Structure- and Context-aware NLP Approaches to Emotion and Subjectivity

by

Hamed Yaghoobian

(Under the Direction of Khaled Rasheed)

Abstract

Incorporating hierarchical structures of language such as trees has been shown to be effective for various Natural Language Processing (NLP) tasks. The utilization of hierarchical and structural information of text can provide insights into context, compensating for data shortage in many situations. In this dissertation, I investigate and propose novel structure- and context-aware models that aim for effective analysis of emotionality, intentionality, and subjectivity at various levels in text. Further, I explore how pre-trained language models and auto-completion generative algorithms built into modern writing tools and word processors alter the practice and experience of writing for the human user. Lastly, I address the gap that exists in pedagogical approaches in computer science education by proposing a framework that cultivates deeper computational thinking skills that have their roots in Human-Centered AI principles, thus fostering more humane, responsible, and critical approaches to computation.

Index words: Sentiment Analysis, Sarcasm Detection, Language Models, Explainable AI

Structure- and Context-aware NLP Approaches to Emotion and Subjectivity

by

Hamed Yaghoobian

B.S., Azad University, Mashhad Branch, Iran, 2011

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2021

Structure- and Context-aware NLP Approaches to Emotion and Subjectivity

by

Hamed Yaghoobian

Major Professor: Khaled Rasheed

Committee: Hamid R. Arabnia
Sheng Li

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2021

تقدیم به مادر و پدرم، فرشته و حسین

Acknowledgments

There were many people who helped me along the way with this dissertation, and most of all, I want to express my gratitude to Dr. Khaled Rasheed, without whom this dissertation might never have been finished. Thank you for all your support and guidance over these past years.

I want to thank all of my committee members, each of whom gave me inspiration during this process. Thank you, Dr. Hamid Arabnia, for generously advising me to "follow the heart" when I first stepped into your office asking for direction. I can not possibly thank you for your encouraging, thoughtful, and wise words. My gratitude to you is immense.

I am grateful to Dr. Sheng Li for giving me insightful advice and for always being willing to offer help and support when I needed it.

Thanks to my undergraduate students at the University of Georgia who participated in my surveys and shared their fresh and not-yet-disciplined insight and ideas. In the end, it is all about learning and being thankful.

Thank you to all my friends at UGA, who helped me on this journey, especially Mohammadreza Davoodi, Saed Rezayi, Arash Aboutorabi, Sahar Voghoei, and

Elliott Kuecker, for their friendship, help, and all wonderful conversations, which have always been a source of comfort, joy, and inspiration. Thank you all for sticking with me all these years.

I want to thank my parents and my sister, Parisa, who have always supported me in all my endeavors. Thanks for bearing with me from one adventure to the next.

Lastly, I want to thank Zohreh, my wife. I could not have done this without you.

کاش می‌شد بنویسم چه نوشتی با چشم
کاش می‌گفت عبارت چه اشارت کردی
- مهدی اخوان ثالث

Contents

Acknowledgments **v**

1 Introduction **1**

1.1 Objectives and Research Motivations 2

1.2 Studied Areas and Methods Used 3

1.3 Contributions 9

1.4 Dissertation Organization 9

2 Experience Affected in The Act of Remembering **11**

2.1 Introduction 12

2.2 Background and Literature Review 14

2.3 Dataset 19

2.4 Method 20

2.5 Experiments and Results 20

2.6 Conclusion 26

3 Sarcasm Detection **27**

3.1 Background and Literature Review 28
3.2 Experiments 43
3.3 Method 48
3.4 Results 51
3.5 Further Experimentation and Conclusion 53

4 Generative Language Models and Co-Creative Writing 56
4.1 Introduction 57
4.2 Contextual Encoders: Architecture 59
4.3 Contextual Embeddings: Pre-training 61
4.4 Findings 66
4.5 Generative Language Models and Co-Creative Writing 67
4.6 Closing Thoughts 73

5 Ethics, Human-Centered AI, and Pedagogical Possibilities In CS Education 76
5.1 Introduction 78
5.2 Foundations of Computational Disciplines 78
5.3 Toward a Human-centered AI Educational Model 81
5.4 Ethics-oriented Critical Engagement in Computer Science Education 83

6 Conclusions and Future Work 88
6.1 Path Forward 91

List of Figures

1.1 Human-Computer Interaction Through The Medium of Language 2

2.1 Summary distribution of political leaning score for three political groups. Political leaning is devised to reveal the political affiliation of participants in 'other' group. 21

2.2 Distribution of discontinuity measure for the two political groups. Discontinuity is defined as the number of shifts in verb tense of a narrative normalized by the total number of verbs. 23

3.1 An example of a discourse tree with 4 EDUs and 3 relations. The actual sentence for this tree is: (Policeman Doesn't Like His Picture Taken: "I'm gonna fucking break your face". EDU1) (This is why we need laws EDU2) (to prevent cops EDU3) (from being filmed. EDU4) 46

3.2 Hierarchical attention mechanism to obtain document-level low rank representation. 50

5.1 Students' Preferred Perspective for Analysis of Topics 86

List of Tables

3.1 Features used for Statistical Classifiers 33

3.2 State-of-the-art NN classifiers and results on Reddit Politics dataset 40

3.3 F1 score comparison chart. Values for other methods are directly reported from their corresponding papers. 52

4.1 Pre-trained Language Models 61

Chapter 1

Introduction

Recent successes in natural language processing have been transformational across a range of applications, from sentiment analysis to natural language generation. However, this has come at the expense of models becoming less interpretable. With an ever-growing complexity, most state-of-the-art machine learning models today deny various levels of explainability and interrogability. Issues range from the disproportional assignment of risk to certain demographics to making predictions based on spurious correlations. This calls for a more successful collaboration between computational models and humans to better formalize some of the nuanced interactions between human-computer criteria and to develop computational tools for inspection and introspection in model development with respect to them.

1.1 Objectives and Research Motivations

Informed by concerns mentioned above, I investigate the spaces that exist between three main components namely, *Human*, *Computer* and *Text*, as in written language. Specifically, I focus on human experience, emotion, and intention reflected in the language in the act of writing. Motivated and inspired by the inter-potentialities between overlapping areas in Natural Language Processing (NLP) and Human-computer Interaction (HCI), my investigation is positioned in the cross-section of sentiment analysis, sarcasm detection, and contextual design of AI systems. I aim to address various gaps that exist when machines[1] are utilized toward understanding human experiences, intentions, and emotions. Figure 1.1 illustrates the conceptualization of the spaces with respect to two main components: Understating of *humans' subjective experience* by machines and understating of of *machines' behavior* by humans through the medium of language use.

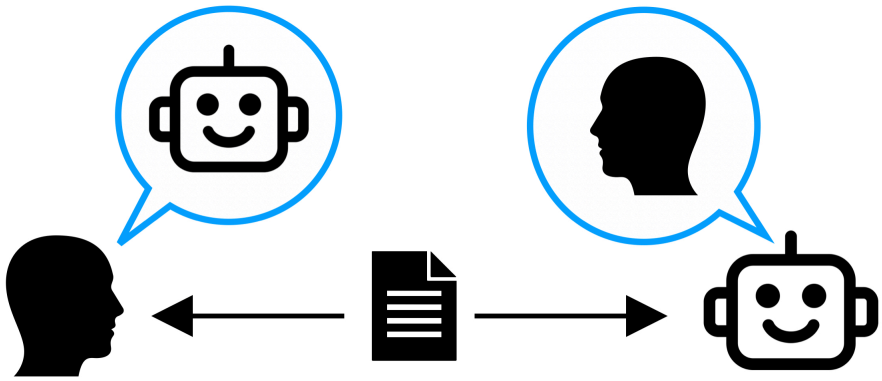

Figure 1.1: Human-Computer Interaction Through The Medium of Language

[1]By machine, I mean a computer powered with computational models and algorithms to understand human-generated content.

In this research, I draw from numerous fields, including my home field of Computer Science and the emerging subfield of Explainable AI (XAI)– a nascent but exciting area that holds much promise for the ideas explored here – as well as design, cultural studies, philosophy— where there is already a substantial history of thinking about humans and tools, with ongoing discussions that inform this work. This research would be impossible from a purely disciplinary perspective, and it is my belief that no technology should be studied outside of an interdisciplinary lens. This raises obvious practical concerns because one cannot achieve expertise in every field that is relevant to topics as far-ranging as computation and language. However, similar to a journalist perhaps who weaves together narratives from a variety of sources, I have tried to present perspectives that capture the essence of such concerns for the future of computational models for natural language understanding and generation. In Section 1.2, I briefly review the different types of methods used in conducting this research.

1.2 Studied Areas and Methods Used

From a computational perspective, the perceived linear and flat surface of natural language manifested in text bellies the underlying structure that is inherently hierarchical. Meaning and grammar operate on multiple levels in text, sometimes beyond the grammatical boundaries at the discourse level. The vast number of studies, in the pre-neural network paradigm in natural language processing, on proposing different tree-like structures is an attestation to the importance of

meaning and grammar, which can be represented as trees. Accordingly, leveraging hierarchical structures of language enables models with more semantic information about the data and highlights the identification of linguistic constructs that convey emotions, sentiments beyond what can be construed from the lexical, syntactic, and semantic features.

Toward enhanced interpretability, in our computational investigation of language use, we utilize and propose models that respect the structural contexts of text and incorporate neural networks and pre-trained language models like BERT for enriching contextual embeddings. Previously existing models are mostly sequence-to-sequence models that are not fit for dealing with hierarchical structures and representations such as trees. We have utilized the architectures that would allow tree-based attention mechanisms and hierarchical embeddings.

In what follows, I overview three overlapping areas of research explored in this dissertation and the gaps each address. Section 1.2 provides a peek into two computational studies of emotionality and intentionality, commonly referred to as sentiment analysis and sarcasm detection within the natural language processing community. In Section 1.2, I shift to the left side of the spectrum illustrated in Figure 1.1, and interrogate human experience and understanding of the behavior of generative language machines. Informed by our analysis and motivated by techniques and methodologies that enable a more interrogative and ethical interaction with machines toward allowing for a more directed and purposeful engagement, in Section 1.2 we briefly explain our educational framework for ethics-oriented computer science.

Sentiment, Emotionality and Intentionality Analysis

Recent years have witnessed a surge of interest in computational methods for affect[2], ranging from opinion mining to subjectivity detection to sentiment and emotion analysis. Sentiment Analysis (SA) is a stream of research in the realm of natural language processing tasked with determining the polarity of a text utterance according to the opinion or sentiment of the speaker or writer, as primarily positive, negative, or neutral. From a technical perspective, SA is essentially a text classification problem. As opposed to the traditional topic-based classification, a challenging aspect of this task is that while topics are often identifiable by keywords alone, sentiment can be expressed in a multitude of subtle and implicit ways. The implicit sentiment or connotative knowledge (i.e., the feeling a concept generally invokes for a person or a group of people) is referred to as prototypical sentiment. Such expressions are devoid of subjective words and rely on common sense shared by the speaker and receiver in an interaction [Joshi et al., 2017].

