However, this specific application, brought to a more general level, underscores how the concept of intellectual autonomy, configuring people's epistemic endeavors as interactive processes, can be used to characterize different communicative strategies. This definition clearly delineates the rhetorical boundaries of persuasion misuse, enhancing the conceptual framework grounding computational research. Such clarity can fosters

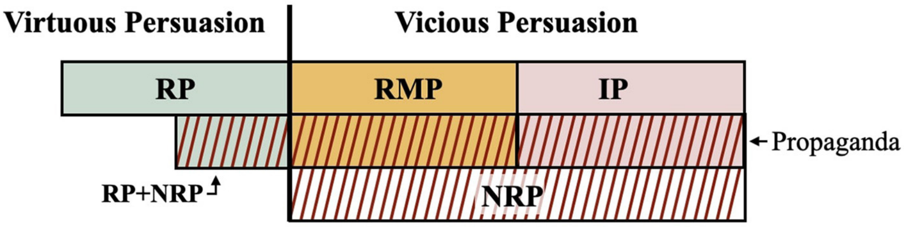

FIGURE 2
Persuasion typologies.

scientific collaboration and addresses the challenges posed by theoretical fragmentation in persuasion research (Druckman, 2022). At the same time, as outlined discussed in Section 4.1, from a computational point of view, developing algorithms able to deal with complex dimension of communication, such as the rhetorical means used, needs to rely on deep learning methods, with critical pitfalls on the side of models accountability.

In this regard, humanities and social sciences in addition to clarifying the concept of persuasion, can also be used to devise symbolic representation of the text, useful to provide linguistic indicators that can be integrated in algorithm development. As we conclude this section, we propose integrating scholarly insights, such as argumentative theory, Rhetorical Structure Theory (RST), and emotional dictionaries, into computational methods. This integration is crucial for distinguishing between fair and unfair persuasion. Furthermore, leveraging these theories helps construct hybrid models that combine deep learning with human symbolic understanding (Panchendrarajan and Zubiaga, 2024). Such models have proven essential in enhancing the explainability and effectiveness of algorithms designed to detect manipulative messages (García-Orosa et al., 2021).

4.2.2 Integrating social science theories into hybrid computational models

Building upon the need to integrate humanities and social sciences with computational methods, we examine specific theories that can enhance persuasive communication analysis. Crucially, these theories enable a deeper characterization of text, allowing us to better differentiate between vicious and virtuous forms of persuasion. This is achieved by providing tools to analyze the rhetorical, argumentative and emotional aspects of communication-elements we identified in the previous section as critical to operate the distinction. Hence, subsequent subsections will discuss the application of "Argumentation Theory," "Rhetorical Structure Theory" (RST), and "Emotion Dictionaries" in crafting hybrid models that combine deep learning with human symbolic understanding. These theories not only enhance the characterization of persuasive elements but also provide a transparent framework that improves the explainability of computational persuasion analysis.

4.2.2.1 Argumentation theory

Wagemans (2023)'s periodic table of arguments presents a systematic framework for classifying text arguments based on the support or refutation modality of a statement. This framework organizes arguments into four categories first-order subject arguments, second-order subject arguments, first-order predicate arguments, and second-order predicate arguments. Additionally, it distinguishes between different forms of argument based on how the premises relate to the conclusion, whether through causality, analogy, generalization, or authority. Implementing this structured approach in annotation schemes, can provide a comprehensive method to analyze the logical and rhetorical structures of persuasive texts, aiding algorithms in detecting persuasive elements, understanding their logical foundations and soundness (Hinton and Wagemans, 2022).

4.2.2.2 Rhetorical structure theory

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) offers a comprehensive method for analyzing the hierarchical organization of texts by identifying rhetorical relationships between various parts of a document. It segments texts into nucleus (holding the core information) and satellite segments (providing supporting details to the nucleus), and describes the types of relationships that connect the two. Applying RST to parse texts can effectively map the relationships among claims, evidence, and counterarguments. This application enables RST-enhanced algorithms with a deeper understanding of how the persuasiveness of texts is built and supported (Seref and Seref, 2019).

