
Deliverable number: D2.1

hybrids

# Technical report on the state of the art of NLP and AI methods for discourse analysis in the political domain

A Comprehensive Survey of Discourse Analysis Approaches in Politics

**Version 1.**

## Project Details

| | |
|---|---|
| **Project Acronym:** | HYBRIDS |
| **Project Title:** | Hybrid Intelligence to monitor, promote and analyse transformations in good democracy practices |
| **Grant Number:** | 101073351 |
| **Call** | HORIZON-MSCA-2021-DN-01 |
| **Topic:** | HORIZON-MSCA-2021-DN-01-01 |
| **Type of Action:** | HORIZON-TMA-MSCA-DN |
| **Project website:** | https://hybridsproject.eu/ |
| **Coordinator** | Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)-Universidade de Santiago de Compostela (USC) |
| **Main scientific representative:** | Prof. Pablo Gamallo Otero, pablo.gamallo@usc.es |
| **E-mail:** | citius.kmt@usc.es, info@hybridsproject.eu |
| **Phone:** | +34 881 816 414 |

## Deliverables Details

| | |
|---|---|
| **Number:** | D2.1 |
| **Title:** | Technical report on the state of the art of NLP and AI methods for discourse analysis in the political domain |
| **Work Package** | WP2: Public Discourse Analysis |
| **Lead beneficiary:** | FBK |
| **Deliverable nature:** | R- Document, report |
| **Dissemination level:** | PU-Public |
| **Due Date (month):** | 31/01/2024 (M13) |
| **Submission Date (month):** | 31/01/2024 (M13) |
| **Keywords:** | Natural Language Processing; Discourse Analysis; Political Domain; Textual Data Processing; Artificial Intelligence |


### Abstract

***D2.1. Technical Report on the State of the Art of NLP and AI methods for discourse analysis in the political domain*** offers a comprehensive overview of existing approaches developed to process textual data in the political domain. It details different tasks and approaches, covering knowledge modelling through ontologies for the political domain, corpus-based analysis and textometry, semantic shift identification in the political domain, argumentation studies, persuasion assessment and the detection of harmful content. The report encompasses approaches that range from word-based distributional analysis to the latest deep learning approach, highlighting for each topic both its limitations and future research directions.

## Deliverable Contributors

| | Name | Institution | E-mail |
|---|---|---|---|
| **Deliverable leader** | Sara Tonelli | FBK | satonelli@fbk.eu |
| **Contributing Authors** | Martial Pastor | RU | martial.pastor@ru.nl |
| | Davide Bassi | CiTIUS-USC | davide.bassi@usc.es |
| | Siddharth Bhargava | FBK | sbhargava@fbk.eu |
| | Erik Marino | UEVORA | erik.marino@uevora.pt |
| | Katarina Laken | FBK | alaken@fbk.eu |
| **Reviewers** | Martín Pereira Fariña | USC | martin.pereira@usc.es |
| | Nelleke Oostdijk | RU | nelleke.oostdijk@ru.nl |
| | Renata Vieira | UEVORA | renatav@uevora.pt |

## History of Changes

| Version | Date | Changes to previous version | Status |
|---|---|---|---|
| 0.1 | 20/12/2023 | First Draft | Draft |
| 0.2 | 16/01/2024 | Consortium Internal review | Review |
| 1.0 | 31/01/2024 | Approved version to be submitted | Final |

## List of Acronyms

| | |
|---|---|
| AAE | African American English |
| AI | Artificial Intelligence |
| CNN | Convolutional Neural Network |
| DA | Discourse Analysis |
| DC | Doctoral Candidate |
| FN | Fake News |
| GRU | Gated Recurrent Unit |
| HS | Hate Speech |
| IDT | Interpersonal Deception Theory |
| LSA | Latent Semantic Analysis |
| LSTM | Long Short-Term Memory Network |
| NLP | Natural Language Processing |
| OCR | Optical Character Recognition |
| PDA | Political Discourse Analysis |
| RNN | Recurrent Neural Network |
| SVM | Support Vector Machine |
| WEM | Word Embedding Model |

# Contents

# 1 Introduction

Researchers in political communication currently face a paradox. Although there has never been so much data available to analyse the different dimensions of political discourse, research in NLP and AI is still needed to exploit the full potential of such resources and gain novel insights into how political discourse is shaped, how its effects can be measured and what are the main risks associated with it.

Political argumentation and knowledge modelling for the political domain have been extensively studied starting from Aristotle's work (Aristotle et al., 1909). Furthermore, existing datasets in this domain have been analysed using corpus-based and textometric approaches including word frequency measures, syntactic patterns and discourse markers. More recently, with the increasing performance achieved by deep learning approaches in several NLP tasks, also research on political discourse has been addressed using novel techniques spanning from word embeddings to large language models, which indeed have reached state-of-the-art performance in several tasks related to political discourse analysis such as stance detection, argument mining or toxic language detection.

In this document, we aim at providing an overview of current research in discourse analysis in the political domain using NLP and AI methods. The topics covered are part of HYBRIDS WP2 and are related to DCs from DC1 through DC5, mainly concerned with public discourse analysis. The overview includes works on ontology and knowledge modelling (Section 2) and on corpus linguistics and textometry for political discourse analysis (Section 3). We further address specific tasks such as diachronic language analysis to identify semantic shifts in the political domain (Section 4) and computational argumentation to capture politicians' rethorical strategies and stance. We then focus on the effects that political discourse can have on its audience. Indeed, we present works on persuasion assessment (Section 6) and on the detection of toxic content in political discourse (Section 7). For each section, we highlight not only current methods, but also existing challenges and possible future research directions. Overall, this overview shows that an effective analysis of political discourse can be performed only through hybrid approaches enabling the combination of deep learning-based methods with domain knowledge, integrating human expertise during learning and decision-making through ontological modelling, human annotation, systems' feedback and more. In this respect, the contribution of human and social sciences is crucial and will need to gain more relevance in NLP and AI-related works for political discourse analysis.

# 2 Ontologies in Political Discourse Analysis

## 2.1 Introduction to Ontologies

Ontologies in computer and information sciences serve as structured frameworks that organize and interpret information (Gruber, 1993). Originating in philosophy, the term "ontology" traditionally referred to the study of existence and the nature of being (Hofweber, 2023). In the digital age, it has evolved to describe a method of representing knowledge in artificial intelligence (Guarino, 1998). As Gruber (1993) elucidates, ontologies define a set of concepts and the relationships between those concepts within a specific domain, facilitating a shared and common understanding that can be communicated between people and application systems (Studer et al., 1998; Borst et al., 1997).

Ontologies play a pivotal role in **Discourse Analysis (DA)**, a field focused on studying the uses and structures of language in texts and spoken or written discourse (Bärenfänger et al., 2008). Ontologies assist in capturing semantic structures, making the underlying meanings in a discourse more accessible (Dou et al., 2015). They also enhance contextual understanding in DA by providing a structured framework that outlines the relations between entities, events, and their attributes (Azzi and Gagnon, 2023). In computational discourse analysis, ontologies enhance the precision and recall of discourse analysis tools (Dou et al., 2015). Ontologies also facilitate interdisciplinary integration, allowing for a richer analysis by combining insights from various disciplines such as linguistics, sociology, and anthropology (Baumann et al., 2021).

## 2.2 Application of Ontologies to Political Discourse

In the intricate field of political discourse, ontologies have become instrumental tools. They allow for a systematic representation of the interconnected beliefs, values, and principles in political ideologies, facilitating, for instance, the tracking of ideological shifts over time (Hanum et al., 2019). Ontologies also enable the dissection of political narratives, uncovering recurrent acts, scenes, nuclei, satellites, topics, agents, semantic roles, and relationships, offering deeper insights into political narratives (Colucci Cante et al., 2023). In political discourse analysis, a key challenge is to go beyond the literal text to grasp the entities and concepts to which the text refers. As Gonzalez-Perez (2023) notes, argument-oriented discourse analysis often adheres closely to the source text. While this faithfulness to the text is essential to avoid bias, it often leads to a lack of attention to the real-world entities and concepts that are being discussed (Gonzalez-Perez, 2020). This limitation necessitates the development of a mental model by the analyst to understand the references, construct meaning, and make sense of the discourse in a broader context.

In this light, ontologies, alongside ontological proxies, play a crucial role (Gonzalez-Perez, 2020). They provide a structured framework that not only captures the literal content of political discourse but also connects it to the underlying entities and ideas in the real world. This approach is particularly beneficial in the analysis of political ideologies and narratives, where it is essential to understand the deeper meanings and references beyond the textual surface. For instance, when a politician discusses "freedom" or "equality", these terms need to be contextualized within specific ideological frameworks to fully understand their implications in the discourse. Further, ontologies assist in elucidating the complex relationships within policy frameworks, aiding in the

comparison and evaluation of policies across different political entities (Kero et al., 2023). In media analysis, these tools help uncover patterns, framing, and potential biases in media reporting, contributing to a better understanding of media influence on public opinion (Piryani et al., 2023).

As an example of the practical application of ontologies in political discourse, Azzi and Gagnon (2023) focused on the development of a new 'Impact of COVID-19 in Canada Ontology' (ICCO) that provides contextualized semantic information on impact in numerous policy areas, building an ontology entirely from Canadian parliamentary debates. This highlights the real-world utility of ontologies in capturing and analyzing specific policy discussions within a given context.

Despite their significant contributions, the application of ontologies in political discourse analysis is not without challenges. The dynamic nature of political discourse requires ontologies to be continuously updated and refined to capture evolving nuances and contexts. Moreover, the creation and maintenance of ontologies demand a significant investment of expertise and time. There is also a risk of over-reliance on ontologies, which may lead to the overlooking of emergent meanings and non-structured elements in the discourse. Plus, as pointed out by Gonzalez-Perez (2023), staying true to the text while simultaneously capturing the entities it refers to requires a delicate balance. Analysts must constantly update and refine their ontological frameworks to keep up with the evolving nature of political discourse and its references to the real world. This task demands not only expertise in ontology development but also a deep understanding of the dynamic political landscape and the various entities and concepts it encompasses.

In conclusion, ontologies, despite their challenges, remain powerful tools in enhancing the depth and breadth of political discourse analysis, enriching academic research, media analysis, and public education.

# 3 Corpus Linguistics and Textometry

## 3.1 Introduction to Textometry

Textometry originated in the 1970s within the domain of lexical statistics (Muller, 1968), primarily aiming to assess vocabulary diversity within a text and perform text analysis exclusively through studying its specific vocabulary. The method's original approach involved using lexical data analysis techniques, such as correspondence analysis and principal component analysis (Benzécri, 1973), to generate visual semantic representations of words within a corpus, including graphs, semantic maps, and clusters. Since 2010, Textometry has become more popular in the social sciences, thanks to the release of open-source software called TXM (Heiden et al., 2010). This user-friendly platform includes the techniques mentioned above, making it easy for users to work with unstructured text. The introduction of this sofware upholds the core principle of textometry, emphasizing the integration of human context analysis alongside statistical data. Indeed with TXM, the interpretation of the calculations is based on numerical indicators as well as the systematic examination of contexts, now facilitated by relevant hyperlinks.

Textometry has gained significant popularity in the fields of humanities and social sciences, spanning from its use in historical archive research (Pincemin et al., 2008; Kastberg Sjöblom and Jacquot, 2016; Thon, 2022) to literary analysis (Boeglin, 2018; Novakova and Siepmann, 2020; Beghini et al., 2023) and finally discourse analysis (Diwersy et al., 2018; Pengam and Jackiewicz,

2022; Carpentieri et al., 2023). This evolution has made it a widely embraced tool for exploring diverse corpora.

A systematic extraction of quantitative information from a written corpus through textometric analysis involves identifying words and text segments with notable frequencies and relationships (Lebart and Salem, 1994). Analytical techniques of this nature have been used to quantitatively identify the primary thematic connections within individual texts as well as within groups of texts. With this methodology, the characterization of the text revolves around the usage of its words within the corpus, while the word itself is defined by its co-occurrences, among other factors (Pincemin, 2012).

In **political discourse analysis**, popular applications of textometry have mainly consisted in identifying unique characteristics of text, groups of text by looking at the repetition of text segments. For instance the authors of a study looking at oral debates of the socialist primaries for the French presidential election (Marchand and Ratinaud, 2012) aim to answer this question in their article “What are the words, phrases, and lexical relationships that characterize each of the debaters?”. In their analysis, it is initially observed that the speakers consistently use the same lexicon throughout the debates. Consequently, the debates did not structurally shape the corpus as much as the debaters themselves. The study reveals that one participant opposes nearly all others, while two others oppose a different set of participants. The focus then shifts to identifying the specificities of the debaters. The textometric analysis gives the authors the ability to achieve a fine-grained analysis of discourse modalities and of the relationship between lexical forms allowing them to identify corpus themes. This suggests that speakers share discussion topics but differ in their approaches.

Other studies move from textometric statistical observations to qualitative explorations of semantic correlates from key phrases. For instance Bouzereau (2022) analyzes Front National discourses focusing on the term “immigration”. Their analysis uses typical textometric methods such as specificity and banality to reveal that the term ”immigration” holds a fundamental place in both the vocabulary of Front National and the construction of far-right ideology.

## 3.2 Specificity, Similarity and Correspondance Analysis

Textometry uses contingency tables for interpreting qualitative data in statistical terms. Some classical methods used in Textometry include *specificity measures*, *factor analysis*, *classification methods*, highlighting their role in understanding complex data. Readers are encouraged to explore nuanced interpretations, departing from elementary descriptive statistics.

**Similarity** analysis uses elements from graph theory (Abric et al., 1962). It constructs a graph illustrating the co-occurrences of forms in segments and the strengths of their associations, following Longhi (2017)’s approach. This graphical analysis visually depicts the local connections authors establish between forms, representing the concepts they employ when discussing a specific subject. Further information on graph theory and the similarity analysis algorithm implemented in Iramuteq can be found in Marchand and Ratinaud (2012). Their report extends the framework to examine the election debates of the French socialist party in 2012. In their brief analysis, they track the lexical similarity among various debaters. Through chronological graphs of both the first and second debates, they illustrate how debaters modify, or refrain from modify-

ing, their vocabulary in response to the election results.

**Specificity** is a lexico-statistical method related to keyword analysis employed in the British tradition of corpus linguistics, as exemplified in works like Rayson (2003). In specificity for corpus linguistics, first introduced by Lafon (1980), the frequency distribution of a linguistic form across a corpus divided into multiple segments is analysed. In contrast to common practices in the field, Lafon advocates using hypergeometric distribution formulas, with the entire corpus as the norm for fragments. These choices result in computing a valid probabilistic indicator across the entire frequency range. Calculating this indicator for each vocabulary form delineates two subsets: specific forms and basic forms, assigning each segment its lexical specificities. In the study conducted by Bouzereau (2022), specificity is used to examine the term "immigration" within the discourse of the Front National. They observe that key phrases centered around "immigration" are not only statistically significant but also pragmatically crucial in shaping the overall discourse. The term consistently carries a negative connotation, introducing new denominations and representations that contribute to shaping a political discourse.

Most textometry softwares such as TXM usually perform clustering using the Reinert analysis, employing a descending hierarchical classification rooted in **correspondence analysis techniques** (Marchand and Ratinaud, 2012). By utilizing frequency tables generated during lemmatization (step 1) and the initially segmented text (step 2), the algorithm consistently partitions the corpus into homogeneous sections. This division is based on the chi-squared correlation between segments and the frequency of occurrence of active forms in comparable segments.

As an example, the work by Diwersy et al. (2018) uses correspondence analysis to investigate French parliamentary debates, emphasizes certain nouns that are particularly associated with the discourse of the right-wing parliamentary group UMP-LR. These encompass nouns related to the nation (e.g., "French") and other traditional components of conservative ideology, both in social aspects (such as "family", "parent", "child") and economic terms (including nouns denoting learned professions like "doctor", "physician", "notary" and "solicitor").

## 3.3 Discursive Process Analysis

In the analysis of discursive processes, the focus typically transitions from individual lexical units to broader discourse units. These discourse units are defined with considerable flexibility, often tailored to the specific domain under examination. They may encompass a single clause or extend across multiple sentences, serving the purpose of articulating a particular stance.

Concerning the identification of discourse units for political discourse analysis, while keyword-based methods remain crucial, there is a notable focus on verbs and their specific inflections based on the syntactic context. Elaborate syntactic patterns are often developed to extract particular phrases from large text corpora.

For instance, Pengam and Jackiewicz (2022) have defined rule-based syntactic patterns to explore the role of causal representations of jihadist radicalization in a corpus of public statements between 2013 and 2018. Their rules mostly depend on the use of causal discourse markers and of specific causal verbs that they have identified. They then perform a qualitative analysis of the extracted segments; they observe that in institutional discourse, the issue of radicalization is primarily presented by creating 'synthetic' connections between distant or different phenomena,

such as linking delinquency and terrorism. By using this method, along with well-known editorial processes (like codification and enunciative smoothing), institutional speeches strengthen an authoritative image, sometimes overlooking scientific or practical on-the-ground knowledge.