Regardless of the method, SA has been approached at three granularity levels: document level, sentence level, and aspect level. A document can be an opinionated product review in which the entire document is considered as the primary information unit with the assumption that the document is known to be opinionated and contain opinions toward a single entity (e.g., an electronic device). The sentence level takes into account the sentimentality of each sentence, although not all sentences in a document can be assumed to contain subjective emotion

[2]Affect, in psychology, refers to the underlying experience of feeling, emotion and is arguably incommensurable with the conception of affect in philosophy.

and opinion. Finally, compared with document- and sentence-level SA, aspect (target)-level is more fine-grained. Its task is to extract and summarize opinions towards entities (data objects) and their aspects (features). For example, from the sentence, "the voice quality of iPhone is great, but its battery sucks" entity extraction should identify "iPhone" as the entity, and aspect extraction should identify that "voice quality" and "battery" are two aspects. Aspect sentiment classification should classify the sentiment expressed on the iPhone's voice quality as positive and on the battery of the iPhone as negative.

While work centering on explicit markers like adjectives as sentiment-bearing words abound in the literature, there are few studies that investigate emotionality beyond lexical, syntactic structures. Identification of implicit emotion and intent heavily relies on contextual information. Chapters 2 and 3.2 present approaches for capturing context toward identification of affect, emotion, and sarcasm at discourse level – any use of language above the clause level.

Subjectivity Reflected in Co-creative Writing

Generative Language Models

NLP systems, traditionally, draw on intrinsically interpretable methods. Such approaches are mostly known as *white box* techniques with varying degrees of transparency and control over the output. More recently, however, substantially advance deep learning models of understanding language have resulted in models being more *black-boxed*, in which the process by which a model arrives at results in obfuscated from the user.

The third version of the Generative Pre-Trained Transformer (GPT-3) and most recently Switch-C [Fedus et al., 2021] have made a splash in the world of Natural Language Processing (NLP) and Machine Learning (ML) as they are able to take unlabeled content as data and use it to create a model that can generate human-like text. Although their ability to generate new textual artefacts has been criticized for failing certain semantic and ethical tests, they show massive improvements on various tasks. For example, medical chat-bots that have utilized GPT-3 have been reported to tell fictitious patients to commit suicide [Korngiebel and Mooney, 2021]. Moreover, chat-bot models have been known to amplify the gender bias that exists in training dialogues – a faithful mirroring of the associations in training data. As a result, the open-ended dialog remains unpredictable, risky, and problematic for various applications. However, through HAI, OpenAI has been able to remedy some of the toxic behavior, and egregious biases in their model[3]. Despite all these concerns, consequences, and proposed remedial corrective and strategies to mitigate the risks, the availability of generative language models marks the arrival of a new era in which we can now mass-produce justifiably good semantic artefacts [Floridi and Chiriatti, 2020]. Studies show that GPT-3 writes better than many people [Elkins and Chun, 2020]. The integration of these "stochastic parrots" [Bender et al., 2021] into writing tools, for the purpose of augmenting human writing and creative capabilities, is the main focus of our study of word processors. With eyes toward the subjective experience of the human user with the computer, I have focused essentially on the co-creative

[3]https://venturebeat.com/2021/06/10/openai-claims-to-have-mitigated-bias-and-toxicity-in-gpt-3/

textual interaction between the human and the writing tool. Mainly how humans in their writing practices rely on machine-generated content and interfaces that allow auto-completion and tonality detection, we ground this analysis in principles of XAI, HCI, and philosophy of technology.

Human-centered AI and Pedagogical Possibilities In CS Education

Computer science holds the unfortunate distinction, at least historically, as a pragmatic and often non-critical domain of practice – in terms of both theory and praxis. Nonetheless, we are witnessing a growing allegiance to *Human-centered AI* and *Data Ethics* in both academia and industry, and purportedly, 200 university curricula claim to have complied with tech ethics. This proliferation is a promising step forward and can be a useful corrective against discriminatory consequences concerning algorithmic systems, particularly in technical fields like CS and AI, where ethical issues are not discursively and normatively foregrounded. However, these efforts are more coordinated with conventional business ethics than more critical traditions of social justice expected to prevail in educational settings. It is required of educational systems to enable students to improve their technological sensibilities and skill sets to generate alternative possibilities that would challenge technocratic ideologies that are uncritical and limit innovation to the measures of efficiency and marketability. Nudged by these concerns and energized by the creative possibilities of a classroom, we realize the integration of the critical and social dimensions of computing technologies into introductory CS courses as a key

area of attention and propose a pedagogical approach largely through concepts that have their legacies in HAI and critical theory.

1.3 Contributions

This dissertation is an exploration of emotion and subjectivity in language from both theoretical and computational perspectives. From a computational perspective, my contributions are fourfold: 1) we propose structure-aware models for understanding human emotions, subjective experience, and intentions (in forms of emotion, affect, and sarcasm) reflected in written language, 2) we introduce a memory narrative dataset of the collected text from surveys of 185 participants, 3) we offer a cross-disciplinary analytical account of the explainability of generative language models built into writing tools, we explore what it would mean for creative activities used to be primarily the province of human capacity, lastly 4) we offer an educational framework for CS courses that complies with AI ethics principles.

1.4 Dissertation Organization

The rest of the dissertation is divided into the forthcoming chapters. Chapter 2 presents a hierarchical approach to emotion and subjectivity reflected in memory narratives of the studies population. We also propose a memory narrative dataset of 185 participants in this chapter. Chapter 3 offers a comprehensive investigation of the existing work on sarcasm detection in which issues and gaps are identified.

Section 3.2 of this chapter is dedicated to our proposed novel structure-aware model for sarcasm detection. In chapter 4, we provide a brief technical account of BERT and Post-BERT Pre-trained Language Models and their features. We then continue to explore the co-constitutive relationship that human user develops with generative writing tools through the lens of philosophy of technology and Explainable AI (XAI). In chapter 5, we go over our proposed CS educational framework that has its legacies in XAI. The dissertation concludes in chapter 6 with a reflection on the broader impact of our work encompassing this dissertation. The chapter ends with future directions for this line of research.

Chapter 2

Experience Affected in The Act of Remembering

Chapter Overview[1]

This article contributes to the empirical understanding of the discursivity of verb morphology and verb tense shifts in memory narratives. Specifically, we explore how the 2016 presidential election result, as a historic and political event of the past decade, is recounted collectively through the lens of language use. In an online survey, 185 undergraduate students in the Computer Science department at the University of Georgia were asked to remember the day they learned about

[1]The content of this chapter is a reprint of the paper:

H. Yaghoobian, S. Rezayi, H. Arabnia and K. Rasheed. "Experience Affected in The Act of Remembering: A Study of Discursivity of Morphosyntactic Verb Tense Shifts in Memory Narrative", Submitted to the SIGNLL Conference on Computational Natural Language Learning (CoNLL) 2021.

the 2016 presidential election results and write a narrative of their experience. The results from our analysis show a distinct correlation between the political leaning of the surveyed population and verb tense shifts in their stories.

2.1 Introduction

People navigate life enduring experiences with different emotional valence varying from negative to positive. Interestingly, bodies of evidence indicate that the feeling of remembering is heightened with emotion [Sharot et al., 2004]; therefore, certain memories, as James et al. [1890] writes, feel as if they have "left a scar upon the cerebral tissues." These memory imprints are not necessarily on an individual level but sometimes on the collective memory of a group or society. For instance, one is more likely to remember what they were doing in the event of an occurrence designated historic by a group they feel affiliated with. This phenomenon underlines the durability of emotional memories and subjective vividness of significant experiences, especially in the course of their expression. The underlying lingual aspect of remembering grants it a discursivity that permeates the psychological aspect to the point that " there is no realm of subjectivity, unconscious feelings, or objective reality, that language does not reach" Edwards [2006]. Foregrounding this social dimension of human memory and its language-dependency, Zerubavel [2012] believes we remember much of what we do only as members of particular communities. Thus through memory as the central faculty of being in time Olick et al. [2011], we define our individual and collective selves. This

understanding of memory narrative as a meaningful cultural and empirical object demands an engrossing examination of the socio-mental and pragmatic nature of language in the discourse of remembrance. In this study, we invoke approaches from the field of memory studies and are interested in the linguistic structures and complexities thereof, as well as in the insights they offer about how attitudes, emotions, and community identification are revealed through patterns of language use. This study explores how both salient and political events of the past are recounted on a collective scale. Specifically, the overall objective is to contribute to the understanding of emotionality in discourse by examining the morphosyntactic constructs and temporal consistency in autobiographical memory narratives.

We ask: Q_1 How is language a gate to emotionality and affect? Q_2 Is there a link between our political identities and the way we remember the past? In answering these two questions, we identify linguistic constructs that convey emotions, sentiments, and attitudes beyond what can be construed from the lexical, syntactic, and semantic features– irrespective of grammatical boundaries [Martin and White, 2003]. Our contribution is two-fold: First, we introduce a memory narrative dataset of the collected text from surveys of 185 participants reflecting on and writing about the 2016 United States presidential election in Spring 2019. Second, we highlight the discursive function of verb tense shifts in memory narrative by demonstrating the correlation between the political leaning of the surveyed population and a proposed metric we dub *discontinuity*.

The rest of this paper is organized as follows: Section 2.2 introduces the theoretical concepts that inform and inspire our research questions. In section 2.2.2,

we briefly review the philosophical concepts of affect– or simply the co-constitutive relation between our bodies and the world– and emotionality, and then in Section 2.2.1 delve into inter-connective theories in psycho- and sociolinguistics and cognitive science surrounding memory narrative, and emotion. Next, we introduce and review our dataset from surveys in Section 2.3 and proceed to our results and findings in Section 2.5. Finally, we provide a conclusion in Section 2.6.

2.2 Background and Literature Review

In this section, we provided an expanded view of theories and concepts on autobiographical memory and the function of emotion and affect enacted in the act of remembering.

2.2.1 Psychoanalytic Affect and Autobiographical Memory Narratives

Telling a story[2] about a past event relies on experience retrieval. It involves reconstruction and drawing on episodic memories, at times with greater cognitive effort for supporting details Hauch et al. [2015]. In this section, we focus on cognitive and memory-centered approaches, supplemented by socio-psychological considerations for the role of emotion and affect in the recall of collective experiences.