4.2.2.3 Emotion dictionaries

Aristotle et al. (1909)'s analysis of persuasive emotions-articulated through dichotomies such as anger vs. calm and hate versus friendship-provides a systematic framework for identifying emotions. This framework delineates the necessary criteria for emotion detection: the subject experiencing the emotion, the target of the emotion, and the provocation of the emotion. Building upon this foundation, we can establish more precise guidelines for identifying the persuasive use of emotional language in texts (Tsinganos et al., 2022).

Furthermore, psycholinguistic theories, which link persuasive effects to specific psychological dimensions, facilitate the identification of key terms that signal emotional content in persuasive contexts. This approach is exemplified by the creation of specialized psycholinguistic dictionaries, such as those discussed by Warriner et al. (2013). These dictionaries, derived from comprehensive psycho-emotional analyses of text, offer invaluable tools for extracting emotional dimensions. Psycholinguistic theories, then, can be used to further refine and expand these dictionaries. For instance, Ta et al. (2022) enhance the LIWC dictionary by incorporating words defined through the Valence-Arousal-Dominance (VAD) circumplex model of emotion (Russell, 1980) and the PAD emotional state model (Bales, 2017). This integration specifically targets the expansion and refinement of linguistic indicators associated with valence, arousal, and dominance, thus underscoring their persuasive effect in textual analysis.

In general, these enriched dictionaries not only improve the detection of persuasion, but also enhance the explainability of the algorithms employed in this process (Goffredo et al., 2022; Vorakitphan et al., 2021; Tsinganos et al., 2022).

4.3 Limitations of the study

One limitation of this study is its non-systematic approach to gathering literature. To ascertain the accuracy of the results returned by Scopus AI, we cross-validated these findings with other databases. Specifically, we conducted a search on Google Scholar using the query "Persuasion Metrics Natural Language Processing Machine Learning" setting the time period from 2018 to 2023, and compared the initial 50 results with those identified by Scopus AI.

As anticipated, the most influential papers identified by Scopus AI (such as Da San Martino et al., 2019; Tan et al., 2016; Pryzant et al., 2018), were also detected through Google Scholar. Conversely, some relevant papers gathered from Scholar, were not included in the Scopus AI results' list. These omissions may be due to Scopus AI's selective aggregation approach, which prioritizes data sources according to specific indexing criteria that, in turn, may not encompass all scholarly outputs. More surprisingly, some papers retrieved by Scopus AI were not found on Google Scholar, a discrepancy that can be attributed to the relational capabilities of the generative artificial intelligence employed by Scopus AI, which enhances the interconnection between concepts through its specialized database (Aguilera Cora et al., 2024).

As outlined in Sections 1, 2.1, the methodology adopted in this study was a strategic choice, driven by the objective to focus on identifying key theoretical references for discussing the study of persuasion through NLP and ML methods. Consequently, while this approach has enabled a focused exploration and critical discussion of the pivotal literature on the detection of persuasive means, it does not provide a systematic empirical substantiation of the findings presented.

As a conclusion from this first approach to a literature review using Scopus AI, it is evident that while the tool is valuable, it cannot currently replace a systematic literature review. It can be added as additional resource for the searching but it has to be complemented with other searching strategies. In this regard we signal (Aguilera Cora et al., 2024), who provides useful insights to increment the comprehensiveness of Scopus AI's results through an iterative process.

5 Conclusions

The computational analysis of persuasion marks a critical frontier at the intersection of natural language processing, humanities, and philosophy. Our review highlights the significant evolution within this field, transitioning from the use of shallow learning algorithms, that analyze explicit linguistic indicators, to the adoption of complex deep learning models capable of discerning nuanced argumentative structures.

In relation to the main objective of our paper—to elucidate how theoretical knowledge on persuasion is transmuted into measurable indicators for computational systems—we have identified a clear link between technological advancements and the theoretical frameworks employed. In fact, the linguistic perspective on persuasion, traditionally tied to shallow learning models, has gradually given way to argumentative approaches, which leverage the advanced capabilities of deep learning technologies. While this "technological boost" holds significant promise for the increasingly nuanced computational harnesses of persuasion, it also presents both challenges and issues.