Another study from Herman (2023) which focuses on the perception of EU intervention in debates, conducts both descriptive and explanatory analyses. The latter explores discursive patterns, emphasizing debates about values and recognizing an accusatory context. In their study, specific labels for discourse analysis are introduced. At Level-1, speakers' positions on intervention are categorized as pro or anti-intervention. Level-2 examines whether arguments are positive or negative, revealing an observed trend of increasing negativity over time. Level-3 further dissects the discourse, categorizing arguments based on type: principles such as human rights, rule of law, and sovereignty, and motives such as corruption, double standards, partisan interest, and manipulation. The legitimacy of the EU as a normative political order is scrutinized through a methodology involving the analysis of 62 debates, coding speakers' affiliations, and using a discourse analysis framework. Findings reveal increasing negativity over time, with pro-intervention statements becoming more frequent. Notably, motive-based arguments, reflecting an accusatory mode, are dominant. The conclusion offers a nuanced understanding of 'Democratic Backsliding' debates, highlighting the role of discursive patterns, values, and an accusatory context in shaping perceptions. Notable trends in voting behavior, agreement types, and argumentation underscore the complexity of the issue. The implications emphasize the study's contribution to future analyses of democratic backsliding, informing policy discussions and recognizing the evolving nature of perceptions and values within the European Parliament.

# 4 Diachronic Semantic Shifts in Political Discourse Analysis

## 4.1 Introduction to Diachronic Language Analysis

Language evolves to mirror transformations in society and culture. For instance, "apple" once exclusively denoted the fruit, but it now also signifies a well-known corporation. "Gay" originally described a state of happiness or a type of personality, but today it predominantly identifies a person's sexual orientation (Hamilton et al., 2016b; Kutuzov et al., 2018). The diachronic analysis of language, which examines the evolution of language across time, traces these shifts in the connotations of words, capturing not just their static dictionary definitions, but also their primary references, usage contexts, related emotions, and characteristic users, all of which collectively express the prevailing political sentiments and cultural norms (Jatowt and Duh, 2014; Deo, 2015; Hamilton et al., 2016b). The fluidity of language mirrors the dynamic landscape of political thought and policy. Diachronic semantic shifts, the changes in word meanings over time, serve as a gauge for cultural and political transformations, shaping public discourse, policy framing, and ideological movements (Li et al., 2021). Advances in NLP and AI have paved the way for unprecedented tools to analyze these shifts, providing deep insights into the evolution of political language and thought (Azarbonyad et al., 2017; Hamilton et al., 2016b). This section explores state-of-the-art NLP and AI methodologies that shed light on diachronic semantic shifts within political discourse. By leveraging these technologies, researchers can decode both subtle and overt changes in political language, offering a perspective on the historical progression of political thought and

communication, and the different viewpoints and ideologies (Azarbonyad et al., 2017; Kutuzov et al., 2018).

## 4.2 Semantic Representation through Word Embeddings

The first approaches to analyse semantic representations in distributional terms were based on word embedding models such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). These models encode words into numerical vectors. Word2Vec utilizes local context, employing neural networks to model word associations, while GloVe aggregates global co-occurrence statistics from a corpus (Jurafsky and Martin, 2023). FastText extends these ideas by incorporating subword information, allowing for better representation of morphologically rich languages and handling of out-of-vocabulary words (Joulin et al., 2017). The fundamental premise of the Word2Vec model is that context information by itself forms a reliable depiction of linguistic entities (Biswas and De, 2022). These models have been instrumental in political NLP, enabling nuanced analysis of political rhetoric, speeches, and documents (Oliveira et al., 2018; Rodman, 2020). For example, the semantic proximity of "freedom" to "security" might evolve in a post-conflict political landscape (Foner, 1999). Additionally, sociolinguistic theories could be applied to understand how social factors and group identities influence the linguistic salience of particular terms within political discourse and vice versa (Goldman and Perry, 2002). This interdisciplinary approach allows for a more nuanced interpretation of the data produced by NLP models, linking linguistic patterns to socio-political phenomena.

The dichotomy between Word2Vec's focus on local context and GloVe's global co-occurrence analysis is not just technical but also interpretative (Jurafsky and Martin, 2023). Word2Vec may pinpoint immediate rhetorical shifts post-political events, whereas GloVe may capture broader semantic evolutions (Dharma et al., 2022). This distinction is crucial for political discourse analysis, as each model may reveal different facets of semantic change within political language. Word2Vec's strength lies in its ability to detect subtle shifts in rhetoric, often immediately following political events, thereby providing insights into the immediate impact of political actions on language use. On the other hand, GloVe's integration of global statistics enables it to trace broader semantic trajectories, offering a macroscopic view of how political discourse evolves over time (Rachman et al., 2018). However, the effectiveness of these models for such specific applications would depend on various factors including the size and representativeness of the corpus used, the specific parameters and preprocessing steps in the model training, and the methodology applied to interpret the word embeddings (Tulu, 2022). The accuracy of deep learning language models like Word2Vec and GloVe increases with the size of the text corpus, despite their common omission of sentiment information, leading researchers to favor pre-trained word embeddings for machine learning inputs (Biswas and De, 2022). Future studies could explore the synergistic application of these models, potentially unveiling the intricate interplay between immediate rhetorical responses and long-term semantic shifts in the ever-evolving landscape of political discourse.

## 4.3 Methods for Semantic Shift Analysis

While semantic representation models are focused on creating word vectors, methods for semantic shift analysis are focused on using these vectors to trace changes over time or across different contexts. The main methods we will mention are Linear Mapping (Mikolov et al., 2013) and the Neighbor-Based Approach (Levy and Goldberg, 2014).

The **linear mapping** technique introduced by Mikolov et al. (2013) enhances vector space models by allowing the comparison of word vectors across different temporal embeddings, effectively charting a word's semantic trajectory over time. This method can illustrate, for example, the shifting connotations of "liberal" across various political epochs.

The **neighbor-based approach** to semantic analysis contends that a word's meaning is intimately connected to its nearest neighbors in the embedding space (Levy and Goldberg, 2014). Tracking the evolving neighborhood of a word can reveal shifts in its semantic field. In political texts, this method could elucidate the changing associations of "democracy" with concepts like "participation", "representation", or "liberty" through different historical periods.

**Combining neighbor co-occurrence with linear mapping** offers a nuanced approach to detecting semantic stability and shift, refining our understanding of the evolution of political terminology, and capturing both immediate and gradual semantic changes. This combination of the previous techniques usually performs better than the single use of any of the two (Azarbonyad et al., 2017).

## 4.4 Historical Analysis of Semantic Shifts using Word Embeddings

The application of word embeddings to historical semantics provides a quantitative approach to language evolution. These models can capture semantic information and are based on the distribution of words in texts, thus allowing historians to study words in relation to other words both synchronically and diachronically (Wevers and Koolen, 2020). By analyzing word embeddings models (WEMs) from historical text corpora, researchers can observe shifts in word meanings over time (Lin et al., 2012; Davies, 2012; Dritsa et al., 2022). Cosine distances between word vectors indicate the degree of change in the semantic context of a word, by measuring the global shift in a word's position in the embedding spaces between different periods (Jurafsky and Martin, 2023). The integration of historical linguistics into NLP-based analysis can provide a crucial temporal dimension, connecting the past with the present. By understanding the historical contexts in which certain political terminologies were used, we can better grasp the catalysts for semantic shifts. Furthermore, incorporating insights from political history can help to elucidate the role of political events and movements in driving language change. When we analyze word embeddings from historical texts, we are not just looking at linguistic evolution but also tracing the intellectual and cultural history that has shaped political discourse. The term "austerity", for example, has shifted from a moral or religious nuance to an economic one, especially after financial crises. This semantic evolution is quantifiable through embedding comparisons across different time frames (Hamilton et al., 2016b). But, as Wevers and Koolen (2020) claim, while WEMs have potential in historical research, especially in the study of semantic change and conceptual history, they require large amounts of training data and are sensitive to the size of the data set, OCR quality, spelling variation, and bias in the data. It is, thus, crucial to conduct intrinsic and extrinsic evalu-

ations. Hamilton et al. (2016b) suggest that a trustworthy model needs a corpus with 100 million words and a vocabulary of 1-2 million words per time period, and for smaller corpora, they recommend co-occurrence matrices over word embeddings. Another concern comes from a study by (Hu et al., 2022) that tested SGNS, GloVe and SVD architectures on a small corpus of medieval and classical Spanish and found out that results can vary significantly depending on the chosen type of embedding.

## 4.5 Advanced Models for Dynamic Semantic Analysis

Techniques such as word2vec and GloVe generate a singular vector representation for every distinct word within the vocabulary. On the other hand, contextual embedding approaches, exemplified by masked language models such as BERT (Devlin et al., 2019), assign a unique vector to each instance of word depending on its specific usage in a sentence. The embeddings produced by masked language models have proven to be exceptionally beneficial (Jurafsky and Martin, 2023). Dynamic word embeddings, which adapt over time to capture the evolution of word meanings, are at the forefront of semantic shift analysis (Giulianelli et al., 2020). Unlike static models that offer a snapshot of language, these embeddings can trace the morphing political rhetoric, providing granular insights into language changes. Kutuzov et al. (2018) and Bamler and Mandt (2017) have been instrumental in advancing this approach, developing models that enforce alignment across different time periods, thus allowing for a coherent analysis of semantic trajectories. The advent of deep learning models like BERT has been a game-changer for semantic analysis. These models generate word representations that are context-sensitive, enabling a nuanced understanding of political discourse. For instance, “wall” would have varying embeddings based on whether it is discussed within the realm of national security or immigration reform, reflecting the dynamic nature of political language (Devlin et al., 2019).

## 4.6 Case Study: The Sociopolitical Impact of Semantic Shifts on Migration-related Terms

Words are not static labels, and the way in which they are used reflects and influences public attitudes and policies. Terms such as “immigrant”, “refugee”, “asylum seeker”, “alien”, “invader”, and “illegal” carry varying connotations that shift with the political winds. Their usage within political discourse provides a mirror to the evolving landscape of immigration policies and public sentiment. Few existing studies have explored migrant-related discourse so far. For example, Yantseva (2023) focused on Facebook, analyzing over 1M posts to compare migrant labels and uncover prevailing narratives about immigration in Swedish-speaking communities. Examining how these terms have been deployed in political rhetoric over time can yield insights into migration policies and the public mood. For instance, a diachronic analysis of parliamentary debates, policy documents, and media coverage can reveal trends in the portrayal of immigrants and refugees, shedding light on the strategies used to galvanize support or dissent for certain policies. Further, the sentiment analysis of social media platforms, where public discourse is vibrant and unfiltered, can offer real-time reflections of public opinion and potentially forecast shifts in policy directions or electoral outcomes. It is essential to note that the semantic shifts in these terms often precede

or coincide with legislative changes, indicating a preparatory phase in public discourse that sets the stage for policy implementation (Yantseva, 2023). Such case studies underscore the power of language in shaping the socio-political narrative. They also raise important questions about the role of media, politicians, and influential public figures in framing these issues. The discourse surrounding migration is a clear example of how language not only describes reality but also has the power to construct it. The findings of Martinc et al. (2020) significantly augment our understanding of semantic shifts in migration terms, particularly in how they reflect and influence socio-political narratives. Their innovative use of BERT to track contextual word representations over time allows for a nuanced analysis of how terms like "immigrant", "refugee", "asylum seeker", and others evolve in response to socio-cultural events. This method, capable of detecting relative semantic changes, could be instrumental in understanding the evolving connotations of migration-related terms in political discourse. For example, the way "immigrant" may become more closely associated with terms like "crisis" or "policy" in specific periods can reveal underlying shifts in public sentiment and political rhetoric.

## 4.7 Challenges and Future Research Directions

From the foundational Word2Vec and GloVe to cutting-edge dynamic embeddings and BERT, a number of tools and approaches has been developed to decode the semantic intricacies of political rhetoric. The integration of machine learning classifiers has further refined our ability to systematically analyze political discourse, providing valuable insights that mirror the ever-changing political landscape. The evolution of political language is not only a linguistic phenomenon but also a reflection of shifting power structures, cultural transformations, and societal values. Political science offers a lens through which we can understand the strategic use of language by political actors, while sociology provides insights into how collective beliefs and societal changes precipitate linguistic evolution. Cultural studies further illuminate how language both influences and is influenced by cultural identity and norms. Together, these disciplines can help us to interpret the complex relationship between language and its wider context, enriching our analysis with a multifaceted perspective on political discourse.

There is a promising potential in **hybrid models** that combine the computational power of LLMs with structured knowledge representation techniques, such as knowledge graphs and ontologies. This approach can provide a richer semantic understanding by connecting linguistic data with structured world knowledge. Such hybrid models would be particularly useful in contextualizing political discourse within a broader socio-political framework. A recent study by Qiang et al. (2023) exemplifies this hybrid approach by integrating the capabilities of LLMs, specifically in the form of two Siamese agents, with the process of ontology matching. This novel methodology, termed "Agent-OM", leverages the generative and comprehension skills of LLMs to enhance the matching of ontologies. This integration allows for a more intuitive and efficient handling of information retrieval and entity matching, while also benefiting from the self-learning and adaptive nature of LLMs. The success of Agent-OM in performing complex matching tasks with limited examples showcases the potential of such hybrid models in not only understanding and connecting linguistic data but also in adapting to various contexts and domains, including the analysis and interpretation of complex socio-political discourses.

Future research should further explore also the **integration of generative LLMs in the analysis of diachronic semantic shifts**. Large Language Models, with their extensive training on diverse datasets, offer an unprecedented understanding of language nuances and complexities. These models can be particularly adept at capturing subtle linguistic changes and can play a significant role in enhancing the granularity of political discourse analysis. The research conducted by Palagin et al. (2023), on the "OntoChatGPT" system contributes to this vision by demonstrating the effective integration of LLMs, specifically ChatGPT, with ontology-driven structured prompts for meta-learning. OntoChatGPT system effectively extracts entities from contexts, classifies them, and generates relevant responses. This integration not only leverages the extensive training and nuanced language understanding of LLMs but also enhances their capability to manage and interpret information with greater context sensitivity. This approach can be particularly useful in analyzing diachronic semantic shifts, as the ontology-based structuring can aid in pinpointing and understanding subtle linguistic changes over time. The methodology developed in "OntoChatGPT" therefore aligns with the future direction of employing LLMs in political discourse analysis, adding an extra layer of sophistication and precision to the detection and interpretation of semantic shifts in languages.

Finally, research should continuously **adapt to the evolving political landscape**. This includes updating models with recent data and refining methodologies to reflect changes in language use, ensuring that the tools remain relevant and effective in capturing the current state of political discourse. As the political landscape continues to evolve, so does the language that defines it. The integration of advanced NLP and AI methodologies, including LLMs and hybrid models, will be pivotal in navigating and understanding these changes, providing researchers and policymakers with powerful tools for exploration and analysis.

# 5 Argumentation Theory in Political Discourse

## 5.1 Introduction to Argumentation Study

Argumentation study is a field of study that is rapidly gaining importance in Artificial Intelligence for its role in understanding and interpreting human reasoning and decision making (Bench-Capon and Dunne, 2007). It draws its inspiration from the times of ancient Greek philosophers and rhetoricians, where it was originally seen as a means of examining errors in logical reasoning and identifying fallacies in statements and opinions. It essentially focused on how assertions are proposed, discussed and resolved in societal settings where diverse opinions and stances are upheld. This approach however ignored the dialectical nature of argumentation, which gained relevance as researchers started studying situations where there is an exchange of ideas and/or positions (such as in political debates, speeches, and dialogues). They started exploring political discourse from an argumentative perspective. Argumentation study has now been applied to the areas of persuasion, negotiation and dispute resolution, finding application in the fields of law, psychology, politics, and sociology. Coupled with linguistic and ontological understanding of the arguments, it has become possible to detect, extract and investigate the arguments in discourse media. Techniques have been developed to derive useful information and knowledge from the extracted argument corpora. This has led to the birth of computational argumentation, which is a

consolidated area of research in the scientific community (Green, 2014).

Much of the traditional interpretations of argumentation involved analyzing how propositions are asserted and inter-related, as in the forms of supporting or attacking a claim, in the context of conflicts or opinions. This requires identifying what constitutes as an argument and its components as well as defining the rules and protocols that guide the argumentation process (Bench-Capon and Dunne, 2007). Over the years, multiple argument component structures and argumentation schemes have been designed and analysed for the purpose of developing efficient argumentation models. However these argumentation schemes are often subjective to the context, the language and the audience. Arguments can be seen then as *argument as a product* (O'Keefe, 1977) and *argument as a process* (Walton, 2003). Likewise, the fields of formal logic and mathematical reasoning have contributed to designing formal deductive argumentation as a powerful mechanism to model and analyse arguments in a logic-based framework (Besnard and Hunter, 2008). These works focus on a monological approach to argumentation, wherein there is a set of conflicting possibilities that are collected by an agent or group of agents. However, arguments are *defeasible* (i.e. could be challenged any time), are *subjective* to the perception and prejudices of the audience, and unlike "mathematical proofs" have no characteristic of "correctness" to themselves. Thus, developing argumentation models in AI requires models that can account for incomplete information and uncertainty, capture defeasibility, and account for the subjective aspect of argumentation (Bench-Capon and Dunne, 2007). Also the logical structure, i.e. relationship between the components of arguments, may be unclear/implicit (known as enthymemes) or not immediately expressed and require further special analysis to extract this information. These challenges can be further compounded when processing unstructured/web data such as social media data (Habernal and Gurevych, 2016).