When events elicit intensities of affect, the valence of the experience (i.e., the polarity of the associated sentiment, positive or negative) can impact the details

[2]The terms 'story' and 'narrative' are used here interchangeably.

remembered Kensinger [2009]. Interestingly, studies on autobiographical memory suggest that emotional self-appraisal of past events tends to be positively biased Walker et al. [1997]. The fading affect bias (FAB) is a tendency for emotions associated with negative or unpleasant-event memories to fade faster than emotions associated with positive-event memories Walker et al. [2003]. Therefore, the affective intensities of extreme and possibly traumatic memories dissipate over time. This understanding is more aligned with a constructivist concept of memory in which past experiences are reconstructed through remembrance such that they fit into a representation of the self.

Although it is difficult and possibly hazardous to undertake the description of a historical event, the privileged rhetorical status of an observer would enable the individual to author a narrative by some means. Biesecker [2002] highlights "what we remember and how we remember an event can tell us something significant about who we are as a people, about the contemporary social and political issues that divide us, and about who we may become." Similarly, Heidegger [2010] believes we are the sort of being that has a concern for what and who it is and is constantly reflecting on its own past Brown and Reavey [2017]. Remembering the past and making sense of the present are clearly intertwined activities. As an example, Zerubavel [2012] underlines the difference between what Americans and Indians tend to recall from wedding ceremonies is a product of their having been socialized into different mnemonic traditions involving altogether different mental filters commonly shared by their respective mnemonic communities.

Evidently, memory narratives allow people and groups to organize and make

sense of complex data of experience in ways that reflect their identities, goals, and values. Halbwachs [1992] contends "a remembrance is a reconstruction of the past achieved with data borrowed from the present." Regardless of the time directionality of these socio-mental representations, past events are communicated following narrative conventions and plot structures acquired during childhood Brown and Kulik [1977]. These conventions necessitate adherence to a temporal structure Bruner [1990]. However, research on autobiographical memory shows that memory narratives may contain abrupt shifts from the past tense to the present tense, even sometimes done intentionally in order to engage the listener or the reader [Pillemer et al., 1998]. In affective situations, the protagonist experiences a presence in the past that enables a vivid account of the perceptual experience through a heightened state of emotion.

This seemingly random switching phenomenon, notably the intrusion of present tense into past narration, has been an object of interest in various genres of narrative in a range of languages, ancient to modern Fleischman [1985]. However, by probing the linguistic foundations in narrative performance, we can minimize the cultural and temporal gap, regardless of the genre and form of narrative.

2.2.2 Affect as a Residue of Lived Experience

The affective histories of lived spatiality and temporality hold agency over the essence of spontaneous emotions and feelings at present. Affect, in this account, is not synonymous with emotion and emerges as necessarily entangled with memories and materials, sensations, and spaces Robinson and Kutner [2019]. Thus, it falls

within a space that is beyond the hermeneutically qualifiable. This understanding of affective experience ensnares and subsumes emotion and is not reducible to singularizable, predictable, capturable, and identifiable feelings. However, this notion of affect does not deny or reject the subjectification of experience and only acknowledges an experiential dispersity that disrupts any fixation or stasis. Alcoff describes lived experience of the body "operating in various ways [that] invokes features of social realities, practices, and discourses and requires analyses that will not lose sight of [such] particularities" Alcoff [2005].

This conceptualization of our first proposed research question can be usefully done through Brian Massumi's account of affect. For Massumi, affect describes an autonomous system of intensity, 'associated with nonlinear processes: resonation and feedback which momentarily suspend the linear progress of the narrative present from past to future' Massumi [2002]. The remembering "I" in a narrative is a fiction composed of multiple connections inside, outside, and through the body; it is "a protagonist that cannot be resolved or recognized as such" Manning [2013].

Arguably, the conception of affect as pre-individual and disperse that possibly decouples affect from emotion and meaning is not incompatible with our collective and distributed understanding of affect that neither emanates from nor belongs to a single individual subject but it is diffused across collective assemblages that encompass both bodies and language.

Motivated and informed by the theoretical concepts of memory narratives, emotion, and affect discussed above, we intend to put affect to work in ways that

fulfill its theoretical potential without either reducing it to psychological conventions that focus on individual feelings and emotions or elevating it to higher levels of abstraction. One way to account for operationalization of the affective intensity in memory narratives lies in the extent to which temporal rhythm is disrupted by a shift in time, like an inadvertent intrusion of present tense into a past narrative. Reflecting on theories, we, therefore, premised our investigation described in detail in Section 2.5 upon the assumption that collective memories are structured linguistically and there exist multiple layers of meaning and affect surrounding the representations of the past and the present. Accordingly, operationalization of the temporal shifts in a memory narrative foregrounds the disconnection of affect from meaning and favors the narrative's affective rhythm as a primary and meaning as a secondary. Empirically, the affective rhythm can be parsed and analyzed by considering verb tense as a temporal marker. Although the function of verb tense in a narrative is not basically that of temporal reference, which in most narrative forms is a priori past tense, tense shifts would push it into pragmatics, ceding its study to discourse narrative. To further highlight the discursivity of morphosyntactic tense shifts, we endeavor to correlate the temporal structure of memory narratives with collective identities of the studied population around social movements and party politics (liberal and conservative ideologies) in the United States.

2.3 Dataset

Here we describe our process for creating the dataset from the conducted surveys.

2.3.1 Survey

The surveys were administered in one non- and two proctored sessions using Online Qualtrics Survey Software[3] in Spring 2019. They consists of 29 questions, including 22 demographic items (e.g., education, job, social and political affiliations) and 7 items particularly related to the 2016 presidential election results. In the two specific items regarding memory narratives, the participants were asked to describe the day they learned about the presidential election results in 2016 using the prompts: *"What is happening around you?"* and *"Describe how feel?"*. To encourage the use of elaborate events, no maximum word limit was imposed on the length of the narratives. Responses to the survey were automatically recorded via Qualtrics. Surveys that were returned less than 100% complete were not considered in the analysis.

2.3.2 Respondent Demographics

The average age of participants (n=185) is 22.34 years old. Participants are predominantly in their fourth and third year of college, 63%, and 33% respectively. Of the participants, 75% identified as male by choosing male pronouns (he/him) as their preferred pronouns, 22% identified as female by choosing female pronouns

[3]https://www.qualtrics.com/

(she/her) and 2% as non-binary by selecting they/them. 70% of the undergraduate students reported the United States as their place of birth. 51.87% reported both parents born in the United States, and 40.11% having non-native parents. 64.17% of the respondents reported both parents having a university degree, and 17.11% reported neither of their parents not holding a university degree. 16.04% and 49.20% self-identified as conservative and liberal, respectively, and 34.76% as "other". 81.82% were eligible to vote in 2016 of which 13.37% voted for Donald Trump, 33.16% Hilary Clinton, 4.8% Gary Johnson, and the other 48.66% preferred not to answer the question[4].

2.4 Method

In this section, we describe our approach to identification of the main verb in memory narratives and calculation of the political leaning of the participants based on their stances concerning social issues.

2.5 Experiments and Results

2.5.1 Political Leaning

Gauging an individual's political leaning is one of the most significant and enduring foci of political science. The results from the survey show levels of inconsistency between participants' self-claimed political leaning and their political

[4]We will open-source the data and the code upon acceptance of the paper.

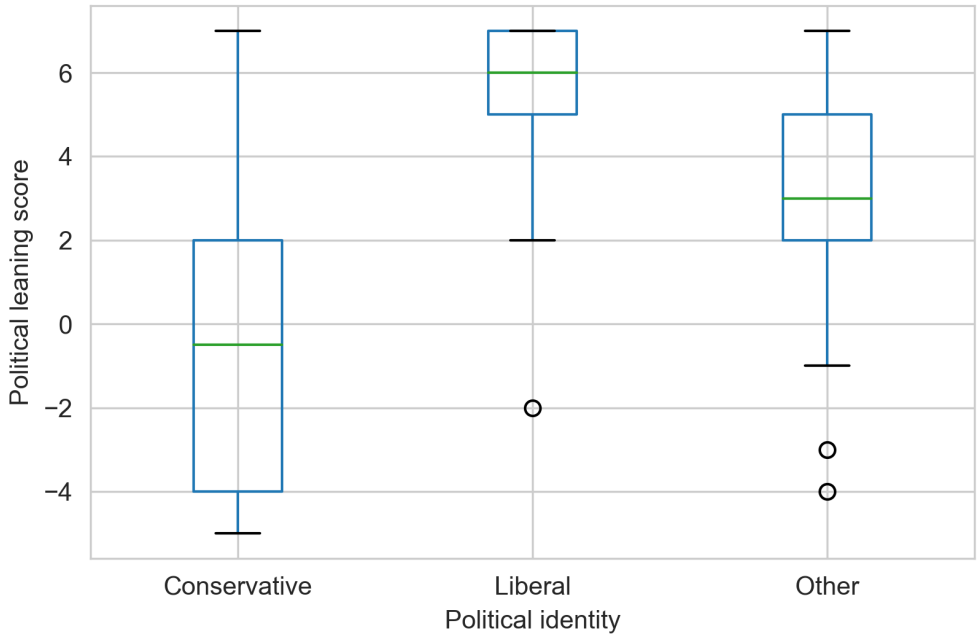

Figure 2.1: Summary distribution of political leaning score for three political groups. Political leaning is devised to reveal the political affiliation of participants in 'other' group.

standings on social issues. Toward gauging the affiliation of the participants who did not pigeonhole their political identities into "liberal" or "conservatives" categories by selecting "other", we devised a procedure that would assign a numerical score to the level of conservativity or liberality of the participants. With this procedure we can find the political affiliation of the "other" group to have a more comprehensive sample. We show that what we define as political leaning is aligned with the political identity that participants claimed. Figure 2.1 demonstrates the

distribution of participants' political leaning scores across the three categories. As evident in the figure, the "other" category is closer to the "liberal" class, while the "conservative" category is more diffuse and incoherent.

To devise the political leaning metric, we asked participants to disclose their opinions on six major social issues, namely Abortion Rights, Black Lives Matter, LGBTQA+ Rights, Public Health Care, Climate Change Denials, Strict Gun Control, and Strict Immigration Laws. The participants could be for, against, or undecided toward these issues. We assign a score to each social issue and a weight to each group (i.e., for or against groups). The overall political leaning score is the weighted sum of score values in both groups:

$$\text{political leaning} = \sum_{s} \sum_{w} s \times w$$

where s is the issue score and w is the group weight. Among all social issues, Abortion Rights, Black Lives Matter, LGBTQA+ Rights, Public Health Care, and Strict Gun Control are assigned a score of 1, and Climate Change Denials and Strict Immigration Laws are assigned a score of -1. Also, for group has a weight of 1 and against group has a weight of -1. For instance, a participant who is for Public Health Care and Strict Gun Control and is against Strict Immigration Laws receives a score of 3.