In fact, the shift toward deep learning models has undoubtedly enhanced our ability to analyze persuasive content at scale, offering unprecedented insights into the mechanics of influence in digital spaces. Yet, this advancement comes at the cost of reduced explainability and interpretability, raising significant concerns about accountability, fairness, and trust in the applications of these technologies.

Moreover, the ethical implications of persuasion detection cannot be overstated. The fine line between "virtuous" and "vicious" persuasion necessitates a more nuanced approach to computational analysis. Simple detection of persuasive elements is insufficient; we must strive to develop systems capable of discerning between fair rhetoric and manipulative tactics.

To address this critical need, we have proposed criteria for distinguishing between virtuous and vicious forms of persuasion. While advanced deep learning models already demonstrated the potential to detect these nuanced shades of language, the pressing issue of explainability remains. It is there that the integration of humanities and social sciences into model development becomes paramount.

By incorporating insights from fields such as argumentation theory, rhetoric, and psycholinguistic (see Section 4.2) we can implement these distinguishing criteria in a more transparent and explainable manner. This interdisciplinary approach not only enhances the effectiveness of models, but also grounds them in established theoretical frameworks, potentially bridging the gap between algorithmic efficiency and ethical considerations. Hence, as we move forward, the challenge lies in developing hybrid systems (Panchendrarajan and Zubiaga, 2024) that leverage the power of deep learning while maintaining the explainability offered by more traditional, shallow-learning approaches.

In light of this, to conclude, we report a list of points regarding the future challenges that the computational study of persuasion will have to address.

5.1 Future challenges and directions

- Despite their potential to increase the explainability of the models, **hybrid approaches** requires dealing with, among the others: technical challenges (such as ensuring an effective communication between deep learning and rule based methods), finding the "right balance" between the portion to cover with the rule-based and deep learning methods (which, in turn, impact, respectively, on the explainability of the model and its performance, adaptability, and scalability) and the necessity for expertise in both rule-based systems and deep learning (posing challenges in terms of finding skilled practitioners and allocating resources).

- The dissemination of persuasion is not confined solely to textual content; the strategic use of images can convey more potent messages than text (Seo, 2020). Consequently, there is a growing imperative to scrutinize diverse data modalities, including images, videos, and speech and the use of a combination of these modalities (multimodal persuasion). This endeavor presents a complex challenge as, while some research has explored the effective comprehension of cross-modal information across diverse domains, limited attention has been devoted to discerning the informative potential of a specific modality in the context of propaganda detection. In this regard, we refer to the work of Dimitrov et al. (2021), which is aimed at detecting persuasion techniques in political memes coming from different social networks. The work, moreover, became a SemEval Shared task for the 2024 edition (Dimitrov et al., 2023).
- Most of the current detectors are assessed solely on a single annotated dataset, based exclusively on the English language. Consequently, we face a deficiency in our capacity to assess how well detectors can extend their performance from controlled environments to real-world and multilingual scenarios. Moving forward, it is important to encourage more scholarships and research initiatives focused on developing multilingual annotated datasets.
- In the context of handling user-generated data, ethical concerns assume a significant role. It is imperative to ensure that any analysis and prospective sharing of datasets strictly adhere to the privacy rights of the individuals involved. An ELSEC (Ethical, Legal, Social, Economic, and Cultural) approach is therefore crucial in AI ethics and data protection. It provides a holistic framework that acknowledges the complex interplay of these factors, ensuring that AI technologies respect diverse societal values, legal requirements, economic considerations, and cultural contexts, thereby fostering responsible and inclusive AI development.
- Finally, recent progress in neural language models has reached a point where distinguishing synthetic text from human-generated text is becoming challenging even for humans. Zellers et al. (2019) demonstrated the effectiveness of a template system in altering the output format of a language model, while Yang et al. (2018) provided insights on transferring the style of a language model to a specific target domain. For some years now, researchers have voiced worries over the potential for the misuse of NLG (Natural Language Generator) models to generate adverse and malicious outputs such as propaganda and disinformation (Schuster et al., 2020; Goldstein et al., 2023). In recent years, these worries have been underscored by demonstrations of the capabilities of current PLMs to generate just such outputs (Bontcheva et al., 2024; Zhou et al., 2023; Vykopal et al., 2024). As the gap between human and machine-written text appears to close, we believe that it will be imperative to expand the scope of analysis beyond textual content alone and delve into the examination of network and dissemination patterns of propaganda and similar forms of vicious persuasion in the future.