Designing good argumentation models depends on several aspects of knowledge: knowing the linguistic constraints, the domain dependence, conceptual relations (like commonsense reasoning), and the discourse structure. This leads to a better discourse awareness, improved explanation of the speaker's intention and beliefs, and the speaker's interaction with other speakers and the world (audience). Argumentation models do not investigate the validity or correctness of an argument but they can be used to differentiate between a true argument and an invalid one, as well as understand the position of the argument with respect to the issue under discussion.

## 5.2 Computational Argumentation

Computational advancements have led to the development of the field of computational argumentation which combines knowledge from symbolic, linguistic and ontological representations of arguments with computational models that can capture and build the knowledge representation. *Argument mining* or argumentation mining aims at automatically extracting argument structures from the natural language text using a combination of NLP and machine learning (ML) techniques (Manfred Stede, 2018). Argument mining is an extremely demanding task in terms of semantics and relies heavily on having proper annotated data. In general due to the lack of an exact definition, most researchers in this field focus on analyzing the discourse at the pragmatic level and apply a certain argumentation theory to the textual data at hand. This also implies that different argument models, focusing on different text genres and having different aims, differ from one an-

other. Some early works such as McGuire et al. (1981) focused on studying non-classical logic on argumentation in AI. They developed a structural model of argumentation incorporating support and attack notions within a graphical structure and applied to textual reasoning. Chesñevar et al. (2000) did a comprehensive survey on designing some of the earliest argumentation models in AI. Some recent and relevant surveys in the domain of computational argumentation include Lippi and Torroni (2016); Reed and Budzynska (2019); Cabrio and Villata (2018); Lawrence and Reed (2019) and Schaefer and Stede (2021). Habernal and Gurevych (2016) focused on studying argumentation mining in web discourse highlighting the differences between structured argument datasets and unstructured argument datasets.

### 5.2.1 Tasks of Argumentation

Computational Argumentation could broadly be divided in four general argument tasks (Manfred Stede, 2018) – the task of structure prediction, the task of evaluation, the task of analysis, and the task of generation. The task of argument structure prediction is to identify and extract arguments and argument components such as premises and claims/conclusions. It may involve performing relation prediction, i.e. understanding how two or more arguments are inter-linked to one another, to study if the arguments are supporting or attacking the claims. In the task of argument analysis, we try to draw information from the argument text itself. This can involve finding implicit premises or conclusions that may need to be made explicit or doing an enthymeme (missing argument) analysis. In the task of evaluation, we determine which argument is strong or weak with respect to a general criteria of evaluation. This can be a relative evaluation that finds application in argument reasoning and quality quantification. Lastly, the task of generation or argument invention can be used to construct new arguments or conclusions from the provided context and finds application in summarizing argumentative texts and essay writing.

### 5.2.2 Argumentation Corpora

Argumentation is a multi-faceted field of study and having well-structured argument-annotated corpora is crucial for training and building computational models that perform the various tasks of argumentation analysis. It can take various forms— spoken, written or graphical. However, it has been observed that most argument annotation projects follow their own assumptions and guidelines, i.e. in terms of the genre of text they focus on and how they apply and analyse the theoretical argumentation model as per their own objectives. The work by Lopes Cardoso et al. (2023) gives a recent and well-documented summary of various argumentative annotation techniques and approaches bridging the discussion between these projects and the theoretical argumentation models. It highlights great variation in the annotation schemes and methodology followed, which has further led to the issues of compatibility and reliability between different argument-annotated corpora and models (Feng and Hirst, 2011; Mochales and Moens, 2011; Walton, 2012; Florou et al., 2013).

One of the earliest works in annotating argumentative discourse was Argumentative Zoning for scientific publications (Teufel et al., 1999). Later, Reed and Rowe (2004) presented Araucaria, a tool for drawing argument maps that supported both convergent and linked arguments, enthymemes and refutations. Stab and Gurevych (2014) annotated 90 essays annotating claims,

major claims, premises and their relations to the claim (support or attack). Online debate portals like the debate.org and debatepedia.org are popular debate portals from where arguments can be collected for various controversial topics (Wachsmuth et al., 2017; Al-Khatib et al., 2016). In Table 1, some datasets of political arguments have been listed.

| Dataset | Document Source | Size |
|---|---|---|
| Lippi and Torroni (2016) | Sky News Debate for UK elections | 9,666 words |
| Duthie et al. (2016) | UK Parliamentary record | 60 sessions |
| Naderi and Hirst (2016) | Canadian Parliament Speeches | 34 sent. (123 paras) |
| Rob Abbott (2016) | Corpora of Political debate on internet forums | 482k posts |
| Haddadan et al. (2019) | 50 years of US presidential campaigns | 39 debates (29k argument components) |
| Ajjour et al. (2019) | Argument framing dataset from online debate portals | 465 topics (12,326 arguments) |
| Menini et al. (2018) | Nixon-Kennedy-Presid. Campaign | 5 topics (1,907 pairs) |
| Visser (2020) | US 2016 Debates and social media discussions | 97,999 words (tokens) |

Table 1: Argumentation datasets pertaining to political topics

## 5.3 Argumentation in the Political Domain

As stated by Aristotle (2000) in his work *Nicomachean Ethics*, we often deliberate, or in a broader term argue, on instances/actions where we own some agency or interest. Often times the outcome in such deliberations is unclear or the right answer is not defined. We seek others in our acts of deliberation when we don't trust our own ability to discern the right answer. Also, we are more interested in deliberating the methods that reach the end rather than the end itself. For instance, a politician would not deliberate about whether s/he will produce a good result but rather state her/his intended goal and then examine on how to achieve the same. Likewise the public expects clear justification on why the policy should be implemented and the politician needs to work on persuading the masses to their cause. We can see the critical role that argumentation plays in studying the field of political science and sociology.

Van Dijk's definition of Political Discourse Analysis (PDA) (Van Dijk, 1997) defines PDA as critical analysis of political discourse focusing on the reproduction and the contestation of political power as perceived from the political discourse. The political discourse, or the political context, refers to the institutional context or platform (parliament/government/Internet discussion forums/social media) that make it possible for the political actors to exert their agency and act on the world in a way that has impact over matters of concern. Many political activities such as reporting, briefing, formulating, summarizing, negotiating, and deliberation require rhetorical and persuasive skills. Consequently, argumentative analysis plays an important role in studying PDA.

In Fairclough (2012), the authors discuss the role that argumentation can play in political discourse analysis, and even argue that practical argumentation is one of the best approaches

that can be undertaken for the same. It has found application in various political activities and studies. These include, but are not limited to, persuasion, negotiation and dispute resolution techniques, frame identification, public opinion research, and hate speech moderation. It has also found many implementations within the field of opinion mining and sentiment analysis, where argument mining can be used to study the polarity of opinions, stance detection and opinion diffusion.

Topoi analysis is a widely used approach, stemming from classical argumentation theory, that has been applied to many dimensions of politics. *topos* (Garssen, 1996) can be defined as, search formulas which tell you how and where to look for arguments. These can be seen as warrants that guarantee the transition from argument to conclusion. In other words, topos are argumentative schemes that can be used to locate arguments in the discourse leading to a conclusion. A well documented explanation of this field can be seen in the research work done by Žagar (2010). An application of this approach can be seen in the work by Reisigl and Wodak (2017), which studied the topoi in discriminatory discourses around migration, collected from various European right-wing populist discourses.

Framing is another tool that employs argumentation to perform discourse analysis on controversial topics. Framing is used to emphasize a certain aspect of the controversy. It is a decisive step to construction of the arguments and affect the outcome of debates. For instance, in Walsh (2017), the author discusses how the rhetorics around climate debate are raised in society. He highlights how the rhetorics and the political framing can shape the conversation or dialogue. Ajjour et al. (2019) highlights how framing can be used in argumentation and presents an unsupervised approach to identifying the frames in argumentative texts. Heinisch and Cimiano (2021) presents a supervised approach to identifying frames in arguments using a multi-task approach that clusters frames from issue-specific, user-provided labels gathered at a variable level of granularity. Haddadan et al. (2023) also do frame identification and topic modelling on political arguments collected from presidential debates. They used simple language models such as BERT and RoBERTa to classify generic frames and show that their approach beats the state of the art.

Public Opinion research focuses on studying the public sentiment or opinion on a specific topic or voting intention. Studying public opinion with regard to the public policy is beneficial in understanding the needs, views and expectations of the public and in turn also ensure engagement from the latter as regards the public agenda (Burstein, 2003). An initiative by COST (European Cooperation in Science and Technology) called APPLY or Argumentation and Public PoLicY (https://publicpolicyargument.eu/about/) focused on researching ways to improve how European citizens understand, evaluate and contribute to public decision-making.

Stance Detection and Opinion mining are another area of research which is very relevant in the political domain. Past research (Hasan and Ng, 2014; Gottipati et al., 2014; Qiu et al., 2013) has used various approaches to recognizing the stances taken by the speakers in online debate platforms. But recently there has been research work that has employed argumentation-motivated features or data to identify the stance or the opinion of people on various controversial topics. Park et al. (2011) focused on dealing with contentious issues in Korean news discourse using ”argument frames” although the formalization on the arguments itself remained unexplained. Walker et al. (2012) used features from the linguistic and argumentative structures such as rhetor-

ical relations, POS generalized dependencies, etc. to improve the performance of their stance classification model. Works by Fang et al. (2012), Bar-Haim et al. (2017) and Körner et al. (2021) have classified the stances of the mined arguments as pro or con towards a given topic. These have found application in argument search, argument synthesis and to study opinion polarization, public opinion and opinion diffusion in the political domain. More on the study of opinion mining and argumentation in controversial issues will be discussed in the Deliverable 2.2 that focuses on discourse analysis techniques in political communication.

Argumentation techniques have also been widely used in persuasion, negotiation and dispute resolution since historical times. This has been further expanded and discussed in Section 6 of this deliverable. In the domain of hate speech moderation, argumentation has been used to improve the performance of existing hate speech algorithms. Some novel works include Domínguez-Armas (2023), where the authors looked into how propagandistic messages play a negative role in public discourse, and the work of Furman et al. (2023), where the authors seek to employ argumentative analysis to produce a more informative textual information. They applied this analysis to the Hateval Corpus and show that some components of hate speech can be reliably identified. More discussions on hate speech detection in the political domain can be found in Section 7.

## 5.4 Challenges and Future Research Directions

One of the main challenges of argumentation in politics is to "find ways of integrating the analysis of the discursive production of reality with the material (social practices) from which the social constructs emerge and in which the speakers (actors) engage." (Aikin, 2014). In other words, one major challenge is applying techniques and tools to the social/public platforms where the voices, opinions and decisions of political actors are communicated with the public.

Another challenge concerning the study of argumentation in the political domain, as raised by Orwell in his essay "Politics and the English Language" (Orwell, 2013), is the challenge of language itself, that is, what if the political language used to discuss makes fair argumentation impossible? Realistically one cannot always expect the ideals of rationality and reason to exist in political debates. Indeed, there is the presence of manipulation, spin and propaganda. In these scenarios, argumentation can be seen as ineffective as it takes a long time to achieve agreement between the arguers. For instance, discussions between Conservatives and Liberals can often be seen as speaking different political languages even though they employ the same terms. This is because the meaning or the employment of the terms themselves differs greatly when used by either side. Studying public argument becomes difficult when dealing with political opponents having different backgrounds and diverse opinions. Thus, when training models on argumentation it is imperative that the data contains views of all possible beliefs and representations, and ensure no bias exists in the data.

Ironic or sarcastic comments (parody), in absence of unmistakable cues, are also quite difficult to distinguish from earnest contributions to the conversation, as seen in the Poe's law (Aikin, 2012). There is also the case of pushover argumentation or straw-man fallacy, where one takes an opponent's views and arguments and manipulates them to make them indefensible, refutes the manipulation and finally presents oneself as having refuted the opponent. These acts of

argumentation provide a disservice to the opponent and also the to fickle-minded audience. A quite popular tool in this regard is the use of memes, which are often used to discredit and ridicule differing political opinions and beliefs. This has also led to the issue of group polarization, which can lead to cognitive contempt, in-cooperativeness and limited perspective between the groups. Also Godwin's law (Godwin, 1994) could apply, which states that the longer a critical discussion continues online, the greater is the likelihood of a Hitler analogy being made. The aim is not to end the debate but to identify a fallacy of relevance. Both Poe's law and Godwin's law invoke different argumentations on the Internet, where Poe's law invocation can be detrimental to argumentation, but Godwin's law can be conducive to good argumentation.

The third possible challenge to online political argumentation is dealing with argument trolls (Cheng, 2017) and Gish Gallops. These are practices that are used to hoard attention toward themselves in an online critical discussion. Argument trolls would employ overblown objections and personal attacks to derail the Internet Argument. They thrive on negative reactions and any interaction with them only makes them more unhinged. "Gish Gallops" refers the people who seek attention by making a big show, by objecting to everything (making a list of complaints) being discussed without the need for a critical assessment. "TL;DR" or *Too Long; Did'nt Read* is an approach to handling the gish gallops but it is considered a rude approach as it invalidates the response from the gish gallop solely because it was too long. Consequently, we are advised to just not participate in polluted argumentative conversations. Clearly, one can see that when it comes to training argumentative models on online platforms, the need to be mindful of these challenges is vital in order to avoid polluting the training and the learning of the AI model.

In general, it is vital for us to make our argumentative practices clear to ourselves, as it can allow for us to understand them better and consequently work on improving them. In the domain of political science, this becomes even more critical as the political platforms are full of polarizing opinions, misinformation, propaganda, spin (framing) and fallacies. Studying these aspects of political discourse can help us understand why and how arguments go wrong and then can help us in learning how we can do better at argument. With the help of Artificial Intelligence as seen in this section, we can show how it is possible to achieve models that can meet near human performance on certain tasks. AI and NLP techniques can potentially enable us to narrow the gap between the theoretical study of argumentation and its application in practice.

# 6 Mixed Methods for Persuasion Assessment in Political Discourses

## 6.1 Introduction to Persuasion Analysis

The systematic study of persuasion boasts an ancient and illustrious lineage, dating back to antiquity, as far as Aristotle (Demirdöğen, 2016). In consonance with their predecessors, contemporary social scientists devote their examinations of the manifold factors governing the degree of success attained by a persuasive endeavor. Nowadays, this scientific enterprise is sustained also through the methodological and instrumental opportunities offered by the world of NLP and AI (Zarouali et al., 2022). In the following sections, we will provide a more comprehensive overview

of how AI and NLP technologies are implemented for the study of persuasion.

The first issue that needs to be addressed is what do we mean with persuasion? Persuasion, in fact, is a fuzzy-edged concept. Accordingly 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. The multifaced nature of this concept generated a proliferation of definitions and methodologies: each one trying to seize a specific feature of the phenomena. This "theoretical-methodological diaspora", in turn, led to confusing if not even contradictory conclusions about persuasion, with detrimental effects on the knowledge production on this theme Druckman (2022). Starting from this "definitory issue", (O'Keefe, 2015) provides a comprehensive definition of persuasion as *"a successful intentional effort at influencing another's mental state through communication in a circumstance in which the persuadee has some measure of freedom"* (O'Keefe, 2015, p.27). In the remainder of this section, we will refer to this last definition since, rather than trying to provide a sharp and definitive characterization of persuasion, it aims at identifying the central core of it, allowing us to bride all the different facets of the phenomena.

## 6.2 Persuasion in Political Discourse

Language plays a crucial role in the process of translating political intent into tangible social action Partington and Taylor (2018). Drawing then on the definition of persuasion offered above, it becomes evident how this particular use of language exhibits a noteworthy affinity with the political sphere. In democratic contexts, indeed, winning an election hinges primarily upon the quantity of individuals whom the candidate has effectively garnered through discursive means. Persuasion, therefore, is an extensively examined topic in political discourse analysis, as it enables diverse stakeholders to finely tailor their messages and maximize their ability to secure public acceptance (Cakanlar and White, 2023; Coppock et al., 2020; Druckman, 2022).