2.5.2 Discontinuity: Verb Tense Shift

The main verb of a sentence is one of the indicators of temporality in a sentence. In order to empirically analyze the structure of a sentence and detect the verb

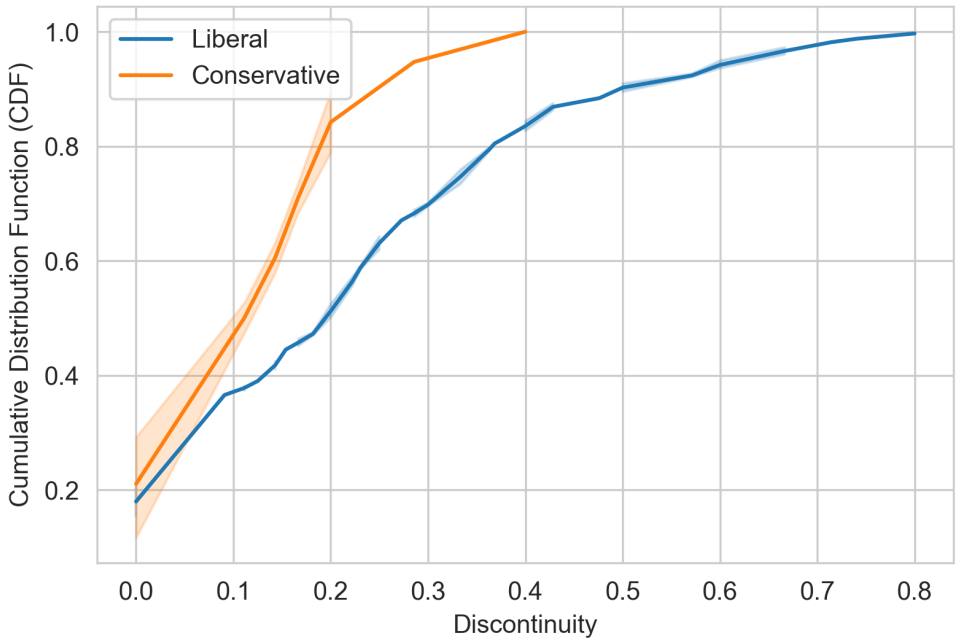

Figure 2.2: Distribution of discontinuity measure for the two political groups. Discontinuity is defined as the number of shifts in verb tense of a narrative normalized by the total number of verbs.

phrases, we utilized constituency parsing to detect the primary, secondary, auxiliary, and modal verbs in the narratives, and then using a rule-based classification on the part-of-speech tags used in the Penn Treebank, we distinguished between past and present tenses in the data. Arguably, the perfective tenses are complex morphosyntactic constructions due to the multiplicity of their semantics and uses. In English, perfects are made of an auxiliary ("have," "be") followed by a past participle. In our categorization, present perfect and present perfect continuous

were classified as "present", and past perfect as "past". Continuous tenses were classified according to their respective auxiliary verbs ("is," "are," "was," "were"). Reducing all verb phrases tags to "present" and "past", we measured the number of verb tense shifts occurred at a clause level in the memory narratives of each participant. We define discontinuity as the normalized version of the number of shifts in verb tenses in a narrative. For example, in "My peers are mixed between joy and distress, although the majority of my friends are in distress. There were cheers in my dorm hall when it was announced that Trump won.", the discontinuity is measured 0.2 since there are five verbs and one tense shift.

2.5.3 Discussion

We try to ground our study upon the rejection of dichotomies and demarcations on which modern social sciences are founded and demonstrate a willingness to search for a new paradigmatic framework for the interpretation of affective memory and its function from an empirically social and linguistic perspective.

Admittedly, the shifts examined in our preliminary study do not appear to be part of a deliberate presentational style used as a rhetorical device. Rather, the occurrence of verb tense shifts is influenced by underlying socio-political and psychological characteristics of the event, namely the 2016 presidential election and the participants. Accordingly, our findings suggest that the temporal discontinuity in narrative characterized by the shifts in verb tense is indicative of the emotional salience of the described experience [Brown and Kulik, 1977, Chafe, 1990, Neisser, 1982] and correlates directly with the calculated political leanings

of the studied collective. The sentiment analysis of the provided adjectives by which the emotionality of the participants' experiences upon learning the presidential election results was expressed corroborates the heightened affect. The intrusion of the present tense in autobiographical memory narratives is suggestive of a lucid description of perceptual experiences of live quality [Pillemer et al., 1998]. Indeed, the inadvertent slips into the present at emotionally and perceptually salient points support the idea of the multiple-leveled, language-based, narrative level of memory representation [Brown and Kulik, 1977, Spence, 1984]. According to Pillemer et al., verb tense shifts in memory narratives point to the existence of "functionally distinct but interacting representational systems". They continue that present tense autobiographical accounts may occur when unusually affective and imagistic representations intrude into ongoing, purposeful narrative processing.

Writing a memory narrative entails not only the expression of affect but the re-experience of it. As Probyn [2010] states, as one writes "affects can seem to get into [their] bodies". This act of remembering, in conjunction with writing, always involves a heightened intensity of experience [Richardson, 2013].

2.5.4 Limitations

It should be noted that this study is not without limitations. The majority of participants are male, and the data is small in size. As with all studies that utilize similar methodologies, small sample sizes may lead to falsity. Also, identifying the full complement of factors that prompts a present tense intrusion into the

recounting of a past episode is a task for future research.

2.6 Conclusion

In this paper, we focus on exploring the interconnectivities of diverse fields and searching for linguistic constructs capable of recruiting empirical and discursive support. In addressing Q_1, we conclude that the intensity of emotions, or subjective experiences, are manifested in language use within and beyond morphosyntactic structures. Correspondingly, in autobiographical memory narratives of the studies group, the verb tense shifts are being used discursively, and the affective dimensions of their subjectivity and objective reality are clearly reflected in their language use. Moreover, as an answer to Q_2, we find a distinct link between the political leaning and the level of discontinuity, indicating higher levels of emotional salience. We believe in the absence of such an account, the representation and analysis of emotionality in the discourse of memory narratives would be negligent of the underlying morphosyntactic constructs in language.

Chapter 3

Sarcasm Detection

Chapter Overview[1]

Sarcasm detection is the task of identifying irony[2] containing utterances in sentiment-bearing text. However, the figurative and creative nature of sarcasm poses a great challenge for affective computing systems performing sentiment analysis. This article compiles and reviews the salient work in the literature of automatic sarcasm detection. Thus far, three main paradigm shifts have occurred in the

[1]The content of this chapter is a reprint of the papers:

H. Yaghoobian, H. Arabnia and K. Rasheed. "Sarcasm Detection: A Comparative Study" arXiv e-prints (2021): arXiv:2107.02276.

H. Yaghoobian, S. Rezayi, and K. Rasheed. "Neural Discourse-level Sarcasm Analysis Using Rhetoric Structure Theory", To be submitted to the 2nd Workshop on Computational Approaches to Discourse (CODI '21).

[2]Irony is considered an umbrella term that also covers sarcasm; distinguishing between these two rhetoric devices is a further challenge for figurative language processing [Farías et al., 2016]. In short, sarcasm often bears an element of scorn and derision that irony does not [Lee and Katz, 1998].

way researchers have approached this task: semi-supervised pattern extraction to identify implicit sentiment, use of hashtag-based supervision, and incorporation of context beyond target text. In this chapter, first, I provide a comprehensive review of the datasets, approaches, trends, and issues in sarcasm and irony detection and how this investigation has helped me identify gaps that need addressing. Next, in Section 3.2, our proposed novel neural discourse-level model for sarcasm detection is presented.

3.1 Background and Literature Review

Sarcasm poses a major challenge for sentiment analysis models [Liu et al., 2010], mainly because sarcasm enables a speaker or writer to conceal their true intention of contempt and negativity under a guise of overt positive representation. Thus, recognizing sarcasm and verbal irony is critical for understanding people's actual sentiments and beliefs [Maynard and Greenwood, 2014]. The figurativeness and subtlety inherent in its sentiment display a positive surface with contemptuous intent (*e.g.*, *"He has the best taste in music!"*) or a negative surface with an admiring tone (*e.g.*, *"She always makes dry jokes!"*) makes the task of its identification a challenge for both humans and machines.

Sarcasm and irony are well-studied phenomena in linguistics, psychology, and cognitive science. In this article, we do not survey the several representations and taxonomies of sarcasm in linguistics [Camp, 2012, Campbell and Katz, 2012, Eisterhold et al., 2006, Ivanko and Pexman, 2003, Wilson, 2006], and focus on a

descriptive account of the computational attempts at automatic sarcasm detection. Empirical studies of this linguistic device refer to the approaches to predict if a given user-generated text is sarcastic or not. From a computational perspective, this task is formulated as a *binary classification* problem. Previous research on automated sarcasm detection has primarily focused on lexical, pragmatic resources [Kreuz and Caucci, 2007] along with interjections, punctuation, sentimental shifts, etc., found in sentences. Nonetheless, sarcasm is often manifested implicitly with no expressed lexical cues. Its identification is reliant on common sense and connotative knowledge that come naturally to most humans but makes machines struggle when extra-textual information is essentially required. Sarcastic utterances are often expressed in such nuanced ways that should be distinguished from a similar phenomenon called *humble-bragging*, which is a self-representational verbal strategy that appears as a complaint concealed within a bragging [Wittels, 2012], as in *"I am a perfectionist at times, it is so hard to deal with"*. To the best of our knowledge, there have been few computational studies that distinguish sarcasm from humble-bragging.

In the following, we split the literature along two main foci, content- 3.1.1 and context-based 3.1.2 models, and then classify empirical approaches to sarcasm detection within each section into rule-based, statistical, and deep learning-based approaches.

3.1.1 Content-based models

Models investigated in this section base their identification of sarcasm on lexical and pragmatic indicators in English [3] language use in social media. There is a myriad of novel and intuitive attempts in the literature that fall in this category. We review and categorize studies in this section based on approaches 3.1.1 (rule-based, semi-supervised and unsupervised), and features 3.1.1 (n-gram, sentiment, pragmatics, and patterns) used. We go over the baselines achieved in the literature in section 3.1.1.

Rule-based

Rule-based attempts look for evidence and indicators of sarcasm and rely on those in forms of rules. Veale and Hao [2010] looks for sarcastic similes (*e.g.*, "as private as a park-bench") in the specific query pattern of *"as * as a *"* on Google and using a nine-step approach reveal that 18% of unique similes are ironical.

Hashtags (or their equivalent, given the social media platform) have been utilized by users to denote sarcasm on Twitter (*e.g.*, #sarcasm, #not) or on Reddit (*e.g.*, /s). Or similarly, if the sentiment of a hashtag does not comply with the rest of the sentence, it is labeled as sarcastic.