Author contributions

DB: Writing – review & editing, Writing – original draft, Data curation, Conceptualization. SF: Writing – review & editing, Conceptualization. MP-F: Writing – review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the European Union's Horizon Europe research and innovation programme (EUHORIZON2021) https://cordis.europa.eu/project/id/101073351/es, with the Marie Skıodowska-Curie Grant No. EUHORIZON2021-101073351 and by the grants PID2020-115482GB-I00 and PID2020-114758RB-I00 both funded by MCIN/AEI/10.13039/501100011033 and by the grant ED431B 2023/24 funded by Consellería de Cultura, Educación, Formación Profesional e Universidades fo Xunta de Galicia and, as appropriate, by "ERDF A way of making Europe," by the "European Union" or by the "European Union NextGenerationEU/PRTR."

Acknowledgments

During the preparation of this work the authors used ChatGPT-4 in order to check the correctness of the English writing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access* 6:52138–52160. doi: 10.1109/ACCESS.2018.2870052

Addawood, A., Badawy, A., Lerman, K., and Ferrara, E. (2019). "Linguistic cues to deception: Identifying political trolls on social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, 15–25. doi: 10.1609/icwsm.v13i01.3205

Adnan, M. (2024). "The importance of interpretability in AI systems and its implications for deep learning: ensuring transparency in intelligent systems," in *Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems* (IGI Global), 41–76. doi: 10.4018/979-8-3693-1738-9.ch003

Aguilera Cora, E., Lopezosa, C., Fernández Cavia, J., and Codina, L. (2024). Accelerating research processes with scopus ai: a place branding case study. *Rev. Panamer. Comunic.* 6, 1–26. doi: 10.21555/rpc.v6i1.3088

Ahmad, S., and Laroche, M. (2015). How do expressed emotions affect the helpfulness of a product review? Evidence from reviews using latent semantic analysis. *Int. J. Electr. Commer.* 20, 76–111. doi: 10.1080/10864415.2016.1061471

Al Khatib, K., Morari, V., and Stein, B. (2020). "Style analysis of argumentative texts by mining rhetorical devices," in *Proceedings of the 7th Workshop on Argument Mining, page*, 106–116.

Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., and Nakov, P. (2022). "Overview of the WANLP 2022 shared task on propaganda detection in Arabic," in *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)* (Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics), 108–118. doi: 10.18653/v1/2022.wanlp-1.11

Alkaraan, F., Albahloul, M., and Hussainey, K. (2023). Carillion's strategic choices and the boardroom's strategies of persuasive appeals: ethos, logos and pathos. *J. Appl. Account. Res.* 24, 726–744. doi: 10.1108/JAAR-06-2022-0134

Allison, H. E. (2011). "The formula of humanity," in *Kant's Groundwork for the Metaphysics of Morals: A Commentary* (Oxford University Press), 204–236. doi: 10.1093/acprof:oso/9780199691531.003.0009

Aristotle, J., Claverhouse, R., and Sandys, J. E. (1909). *The Rhetoric of Aristotle: a Translation*. Washington, DC: The University Press.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science* 355, 483–485. doi: 10.1126/science.aal4321

Bales, R. (2017). *Social Interaction Systems: Theory and Measurement*. London: Routledge. doi: 10.4324/9781315129563

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., et al. (2024). "Generative AI and disinformation: recent advances, challenges, and opportunities," in *European Digital Media Observatory*.