The study of persuasion in political discourse analysis, thus, can effectively be implemented to nourish the physiology of politics. At the same time, however, it can be employed for harmful political uses, threatening the public sphere. There is in fact an increasing awareness regarding the detrimental effects of persuasion misuse in politics: ranging from political distrust to the manipulation of the election outcome (Goovaerts and Marien, 2020; House of Commons, 2019). The adverse facets of persuasion, particularly in the digital age, have garnered significant attention in the scientific community. The rise of the Internet and data digitization has led to an unprecedented surge in data creation, aggregation, and transformation (Haq et al., 2020). 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). Political entities have adeptly leveraged these platforms to propagate their ideologies (Zarouali et al., 2022; Haq et al., 2020). Intensifying the reach and the efficacy of the efforts to shape the public discourse, however, represents also a threat for transparent democratic deliberations, as testified by the propagandist campaigns carried out by authoritarian regimes in the recent years (Feldstein, 2023).

Given this context, there is a compelling demand for methodologies that, exploiting human expertise, can develop algorithms to autonomously scrutinize extensive volumes of online data, providing institutions and citizens with tools and information to deal with the risks deriving from

online persuasion. Subsequently, the forthcoming sections will provide an overview of the artificial intelligence (AI) and natural language processing (NLP) techniques employed in the automated analysis of online persuasion.

## 6.3 Mixed Methods for the Observation of Political Persuasion

As stated above, the study of persuasion is characterized by a pluralism of perspectives. Moreover, the integration of computational techniques has expanded the potential for a diverse array of novel methodologies and strategies in the assessment and analysis of persuasive communication within the political sphere. In the following sections we distinguished the collected literature in two main categories to create a comprehensive overview. The first one *Persuasion as a set of Linguistic Style Units* contains all the research conceiving 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. This approach, thus, focuses 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 rhetorical-argumentative structure. The studies falling inside 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 materials used to realize it.

### 6.3.1 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 this 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, 2018), training three log-linear probability classifier on a corpus of political debates, obtaining promising results (Chiasmus F1=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 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 applied also to dataset 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 studied rhetorical figure is the one of “hyperbole”, also known as “exaggeration”: a rhetorical figure implemented mainly to create amusement, express emotions and draw attention. Regarding the political domain, 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 traditional models. Depending on the particular rhetorical figure, models exploited different linguistic features, such as: punctuation, sentence size, similarity and lexical structures; combined with different embeddings. The best results in this classification task were obtained using the most interpretable 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), Kong et al. (2020) furthered the exploration by developing a Chinese dataset of hyperboles (HYPO-cn), which comprises 4762 sentences, including 2680 hyperbolic ones. This focus on a Chinese dataset 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 accuracy of 0.85 ($\approx$+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 rhetorical devices (such as “pysma”, “epizeuxis” or “polysyndeton”). In this case, using Apache Ruta (Kluegl et al., 2016)(a rule-based script language designed to enable rapid development of text processing), the authors created an algorithm based on the formalized syntactic structure of the rhetorical figures. 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. Subsequently, the authors implemented the model to analyze the use of these rhetorical 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, as well as to assess the different use of rhetorical figures between, for example, monologues (such as newspapers articles) and dialogues (such as political debates).

To identify persuasive arguments, 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 about it in order to try to change the perspective of the publisher. The convincing arguments are tagged with a $\triangle$. 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 assumed that the level of arousal (the intensity of an emotion), concreteness (denoting something perceptible), dominance (expressions of control) and valence (words' pleasantness) 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, POS, Number of Words and a combination of all of them). The study confirmed the relation between some of the linguistic features and level of persuasiveness. Moreover, thanks to the theory-driven approach adopted, the researcher managed to 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 problem of identifying the messages generated by the trolls during the 2016 USA president elections, the authors referred to the Interpersonal Deception Theory (IDT) (Buller and Burgoon, 1996) to identify 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 software designed to connect the extracted linguistic features of a text to 80 different psychological categories (e.g. anger, anxiety etc.). 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 trolls. Using these tools, thus, the utilization of persuasive language cues was quantified by analyzing tweets from suspected political trolls 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 s/he 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 having some limitations connected to the extent of understanding of 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 is concrete and information are 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 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.

***Lexicon Inducted Persuasive Features*** We have described how social science theories can inform the building and the use of computational tools. Nevertheless, this relation can also be structured the other way around. A branch of research in persuasion, in fact, adopted an approach based more on "real-world-related" social interactions. This approach is called *lexicon induction* (Hamilton et al., 2016a; 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 follow. 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 were employed.

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. Subsequently 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. They compared the performance of the proposed algorithm in detecting words correlated with successful outcomes with other traditional learning methods. The results showed a general trend: "deep learning approaches" outperform the "traditional 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 to the chosen dataset: with critical pitfalls for the generalizability of the the results and the emergence of possible biases characteristic of the dataset.

Another 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 distinguish this study from Tan et al. (2016), is that the authors didn't 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 run 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 surfaced-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.

This inductive approach offers different upsides. It grounds the produced knowledge in real-world situations, providing data that are directly connected to the everyday experience of people. Moreover, identifying features that are indicative of a certain outcome and decorrelating them with confounds, promotes a better understanding and interpretability of machine learning models in NLP. On this regards, 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 ground. Finally, we highlight some criticalities that it is possible to anticipate. This approach, strongly relying on the dataset features to define what persuasion is, can be subject of some biases. (A) The persuasive linguistic features for a topic or a certain group of people, could be not effectively persuasive if applied in a different context or theme. Another problem is related to the platform used for the dataset building. The community of the Reddit platform `\r\Changemyview`, is composed by a set of people who start premising an openness to changing one's point of view: vitiating 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 the introduction, see also Druckman (2022).

### 6.3.2 Persuasion 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 San Martino et al. (2019). It elaborated an algorithm to perform a fine-grained analysis of propaganda[1] 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 recur 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. After annotating the corpus, they fine-tuned a BERT-based model 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 (San Martino et al., 2020). Finally, software has been created (Prta – Propaganda Persuasion Techniques Analyzer) (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.

Starting from the work in San Martino et al. (2019), Vorakitphan et al. (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 as-

[1] We included this study on *propaganda* since, despite the different term, the authors defined it as a persuasive effort, aimed at "*influencing people's mindset with the purpose of advancing a specific agenda*" (San Martino et al., 2019). This definition, in turn, is in line with the one described in the introduction.

sumed 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 SOTA models for propaganda detection (retrieved from San Martino et al. (2019); Yoosuf and Yang (2019); Jurkiewicz et al. (2020)) and observed that the implementation of the features generated a substantial improvement in accuracy ($+0.10$ of F1). This study exemplifies how, to tackle the complex task of detecting propaganda techniques in texts, the argumentative approach can be combined with the linguistic one to improve the performance of the algorithm.

Given the promising results obtained through this multi-methods approach, this last 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 did not employ argumentative information. Interestingly, contrary to the results obtained by Vorakitphan et al. (2021) (see above), these contextual information helped in substantially improving the performance of the model, which reached an average $F1 = 0.84$ ($\approx +0.2$). Finally, we highlight how, as the database consists of debates collected from many different politicians in a 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 over the time.

A similar work to San Martino et al. (2019), is Jin et al. (2022). Starting from the assumption that 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 then serves 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 (determined through grid search on the development dataset), the words were classified as similar. This "masked data" were then used to train a deep learning model aimed at detecting 13 different persuasive techniques. 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, they 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 chat generated) tagging the posts containing "ad hominem" attacks. The dataset was used to fine-tuned a deep learning model (BERT based) model 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 hominems" attacks for discussions around marginalized communities. Moreover, they observed that different quantities of "ad hominems" in the training data can influence the likelihood of generating "ad hominems" in 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 applications 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.

Starting from the problem of persuasion theoretical fragmentation, Pauli et al. (2022) 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.

## 6.4 Trends in the Analysis of Political Persuasion

In this section, we discuss 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.

We observe that there has been over the years a decrease in studies conducted with reference to "traditional" learning models, i.e. models characterized by easily explainable and interpretable

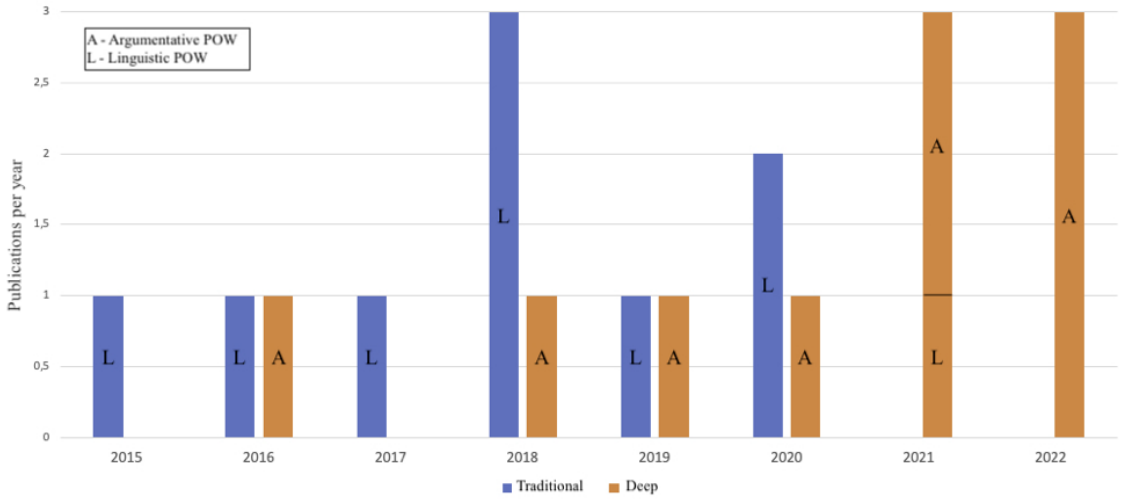


Figure 1: Studies on automated persuasion over time based on traditional (blue) or deep learning (orange) approaches

learning processes, in favor of deep learning models. These latest methods, in turn, can handle a higher level of complexity by addressing intricate tasks with high-dimensional data, automatically learning from raw data, and possessing significant scalability potential, thereby facilitating generalization. Nevertheless, precisely for this increased complex nature, these algorithms often rely on non-linear and non-intuitive interactions between their features, decreasing this way the explainability of how each input contribute and affect the final prediction. The lack of a consistent understanding of the functioning of the algorithms, in turn, could generate critical pitfalls. Imagining of wanting to use a computational persuasive system to, for example, detect propagandist contents to eliminate or to create personalized messages to increase citizens engagement in political processes, the stakeholder would have to face the following issues:

- *accountability*: when the model makes a decision (which account to ban or which persuasive message to write) it might be challenging to trace back the rationale behind that decision.
- *fairness*: models may inadvertently learn biases present in the training data, perpetuating them or even amplifying them, generating unfair or discriminatory outcomes.
- *trust*: as consequence from the two previous points, stakeholders might be hesitant to trust those tools, reducing the virtuous impact they can play on society.

Another evident 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 (see overview in Figure 2). All the research in Section 6.2, in fact, have been conducted resorting to deep learning approaches. Given the complexity of understanding the argumentative processes underlying persuasion, this change can be definitely linked to the higher availability of

| | | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|---|---|---|
| Linguitic | Traditional | Ahmad & Laroche | Tan et al. | Khazaei et al. | Dubremetz & Nivre; Troiano et al.; Pryzant et al. | | Kong et al.; Al Khatib et al. | | |
| | Deep | | | | Pryzant et al. | Addawood et al. | | Vorakitphan et al. | |
| Argumentative | Traditional | | | | | | | | |
| | Deep | | Habernal & Gurevych | | | Da San Martino et al. | Kong et al. | Vorakitphan et al.; Sheng et al. | Goffredo Jin et al. Pauli et al. |

Figure 2: Studies published each year since 2015 based on linguistic or argumentative approaches

deep learning systems. This, on one side, represents an important step forward for the computational study of persuasion, since it allows researcher to employ also all the theories elaborated in the fields of argumentation or rhetoric. At the same time, however, it is still exposed to the risks outlined above related to deep learning systems.

Finally, we highlight how $90\%$ 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 founding this construct and the methodological possibilities generated by the technological development. On this regard, it is possible to observe how, although argumentative theories on persuasion exist 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, thus, emerges the necessity to adopt more interpretable systems. With respect to this last aspect, we stress how, from the material collected and analyzed, a low (if not absent) use of hybrid methodologies has been observed. 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. therefore contributing to more responsible and effective computational persuasion systems.

## 6.5 Challenges and Future Research Directions

Research that merges studies on persuasion and NLP methods holds significant promise for various applications. On one side, for example, we showed how persuasion can be used for propagandists means and how detecting persuasive devices can be used to reduce the online diffusion of these contents. On the other side, it is necessary to stress how persuasion can also be used for more “virtuous aims”, such as fostering civic engagement in political matters or encouraging philanthropic contributions to charitable causes (Wang et al., 2019).

We also discussed the limitations impacting this field. In the introduction we mentioned the limits imposed by the existing theoretical fragmentation in this field. Indeed, the disparate nature of theories and approaches in persuasion studies can pose a challenge in achieving a unified and

comprehensive understanding. Therefore, future endeavors in this area should strive to bridge these theoretical gaps and establish a more cohesive framework for leveraging computational methods for persuasive communication. In addition, we have discussed the technical challenges posed by the latest ML approaches, related especially to the explainability and the generalizability of the built methods. Hybrid approaches could definitely play an important role in dealing with these issues. More precisely, hybrid approaches, by integrating deep learning models with rule-based methods, could:

- Contribute to enhancing the transparency of the model thanks to their rule-based models, allowing users to understand how specific decisions are influenced by predefined rules. This means that features derived from rule-based analyses can be more easily interpreted and correlated with persuasive outcomes.
- Encode domain-specific knowledge and expert insights to improve the system performance in handling edge cases. This capability is particularly important in addressing nuanced or context-specific aspects of persuasion, as well as all the cases that may not be well-represented in the training data.
- Enhance trust and acceptance of the algorithms thanks to this increased transparency about their functioning.

Overall, there are a number of future challenges that the computational study of persuasion still has to address. Despite their potential to increase the explainability of the models, employing hybrid approaches requires dealing with, among the others: technical challenges (such as ensuring an effective communications 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, impacts, respectively, on the interpretability 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; at times, images can convey more potent messages than text. 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, i.e. **multi-modal 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. On this regard we signal the work of Dimitrov et al. (2021) 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[2].

Most of the current detectors are assessed solely on a single annotated dataset and, as we outlined before, using in the most cases 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 multi-lingual scenarios**. In the future, it is recommended to allocate more resources to the creation of multi-lingual annotated datasets. In the context of

[2] https://propaganda.math.unipd.it/semeval2024task4/

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 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. (2018b) provided insights into transferring the style of a language model to a specific target domain. With these foundational elements in position, there is a high likelihood that automatically generated propaganda will emerge in the near future. For this reason there is an increasing impetus to **address both the textual and network dimensions of propaganda detection** concurrently, recognizing that relying solely on a single paradigm is likely to lead to inadequacies. Consider, for instance, the use of a pre-trained language model like GPT-4 as an automated tool for generating propaganda. In such cases, emphasizing linguistic features alone for propaganda detection may prove ineffective, as the generation of propaganda may outpace its detection significantly. Consequently, in the future, it will be imperative to expand the scope of analysis beyond textual content and delve into the examination of network nodes and connectivity patterns that facilitate the dissemination of propaganda.

# 7 Detection of Harmful Content in Political Discourse

One of the biggest problems modern democracies face is the lightning-fast spread of undesired content, such as fake news and hate speech on the Internet. This has sparked a wealth of research on the detection of these types of harmful content. This chapter aims to give an overview of the research going on in both and discuss its opportunities and shortcomings.

*Fake news* (FN) is a controversial term (Colomina et al., 2021) that refers to a very broad category of false and misleading information. It includes misinformation (all types of false information, from honest reporter mistakes to propaganda), disinformation (fake information that aims to do harm) and malinformation (true information that is framed to make it support a false and/or harmful narrative). An author can lead the reader to a wrong conclusion without bluntly lying by misrepresenting true facts (Watts et al., 2021). However, most research on text mining for fake news detection takes a rather broad approach, defining FN as any false information that resembles news content (Zhang et al., 2018).

Shu et al. (2019b) state that 'detection of fake news is a difficult task as it is intentionally written to falsify information.' (p. 62). However, this is certainly not the case for all fake news; in fact, it seems that a lot of fake news is created in order to resemble *viral content*, rather than *true content*, as much as possible (Hughes and Waismel-Manor, 2021). It is closely related to the concept of *junk news*, defined by Venturini (2019) as provoking information that aims to grab attention, either for monetary goals or for the obfuscation of public/political discourse. There is no

clear line between FN meant to reach some political goal and commercial FN that is just meant to gather clicks (Burger et al., 2019).

Hate speech (HS) is defined as abusive speech that incites hate against a group based on a certain protected characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, or religion (Aluru et al., 2020). Many definitions include that the target should be a disadvantaged group (Waseem and Hovy, 2016; Davidson et al., 2017). However, not all research on hate speech takes this sociological context into account. Some do consider the protected characteristics mentioned above, but do not differentiate between disadvantaged and privileged social groups. Others confuse HS with cyberbullying or other types of abusive language that target individuals or groups based on not-protected characteristics (MacAvaney et al., 2019; Zhang and Luo, 2019; Bohra et al., 2018). Other papers, instead, do not give a definition of HS at all (Poletto et al., 2021). But even with an exact definition, it is not always clear-cut to distinguish between HS and merely offensive language. HS detection is a difficult task even for humans. This is illustrated by inter-annotator agreement for HS datasets, that is typically low (Fortuna et al., 2020).