Bharti et al. [2015] use a combination of two approaches in their study of sarcasm. They propose a parsing algorithm that looks for sentiment-bearing sit-

[3] Most research in sarcasm detection exists for English. However, some research in the following languages has also been reported: Chinese, Italian, Czech, Dutch, Greek, Indonesian, and Hindi.

uations and identifies sarcasm in forms of a contradiction of negative (or positive) sentiment and positive (or negative) situation. They also look for the co-occurrence of interjection hyperbolic words like *"wow"*, *"yay"*, etc. at the start of tweets, and intensifiers like *"absolutely", "huge" e.g.*, "Wow, that's a huge discount, I'm not buying anything!! #sarcasm."

Similarly, [Riloff et al., 2013] identify a positive/*negative* contrast between a sentiment and a situation helpful and indicative of sarcasm, *e.g.*, "I'm so pleased mom *woke me up* with vacuuming my room this morning. :)". Likewise, Van Hee et al. [2018b] speculated that sentiment incongruity within an utterance signifies sarcasm. To this end, they gathered all real-world concepts that carried an implicit sentiment and labeled them with either a "positive" or "negative" sentiment label *e.g.*, "going to the dentist" is often associated with a negative sentiment. Although their model did not surpass the baseline, they highlighted the difficulty and importance of incorporating sarcasm detection into sentiment classifiers. They view their efforts in the extension of seminal work by Greene and Resnik [2009] to use a concept called *syntactic packaging* to demonstrate the influence of syntactic choices on the perceived implicit sentiment of news headlines.

One of the earliest work is Tepperman et al. [2006]'s that identifies sarcasm in spoken dialogues and relies heavily on cues like laughter, pauses, speaker's gender, and spectral features; their data is restricted to sarcastic utterances that contain the expression 'yeah-right'. Carvalho et al. [2009] improve the accuracy of their sarcasm model by using oral or gestural clues in user comments, such as emoticons, onomatopoeic expressions (*e.g.*, *achoo, haha, grr, ahem*) for laughter,

heavy punctuation marks, quotation marks, and positive interjections. Davidov et al. [2010] and Tsur et al. [2010] utilize syntactic and pattern-based linguistic features to construct their feature vectors. Barbieri et al. [2014] take a similar approach and extend previous work by relying on the inner structure of utterances such as unexpectedness, the intensity of the terms, or imbalance between registers.

Feature sets

In this section, we go over the salient textual features utilized toward the detection of sarcasm. Most studies use bag-of-words to an extent. Nonetheless, in addition to these, the use of several other sets of features have been reported. Table 3.1 summarizes the main content-based features most commonly used in the literature. We discuss contextual features (*i.e.*, features reliant on the codification of information presented beyond text) in Section 3.1.1.

Reyes et al. [2012] introduce a set of humor-dependent or irony-dependent features related to ambiguity, unexpectedness, and emotional scenario. Ambiguity features cover structural, morphosyntactic, semantic ambiguity, while unexpectedness features gauge semantic relatedness. As we discussed in Section 3.1.1, Riloff et al. [2013], in addition to a rule-based classifier, use a set of patterns, specifically positive verbs and negative situation phrases, as features. Liebrecht et al. [2013] use bigrams and trigrams and similarly, Reyes et al. [2013] look into skip-gram and character-level features. In a kindred effort Ptáček et al. [2014] use word-shape and pointedness features. Barbieri et al. [2014] includes seven sets of features such as maximum/minimum/gap of intensity of adjectives and adverbs,

max/min/average number of synonyms and synsets for words in the target text, and so on. Buschmeier et al. [2014] incorporates ellipsis, hyperbole, and imbalance in their set of features. Joshi et al. [2015] uses features corresponding to the linguistic theory of incongruity. The features are classified into two sets: implicit and explicit incongruity-based features.

Studies	**Features Used**
Reyes et al. [2012]	Structural, morphosyntactic and semantic ambiguity features
Tsur et al. [2010]	Internal syntactic patterns and punctuations
González-Ibánez et al. [2011]	User mentions (replies), emoticons, N-grams, dictionary- and, sentiment-lexicon-based features
Liebrecht et al. [2013]	N-grams, emotion marks, intensifiers
Hernández-Farías et al. [2015]	Length of tweet, capitalization, punctuation marks, and emoticons
Farías et al. [2016]	Lexical markers and structural features,
Mishra et al. [2016]	Cognitive features extracted from eye-movement patterns of human readers
Joshi et al. [2016]	Features based on word embedding similarity

Table 3.1: Features used for Statistical Classifiers

Mishra et al. [2016] propose a novel approach for investigating the salient features of sarcasm in text. They designed a set of gaze-based features such as average fixation duration, regression count, skip count, etc., based on annotations from their eye-tracking experiments. In addition, they also utilize complex gaze features based on saliency graphs, created by treating words as vertices and saccades (*i.e.*, quick jumping of gaze between two positions of rest) between a pair of words as edges.

Learning-based methods

In the following, we delve more into supervised learning, semi-supervised learning, unsupervised learning, structural and hybrid learning. A brief descriptive account of these approaches toward predictive sarcasm identification in text is given below.

Supervised learning In traditional machine learning approaches, most work on statistical detection of sarcasm has relied on various combinatory forms of Random Forests (RF), Support Vector Machines (SVM), Decision trees (DT), Naïve Bayes (NB) and Neural Networks (NN) Davidov et al. [2010], Joshi et al. [2015, 2016], Kreuz and Caucci [2007], Reyes and Rosso [2012], Tepperman et al. [2006], Tsur et al. [2010]. For instance, González-Ibánez et al. [2011] uses SVM with sequential minimal optimization (SMO) and Logistic Regression (LogR), which are usually used toward sentiment analysis, to identify discriminating features. Riloff et al. [2013] utilized a hybrid SVM system that outperformed the SVM classifier. Similarly, the use of balanced winnow algorithms to determine high-ranking features Liebrecht et al. [2013] and Naive Bayes and Decision Trees for multiple pairs of labels among irony, humor, politics, and education [Reyes et al., 2013] and fuzzy Clustering for sarcasm detection [Mukherjee and Bala, 2017] are reported. Bamman and Smith [2015] present the use of binary Logistic Regression and SVM-HMM toward incorporating the sequential nature of output labels into a conversation. Likewise, Joshi et al. [2015] reports that sequence labeling algorithms are more useful for conversational data as opposed to classification methods. They use SVM-HMM and SEARN as the sequence labeling algorithms. Liu et al. [2014] present a multi-strategy ensemble learning approach (MSELA) including Bagging, Boosting, etc., to handle the imbalance between sarcastic and non-sarcastic samples.

While rule-based approaches mostly rely upon lexical information and require no training, machine learning invariably makes use of training data and exploits

different types of information sources (or features), such as bags of words, syntactic patterns, sentiment information or semantic relatedness. Earliest attempts in this line use similarity between word embeddings as features for sarcasm detection. For instance, Joshi et al. [2016] in which word embeddings are then augmented based on their similarity. Ghosh and Veale [2016] use a combination of convolutional neural networks, LSTM followed by a DNN. Van Hee et al. [2018a] propose a model that identifies sarcastic tweets and subsequently differentiate the type (out of four classes) of expressed sarcasm. The systems that were submitted for both subtasks represent a variety of neural-network-based approaches (*i.e.*, CNNs, RNNs, and (bi-)LSTMs) exploiting word and character embeddings as well as handcrafted features.

Semi-supervised learning This form of machine learning, which falls between unsupervised learning and supervised learning, uses a minimal quantity of annotated (labeled) data and a large amount of un-annotated (unlabelled) data during training [Tsur et al., 2010]. The presence of the unlabelled datasets and the open access to the unlabelled datasets is the feature that differentiates the semi-supervised from supervised learning. Davidov et al. [2010] employed a semi-supervised learning approach for automatic sarcasm identification using two different forms of text, tweets from Twitter, and product reviews from Amazon. A total number of 66,000 products and book reviews were collected in their study, and both syntactic and pattern-based features were extracted. The sentiment polarity of 1 to 5 was chosen on the training phase for each training data. The authors reported a performance of %77 precision.

Unsupervised learning Unsupervised learning in automatic sarcasm identification is still in its infancy, and most approaches are clustering-based, which are mostly applicable to pattern recognition. Nudged by the limitations and difficulties inherent in labeling the datasets (*i.e.*, time-, and labor-intensivity) in supervised learning methods, researchers seek to eliminate such exertions by focusing on the development of unsupervised models. Nozza et al. [2016] propose an unsupervised framework for domain-independent irony detection. They build on probabilistic topic models originally defined for sentiment analysis. These models are extensions of the well-known Latent Dirichlet Allocation (LDA) model [Blei et al., 2003]. They propose Topic-Irony model (TIM), which is able to model irony toward different topics in a fully unsupervised setting, enabling each word in a sentence to be generated from the same irony-topic distribution. They enrich their model with a neural language lexicon derived through word embeddings. In a similar attempt, Mukherjee and Bala [2017] utilize both supervised and unsupervised settings. They use Naïve Bayes for supervised and Fuzzy C-means (FCM) clustering for unsupervised learning. Justifiably, FCM does not perform as effectively as NB.

3.1.2 Context-based models

Making sense of sarcastic expressions is heavily reliant on the background knowledge and contextual dependencies that are formally diverse. As an example, a sarcastic post from Reddit, "I'm sure Hillary would've done that, lmao." requires prior knowledge about the event, *i.e.*, familiarly with Hillary Clinton's perceived

habitual behavior at the time the post was made. Similarly, sarcastic posts like "But atheism, yeah *that's* a religion!" require background knowledge, precisely due to the nature of topics like *atheism* which is often subject to extensive argumentation and is likely to provoke sarcastic construction and sarcastic interpretation. The proposed models in this section utilize both content and contextual information required for sarcasm detection. In addition, there has been a growing interest in using neural language models for pre-training for various tasks in natural language processing. We go over the utilization of existing language models *e.g.*, BERT, XLNet, etc. toward sarcasm detection in section 3.1.2.

Wallace et al. [2014] claim that human annotators consistently rely on contextual information to make judgments regarding sarcastic intent. Accordingly, recent studies attempt to leverage various forms of contextual information mostly external to the utterance, toward more effective sarcasm identification. Intuitively, in the case of Amazon product reviews, knowing the type of books an individual typically likes might inform our judgment: someone who mostly reads and reviews Dostoevsky is statistically being ironic if they write a laudatory review of Twilight. Evidently, many people genuinely enjoy reading Twilight, and so if the review is written subtly, it will likely be difficult to discern the author's intent without this preferential background. Therefore, Mukherjee and Bala [2017] report that including features independent of the text leads to ameliorating the performance of sarcasm models. To this end, studies take three forms of context as feature: author context [Bamman and Smith, 2015, Hazarika et al., 2018], conversational context [Wang et al., 2015], and topical context [Ghosh and

Veale, 2017]. Another popular line of research utilizes user embedding techniques that encode users' stylometric and personality features to improve their sarcasm detection models Hazarika et al. [2018]. Their model, CASCADE, utilizes user embeddings that encode stylometric and personality features of the users. When used along with content-based feature extractors such as Convolutional Neural Networks (CNNs), a significant boost in the classification performance on a large Reddit corpus is achieved. Similarly to how a user controls the degree of sarcasm in a comment, they extrapolate that the ensuing discourse of comments belonging to a particular discussion forum contains contextual information relevant to the sarcasm identification. They embed topical information that selectively incurs bias towards the degree of sarcasm present in the comments of a discussion. For example, comments on political leaders or sports matches are generally more prone to sarcasm than natural disasters. Contextual information extracted from the discourse of a discussion can also provide background knowledge or cues about that discussion topic. To extract the discourse features, they take a similar approach of document modeling performed for stylometric features.