Buller, D. B., and Burgoon, J. K. (1996). Interpersonal deception theory. *Commun. Theory* 6, 203–242. doi: 10.1111/j.1468-2885.1996.tb00127.x

Carter, J. A. (2020). Intellectual autonomy, epistemic dependence and cognitive enhancement. *Synthese* 197, 2937–2961. doi: 10.1007/s11229-017-1549-y

Chung, C. K., and Pennebaker, J. W. (2012). "Linguistic inquiry and word count (liwc): pronounced "luke,"... and other useful facts," in *Applied Natural Language Processing: Identification, Investigation and Resolution* (IGI Global), 206–229. doi: 10.4018/978-1-60960-741-8.ch012

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., and Nakov, P. (2020). "Semeval-2020 task 11: detection of propaganda techniques in news articles," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1377–1414. doi: 10.18653/v1/2020.semeval-1.186

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. (2019). "Fine-grained analysis of propaganda in news article," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5636–5646. doi: 10.18653/v1/D19-1565

Demirdöğen, Ü. D. (2016). The roots of research in (political) persuasion: ethos, pathos, logos and the yale studies of persuasive communications. *Int. J. Soc. Inquiry* 3, 189–201. Available at: https://dergipark.org.tr/en/pub/ijsi/issue/17732/185728#article_cite

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., et al. (2021). "Detecting propaganda techniques in memes," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 6603–6617. doi: 10.18653/v1/2021.acl-long.516

Dimitrov, D., Da San Martino, G., Nakov, P., Alam, F., Hasanain, A., and Silvestri, F. (2023). "SemEval2024 shared task on "Multilingual Detection of Persuasion Techniques in Memes"," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 2009–2026.

Druckman, J. (2022). A framework for the study of persuasion. *Ann. Rev. Polit. Sci.* 25, 65–88. doi: 10.1146/annurev-polisci-051120-110428

Dubremetz, M., and Nivre, J. (2015). "Rhetorical figure detection: the case of chiasmus," in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (Denver, Colorado, USA. Association for Computational Linguistics), 23–31. doi: 10.3115/v1/W15-0703

Dubremetz, M., and Nivre, J. (2018). Rhetorical figure detection: chiasmus, epanaphora, epiphora. *Front. Digital Hum.* 5:10. doi: 10.3389/fdigh.2018.00010

Duffy, M., and Thorson, E. (2016). *Persuasion Ethics Today*. London: Routledge and CRC Press. doi: 10.4324/9781315651309

Egami, N., Fong, C., Grimmer, J., Roberts, M., and Stewart, B. (2022). How to make causal inferences using texts. *Sci. Adv.* 8:2652. doi: 10.1126/sciadv.abg2652

Falk, E., Rameson, L. T., Berkman, E., Liao, B., Kang, Y., Inagaki, T. K., et al. (2010). The neural correlates of persuasion: *a* common network across cultures and media. *J. Cogn. Neurosci.* 22, 2447–2459. doi: 10.1162/jocn.2009.21363

Feldstein, S. (2023). "The consequences of generative ai for democracy, governance and war," in *Survival: October-November 2023* (London: Routledge), 117–142. doi: 10.4324/9781003429388-13

Floridi, L. (2013). *The Ethics of Information*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199641321.001.0001

Fogg, B. J. (2008). "Mass interpersonal persuasion: an early view of a new phenomenon," in *Persuasive Technology: Third International Conference, PERSUASIVE 2008, Oulu, Finland, June 4-6, 2008. Proceedings 3* (Springer), 23–34. doi: 10.1007/978-3-540-68504-3_3

García-Orosa, B., Gamallo, P., Martín-Rodilla, P., and Martínez-Castaño, R. (2021). Hybrid intelligence strategies for identifying, classifying and analyzing political bots. *Soc. Sci.* 10:357. doi: 10.3390/socsci10100357

Gardikiotis, A., and Crano, W. D. (2015). Persuasion theories. *Int. Encycl. Soc. Behav. Sci.* 2015, 941–947. doi: 10.1016/B978-0-08-097086-8.24080-4