There is a close relation between HS and FN. Both capitalize on emotions and lingering feelings of dissatisfaction in a population, 'mobilizing feelings of mass anxiety' (Cheema et al., 2019). FN is often weaponized against scapegoated social groups (see for example Liebowitz (2019)). Not all FN is hateful, but it does often evoke emotions to gain more clicks (Ajao et al., 2019). Hate speech is also emotional. It is thus not surprising that research on both FN and HS has used sentiment analysis for their detection (for an overview, see Alonso et al. (2021)).

Some research on HS detection uses data on the intersection of HS and FN by studying the automatic detection of hateful news. Berk and Filatova (2019) and Hüsünbeyi et al. (2022) focused their research on *incendiary news*: a subtle form of hate speech that evokes ethnic hatred, but does not contain explicit slurs or threats. Bourgeade et al. (2023) analyzed racial hoaxes in Twitter threads from Spain, Italy, and France, showing how fake stories are used to spread hate against immigrants. However, the bodies of work on FN and HS are clearly two separate worlds, with authors rarely citing papers on the other topic.

## 7.1 Fake news detection

FN detection is often framed in the NLP research community as a text classification task. Three types of features are commonly used: message content, social/network propagation features, and user-based features (Zhang and Ghorbani, 2020). Some studies focus on domain-general fake news detection, whereas others only take a specific domain, most commonly Covid (Lee and Kim, 2022; Du et al., 2021; Ghayoomi and Mousavian, 2022). This section discusses challenges and strategies in the FN detection field.

### 7.1.1 Content-based approaches

There is an extensive body of research on the use of traditional ML techniques for FN detection (Della Vedova et al., 2018; Helmstetter and Paulheim, 2018; Ozbay and Alatas, 2020; Khanam et al., 2021), but recent studies find that deep learning approach outperform traditional methods (Kaliyar et al., 2021; Sahoo and Gupta, 2021). CNNs are the most commonly used architecture in DL approaches (Thompson et al., 2022; Goldani et al., 2021; Samadi et al., 2021), with good

results (Kaliyar et al., 2020; Saleh et al., 2021), although RNNs such as LSTM's (Chauhan and Palivela, 2021) and biLSTM's (Bahad et al., 2019; Aslam et al., 2021) have also been succesfully used. Ensemble approaches also work well, such as CNN+biLSTM (Kumar et al., 2020) or voting classifiers (Ahmad et al., 2020). A hybrid approach by Nasir et al. (2021) used the output of a CNN layer as the input for an RNN layer, allowing their model to learn both scenic features (CNN) and sequential features (RNN).

Nowadays, transformers are quickly winning ground also in the field of fake news detection. Kaliyar et al. (2021) used transformer-based sentence embeddings to represent fake news articles from the Kaggle dataset. These are then fed to several blocks of CNNs for classification. A similar approach was implemented by Raza and Ding (2022), who used both news content and social content. Verma et al. (2021) used word embeddings to represent manually crafted linguistic features, feeding them to different classifiers. They got good results over different datasets. Tariq et al. (2022) used adversarial training for binary fake news detection. Fine-tuning models for fake news detection using clean and adversarial examples simultaneously (made by perturbing the word embedding matrix) was found to improve performance.

Other studies modeled textual content using graphs. Vaibhav et al. (2019) represented articles as fully connected graphs where the embedding of each sentence was represented as a node in a graph. The classification was performed on the graph as a whole. Their model overfitted, but still outperformed the other models they tested. Karnyoto et al. (2022) used a similar approach, representing tweets as graphs, and improve results with an elementary data augmentation strategy (random swap/deletion/insertion, and synonym replacement). Song et al. (2021b) propose temporal graph NNs for FN detection. They built their network over time, using a temporal memory module. A drawback of this approach is that it is very computationally heavy; an advantage is that it is suitable for real-time FN detection. Shahsavari et al. (2020) aimed to analyze conspiracy theories and monitor upcoming ones by automatically creating clusters in narrative graphs based on co-occurrences of named entities.

Various studies deal with the use of sentiment analysis for FN detection (Alonso et al., 2021). Ajao et al. (2019) found that emotionality features improved the performance of their SVM classifier on FN on Twitter, indicating that FN are more emotional than truthful news. Similar findings are reported by Bhutani et al. (2019). Choudhary and Arora (2021) extracted sentiment, syntactic, and readability features from fake news articles and fed those to a neural model. Their model performed on-par with embedding based neural approaches, but was significantly faster and lighter. It is not surprising that FN is characterized by emotionality. Bakir and McStay (2018) connected the way FN spreads to the concept *economics of emotion*: emotions generate attention and viewing time, which translates to advertising income. Social media thus favour affective content, which not only spreads faster, it is also designed to create a more inflammable social atmosphere, posing a direct threat to the democratic public sphere.

Other studies take a multimodal approach to FN detection. Text and image can either be projected in the same vector space (Yang et al., 2018a), or in different vector spaces and consequently concatenated (Singhal et al., 2019; Qian et al., 2021). Song et al. (2021a) used a transformer architecture with a cross-modal attention residual layer in order to capture the complex interactions between text and image.

Several papers approach FN detection as a stance detection task. On the one hand, headlines

or article claims can be matched to a pre-fact-checked database of claims. Other studies attempt to determine the relative position of an article to its headline, where a mismatch is considered an indication that an article is fake. This is based on research by Dong et al. (2019), who found that an article that contradicts its headline is a strong indicator of clickbait.

Many authors used data from the 2017 FNC-1 task, see for example Hanselowski et al. (2018). Thota et al. (2018) represented headline and article with concatenated TF-IDF vectors and fed them to a CNN. Borges et al. (2019) concatenated the relative stance of the headline with representations of the article and headline. Umer et al. (2020) and Aljrees et al. (2023) found that using a feature reduction algorithm (PCA) boosts both training speed and performance in a neural architecture combining CNN and LSTM. Slovikovskaya (2019) and Kasnesis et al. (2021) used pre-trained Transformer models for the same task and data, outperforming earlier models. Karande et al. (2021) found that including stance detection features (i.e. the cosine similarity of embeddings; article vs. article headline) as a feature marginally improved model performance.

Other papers detect FN by fact-checking papers against an external knowledge base (KB). Hu et al. (2021) classifies news as fake or real through a GNN that links entities with their Wikipedia pages and compares the claims made about that entity in the article to the claims made on Wikipedia. A more abstract approach was taken by Whitehouse et al. (2022), who use fine-tuned LLMs that encode real-world knowledge. These were found to outperform 'plain' LLMs, but not consistently and not by a large extent. Seddari et al. (2022) combines linguistic-content based features with features that estimate the reliability of the source (reputation, as assessed by journalists), and the news (coverage and whether it figures in a fact-check database).

Glockner et al. (2022) criticizes fact checking for FN detection. They posit that the way human fact checkers use sources for fact checking is fundamentally different from fact checking NLP. Moreover, most evidence that datasets for fact-checking include is either leaked (from fact-checking sites) or insufficient. These studies therefore lack actual power or usefulness in the real world. A good alternative would be the detection of previously fact-checked claims, as suggested by Shaar et al. (2020). Such a tool would be fast, explainable, and a reliable resource for journalists to fact check politicians real time.

Various studies deal with the generalizability of FN detection algorithms across domains and topics. A model that generalizes well over events and topics will also be better at the detection of newly emerging fake narratives. However, being able to use event-specific features will help classification later on, when more data is available about an event or topic. Wang et al. (2018) address this issue by introducing a multimodal CNN with a separate event discriminator module. Nan et al. (2021) introduce a domain gate to aggregate multiple representations; this is like an attention mechanism based on a domain embedding that determines which experts (i.e. decision networks) will be employed, making their model better at cross-topic FN detection. Yuan et al. (2021) implement a model that jointly learns to detect domain-dependent and domain-independent features in fake news (through adversarial learning). It then models the news through a graph-attention-based classifier that uses interactions between both users and words/topics.

There is also a large body of work on FN detection across languages. Abonizio et al. (2020) and Faustini and Covoes (2020) both tried to extract language-independent features (text statistics) and use them for text classification. Training a model on multilingual word or sentence

embeddings is a popular approach. Ghayoomi and Mousavian (2022) used transfer learning with multilingual word embeddings to leverage English training data for the classification of Persian tweets regarding Covid-19. Mohawesh et al. (2023) used multilingual embeddings to place news items in a semantic graph. Other papers use translated data (Nassif et al., 2022; Schütz et al., 2022). Hammouchi and Ghogho (2022), for instance, used translated data for a cross-verification using the Google search engine.

### 7.1.2 Propagation- and user based approaches

The use of social/user/network propagation features is very popular in the field of FN detection. Reis et al. (2019) conducted experiments with various traditional ML techniques and found that social and user-based features outperform content-based features. Another advantage is that they are language-independent and more stable over time than content-based approaches (Monti et al., 2019).

Many papers combine social features with content features. Shu et al. (2017) looked at features of the relationships between the news, the news user/reader, and the publisher of the news. An algorithm based on the same idea was implemented by Zhang et al. (2018) and Zhang et al. (2020). They used a deep diffusive network model to model the relationships between tweets, users and spreaders.

Tacchini et al. (2017) aimed to find hoaxes on Facebook by using users as features. They compared a logistic regression (where each user is a feature that gets assigned a weight) to a harmonic Boolean Label Crowdsourcing (hBLC) approach (where the Facebook users interacting with the post were modeled as the 'crowd' and a 'like' is considered a 'vote'). Their approach was further elaborated upon by Della Vedova et al. (2018), who combined social features with content features order to make their approach more robust in real-life scenarios where social features might not always be available. They found that a simple approach using either the content-based or the social-based classifier based on the number of likes a post had received outperformed other more sophisticated methods they implemented.

Liu and Wu (2018) classified the *propagation path* of a piece of information, represented as a multivariate time series using user features. This time series was then classified using RNN and CNN layers in order to effectively capture both local and global information. This idea was further developed in Liu and Wu (2020), who incorporated a *status-sensitive crowd response feature extractor*: a feature extracting algorithm that uses both user responses (in the form of text messages written in response to a piece of news) and user statuses (a typology of the user at that given point in time). These approaches allow for real-time detection of FN.

Monti et al. (2019) used *geometric deep learning*, an approach that models CNNs as graphs. This allows for the processing of very heterogeneous data, which makes it easy to incorporate different types of features. Another graph-based approach is implemented by Dou et al. (2021), who used joint graph and content modeling. They used a graph NN to generate a user engagement embedding for each piece of news, which was concatenated with a text embedding; this was classified by a neural classifier, outperforming other neural models. Ni et al. (2021) perform a similar study, but they represent social propagation to a graph NN; this graph structure is concatenated with the text embedding and classified through a FC layer.

### 7.1.3 Datasets for FN detection

Since FN detection is often framed as a supervised binary classification task, it depends on the availability of labeled datasets (D'Ulizia et al., 2021). The datasets available for the task include a large variety of types of data such as claims (Politifact) (Vlachos and Riedel, 2014), entire articles (FakeNewsNet) (Shu et al., 2020), Facebook posts (Buzzface) (Santia and Williams, 2018) or tweets (Credbank and PHEME) (Mitra and Gilbert, 2015; Zubiaga et al., 2016). However, the manual labeling of such datasets is very laborious. The availability of high-quality, up-to-date, diverse test and train data is a significant problem for the field of FN detection.

Many datasets use the trustworthiness of the spreader of the news as a proxy for truthfulness and vice versa (Tacchini et al., 2017; Garcia et al., 2022). This might not necessarily be a problem. Helmstetter and Paulheim (2018) trained a classifier on tweets by users that were labeled by journalist fact-checkers (as trustworthy/untrustworthy) and then used this classifier to classify tweets that were individually labeled as true or fake. They still reached an accuracy of 0.9, indicating that source-based datasets can still be a stand-in for datasets in which each instance is manually labeled. A possible explanation is that the majority of FN in commonly used datasets *is*, in fact, clickbaity, so a FN classifier trained on clickbait will reach a high accuracy/F1 on most datasets.

The situation is thus as follows: most datasets (and all large ones) contrast news from reputable journalists with low-quality clickbait articles from questionable websites. This begs the question *what it is* these models are classifying. Although all papers on FN detection engage in some political and sociological discourse on the spread, types and danger of FN in their introduction, very few papers actually come back to this theory when creating or choosing their datasets or when evaluating their models. Error analysis is often lacking, making it unclear what types of fake news and clickbait can or cannot be detected (Miró-Llinares and Aguerri, 2023). Differentiating between more subtle forms of fake news could be a promising avenue for future research.

One way to deal with the data problem is by exploring alternative ways of training models. Yang et al. (2019) proposed unsupervised learning for FN detection. They modeled the truthfulness of news and users' credibility based on the way verified and non-verified Twitter users interact with news content. Liu and Wu (2020) successfully explored a positive/unlabeled learning approach, in which a neural network is fed with positive and unlabeled examples of FN content on Twitter and Weibo. This allowed them to train a model using relatively few learning examples. Raza and Ding (2022) proposed a weak supervision method: they used source level labels (trustworthy/untrustworthy news spreader) and combine them with 'weak', article-level labels based on the response of the public. Samadi et al. (2021) proposed the use of a Gaussian noise layer (that propagates noise throughout the network). Not only does this work as a way of data augmentation, it also helps prevent overfitting. Other approaches (Silva et al., 2021) tried to minimize annotation costs by having a model identify instances that should be labeled by humans, which were then used to further train the model. Tschiatschek et al. (2018) created an active learning algorithm were user flags are utilized in order to decide what news to send to experts to be checked. They learn user flagging behaviour over time, making their approach robust to bad faith flaggers.

## 7.2 Hate speech detection

Like FN detection, HS recognition is often framed as a text classification task. This classification can be binary, i.e. hate/non-hate (Gitari et al., 2015) or multi-class for different targets (for example sexism, racism, homophobia etc.) (Waseem and Hovy, 2016; Zhang and Luo, 2019). ElSherief et al. (2018) differentiate between generalized and directed HS. Other popular datasets make a three-way distinction between hate speech, offensive speech, and normal speech (Davidson et al., 2017; Watanabe et al., 2018). Del Vigna et al. (2017) distinguishes between no hate, weak hate, and strong hate.

Gröndahl et al. (2018) tested various binary (HS vs. no HS) models on offensive data (that was not HS) and found that they all classified offensive language as HS. Generally, it is easy for models to distinguish between 'some kind of explicit toxicity' and 'no toxicity', but scores drop when more fine-grained distinctions are introduced. Kapil and Ekbal (2020) tried to leverage the variety of existing subtasks and definitions of HS by implementing a multitask learning strategy. This allowed them to use features and strategies of multiple related tasks (sexism detection, offensive language identification etc) while also having very specific classifiers.

Early research mainly focused on traditional ML methods, such as SVM (Del Vigna et al., 2017; Malmasi and Zampieri, 2017; Vidgen and Yasseri, 2020), decision trees (Watanabe et al., 2018), or logistic regression (Waseem and Hovy, 2016). Gitari et al. (2015) started from sentiment analysis and subjectivity detection to locate checkworthy utterances, which were then classified with a rule-based classifier. Later research started implementing RNNs such as LSTMS (Pitsilis et al., 2018; Corazza et al., 2020; Pereira-Kohatsu et al., 2019). Corazza et al. (2020) found that biLSTMS were not very useful, possibly due to the short length of tweets, the most common data type in HS detection; on the other hand, a systematic comparison of models by Toktarova et al. (2023) found that biLSTM outperformed a range of other ML and DL techniques.

Around 2017, the SOTA for HS detection were SVM and biLSTM (Bosco et al., 2018). An important early paper using DL for HS detection was written by Badjatiya et al. (2017). They used LSTM to classify the dataset of Waseem and Hovy (2016), reaching an impressive F1 of over .9, at a time when most other approaches had F1 scores of around .7 (Del Vigna et al., 2017). However, Arango et al. (2019) found that their approach leaked data between test and training set, and that their results were not generalizable to other datasets.