Agrawal et al. [2020] formulate the task of sarcasm detection as a sequence classification problem by leveraging the natural shifts in various emotions over the course of a piece of text. Li et al. [2020] propose a semi-supervised method for contextual sarcasm detection in online discussion forums. They adopt author and topic sarcastic prior preference as context embedding that provides a simple yet representative background knowledge. Nimala et al. [2020] also propose an unsupervised probabilistic relational model to identify common sarcasm topics

based on the sentiment distribution of the words in the tweets.

Sarcasm detection using pre-trained language models

Given the highlighted importance of context to capture figurative language phenomena and the difficulties of data annotation, transfer learning approaches are gaining attention in various domain adaptation problems. In particular, the utilization of pre-trained embeddings such as Global Vectors (GloVe) [Pennington et al., 2014], and ELMo [Peters et al., 2018] or leveraging Transformer seq2seq methods such as BERT (Bidirectional Encoder Representations from Transformers Devlin et al. [2019], RoBERTa [Liu et al., 2019b], and XLNet [Yang et al., 2019], etc. are witnessing a surge.

Potamias et al. [2020] propose Recurrent CNN RoBERTA (RCNN-RoBERTa), a hybrid neural architecture building on RoBERTA architecture, which is further enhanced with the employment and devise of a recurrent convolutional neural network. They report a performance with an accuracy of %79 on SARC dataset [Khodak et al., 2018]. Similarly, Dadu and Pant [2020] use an ensemble of RoBERTa and ALBERT [Lan et al., 2019] on *Get it #OffMyChest* dataset [Jaidka et al., 2020] achieve a performance of %85 accuracy with $F1$ score of 0.55. Javdan et al. [2020] use BERT along with aspect-based sentiment analysis to extract the relation between context dialogue sequence and response. They obtain an F1 score of 0.73 on the Twitter dataset and 0.73 over the Reddit dataset[4]. We expect to

[4]Twitter and Reddit datasets used for in this study were provided in the shared task on Sarcasm Detection, organized at Codalab.

see more studies geared toward leveraging pre-trained contextual embeddings and transformers toward sarcasm detection in the upcoming years.

System	**Acc**	**F1**
ELMo	0.70	0.70
Wang and Manning [2012] (NBSVM)	0.65	0.65
XLnet	0.76	0.76
BERT-cased	0.76	0.76
RoBERTa	0.77	0.77
Hazarika et al. [2018] (CASCADE)	0.74	0.75
Ilic et al. [2018]	0.79	-
Khodak et al. [2018]	0.77	-
Potamias et al. [2020] (RCNN-RoBERTa)	0.79	0.78

Table 3.2: State-of-the-art NN classifiers and results on Reddit Politics dataset

3.1.3 Datasets

This section outlines the datasets used for computational studies on sarcasm detection. Commonly, they are divided into three categories short text (*e.g.*, Tweets, Reddits), long text (*e.g.*, discussions on forums), transcripts (*e.g.*, conversational transcripts of a TV show or a call center). Short text can contain only one (possibly sarcastic) utterance, whereas long text may contain a sarcastic sentence among other non-sarcastic sentences that could potentially function as context.

Short text

This category of data is the dominant form of expression on Social Media, mostly as a direct result of restriction on text length. Consequently, this type of text is rife with abbreviations to make efficient use of space on platforms such as Twitter. Two main approaches are utilized toward annotation of tweets: Manual and hashtag-based. Maynard and Greenwood [2014], Mishra et al. [2016], Ptáček et al. [2014], Riloff et al. [2013] introduce a manually annotated dataset of sarcastic utterances. Most annotation approaches in the literature are conducted using hashtags to create labeled datasets. Sarcastic intent in English is commonly and culturally communicated using hashtags such as #sarcasm, #sarcastic, #not. Davidov et al. [2010], González-Ibánez et al. [2011], Reyes et al. [2012] use hashtag-based datasets of tweets. Liebrecht et al. [2013] only uses #not to collect and label their tweets. While collecting sarcastic tweets using this method is undemanding, the inclusion of non-sarcastic tweets can be challenging since tweets containing #notsarcastic may not represent a general non-sarcastic text [Bamman and Smith, 2015]. Another approach is to collect the non-sarcastic tweets of users whose sarcastic tweets are also present in the dataset. To ensure collection of true sarcasm, some studies like Fersini et al. [2015] manually verified the initial hashtag-based tweets using annotators.

Reddit is the other popular platform for researchers to collect sarcasm using hashtag "/s" (Reddit's equivalent of "#sarcasm" on Twitter). Khodak et al. [2018] present SARC, a large-scale self-annotated corpus for sarcasm that contains more than a million examples of sarcastic/non-sarcastic statements made on Reddit.

Long text

Lukin and Walker [2013] uses the Internet Argument Corpus (IAC) [Walker et al., 2012] which contains a set of 390, 704 posts in 11, 800 discussions extracted from the online debate site 4forums.com, annotated for several dialogic and argumentative markers, one of them being sarcasm. Reyes and Rosso [2014] collect a dataset of movie and book reviews, along with news articles marked with sarcasm and sentiment. In an earlier study, Reyes and Rosso [2012] garner 11,000 reviews of products with sarcastic expressions. Filatova [2012] present a corpus generation experiment where they collect regular and sarcastic Amazon product reviews. This resulting corpus can be used for identifying sarcasm on two levels: a document and a text utterance, where a text utterance can be as short as a sentence and as long as a whole document.

Transcripts & dialogues

Sarcasm is often expressed in the context of a conversation, as a response projecting contemptuous intent. Tepperman et al. [2006] uses 131 call center transcripts to look for occurrences of "yeah right" as a marker of sarcasm. Similarly, Rakov and Rosenberg [2013] through crowd-sourcing collect sentences from an MTV show called "Daria." Similarly, Joshi et al. [2016] present a manually annotated transcript of the popular sitcom "Friends."

3.1.4 Discussion

Sarcasm detection research has seen a significant surge in interest in the past few years, which justifies a thorough investigation. This article centers on automatic approaches sarcasm detection in text. We identify three discernible paradigms shifts in the history of sarcasm detection research: 1) the use of hashtag-driven supervised learning toward building annotated datasets, 2) semi-supervised pattern extraction to identify implicit sentiment, and 3) the utilization of extra-textual information as context (*e.g.*, user's characteristic profiling). Whilst rule-based approaches attempt to capture any indication of sarcasm in the form of rules, statistical methods use features like shifts in sentiment, specific semi-supervised patterns, etc. Deep learning techniques have also been used to incorporate context, *e.g.*, additional stylometric features of authors in conversations and the nature of discussion topics. An underlying theme of these past approaches (either in terms of rules or features) is predicated on sarcasm's contemptuous nature. Novel techniques to incorporate contextual insight have also been explored, mostly centered on the emerging direction toward utilizing language models.

3.2 Experiments

Confronted with the challenges the figurative and creative nature of sarcastic utterances poses affective systems performing sentiment analysis. This paper presents a truly discourse-level model for sarcasm detection that functions autonomously from a user's stylometric and personality features. Our model ex-

tracts the contextual discursive information in a fully impersonal fashion with no learned bias toward a specific user and aims to serve as a useful corrective to the currently existing sarcasm classification models that rely heavily on user embeddings. The proposed approach consists of three stages: First, Elementary Discourse Units (EDUs) and discourse relations among them are obtained by a discourse parser. Second, distributed representations are obtained using attentive recurrent neural networks. Finally, we apply the proposed model to a benchmark dataset, namely SARC, to perform document-level binary classification and show the effectiveness of the proposed model.

3.2.1 Motivation

Recognizing sarcasm and the verbal irony is critical for understanding people's actual sentiments and beliefs Maynard and Greenwood [2014]. The figurativeness and subtlety inherent in its sentiment that displays a positive surface yet with contemptuous intent make its identification a challenge for both humans and machines. Empirical studies of this linguistic phenomenon refer to the computational approaches to predict if a given user-generated text is sarcastic. Previous research in automated sarcasm detection has primarily focused on lexical, pragmatic resources Kreuz and Caucci [2007] along with interjections, punctuation, sentimental shifts, etc., found in sentences. Nonetheless, sarcasm is often manifested in an implicit manner with no expressed lexical cues. Making sense of such expressions is heavily reliant on background knowledge and contextual dependencies, which

are formally diverse. As an example, a sarcastic post from Reddit[5], *"I'm sure Hillary would've done that, lmao."* requires prior knowledge about the event, i.e., a familiarly with Hillary Clinton's perceived habitual behaviour at the time the post was made. Similarly, sarcastic posts like *"But atheism, yeah *that's* a religion!"* requires the knowledge that topics like *atheism* which is often subject to extensive argumentation and likely to provoke sarcastic construction and sarcastic interpretation.

Motivated by the potential usefulness of the contextual information, we propose a neural discourse-level model using Rhetoric Structure Theory (RST) for the task of sarcasm classification in online discussion forums. Our proposed model utilizes both *content* and *contextual* information required for sarcasm detection. The latent semantics of a document is learned through a hierarchical recurrent attention network that models representation of word sequences at both word and discourse unit level. In Rhetorical Structure Theory (RST) Mann and Thompson [1988], the semantics of a text is formed from smaller segments of the text and their discursive roles such as concession, conjunction, etc. These roles, also referred to as discourse relations, are paratactic or hypotatic relations that hold across two or more text spans, each of which is considered as nucleus or satellite by its role in the discourse relation. In this manner, the discourse structure built from discourse relations among sentences plays the key role in determining the semantics of a document, *i.e.*, the discourse. An example of an RST tree is shown in Figure 3.1.

[5]https://www.reddit.com/

The main thrust of the work presented here lies in the incorporation of an RST into a hierarchical recurrent attention network that learns the leaf node embedding of the discourse tree, known as Elementary Discourse Units (EDUs), for the purpose of learning the semantics of a document. We examine the performance of the model by applying it to document-level sarcasm detection task and compare the results against the state-of-the-art solutions discussed earlier in Section 3.