Gavenko, S. V. (2001). "Analysis of the argumentative effect of evaluative semantics in natural language," in *International Conference on Computational Science* (Springer), 979–988. doi: 10.1007/3-540-45545-0_110

Godber, A., and Origgi, G. (2023). Telling propaganda from legitimate political persuasion. *Episteme* 20, 778–797. doi: 10.1017/epi.2023.10

Goffredo, P., Haddadan, S., Vorakitphan, V., Cabrio, E., and Villata, S. (2022). "Fallacious argument classification in political debates," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, 4143–4149. doi: 10.24963/ijcai.2022/575

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., and Tomz, M. (2024). How persuasive is ai-generated propaganda? *PNAS Nexus* 3:pgae034. doi: 10.1093/pnasnexus/pgae034

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Goovaerts, I., and Marien, S. (2020). Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Polit. Commun.* 37, 768–788. doi: 10.1080/10584609.2020.1753868

Hamilton, W., Clark, K., Leskovec, J., and Jurafsky, D. (2016). "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 595–605. doi: 10.18653/v1/D16-1057

Haq, E., Braud, T., Kwon, Y., and Hui, P. (2020). A survey on computational politics. *IEEE Access* 8, 197379–197406. doi: 10.1109/ACCESS.2020.3034983

Hinton, M., and Wagemans, J. H. (2022). Evaluating reasoning in natural arguments: a procedural approach. *Argumentation* 36, 61–84. doi: 10.1007/s10503-021-09555-1

House, O. C. (2019). *Disinformation and 'Fake News': Final Report*. London: House of Commons.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electron. Mark.* 31, 685–695. doi: 10.1007/s12525-021-00475-2

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., et al. (2022). "Logical fallacy detection," in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Abu Dhabi, United Arab Emirates: Association for Computational Linguistics), 7180–7198. doi: 10.18653/v1/2022.findings-emnlp.532

Jowett, G. S., and O'donnell, V. (2018). *Propaganda Persuasion*. New York: Sage publications.

Jurkiewicz, D., Borchmann, L., Kosmala, I., and Graliński, F. (2020). "ApplicaAI at SemEval-2020 task 11: on RoBERTa-CRF, span CLS and whether self-training helps them," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (Barcelona: International Committee for Computational Linguistics), 1415–1424. doi: 10.18653/v1/2020.semeval-1.187

Khazaei, T., Xiao, L., and Mercer, R. (2017). "Writing to persuade: analysis and detection of persuasive discourse," in *iConference 2017 Proceedings*.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures Their Consequences*. New York: Sage. doi: 10.4135/9781473909472

Klemp, N. (2010). "When rhetoric turns manipulative: disentangling persuasion and manipulation," in *Manipulating Democracy* (Routledge), 77–104.

Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., and Puppe, F. (2016). Uima ruta: Rapid development of rule-based information extraction applications. *Nat. Lang. Eng.* 22, 1–40. doi: 10.1017/S1351324914000114

Kong, L., Li, C., Ge, J., Luo, B., and Ng, V. (2020). "Identifying exaggerated language," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7024–7034. doi: 10.18653/v1/2020.emnlp-main.571

Lauriola, I., Lavelli, A., and Aiolli, F. (2022). An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing* 470, 443–456. doi: 10.1016/j.neucom.2021.05.103

Mann, W. C., and Thompson, S. A. (1988). Rhetorical structure theory: toward a functional theory of text organization. *Text-Interdisc. J. Study Discour.* 8, 243–281. doi: 10.1515/text.1.1988.8.3.243

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. doi: 10.3115/v1/P14-5010

Manson, N. C. (2012). Making sense of s pin. *J. Appl. Philos.* 29, 200–213. doi: 10.1111/j.1468-5930.2012.00566.x

Mercier, H. (2020). *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton, NJ: Princeton University Press. doi: 10.1515/9780691198842

Mercier, H., and Sperber, D. (2017). *The Enigma of Reason*. Cambridge: Harvard University Press. doi: 10.4159/9780674977860

Miller, C. R. (1939). *How to Detect and Analyze Propaganda*. Town Hall, Incorporated.