CNNs are a popular architecture for HS detection. Zhang and Luo (2019) experimented with a CNN that learned to skip words in the classification, in order to capture relations between words that are further away in the text. Their proposed model marginally but consistently outperformed a traditional approached based on word embeddings + SVM and approaches with multiple concatenated CNNs. Zimmerman et al. (2018) used an ensemble model of multiple CNNs with different parameter settings/weight initializations. They averaged the softmax score over models and took this as their final prediction. Mozafari et al. (2020) were among the first to combine BERT embeddings with a CNN. This was the beginning of a turn towards transformers in the HS detection field (like most NLP fields), with many other studies using transformers for HS detection (for example Beyhan et al. (2022); Pérez et al. (2023); Alatawi et al. (2021); Lemmens et al. (2021)). Frenda et al. (2022) focused on learning implicit abuse/HS experimenting with transformers. They used multi-task learning, and explicit linguistic features to reveal the elements involved in the implicit

and explicit manifestations of abuse. They found that HS detection is enhanced when the classifier is also taught to detect stereotypes as a separate task. Koufakou et al. (2020) enriched a BERT model by adding lexical features. They appended them to BERT embeddings and fed this concatenation to a dense classification layer. They experimented with adding encodings (of the category of the offensive words in the comments) or embeddings (of the comment with its specific offensive words). They found that the embedding-enriched model performed better than the encoding-based model, but both outperformed the baseline.

HS detection algorithms seem to benefit from models/embeddings that were task-agnostically pretrained on corpora of abusive language. Possibly, this is due to the large amount of misspellings and non-standard words typically present in hateful SM data (Alatawi et al., 2021). For example, Vidgen and Yasseri (2020) found that the embeddings they trained themselves outperformed general pre-trained embeddings. This is in line with findings by Corazza et al. (2020), who recommend the use of domain/platform-specific embeddings over general embeddings were possible. Alatawi et al. (2021) also found that an LSTM with domain-specific embeddings outperforms LSTM based models with domain-general embeddings (although BERT outperforms them all). They noted that domain-specific embeddings are better at handling misspellings and domain-specific words/meanings of words. Glavaš et al. (2020) added additional MLM on a corpus of abusive language as a pretraining step for cross-lingual HS detection. They also found that this improves results.

Finally, various papers deal with multilingal HS detection. Aluru et al. (2020) conducted an analysis of HS detection in 9 languages and found that the best method depends on the availability of resources for a language: BERT-based models were the best, but only when there was a lot of data available. Low resource settings were better served using multilingual LASER embeddings with LR. The automatic translation of data also worked well. Plaza-del Arco et al. (2021) compared various pre-trained transformers for HS detection in Spanish found that a monolingual Spanish model outperformed both multilingual models, probably due to vocabulary size. Corazza et al. (2020) aimed to build a HS detection model that produces robust results for three languages (Italian, English, German). Their best performing methods are LSTM and fasttext embeddings (English), LSTM and tweet-based embeddings (Italian), and a GRU network with fasttext embeddings (German). The usefulness of features also differed between languages. Glavaš et al. (2020) disentangled the effect of domain shift and language shift. They created a corpus from three English datasets from different sources (fox news news comments, social media, and Wikipedia) and manually translated those into 5 target languages. They then performed various experiments using mono- and multilingual BERT and RoBERTa. They found that adding training data from a different domain worsens performance in cross-language transfer learning for HS detection compared to cross-lingual in-domain HS detection. Errors in cross-lingual transfer learning often stem from idioms/compounds that are lost in translation or that require extensive world knowledge. De la Peña Sarracén and Rosso (2022) represented texts containing HS in a graph using Graph Auto-Encoders, which allowed them to make multilingual embeddings in an unsupervised way; these were then used for HS classification using transformer and convolutional layers.

### 7.2.1 Use of context for HS detection

A lot of HS cannot be recognized as such without the use of additional content, such as an image (Perifanos and Goutsos, 2021), earlier messages (Pavlopoulos et al., 2020), or the profile of the author (Mosca et al., 2021) or target of the message (Dinakar et al., 2012). There is plenty of research that uses context outside of the text for HS detection, most notably user profiling. Generally speaking, these approaches tend to use them to enrich textual representation/additional features, whereas for FN detection they are often the main focus of the research. An exception is Mathew et al. (2019). They tried to detect hateful users of Gab by classifying propagation path, user features, and network features.

Some authors argue for the incorporation of outside world knowledge in the model (Yin and Zubiaga, 2021). An early example is Dinakar et al. (2012), who used a knowledge base of stereotypes to detect HS against the LGBTQ community. Lemmens et al. (2021) annotated hateful metaphors in a corpus of Dutch Facebook comments to enrich the text with explicit/manual information before passing it to an SVM or BERT transformer, which was found to improve F1 for the detection 1) type of HS and 2) target of HS. Gao and Huang (2017) created a context-aware corpus of hate speech in comments from Fox News and studied how the incorporation of context (news article headline + username of the comment author) in a HS classification model could improve classifier performance. They tested a logistic regression model and a bi-LSTM. They found that the incorporation of context features significantly improved F1 on both tasks; however, it stayed low for both models (.54 and .55 respectively). Pavlopoulos et al. (2020) focused on how context affected human judgements of toxicity, and how the addition of context could improve the performance of HS detection algorithms. They found that the perceived toxicity of a Wikipedia post changes in around 5.2% of cases when human annotators are provided with context, but they surprisingly found no difference in performance between context-aware and context-unaware classifiers.

Recent findings by Pérez et al. (2023) affirm the opposite. These authors found that adding the parent tweet as context for HS classification on Twitter improved a simple BERT model for sequence classification. This held true for both binary and more fine-grained (racism, sexism, homophobia etc.) HS classification. Especially HS targeting LGTBQI people benefited from added context. Similarly, Markov and Daelemans (2022) focused on the role of context when detecting the *target* of HS. They found that adding relevant contextual information (i.e. relevant comments) to a transformer-based classification of HS helps significantly in determining the target of the HS for Dutch Facebook comments.

Another way of incorporating context, other than looking at parent comments/textual features of surrounding comments/news articles, is by profiling the author of the comment. Mosca et al. (2021) used the social network of each user (i.e. how they are connected to other users) as a feature for hate speech recognition in two different Twitter corpora using a multilayer perceptron. Classifier performance improved for all classes (racism, sexism, and neither) when social features were included. Mishra et al. (2019) created graphs with both tweets and users as nodes and used a graph CNN to create author profiles. A LR classifier trained on these profiles outperformed their baselines. Yu et al. (2022) attempted to identify whether a Twitter user spreads sarcasm (an often-used strategy to package HS as more socially acceptable 'dark humour'). Their best model was

a soft voting ensemble approach based on BERTweet models with different loss functions and a feature-based CNN.

### 7.2.2 Datasets for HS detection

A problem in the field of HS detection is the real-world validity of many approaches. Arango et al. (2019) found that many models overfit considerably. Model performance depends more on dataset than on model architecture, and models do not perform well across datasets. Gröndahl et al. (2018) tried various models that were promoted as SOTA and found that they all performed more or less equally well when retrained on different datasets (see also MacAvaney et al. (2019)). Zhang and Luo (2019) did a meta-analysis of HS detection models and found that 'our experiments could not identify a best performing candidate among the three state of the art methods on all datasets, by all measures'.

Datasets use different labels/categorizations, making it difficult to compare systems across datasets Zhang and Luo (2019); Yin and Zubiaga (2021). Annotation protocols can slightly differ or even contradict each other. For example, some consider the presence of profanities an indicator for abuse or HS (Sigurbergsson and Derczynski, 2020), whereas others explicitly mention that they don't (Davidson et al., 2017; Gröndahl et al., 2018). There is also a difference in sample methods, meaning that the data and distribution of the content in the datasets can be rather different across datasets (Fortuna et al., 2020). One of the most used datasets by Waseem and Hovy (2016) has a very strong user bias: %65 of all racist and sexist tweets in this corpus were produced by only two users. It is thus not a surprise that the performance of models trained on this dataset drops quite dramatically when tested in a real-life setting, or even just on a different dataset Arango et al. (2019). However, this bias is not self evident from the dataset description.

The majority of HS datasets come from Twitter (Gröndahl et al., 2018; Poletto et al., 2021). This is a problem because many of these datasets are now no longer available for research. Moreover, most Twitter users have not given informed consent to their data being used, and there is no dialogue possible between users and the researchers studying their posts (Matamoros-Fernández and Farkas, 2021). There are exceptions such as Mathew et al. (2019), who collected a dataset of posts on Gab, a social network known for its libertarian approach to free speech that has made it a free haven of hate and bigotry. Gao and Huang (2017) collected user comments from the Fox News website. Moreover, although most datasets are in English, there are many datasets available in other languages; an overview is maintained by Vidgen and Derczynski (2020) at www.hatespeechdata.com.

Human annotators also make lots of mistakes and inter-annotator agreement is typically low. Many papers mention annotation mistakes as a source of wrong predictions in the error analysis (Aluru et al., 2020; Corazza et al., 2020; Plaza-del Arco et al., 2021). Davidson et al. (2017) mention that human coders are bad at identifying misogyny, considering hate speech against women merely offensive. Moreover, most annotators and researchers are not actually part of the minorities affected by the HS they research (Yin and Zubiaga, 2021), but few authors are self aware of the bias this might bring along (Matamoros-Fernández and Farkas, 2021). Sap et al. (2019) found that annotators of HS datasets are often not familiar with African American English (AAE), so they wil often mislabel tweets in AAE as toxic or abusive, even when this is not the

case. This bias is propagated in models trained on these datasets. A step towards a solution (apart from the obvious - having more experienced and diverse annotators) could be the use of more fine-grained categories, that allow for more precise annotation guidelines and therefore a higher inter-annotator agreement (Assimakopoulos et al., 2020).

Finally, as for FN datasets, HS datasets greatly differ in how carefully and ethically they were sourced. For example, the Gab dataset by Mathew et al. (2019) only tags users as hateful/not hateful, not individual posts, and the way they collected their data has some procedural flaws: their paper has no definition of HS, the annotation is done by students of CS with, as far as we know, no background in social science/HS, and they assume that randomly sampled users are not hateful, which is dubious on a network like Gab.

## 7.3 Challenges and Future Research Direction

The detection of harmful content online, especially in political communication, still poses several challenges. As regards the types of errors made by HS detection systems, it is difficult to assess where they stem from (Corazza et al., 2020): sometimes the classification models wrongly classified examples that seemed to be straightforward, with no clear explanation, but most models show similar error patterns (Aluru et al., 2020).

Models tend to struggle with implicit HS (Yin and Zubiaga, 2021). Some HS can only be correctly classified if one is aware of context, for example the gender of the target (Dinakar et al., 2012). Furthermore, domain generalisation is still a challenge, as well as the detection of some specific hate targets. For example, transphobia is often more difficult than other types of HS, maybe because it is often very subtle. For example, misgendering someone on purpose is very hard to detect; not only does one need to know the gender of the target of the HS, it is also practically impossible to figure out if someone is doing this on purpose or making an honest mistake (Pérez et al., 2023). Another example of implicit HS is hateful language that does not contain slurs, profanities, or outright calls to violence. This is exacerbated by the fact that hateful communities often use seemingly innocent words in a hateful/bigoted way, which is hard for HS detectors to pick up on (Yin and Zubiaga, 2021). Taylor et al. (2017) aim to automatically detect these 'code words'. First, they use a network based approach to detect hateful communities on Twitter. Their posts were collected and used to train new word embeddings, that are used to model the candidacy of words to be a 'code word'. A similar approach is adopted by Magu and Luo (2018). Social media users also deliberately misspell words or post HS as image rather than text to evade HS detection. Vocabulary size can be a large obstacle to harmful content detection due to the presence of many non-standard words and spellings when dealing with social media data. This is a problem, since HS detection models are not very robust to adversarial attacks Gröndahl et al. (2018). This problem is partly mitigated by the use of sentence embeddings, rather than word embeddings (Plaza-del Arco et al., 2021).

Models can classify false positives by relying on neutral words that describe certain groups that are often the target of HS, such as 'gay' or 'woman' (Yin and Zubiaga, 2021). This is problematic, because if using these words triggers censorship, HS detection might actually be hurting communities by censuring their speech, or any non-hateful conversation about them. Another source of false positives are abusive words when used in a non-hateful context (for example

reclaimed slurs or descriptions of racism).

It is not at all straightforward to actually improve the democratic public sphere using FN and HS detection algorithms. The problem is diffused across platforms and perpetrates mainstream media, warranting a large-scale approach (Watts et al., 2021). Bakir and McStay (2018) argue that advertising agencies have an economic incentive to identify pages that spread hateful and fake content, as having ads published on these websites damages a company's brand. Advertising agencies have the data and technical resources to quickly identify undesirable content. Governments should thus collaborate with the advertising world to fight undesirable content by making it less economically attractive.

Another big issue related to the automatic detection of harmful content online concerns possible ethical risks. HS detection in general opens up a complicated discussion about free speech and censorship. Islamophobia detection is an illustrative example, because there is not always a clear boundary between islamophobia and religious criticism. Vidgen and Yasseri (2020) choose not to distinguish between criticism of Islam and criticism of Muslims, as both can have detrimental effects to the social situation of Muslims in the UK. However, the practical use of a classifier that lumps negativity towards Islam as a religion/ideology together with negativity towards Muslims as a group would arguably infringe on individuals' rights to criticize religious institutions.

Having **black-box models that take politically sensitive decisions** based on enormous amounts of uninterpretable parameters is a problem. Many studies explicitly focus on making their model explainable. Shu et al. (2019a,b); Shaar et al. (2020) focus on explainable FN detection by having their model output check-worthy sentences, rather than truthfulness judgments. Szczepański et al. (2021) use LIME and Anchors[3] to gain insight into the decision making process of a BERT-based FN classifier. A similar approach was implemented by Mehta and Passi (2022) for HS detection. Pereira-Kohatsu et al. (2019) modeled HS on social media as a network of concepts and actors (users). It gives the estimated probability that a tweet is hateful, shows semantic maps, neighbours, and allows the user to filter and inspect specific users, topics, and words. Mathew et al. (2021) introduce a dataset aimed specifically at explainable HS detection. It also highlights the *rationales* in the data, i.e. the data that made annotators reach a certain conclusion.

Another problem is implicit **model bias**. Thiago et al. (2021) found that the toxicity detection software 'Perspective API'[4] deems drag queens to be more toxic than white nationalists. One way to deal with it is trying to make the annotation less biased. Sap et al. (2019) found that annotators of HS datasets are often not familiar with African American English (AAE), so they wil often mislabel tweets in AAE as toxic or abusive, even when this is not the case. This bias is propagated in models trained on these datasets. Explicit instruction was found to alleviate this bias. Dixon et al. (2018) aim to reduce bias by balancing their training data by introducing synthetic balancing data. A post hoc approach is adopted by Mozafari et al. (2020), who finetune BERT for HS detection and mitigate its bias by re-weighting the samples.

The fields of HS and FN recognition as a whole also may enforce a biased perspective. Most of the research focused on HS and FN detection focuses on the global North, even though these

[3] Both algorithms use the sampling of perturbations around a prediction in order to explain model behaviour without having to know its internal parameters

[4] https://perspectiveapi.com/

phenomena are by no means a Western/rich nation problem. Researchers and dataset annotators are usually not part of the minorities their HS is targeted against, but this lack of diverse perspectives is not often recognized (Matamoros-Fernández and Farkas, 2021). Especially studies on FN detection are remarkably apolitical and most papers do not seem interested in knowing what kind of errors their models make, what their practical usefulness is, or what kinds of data are in their datasets.

Many papers on HS detection use **user profiling** features for HS classification (Schmidt and Wiegand, 2017; Pitsilis et al., 2018; Mathew et al., 2019). Others, like Mathew et al. (2019); Huang et al. (2020) and Mossie and Wang (2020), even have this as the main objective of their research. There is a tension between HS recognition, that might warrant profiling sensitive personal characteristics, and the inherent problematic of automatic profiling these types of features, that can lead to stigmatization and discrimination (Ploug, 2023). A solution is proposed by Allein et al. (2023), who aimed to leverage user features for the recognition of FN, without actually profiling users in the classification process. They employ user features to guide the latent vector space, but do not use them as actual input during classification.

**Datasets** for HS and FN are not always collected ethically and transparently. For example, Sigurbergsson and Derczynski (2020) created a HS corpus for Danish, but when the Facebook API would not let them collect data, they just manually scraped it. Sahoo and Gupta (2021) crawl a FN corpus from Facebook using publicly available user information from 5,000 users. Their data file automatically updates the moment a user visits another page and they do not address privacy issues at all. An often-used dataset is the Kaggle Fake News dataset (Lifferth, 2018), which was not peer reviewed and does not contain any explanation on how it was collected and almost no metadata. No real-world application for such a politically sensitive task should be based on such data. Other studies, like Aldwairi and Alwahedi (2018), do not even mention which dataset they use. Moreover, many papers feature tweets or other social network posts verbatim. A simple Google search would be enough to find the message, and thereby the user, possibly exposing them to doxing/revealing their privacy (Matamoros-Fernández and Farkas, 2021). More privacy-preserving methods and best practices are therefore needed to fight the propagation of fake news and hateful content in an ethical way.

# 8 Conclusions

In this document, we cover different research topics related to WP2 “Public Discourse Analysis”, focusing in particular to NLP and AI approaches to better understand political communication. The document encompasses the work of five DCs in the doctoral network (DC 1 – 5), i.e. those that are mainly focused on the application of human and social sciences to study different aspects of political discourse on social networks. Indeed, this deliverable complements the content of D2.2. “Technical report on the state of the art of political communication in social networks”.