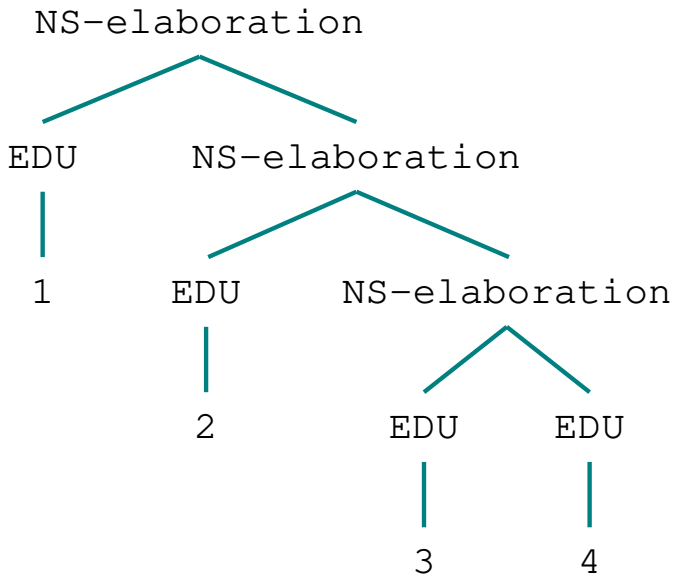

Figure 3.1: An example of a discourse tree with 4 EDUs and 3 relations. The actual sentence for this tree is: (Policeman Doesn't Like His Picture Taken: "I'm gonna fucking break your face". EDU1) (This is why we need laws EDU2) (to prevent cops EDU3) (from being filmed. EDU4)

In what follows, we review the previous works in two related empirical areas of document classification using RST and sarcasm detection. Section 3.3 shows how the generated discourse trees from the RST parser serve as an input to our novel methods for sarcasm detection. In the next section 3.4, we explain the used dataset and demonstrate the results. Finally, section 3.5 provides an explanatory

account of our failed attempts and future direction ahead.

In this section, we only introduce the previous efforts on the downstream task of text classification using RST parser since an extensive account of the previous efforts on sarcasm detection has already been provided in previous sections.

3.2.2 Document Classification Using RST

RST structures documents hierarchically by splitting the content into (sub-)clauses called Elementary Discourse Units (EDUs). The EDUs are then connected to form a binary discourse tree. RST discriminates between a nucleus, which conveys primary, and satellite, which conveys ancillary information. In this part, we go over the studies in which the tree structure of discourse is utilized towards improvement of classification model performance. Taboada et al. [2008] are of the earliest attempt at incorporating rhetorical structure theory and sentiment analysis, whose model wights up adjectives in a nucleus more heavily than those in a satellite. Others experiments with relation type Hogenboom et al. [2015] and depth Märkle-Huß et al. [2017] of the discourse tree. While these references provide a reductionist representation of discourse tree by dropping partial information from the tree, Kraus and Feuerriegel [2019] offer *Discourse-LSTM* by building upon the work of Tai et al. [2015]. Their model, a tensor-based, tree-structured deep neural network, processes the discourse trees in their entirety without recourse to pruning at certain thresholds to yield a tree of fixed depth.

3.3 Method

The objective of this work is to develop a discourse-aware model for sarcasm detection that is capable of identifying the subtleties and differences in salience between individual sub-ordinate clauses and discriminating the relevance of sentences based on their function (e.g., whether it introduces a new fact or elaborates upon or contradict an existing one). To this end, we consider the following criteria to design our model. We want our model to be:

1. considerate of the sequential nature of the input: it is a well-established fact that there is a short term dependency in textual data that can be exploited to improve language modeling performance. In the neural setting this dependency can be captured by Long Short Term Memory (LSTM) networks.

2. attentive to discourse units: attention mechanism is another powerful technique in deep learning that learns a score for each unit of text based on its importance in the entire document. Word level and sentence level attention has been explored in the literature Yang et al. [2016], here we argue that EDU-level attention could inform the way the representation of a document, *i.e.*, a discourse tree, is constructed.

3. mindful of latent tree structure of documents: learning the latent tree structure of sentences has been proven to be effective Kraus and Feuerriegel [2019], Tai et al. [2015]. As the target of our study is sarcastic documents,

we argue that we can manipulate the discourse tree, instead of constituency tree, to successfully learn this hidden variable.

4. heedful of relations among discourse units: in addition to the tree structure of the sentences, nodes in the discourse tree are related to each other by relations which carry the semantic properties between segments (EDUs) of the sentence. These relations are common to most theories of discourse structure such as RST.

In this study, we design a hierarchical attention-based recurrent model to address the first and the second criteria, see subsection **??**. We also discuss our attempts and ideas for the last two criteria in Section 3.5.

3.3.1 Model

We use DPLP discourse parser as a tool Ji and Eisenstein [2014] to produce RST trees. DPLP itself employs Stanford CoreNLP package Manning et al. [2014] to perform syntactic parsing. After generating discourse trees, we group sentences into word-level and EDU-level structures and use a hierarchical attention model to obtain document-level representation. An illustration of the model is shown in Figure 3.2.

Our model uses GloVe Pennington et al. [2014] to obtain initial embeddings for each word, then a GRU layer is employed to capture short term dependency among words, $h_{it} = \text{GRU}(x_{it})$. Next, h_{it} is fed to a fully connected layer to obtain another low-dimensional representation,

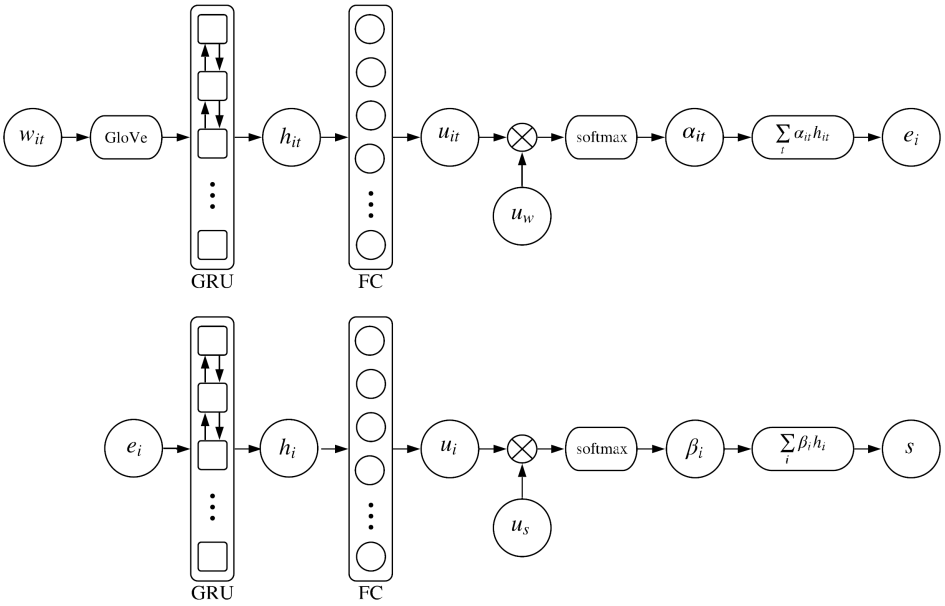

Figure 3.2: Hierarchical attention mechanism to obtain document-level low rank representation.

$$u_{it} = \tanh(W_w h_{it} + b_w)$$

Similar to Yang et al. [2016], we introduce a context vector, u_w that will be learned and updated during the training and is used to calculate the attention score per token:

$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum \exp(u_{it}^\top u_w)}$$

Finally, we obtain the EDU-level embedding as a weighted sum of all the word embeddings:

$$e_i = \sum_t \alpha_{it} h_{it}$$

Similarly, we repeat this process one more time to obtain the document-level embedding and use them in binary classification. This process is demonstrated as follows:

$$\begin{aligned} h_i &= \text{GRU}(x_i) \\ u_i &= \tanh(W_e h_i + b_e) \\ \alpha_i &= \frac{\exp(u_i^\top u_s)}{\sum \exp(u_i^\top u_s)} \\ s &= \sum_i \alpha_t h_i \end{aligned}$$

where s is the final embedding per document.

3.4 Results

Dataset: We use the Self-Annotated Reddit Corpus (SARC) Khodak et al. [2018] which contains 1.3 million sarcastic statements. The sentences in this corpus are self annotated by their authors, meaning they are tagged with a special character, "/s", which implies expressions with sarcastic intention. In this study, we use a subset of the dataset which belongs to the politics subreddit and contains 140K comments. In terms of labeling, the dataset is balanced with 50% sarcastic and 50% non-sarcastic comments.

Method	Politics
BoW	0.60
Poria et al., 2016	0.67
Amir et al., 2016	**0.69**
CASCADE	**0.69**
CASCADE (with personal info)	0.75
SVM	0.58
Our model (GRU)	0.65
Our model (GRU+Attn)	**0.69**

Table 3.3: F1 score comparison chart. Values for other methods are directly reported from their corresponding papers.

Experimental Setup: We use 80% of the dataset to train the model, and tune the hyperparameters on 10% of the dataset and test the trained model on the remaining 10%. We implement our code using the Keras library. We use Mean Squared Error (MSE) loss function and optimize it using Adam optimizer in 100 epochs, and we obtain our best result after 10 epochs.

Results: The result of our experiment is presented in Table 3.3. As this table reveals, we obtain the state-of-the-art performance (F1 score = 69%). Also, we see that EDU-level attention improves the F1 measure by 4%.

Discussion: Table 3.3 also shows that CASCADE Hazarika et al. [2018] obtains $F1 = 75\%$ when considering personal information. In order to obtain personal information, they consider all comments of a user and compute a sarcastic score per user; they further incorporate this score in their comment-level classification. We claim that this approach leads to a bias toward the user.

3.5 Further Experimentation and Conclusion

In this section, we discuss what other approaches we implemented and what areas of improvement are available to further improve the performance and quality of this work.

As discussed in Section 3.3, we want our model to learn the latent tree structure of the sentences. To this end, we construct a graph data structure per discourse tree as follows (Note that in the original discourse tree, relations are not nodes, rather they are links between two EDUs or one EDU and a collection of multiple EDUs): We consider all EDUs and relations as nodes. All EDUs are directly connected to their immediate relations, and relations are either connected to two EDUs or one EDU and one relation. In terms of shape, Figure 3.1 is an intuitive visualization. After constructing these trees, we feed them to a Graph Convolutional Network (GCN) Kipf and Welling [2017] unit to obtain embeddings per EDU and relation. Finally, we take an average across all node embeddings to obtain the embedding of the document and use these embeddings to define an l2-norm loss as follows:

$$\mathcal{L}_{\text{EMB}} = ||\mathbf{Z}_S - \mathbf{Z}_G||^2$$

where Z_S and Z_G are the embedding matrices obtained from sequential and graph convolutional models, respectively. Finally we minimize the total loss, $\mathcal{L}$ which is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{GCN}} + \lambda_1 \mathcal{L}_{\text{GRU}} + \lambda_2 \mathcal{L}_{\text{EMB}}$$

where $\mathcal{L}_{GCN}$ and $\mathcal{L}_{GRU}$ are the default loss functions for GCN and GRU models, respectively, and λ_1 and λ_2 are coefficients that control the contribution of each loss to the total loss. This approach was not successful as GCN treats all nodes (and relations) the same while they have different purposes. Also, the way we applied GCN might be problematic because we input each small graph to GCN in a batch mode, while default GCN takes one huge graph as input and calculates the low dimensional representation per node. In fact, this is a graph classification problem, and one way to tackle this issue is to generate a large block diagonal adjacency matrix where each block is an adjacency matrix of a discourse tree. Besides, averaging across all embeddings while it is practical might not be a sound approach.