Mohammad, S. (2018). "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 174–184. doi: 10.18653/v1/P18-1017

Mutlu, E. Ç., Yousefi, N., and Ozmen Garibay, O. (2022). "Contrastive counterfactual fairness in algorithmic decision-making," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 499–507. doi: 10.1145/3514094.3534143

Nannini, L., Balayn, A., and Smith, A. L. (2023). "Explainability in ai policies: a critical review of communications, reports, regulations, and standards in the EU, US, and UK," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1198–1212. doi: 10.1145/3593013.3594074

Nannini, L., Bonel, E., Bassi, D., and Joshua, M. M. (2024). Beyond phase-in: Assessing impacts on disinformation of the EU digital services act. *AI Ethics* 2024, 1–26. doi: 10.1007/s43681-024-00467-w

Nettel, A. L., and Roque, G. (2012). Persuasive argumentation versus manipulation. *Argumentation* 26, 55–69. doi: 10.1007/s10503-011-9241-8

Novelli, C., Taddeo, M., and Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI Soc.* 4, 1–12. doi: 10.2139/ssrn.4180366

O'Keefe, D. J. (2009). "Theories of persuasion," in *The SAGE Handbook of Media Processes and Effects*, 269–282.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds Mach.* 29, 441–459. doi: 10.1007/s11023-019-09502-w

Panchendrarajan, R., and Zubiaga, A. (2024). Synergizing machine learning symbolic methods: a survey on hybrid approaches to natural language processing. *arXiv preprint arXiv:2401.11972.*

Partington, A., and Taylor, C. (2018). *The Language of Persuasion in Politics: An Introduction*. London: Routledge and CRC Press. doi: 10.4324/9781315177342

Pauli, A., Derczynski, L., and Assent, I. (2022). "Modelling persuasion through misuse of rhetorical appeals," in *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, 89–100. doi: 10.18653/v1/2022.nlp4pi-1.11

Perelman, C., and Olbrechts-Tyteca, L. (1971). *The New Rhetoric: A Treatise on Argumentation*. Notre Dame: University of Notre Dame Press.

Pessach, D., and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Comput. Surv.* 55, 1–44. doi: 10.1145/3494672

Petrova, M., and Yanagizawa-Drott, D. (2016). "Media persuasion, ethnic hatred, and mass violence," in *Economic Aspects of Genocides, Other Mass Atrocities, and Their Prevention*, 274. doi: 10.1093/acprof:oso/9780199378296.003.0012

Petty, R., and Cacioppo, J. (1986). "The elaboration likelihood model of persuasion," in *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (Springer), 1–24. doi: 10.1007/978-1-4612-4964-1_1

Piskorski, J., Stefanovitch, N., Nikolaidis, N., Da San Martino, G., and Nakov, P. (2023). "Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 3001–3022. doi: 10.18653/v1/2023.acl-long.169

Pryzant, R., Shen, K., Jurafsky, D., and Wagner, S. (2018). "Deconfounded lexicon induction for interpretable social science," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1615–1625. doi: 10.18653/v1/N18-1146

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. (2020). "Stanza: a python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. doi: 10.18653/v1/2020.acl-demos.14

Reimers, N., and Gurevych, I. (2019). "Sentence-bert: sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. doi: 10.18653/v1/D19-1410

Roberts, R. C., and Wood, W. J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199283675.001.0001

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39:1161. doi: 10.1037/h0077714

Schuster, T., Schuster, R., Shah, D. J., and Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Comput. Ling.* 46, 499–510. doi: 10.1162/coli_a_00380

Seo, K. (2020). Meta-analysis on visual persuasion-does adding images to texts influence persuasion. *Athens J. Mass Media Commun.* 6, 177–190. doi: 10.30958/ajmmc.6-3-3

Seref, M. M. H., and Seref, O. (2019). "Rhetoric mining for fake news: identifying moves of persuasion and disinformation," in *25th Americas Conference on Information Systems, AMCIS 2019, Cancún, Mexico, August 15-17, 2019* (Association for Information Systems).