We cover six topics of major interest that have been addressed with NLP and AI methods in the domain of political discourse analysis: ontologies and knowledge modelling (Section 2), corpus linguistics and textometry (Section 3), diachronic language analysis (Section 4), computational argumentation (Section 5), persuasion assessment (Section 6) and the detection of toxic

content (Section 7). This document represents the foundations upon which future activities by the HYBRIDS PhD fellows will be based, leading not only to the development of novel methodologies for political discourse analysis (D2.3) but also to the release of text corpora and annotated datasets (D2.4) as well as of prototypes and tools to process them (D2.5).

# References

Abonizio, H. Q., de Morais, J. I., Tavares, G. M., and Barbon Junior, S. (2020). Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5):87.

Abric, J., Bardin, L., Barreiro, A., Berti, C., Molinari, L., Speltini, G., Campbell, T., Clémence, A., Doise, W., Darley, J., et al. (1962). L'analyse de similitude. *Cahiers du Centre de Recherche Opérationnelle*, 4:63–97.

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*, volume 13, page 15–25.

Ahmad, I., Yousaf, M., Yousaf, S., and Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, 2020:1–11.

Ahmad, S. and Laroche, M. (2015). How do expressed emotions affect the helpfulness of a product review? evidence from reviews using latent semantic analysis. *International Journal of Electronic Commerce*, 20(1):76–111.

Aikin, S. F. Talisse, R. B. (2014). *Why We Argue and How We Should: A Guide to Political Disagreement*. New York: Routledge.

Aikin, S. (2012). Poe's law, group polarization, and argumentative failure in religious and political discourse. *Social Semiotics*, 23:1–17.

Ajao, O., Bhowmik, D., and Zargari, S. (2019). Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511. IEEE.

Ajjour, Y., Alshomary, M., Wachsmuth, H., and Stein, B. (2019). Modeling frames in argumentation. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932, Hong Kong, China. Association for Computational Linguistics.

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.

Al-Khatib, K., Wachsmuth, H., Kiesel, J., Hagen, M., and Stein, B. (2016). A news editorial corpus for mining argumentation strategies. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Alatawi, H. S., Alhothali, A. M., and Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

Aldwairi, M. and Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, 141:215–222.

Aljrees, T., Cheng, X., Ahmed, M. M., Umer, M., Majeed, R., Alnowaiser, K., Abuzinadah, N., and Ashraf, I. (2023). Fake news stance detection using selective features and fakenet. *PloS one*, 18(7):e0287298.

Alkaraan, F., Albahloul, M., and Hussainey, K. (2023). Carillion's strategic choices and the boardroom's strategies of persuasive appeals: Ethos, logos and pathos. *Journal of Applied Accounting Research*, 24(4):726–744.

Allein, L., Moens, M.-F., and Perrotta, D. (2023). Preventing profiling for ethical fake news detection. *Information Processing & Management*, 60(2):103206.

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., and Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics*, 10(11):1348.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.

Aristotle (2000). *Aristotle: Nicomachean Ethics*. Cambridge University Press.

Aristotle, J., Claverhouse, R., and Sandys, J. E. (1909). The rhetoric of aristotle: a translation. *(No Title)*.

Aslam, N., Ullah Khan, I., Alotaibi, F. S., Aldaej, L. A., and Aldubaikil, A. K. (2021). Fake detect: A deep learning ensemble model for fake news detection. *complexity*, 2021:1–8.

Assimakopoulos, S., Muskat, R. V., Van Der Plas, L., and Gatt, A. (2020). Annotating for hate speech: The maneco corpus and some input from critical discourse analysis. *arXiv preprint arXiv:2008.06222*.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., and Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Azzi, S. and Gagnon, S. (2023). Ontology-driven parliamentary analytics: Analysing political debates on covid-19 impact in canada. In *Electronic Government and the Information Systems Perspective: 12th International Conference, EGOVIS 2023, Penang, Malaysia, August 28–30, 2023, Proceedings*, page 89–102, Berlin, Heidelberg. Springer-Verlag.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Bahad, P., Saxena, P., and Kamal, R. (2019). Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, 165:74–82.

Bakir, V. and McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2):154–175.

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *International conference on Machine learning*, pages 380–389. PMLR.

Bar-Haim, R., Edelstein, L., Jochim, C., and Slonim, N. (2017). Improving claim stance classification with lexical knowledge expansion and context utilization. In Habernal, I., Gurevych, I., Ashley, K., Cardie, C., Green, N., Litman, D., Petasis, G., Reed, C., Slonim, N., and Walker, V., editors, *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark. Association for Computational Linguistics.

Bärenfänger, M., Hilbert, M., Lobin, H., and Lüngen, H. (2008). Owl ontologies as a resource for discourse parsing. *Journal for Language Technology and Computational Linguistics*, 23(1):17–26.

Baumann, K., Bertino, A., Rettig, L., Sigloch, S., Subotic, D., and Subotic, I. (2021). The rescs ontology: Linking open research data from multiple sources to support interdisciplinary investigations. *Semantic Web*, 0(0):1–22.

Beghini, F. et al. (2023). Étude textometrique de l'œuvre de milan kundera. à la recherche de la pepite d'or.

Bench-Capon, T. and Dunne, P. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, 171.

Benzécri, J.-P. (1973). L'analyse des correspondances. l'analyse des données, vol. 2. *Dunod. Paris*.

Berk, E. A. and Filatova, E. (2019). Incendiary news detection. In *The Thirty-Second International Flairs Conference*.

Besnard, P. and Hunter, A. (2008). *Elements of Argumentation*. The MIT Press.

Beyhan, F., Çarık, B., Arın, İ., Terzioğlu, A., Yanikoglu, B., and Yeniterzi, R. (2022). A turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185.

Bhutani, B., Rastogi, N., Sehgal, P., and Purwar, A. (2019). Fake news detection using sentiment analysis. In *2019 twelfth international conference on contemporary computing (IC3)*, pages 1–5. IEEE.

Biswas, R. and De, S. (2022). A comparative study on improving word embeddings beyond word2vec and glove. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 113–118. IEEE.

Boeglin, N. (2018). *Représentations romanesques de la modernité parisienne dans le" Grand XIXème siècle", 1830-1913*. PhD thesis, Lyon.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Borges, L., Martins, B., and Calado, P. (2019). Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26.

Borst, P., Akkermans, H., and Top, J. (1997). Engineering ontologies. *International Journal of Human-Computer Studies*, 46(2–3):365–406.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., Maurizio, T., et al. (2018). Overview of the evalita 2018 hate speech detection task. In *Ceur workshop proceedings*, volume 2263, pages 1–9. CEUR.

Bourgeade, T., Cignarella, A. T., Frenda, S., Laurent, M., Schmeisser-Nieto, W., Benamara, F., Bosco, C., Moriceau, V., Patti, V., and Taulé, M. (2023). A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 674–684.

Bouzereau, C. (2022). Complémentarité du quantitatif et du qualitatif en adt. l'immigration dans les discours du front national. In *JADT 2022: 16th International Conference on Statistical Analysis of Textual Data*.

Buller, D. B. and Burgoon, J. K. (1996). Interpersonal deception theory. *Communication theory*, 6(3):203–242.

Burger, P., Kanhai, S., Pleijter, A., and Verberne, S. (2019). The reach of commercially motivated junk news on facebook. *PloS one*, 14(8):e0220446.

Burstein, P. (2003). The impact of public opinion on public policy: A review and an agenda. *Political Research Quarterly*, 56(1):29–40.

Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Cakanlar, A. and White, K. (2023). A systematic review on political ideology and persuasion. *Psychology and Marketing. Scopus*.

Carpentieri, G., Guida, C., and Sgambati, S. (2023). Textometric analysis on the ongoing academic spatial planning debate. *TeMA-Journal of Land Use, Mobility and Environment*, pages 197–223.

Chauhan, T. and Palivela, H. (2021). Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, 1(2):100051.

Cheema, A., Chacko, J., and Gul, S. (2019). Mobilising mass anxieties: Fake news and the amplification of socio-political conflict in pakistan. *Fake News*, 17.

Cheng, J., B. M. D.-N.-M. C. . L. J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *CSCW : proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, page 1217–1230.

Chesñevar, C. I., Maguitman, A. G., and Loui, R. P. (2000). Logical models of argument. *ACM Comput. Surv.*, 32(4):337–383.

Choudhary, A. and Arora, A. (2021). Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.

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*, pages 206–229. IGI Global.

Colomina, C., Margalef, H. S., Youngs, R., and Jones, K. (2021). The impact of disinformation on democratic processes and human rights in the world. *Brussels: European Parliament*.

Colucci Cante, L., Di Martino, B., and Graziano, M. (2023). *A Comparative Analysis of Formal Storytelling Representation Models*, pages 327–336.

Coppock, A., Hill, S., and Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*, 6(36):4046.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157.

De la Peña Sarracén, G. L. and Rosso, P. (2022). Unsupervised embeddings with graph autoencoders for multi-domain and multilingual hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2196–2204.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M., and De Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. In *2018 22nd conference of open innovations association (FRUCT)*, pages 272–279. IEEE.

Demirdöğen, (2016). The roots of research in (political) persuasion: Ethos, pathos, logos and the yale studies of persuasive communications. *International Journal of Social Inquiry*, 3(1).

Deo, A. (2015). Diachronic semantics. *Annu. Rev. Linguist.*, 1(1):179–197.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dharma, E. M., Gaol, F. L., Warnars, H., and Soewito, B. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *J Theor Appl Inf Technol*, 100(2):31.

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P., and San Martino, G. (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 (Volume 1: Long Papers*, page 6603–6617.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

Diwersy, S., Frontini, F., and Luxardo, G. (2018). The parliamentary debates as a resource for the textometric study of the french political discourse. In *Proceedings of the ParlaCLARIN@ LREC2018 workshop*.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Domínguez-Armas, Á., S.-R. A. . L. M. (2023). Provocative insinuations as hate speech: Argumentative functions of mentioning ethnicity in headlines. *Topoi*, 42:419–431.

Dong, M., Yao, L., Wang, X., Benatallah, B., and Huang, C. (2019). Similarity-aware deep attentive model for clickbait detection. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23*, pages 56–69. Springer.

Dou, D., Wang, H., and Liu, H. (2015). Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*, pages 244–251. IEEE.

Dou, Y., Shu, K., Xia, C., Yu, P. S., and Sun, L. (2021). User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055.

Dritsa, K., Thoma, A., Pavlopoulos, I., and Louridas, P. (2022). A greek parliament proceedings dataset for computational linguistics and political analysis. *Advances in Neural Information Processing Systems*, 35:28874–28888.

Druckman, J. (2022). A framework for the study of persuasion. *Annual Review of Political Science*, 25:65–88. Scopus.

Du, J., Dou, Y., Xia, C., Cui, L., Ma, J., and Philip, S. Y. (2021). Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE.

Dubremetz, M. and Nivre, J. (2018). Rhetorical figure detection: Chiasmus, epanaphora, epiphora. *Frontiers in Digital Humanities*, 5.

Duffy, M. and Thorson, E. (2016). *Persuasion Ethics Today*. Routledge CRC Press.

Duthie, R., Budzynska, K., and Reed, C. (2016). *Mining Ethos in Political Debate*, volume 287 of *Frontiers in Artificial Intelligence and Applications*, pages 299–310. IOS Press, Netherlands. This research was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National Science Centre under grant 2015/18/M/HS1/00620.

D'Ulizia, A., Caschera, M. C., Ferri, F., and Grifoni, P. (2021). Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Egami, N., Fong, C., Grimmer, J., Roberts, M., and Stewart, B. (2022). How to make causal inferences using texts. *Science Advances*, 8(42):2652.

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., and Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Fairclough, I., . F.-N. (2012). *Political Discourse Analysis: A Method for Advanced Students (1st ed.)*. Routledge.

Fang, Y., Si, L., Somasundaram, N., and Yu, Z. (2012). Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 63–72, New York, NY, USA. Association for Computing Machinery.

Faustini, P. H. A. and Covoes, T. F. (2020). Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Feldstein, S. (2023). The consequences of generative ai for democracy, governance and war. In *Survival: October–November 2023*, pages 117–142. Routledge.

Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Florou, E., Konstantopoulos, S., Koukourikos, A., and Karampiperis, P. (2013). Argument extraction for supporting public policy formulation. In Lendvai, P. and Zervanou, K., editors, *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria. Association for Computational Linguistics.

Foner, E. (1999). *Story of American freedom*. WW Norton & Company.

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Frenda, S., Patti, V., and Rosso, P. (2022). Killing me softly: Creative and cognitive aspects of implicitness in abusive language online. *Natural Language Engineering*, pages 1–22.

Furman, D. A., Torres, P., Rodríguez, J. A., Alonso Alemany, L., Letzen, D., and Martínez, V. (2023). Which argumentative aspects of hate speech in social media can be reliably identified? In Bonn, J. and Xue, N., editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 136–153, Nancy, France. Association for Computational Linguistics.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395*.

Garcia, G. L., Afonso, L. C., and Papa, J. P. (2022). Fakerecogna: A new brazilian corpus for fake news detection. In *International Conference on Computational Processing of the Portuguese Language*, pages 57–67. Springer.

Garssen, B. Kienpointner, M. (1996). Vernünftig argumentieren. regeln und techniken der diskussion [reasonable argumentation. rules and techniques of discussion]. *Argumentation*, 16:259–262.

Ghayoomi, M. and Mousavian, M. (2022). Deep transfer learning for covid-19 fake news detection in persian. *Expert Systems*, (e13008).

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Giulianelli, M., Tredici, M. D., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R.,

editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3960–3973. Association for Computational Linguistics.

Glavaš, G., Karan, M., and Vulić, I. (2020). Xhate-999: Analyzing and detecting abusive language across domains and languages. Association for Computational Linguistics.

Glockner, M., Hou, Y., and Gurevych, I. (2022). Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Godwin, M. (1994). Meme, counter-meme. *Wired*, 2(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*, page 4143–4149.

Goldani, M. H., Safabakhsh, R., and Momtazi, S. (2021). Convolutional neural network with margin loss for fake news detection. *Information Processing & Management*, 58(1):102418.

Goldman, M. and Perry, E. J. (2002). *Changing meanings of citizenship in modern China*, volume 13. Harvard University Press.

Gonzalez-Perez, C. (2020). Connecting discourse and domain models in discourse analysis through ontological proxies. *Electronics*, 9(11).

Gonzalez-Perez, C. (2023). What archaeological texts argue about: Denotations and ontological proxies. In *Discourse and Argumentation in Archaeology: Conceptual and Computational Approaches*, pages 93–114. Springer.

Goovaerts, I. and Marien, S. (2020). Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communication*, 37(6):768–788.

Gottipati, S., Qiu, M., Yang, L., Zhu, F., and Jiang, J. (2014). An integrated model for user attribute discovery: A case study on political affiliation identification. In *Advances in Knowledge Discovery and Data Mining: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part I 18*, pages 434–446. Springer.

Green, N. (2014). Towards creation of a corpus for argumentation mining the biomedical genetics research literature. *In Proceedings of the First Workshop on Argumentation Mining*, page 11–18.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Guarino, N. (1998). Formal ontologies and information systems.

Habernal, I. and Gurevych, I. (2016). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Haddadan, S., Cabrio, E., Soto, A. J., and Villata, S. (2023). Topic modelling and frame identification for political arguments. In *AIxIA 2022 – Advances in Artificial Intelligence: XXIst International Conference of the Italian Association for Artificial Intelligence, AIxIA 2022, Udine, Italy, November 28 – December 2, 2022, Proceedings*, page 268–281, Berlin, Heidelberg. Springer-Verlag.

Haddadan, S., Cabrio, E., and Villata, S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Hamilton, W., Clark, K., Leskovec, J., and Jurafsky, D. (2016a). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 595–605.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hammouchi, H. and Ghogho, M. (2022). Evidence-aware multilingual fake news detection. *IEEE Access*, 10:116808–116818.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., and Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*.

Hanum, S. L., Rusmawati, Y., and Arzaki, M. (2019). Construction of the ontology design for political parties' ideological characteristics. In *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, pages 381–388.

Haq, E., Braud, T., Kwon, Y., and Hui, P. (2020). A Survey on Computational Politics (arXiv:1908.06069). arXiv.

Hasan, K. S. and Ng, V. (2014). Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 751–762.

Heiden, S., Magué, J.-P., and Pincemin, B. (2010). Txm: Une plateforme logicielle open-source pour la textométrie-conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

Heinisch, P. and Cimiano, P. (2021). A multi-task approach to argument frame classification at variable granularity levels. *it - Information Technology*, 63(1):59–72.

Helmstetter, S. and Paulheim, H. (2018). Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE.

Herman, L. (2023). *Democratic Partisanship: Party Activism in an Age of Democratic Crises*. Edinburgh University Press.

Hofweber, T. (2023). Logic and Ontology. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition.

House of Commons (2019). Disinformation and 'fake news': Final report. *London: House of Commons*.

Hu, H., Amaral, P., and Kübler, S. (2022). Word embeddings and semantic shifts in historical spanish: Methodological considerations. *Digital Scholarship in the Humanities*, 37(2):441–461.

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., and Zhou, M. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Huang, X., Xing, L., Dernoncourt, F., and Paul, M. J. (2020). Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Hughes, H. C. and Waismel-Manor, I. (2021). The macedonian fake news industry and the 2016 us election. *PS: Political Science & Politics*, 54(1):19–23.

Hüsünbeyi, Z. M., Akar, D., and Özgür, A. (2022). Identifying hate speech using neural networks and discourse analysis techniques. In *Proceedings of the first workshop on language technology and resources for a fair, inclusive, and safe society within the 13th language resources and evaluation conference*, pages 32–41.

Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. In *IEEE/ACM joint conference on digital libraries*, pages 229–238. IEEE.

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., and Schoelkopf, B. (2022). Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 7180–7198.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 427–431. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Draft, Stanford University; University of Colorado at Boulder, 3 edition. Draft of January 7, 2023. Comments and typos welcome!

Jurkiewicz, D., Borchmann, Ł., 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*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. (2020). Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Kapil, P. and Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Karande, H., Walambe, R., Benjamin, V., Kotecha, K., and Raghu, T. (2021). Stance detection with bert embeddings for credibility analysis of information on social media. *PeerJ Computer Science*, 7:e467.

Karnyoto, A. S., Sun, C., Liu, B., and Wang, X. (2022). Augmentation and heterogeneous graph neural network for aaai2021-covid-19 fake news detection. *International journal of machine learning and cybernetics*, 13(7):2033–2043.

Kasnesis, P., Toumanidis, L., and Patrikakis, C. Z. (2021). Combating fake news with transformers: a comparative analysis of stance detection and subjectivity analysis. *Information*, 12(10):409.

Kastberg Sjöblom, M. and Jacquot, S. (2016). Le livret d'opéra : établissement et exploration textométrique d'un corpus patrimonial de l'époque classique. In *Le livret d'opéra : établissement et exploration textométrique d'un corpus patrimonial de l'époque classique*.

Kero, A. A., Demissie, D., Kekeba, K., et al. (2023). An ontology driven machine learning applications in public policy analysis: A systematic literature review. *PREPRINT*. Version 1.

Khanam, Z., Alwasel, B., Sirafi, H., and Rashid, M. (2021). Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing.

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

Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., and Puppe, F. (2016). Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40.

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)*, page 7024–7034.

Körner, E., Wiedemann, G., Hakimi, A. D., Heyer, G., and Potthast, M. (2021). On classifying whether two texts are on the same side of an argument. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 10130–10138.

Koufakou, A., Pamungkas, E. W., Basile, V., Patti, V., et al. (2020). Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2):e3767.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, 1(1):127–165.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Lebart, L. and Salem, A. (1994). Statistique textuelle. *Paris: Dunod*.

Lee, J.-W. and Kim, J.-H. (2022). Fake sentence detection based on transfer learning: Applying to korean covid-19 fake news. *Applied Sciences*, 12(13):6402.

Lemmens, J., Markov, I., and Daelemans, W. (2021). Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the fourth workshop on NLP for internet freedom: censorship, disinformation, and propaganda*, pages 7–16.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Li, R., Tian, P., and Wang, S. (2021). Study concept drift in 150-year english literature. In *AII@ iConference*, pages 153–163.

Liebowitz, J. (2019). Hate speech, disinformation and political violence in myanmar. *FAKE NEWS*, page 1.

Lifferth, W. (2018). Fake news. https://kaggle.com/competitions/fake-news.

Lin, Y., Michel, J.-B., Lieberman, E. A., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Lippi, M. and Torroni, P. (2016). Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Liu, Y. and Wu, Y.-F. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Liu, Y. and Wu, Y.-F. B. (2020). Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–33.

Longhi, J. (2017). Humanities, digital: From corpora to meaning, from meaning to corpora. *Questions de communication*, 31(1):7–17.

Lopes Cardoso, H., Sousa-Silva, R., Carvalho, P., and Martins, B. (2023). Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Magu, R. and Luo, J. (2018). Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 93–100.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. In Mitkov, R. and Angelova, G., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Manfred Stede, J. S. (2018). *Argumentation Mining*. Springer Cham.

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*, page 55–60.

Marchand, P. and Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT*, 2012:687–699.

Markov, I. and Daelemans, W. (2022). The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42.

Martinc, M., Novak, P. K., and Pollak, S. (2020). Leveraging contextual embeddings for detecting diachronic semantic shift. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4811–4819. European Language Resources Association.

Matamoros-Fernández, A. and Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

McGuire, R., Birnbaum, L., and Flowers, M. (1981). Opportunistic processing in arguments. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc. Proceedings of the Seventh International Joint Conference on Artificial Intelligence ; Conference date: 01-01-1981.

Mehta, H. and Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8):291.

Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2018). Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Miller, C. R. (1939). *How to detect and analyze propaganda*. Town Hall, Incorporated.

Miró-Llinares, F. and Aguerri, J. C. (2023). Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. *European Journal of Criminology*, 20(1):356–374.

Mishra, P., Del Tredici, M., Yannakoudakis, H., and Shutova, E. (2019). Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*.

Mitra, T. and Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 258–267.

Mochales, R. and Moens, M. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Mohawesh, R., Liu, X., Arini, H. M., Wu, Y., and Yin, H. (2023). Semantic graph based topic modelling framework for multilingual fake news detection. *AI Open*, 4:33–41.

Monti, F., Frasca, F., Eynard, D., Mannion, D., and Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Mosca, E., Wich, M., and Groh, G. (2021). Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102.

Mossie, Z. and Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020). A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Muller, C. (1968). *Initiation à la statistique linguistique*. Langue et langage. Larousse.

Naderi, N. and Hirst, G. (2016). Argumentation mining in parliamentary discourse. In Baldoni, M., Baroglio, C., Bex, F., Grasso, F., Green, N., Namazi-Rad, M.-R., Numao, M., and Suarez, M. T., editors, *Principles and Practice of Multi-Agent Systems*, pages 16–25, Cham. Springer International Publishing.

Nan, Q., Cao, J., Zhu, Y., Wang, Y., and Li, J. (2021). Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Nasir, J. A., Khan, O. S., and Varlamis, I. (2021). Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Nassif, A. B., Elnagar, A., Elgendy, O., and Afadar, Y. (2022). Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18):16019–16032.

Ni, S., Li, J., and Kao, H.-Y. (2021). Mvan: Multi-view attention networks for fake news detection on social media. *IEEE Access*, 9:106907–106917.

Novakova, I. and Siepmann, D. (2020). Literary style, corpus stylistic, and lexico-grammatical narrative patterns: Toward the concept of literary motifs. *Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives*, pages 1–15.

O'Keefe, D. J. (1977). Two concepts of argument. *The Journal of the American Forensic Association*, 13(3):121–128.

Oliveira, L., de Melo, P. V., Amaral, M., and Pinho, J. A. (2018). When politicians talk about politics: Identifying political tweets of brazilian congressmen. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Orwell, G. (2013). Politics and the english language. *Penguin Classics*.

Ozbay, F. A. and Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: statistical mechanics and its applications*, 540:123174.

O'Keefe, D. (2015). *Persuasion: Theory and Research*. SAGE Publications.

Palagin, O., Kaverinskiy, V., Litvin, A., and Malakhov, K. (2023). Ontochatgpt information system: Ontology-driven structured prompts for chatgpt meta-learning. *arXiv preprint arXiv:2307.05082*.

Park, S., Ko, M., Kim, J., Liu, Y., and Song, J. (2011). The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 113–122.

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

Pauli, A., Derczynski, L., and Assent, I. (2022). Modelling persuasion through misuse of rhetorical appeals.

Pavlopoulos, J., Sorensen, J., Dixon, L., Thain, N., and Androutsopoulos, I. (2020). Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*.

Pengam, M. and Jackiewicz, A. (2022). Les représentations causales de la radicalisation. analyse sémantico-discursive des discours institutionnels français (2013-2018). In *SHS Web of Conferences*, volume 138, page 01008. EDP Sciences.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Pérez, J. M., Luque, F. M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P. S., Zajac, J., Miguel, P., Debandi, N., Gravano, A., et al. (2023). Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Perifanos, K. and Goutsos, D. (2021). Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Pincemin, B. (2012). Sémantique interprétative et textométrie. *Texto! Textes et cultures*, 17(3):1–21.

Pincemin, B., Guillot, C., Heiden, S., Lavrentiev, A., and Marchello-Nizia, C. (2008). Usages linguistiques de la textométrie: analyse qualitative de la consultation de la base de français médiéval via le logiciel weblex. *Syntaxe et sémantique*, pages 87–110.

Piryani, R., Aussenac-Gilles, N., and Hernandez, N. (2023). Comprehensive survey on ontologies about event. *CEUR Workshop Proceedings*, SEMMES'23: Semantic Methods for Events and Stories co-located with ESWC, May 29:15.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48:4730–4742.

Plaza-del Arco, F. M., Molina-González, M. D., Urena-López, L. A., and Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Ploug, T. (2023). The right not to be subjected to ai profiling based on publicly available data—privacy and the exceptionalism of ai profiling. *Philosophy & Technology*, 36(1):14.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

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*, volume 1. Long Papers), 1615–1625.

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*, page 101–108.

Qian, S., Wang, J., Hu, J., Fang, Q., and Xu, C. (2021). Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Qiang, Z., Wang, W., and Taylor, K. (2023). Agent-om: Leveraging large language models for ontology matching. *arXiv preprint arXiv:2312.00326*.

Qiu, M., Sim, Y., Smith, N. A., and Jiang, J. (2013). Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*, pages 855–863.

Rachman, G. H., Khodra, M. L., and Widyantoro, D. H. (2018). Word embedding for rhetorical sentence categorization on scientific articles. *Journal of ICT Research & Applications*, 12(2).

Rayson, P. E. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Lancaster University (United Kingdom).

Raza, S. and Ding, C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Reed, C. and Budzynska, K. (2019). Advances in argument mining. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 39–42.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 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*, page 3982–3992.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Reisigl, M. and Wodak, R. (2017). The discourse-historical approach (dha). *The Routledge Handbook of Critical Discourse Studies*.

Rob Abbott, Brian Ecker, P. A. M. A. W. (2016). Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. *Language Resources and Evaluation Conference (LREC)*.

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111.

Sahoo, S. R. and Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100:106983.

Saleh, H., Alharbi, A., and Alsamhi, S. H. (2021). Opcnn-fake: Optimized convolutional neural network for fake news detection. *IEEE Access*, 9:129471–129489.

Samadi, M., Mousavian, M., and Momtazi, S. (2021). Deep contextualized text representation and learning for fake news detection. *Information processing & management*, 58(6):102723.

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*, page 1377–1414.

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*, page 5636–5646.

Santia, G. and Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Schaefer, R. and Stede, M. (2021). Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Schütz, M., Böck, J., Andresel, M., Kirchknopf, A., Liakhovets, D., Slijepčević, D., and Schindler, A. (2022). Ait fhstp at checkthat! 2022: cross-lingual fake news detection with a large pre-trained transformer. *Working Notes of CLEF*.

Seddari, N., Derhab, A., Belaoued, M., Halboob, W., Al-Muhtadi, J., and Bouras, A. (2022). A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. *IEEE Access*, 10:62097–62109.

Shaar, S., Babulkov, N., Da San Martino, G., and Nakov, P. (2020). That is a known lie: Detecting previously fact-checked claims. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., and Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

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*, page 750–767.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. (2019a). defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Shu, K., Mahudeswaran, D., and Liu, H. (2019b). Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25:60–71.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Shu, K., Wang, S., and Liu, H. (2017). Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 8.

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive language and hate speech detection for Danish. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Silva, A., Luo, L., Karunasekera, S., and Leckie, C. (2021). Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Slovikovskaya, V. (2019). Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. *arXiv preprint arXiv:1910.14353*.

Smith, C. A. and Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Song, C., Ning, N., Zhang, Y., and Wu, B. (2021a). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

Song, C., Shu, K., and Wu, B. (2021b). Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, 58(6):102712.

Sridhar, D. and Blei, D. (2022). Causal inference from text: A commentary. *Science Advances*, 8(42):6585.

Stab, C. and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In Tsujii, J. and Hajic, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1–2):161–197.

Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. (2021). New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., and De Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. In *Proceedings of the Second Workshop on Data Science for Social Good (SoGood), Skopje*, volume 1960, pages 1–12.

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*, page 613–624.

Tariq, A., Mehmood, A., Elhadef, M., and Khan, M. U. G. (2022). Adversarial training for fake news classification. *IEEE Access*, 10:82706–82715.

Taylor, J., Peignon, M., and Chen, Y.-S. (2017). Surfacing contextual hate speech words within social media. *arXiv preprint arXiv:1711.10093*.

Teufel, S., Carletta, J., and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In Thompson, H. S. and Lascarides, A., editors, *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Thiago, D. O., Marcelo, A. D., and Gomes, A. (2021). Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture*, 25(2):700–732.

Thompson, R. C., Joseph, S., and Adeliyi, T. T. (2022). A systematic literature review and meta-analysis of studies on online fake news detection. *Information*, 13(11):527.

Thon, V. (2022). Textometry as a method for analysing medieval emotions? the case of peter damian (1007-1072/73) and his combat letters. In *Feeling Medieval: The Inaugural Conference of the Society for the Study of Medieval Emotions (Université de St. Andrews, Ecosse)*.

Thota, A., Tilak, P., Ahluwalia, S., and Lohia, N. (2018). Fake news detection: a deep learning approach. *SMU Data Science Review*, 1(3):10.

Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., Suieuova, N., and Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. *International Journal of Advanced Computer Science and Applications*, 14(5).

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*, page 3296–3304.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018*, pages 517–524.

Tulu, C. N. (2022). Experimental comparison of pre-trained word embedding vectors of word2vec, glove, fasttext for word level semantic text similarity measurement in turkish. *Advances in Science and Technology. Research Journal*, 16(4).

Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., and On, B.-W. (2020). Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, 8:156695–156706.

Vaibhav, V., Annasamy, R. M., and Hovy, E. (2019). Do sentence interactions matter? leveraging sentence level representations for fake news classification. *arXiv preprint arXiv:1910.12203*.

Van Dijk, T. A. (1997). What is political discourse analysis. *Belgian journal of linguistics*, 11(1):11–52.

Venturini, T. (2019). From fake to junk news: The data politics of online virality. In *Data politics*, pages 123–144. Routledge.

Verma, P. K., Agrawal, P., Amorim, I., and Prodan, R. (2021). Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Vidgen, B. and Yasseri, T. (2020). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Visser, J., K. B. D. R. e. a. (2020). Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Lang Resources Evaluation*, 54:123–154.

Vlachos, A. and Riedel, S. (2014). Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

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*, page 1498–1507.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T., Hirst, G., and Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, page 176–187, Long Papers.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J. (2012). That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Walsh, L. (2017). Understanding the rhetoric of climate science debates. *Wiley Interdisciplinary Reviews: Climate Change*, 8(3):e452.

Walton, D. (2012). Argument mining by applying argumentation schemes. *Studies in Logic*, 4:38–64.

Walton, D. (October 2003). *Relevance in Argumentation*. Routledge.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. (2019). Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 5635–5649.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Watts, D. J., Rothschild, D. M., and Mobius, M. (2021). Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15):e1912443118.

Weston, A. (2018). *A rulebook for arguments*. Hackett Publishing.

Wevers, M. and Koolen, M. (2020). Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(4):226–243.

Whitehouse, C., Weyde, T., Madhyastha, P., and Komninos, N. (2022). Evaluation of fake news detection with knowledge-enhanced language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1425–1429.

Wilson, T., Wiebe, J., and Cardie, C. (2017). Mpqa opinion corpus.

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., and Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., and Yu, P. S. (2018a). Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

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

Yantseva, V. (2023). Discursive construction of migrant otherness on facebook: A distributional semantics approach. *Discourse & Society*, 34(2):236–254.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

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*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

Yu, W., Boenninghoff, B., and Kolossa, D. (2022). Bert-based ironic authors profiling.

Yuan, H., Zheng, J., Ye, Q., Qian, Y., and Zhang, Y. (2021). Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, 151:113633.

Zarouali, B., Boerman, S., Voorveld, H., and Noort, G. (2022). The algorithmic persuasion framework in online communication: Conceptualization and a future research agenda. *Internet Research*, 32(4):1076–1096. Scopus.

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

Zhang, J., Cui, L., Fu, Y., and Gouza, F. (2018). Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*, 3.

Zhang, J., Dong, B., and Philip, S. Y. (2020). Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 1826–1829. IEEE.

Zhang, X. and Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Zimmerman, S., Kruschwitz, U., and Fox, C. (2018). Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Zubiaga, A., Wong Sak Hoi, G., Liakata, M., and Procter, R. (2016). Pheme dataset of rumours and non-rumours.

Žagar, I. (2010). Topoi in critical discourse analysis. *Lodz Papers in Pragmatics*, 6:3–27.