Sethumadhavan, A. (2018). Trust in artificial intelligence. *Ergon. Des.* 27, 34–34. doi: 10.1177/1064804618818592

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2021). "Nice try, kiddo: investigating ad hominems in dialogue responses," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 750–767. doi: 10.18653/v1/2021.naacl-main.60

Smith, C. A., and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *J. Pers. Soc. Psychol.* 48:813. doi: 10.1037//0022-3514.48.4.813

Sridhar, D., and Blei, D. (2022). Causal inference from text: a commentary. *Sci. Adv.* 8:6585. doi: 10.1126/sciadv.ade6585

Ta, V., Boyd, R., Seraj, S., Keller, A., Griffith, C., Loggarakis, A., et al. (2022). An inclusive, real-world investigation of persuasion in language and verbal behavior. *J. Computat. Soc. Sci.* 5, 883–903. doi: 10.1007/s42001-021-00153-5

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions," in *Proceedings of the 25th International Conference on World Wide Web*, 613–624. doi: 10.1145/2872427.2883081

Tindale, C. W. (2007). *Fallacies and Argument Appraisal, Chapter An Introduction to the Study of Fallaciousness*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511806544

Troiano, E., Strapparava, C., Özbal, G., and Tekiroğlu, S. (2018). "A computational exploration of exaggeration," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3296–3304. doi: 10.18653/v1/D18-1367

Tsinganos, N., Mavridis, I., and Gritzalis, D. (2022). Utilizing convolutional neural networks and word embeddings for early-stage recognition of persuasion in chat-based social engineering attacks. *IEEE Access* 10, 108517–108529. doi: 10.1109/ACCESS.2022.3213681

Villarán, A. (2017). Irrational advertising and moral autonomy. *J. Bus. Ethics* 144, 479–490. doi: 10.1007/s10551-015-2813-z

Vorakitphan, V. (2021). *Fine grained classification of polarized and propagandist text in news articles and political debates* (Doctoral dissertation, Université Céte d'Azur).

Vorakitphan, V., Cabrio, E., and Villata, S. (2021). "Don't discuss: Investigating semantic and argumentative features for supervised propagandist message detection and classification," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1498–1507. doi: 10.26615/978-954-452-072-4_168

Vykopal, I., Pikuliak, M., Srba, I., Moro, R., Macko, D., and Bielikova, M. (2024). Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838.*

Wagemans, J. H. (2023). How to identify an argument type? On the hermeneutics of persuasive discourse. *J. Pragmat.* 203, 117–129. doi: 10.1016/j.pragma.2022.11.015

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behav. Res. Methods* 45, 1191–1207. doi: 10.3758/s13428-012-0314-x

Weston, A. (2018). *A Rulebook for Arguments*. Indianapolis: Hackett Publishing.

Wilson, T., Wiebe, J., and Cardie, C. (2017). *MPQA Opinion Corpus*. doi: 10.1007/978-94-024-0881-2_29

Xu, Z. (2023). Research on deep learning in natural language processing. *Adv. Comput. Commun*. 6:18. doi: 10.26855/acc.2023.06.018

Yang, Z., Hu, Z., Dyer, C., Xing, E., and Berg-Kirkpatrick, T. (2018). "Unsupervised text style transfer using language models as discriminators," in *Advances in Neural Information Processing Systems*, 31.

Yoosuf, S., and Yang, Y. (2019). "Fine-grained propaganda detection with fine-tuned BERT," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, 87-91, Hong Kong, China* (Association for Computational Linguistics). doi: 10.18653/v1/D19-5011

Zarouali, B., Boerman, S., Voorveld, H., and Noort, G. (2022). The algorithmic persuasion framework in online communication: conceptualization and a future research agenda. *Internet Res*. 32, 1076–1096. doi: 10.1108/INTR-01-2021-0049

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., et al. (2019). "Defending against neural fake news," in *Advances in Neural Information Processing Systems*, 32.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M. (2023). Synthetic lies: understanding AI-generated misinformation and evaluating algorithmic and human solutions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany: ACM), 1–20. doi: 10.1145/3544548.3581318