Another area of improvement is to customize GCN to our own problem. We know that GCN uses neighborhood information, adjacency matrix denoted as $\mathbf{A}$ and initial feature matrix denoted by $\mathbf{X}$ to calculate the output as follows:

$$\text{GCN}(\mathbf{A}, \mathbf{X}) = \tanh(\hat{\mathbf{A}}\mathbf{X}\mathbf{W} + \mathbf{b})$$

In our experiment, we initialized $\mathbf{X}$ as identity matrix, *i.e.*, $\mathbf{X} = \mathbf{I}$, while we can use textual information present in each node to enhance the feature matrix. Besides, the notion of Nucleus and Satellite in RST already employs different segments that receive different weights depending on their role. For instance,

in Figure 3.1, we see EDU3 and EDU4 are related to each other via *elaboration* relation. In this example, EDU3 is the Nucleus, and EDU4 is the Satellite, meaning that EDU3 is the salient segment. The proposed GCN model is ignorant of such information, and we believe incorporating this knowledge could boost performance.

To generalize our model, we plan to experiment with other sarcasm- and sentiment-annotated datasets (both short and long texts), such as Lukin and Walker [2013] that present the Internet Argument Corpus, Reyes and Rosso [2014] a dataset of reviews and news articles marked with sarcasm and sentiment and Liu et al. [2014] that use a dataset from multiple sources.

Chapter 4

Generative Language Models and Co-Creative Writing

Chapter Overview[1]

This chapter provides a comparative study of the most recent language representation models, and their achieved performances on diverse language modeling benchmarks. We focus on the notable work of Devlin et al. [2019] known as **BERT** (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) and three of its main successors namely, RoBERTa [Liu et al., 2019b], ALBERT [Lan et al., 2019], and

[1]The content presented in this chapter is based on and in part a reprint of the papers:

H. Yaghoobian, S. Voghoei, H. Arabnia and K. Rasheed. "A Short Review of BERT and Post-BERT Methods: RoBERTa, ALBERT and XLNet" Submitted to Computer Science, Computer Engineering, & Applied Computing, 2021.

E. Kuecker, and H. Yaghoobian. "When Word Processing was a Focal Practice", Humanist Studies & the Digital Age Journal, Under review.

XLNet [Yang et al., 2019]. From a historical and philosophical perspective, we also discuss issues surrounding the integration of these generative language models into model writing tools.

4.1 Introduction

Language modeling is a central area of focus in Natural Language Processing (NLP) and Understanding (NLU), which has witnessed significant advancements over the past decade and has undergone paradigm shifts in the formulation of the problem since the emergence of neural net models. In the new realm of practice and prior to ELMo [Peters et al., 2018] and BERT and other Mupperwares[2], global word representation techniques were limited to Word2Vec and GloVe [Pennington et al., 2014]. These distributional word representations trained on large-scale corpora in an unsupervised fashion were considered a major leap toward modeling the relationship between words, a feature that was not previously taken into account in known context-independent approaches such as Bag of Words (BoW) and TF-IDF. Nonetheless, Word2Vec and GloVe only capture a single global representation for each word and are dismissive of the contextual surroundings.

Neural models diminish the issue with *feature engineering* to a great extend. While non-neural language models are heavily reliant on hand-crafted features, neural models take advantage of low-dimensional and dense word vector representations to capture the language's implicit semantic and syntactic features. How-

[2]https://www.theverge.com/2019/12/11/20993407/ai-language-models-muppets-sesame-street-muppetware-elmo-bert-ernie

ever, at least up to a point in time, neural models were prone to overfitting largely due to the insignificance of the scale of training data and hence failed to generalize well. Substantial work has been put into these models over the years and development of computational power has made training on large bodies of text feasible. As a result, the first generation of pre-trained language models utilize the work of Vaswani et al. [2017] and aim to learn better word embeddings.

ELMo and BERT assign a representation based on the various contexts that a word has appeared within, across different domains and languages (*e.g.*, to model polysemy). BERT [Devlin et al., 2019] is an extension and a bidirectional variant of Transformer networks [Vaswani et al., 2017] trained to jointly predict a masked word from its context and to classify whether two sentences are consecutive or not. The trained models can be fine-tuned for downstream NLP tasks (both sentence-level and token-level) such as Semantic Analysis, Named Entity Recognition (NER), Word Sense Disambiguation (WSD), Question Answering (QA), and Language Inference (LI) without substantial modification.

Among the pre-training approaches, two of the most successful ones are autoregressive (AR) and autoencoding (AE) language modeling. BERT is a notable example of AE, which is based on denoising autoencoding. As an AE-based model, BERT has the capability of modeling bidirectional contexts by reconstructing the original text from corrupted input. However, as we explain further in the following sections, due to the long-range dependency characteristics in natural language, BERT oversimplifies the task of predicting the next token by assuming they are independent of each other. On the other hand, XLNet leverages the best of both

autoregressive (AR) language modeling and autoencoding (AE), while managing to avoid their limitations. AR-based models learn context unidirectionally, either in the forward or backward direction in a text sequence. However, AR-based models fall short when bidirectional context is required to be utilized simultaneously. The power of XLNet lies in the concomitant utilization of both AR and AE.

The rest of the article is organized as follows: In Section 4.2, we provide a brief account of the conceptual foundations and empirical frameworks that language models (LM) have utilized and built upon, and then Section 4.3 delves into pretraining methods for contextual embeddings. In Section 4.4 we discuss a few evaluative studies into the interpretability of BERT and its claims. Lastly, in Section 4.5 we offer our ethics-oriented analysis of large language models and their shortcomings.

4.2 Contextual Encoders: Architecture

The majority of neural contextual encoders are either sequence models or graph-based models. We succinctly provide a technical account of BERT and post-BERT models in this section.

4.2.1 Sequential Models

Sequence models essentially capture the local context of a given token in sequential order. They can be categorized into convolutional and recurrent models. Convolutional models take the word embeddings of the input sentence and capture a

representation of words by aggregating the local contextual information from its neighbors by convolution operations [Kim, 2014].

Recurrent models excel at capturing the contextual representations of words with short memory (LSTM) and gated recurrent units (GRU) [Chung et al., 2014]. Most LM models have utilized bi-directional LSTMs or GRUs to collect information from both sides of a word but remain strained because of the limitations long-term dependency imposes.

4.2.2 Graph-based and Non-sequential Models

Non-sequence models learn the contextual representation with a pre-defined tree or graph structure between words, such as the syntactic structures or semantic relations. Some popular non-sequence models include Recursive NN [Socher et al., 2013], TreeLSTM [Tai et al., 2015, Zhu et al., 2015], and GCN [Kipf and Welling, 2016]. Although the linguistic-aware graph structure can provide useful inductive bias, how to build a good graph structure is also a challenging problem. Besides, the structure depends heavily on expert knowledge or external NLP tools, such as the dependency parser.

In practice, a more straightforward way is to use a fully-connected graph to model the relation of every two words and let the model learn the structure by itself. Usually, the connection weights are dynamically computed by the self-attention mechanism, which implicitly indicates the connection between words. A successful instance of a fully-connected self-attention model is the Transformer [Vaswani et al., 2017], which also needs other supplement modules, such as po-

sitional embeddings, layer normalization, residual connections, and position-wise feed-forward network (FFN) layers. Table 4.1 provides a summary of pre-trained language models discussed in this article.

PLMs	Architecture	Task	Corpora	# Parameters	GLUE	Fine-tuning
ELMo	LSTM	BiLM	WikiText-103			No
BERT	Transformer Encoder	MLM+NSP	WikiEn+BookCorpus	110∼ 340M	81.9	Yes
RoBERTa	Transformer Encoder	MLM	BookCorpus+CCNews +OpenWebText+ STORIES	355M	88.5	Yes
XLNet	Two-Stream Transformer Encoder	PLM	WikiEn+ BookCorpus+Giga5 +ClueWeb+Common Crawl	110∼340M	90.5	Yes

Table 4.1: Pre-trained Language Models

4.3 Contextual Embeddings: Pre-training

Classic word embedding methods such as Word2vec [Mikolov et al., 2013] and Glove [Pennington et al., 2014] aim to learn a *global* word embedding matrix $E \in \mathbb{R}^{V \times d}$, where V is the vocabulary size and d is the number of dimensions. Therefore, the obtained non-contextual representations do not reflect the surrounding words (*i.e.*, context), making them not well-suited for context-dependent NLP tasks that require sequence-level semantics (*e.g.*, polysemy). In this section, we look into pre-training methods for learning contextual embeddings and divide them into Unsupervised methods 4.3.1 *e.g.*, language modeling and Supervised methods 4.3.2 *e.g.*, machine translation.

4.3.1 Unsupervised methods

In general, language modeling is a probability distribution over a sequence of tokens or words. Maximum Likelihood Estimation (MLE), penalized with regularization terms, is utilized toward estimating model parameters. Since language models are mostly trained on large-scale unlabelled data using neural conditional probabilistic models [Bengio et al., 2003], the obtained representations are transferable to downstream tasks.

In the literature of language modeling, the work of [Dai and Le, 2015] is one of the earliest attempts to use a sequence auto-encoder to improve sequence learning of recurrent neural networks. In a similar attempt, Ramachandran et al. [2017] build on Dai and Le [2015]'s work and propose an approach that purportedly improves the performance of the sequence to sequence model. The encoder and decoder of their presented seq2seq model are initialized with pre-trained weights of two language models, which are distinctly trained on the News Crawl English or German corpora for machine translation. They fine-tune their model on labeled data from WMT English to German and the CNN/Daily Mail corpus. During the fine-tuning, they jointly train the seq2seq objective with the language modeling objectives to prevent overfitting. More recent architectures such as Transformers are the successors of these two seminal works. For instance, ELMo (Embeddings from Language Models) [Peters et al., 2018] generalize previous approaches by extracting deep context-dependent representations from a bidirectional LSTM coupled with an LM, hence biLM. They called it deep in the sense that their representations are a function of all of the internal layers of the biLM. In order to

enhance global representations x_k for a task-specific model, they can be concatenated with their corresponding context-dependent representations ${ELMo_k}^{\text{task}}$ to obtain $[x_k; {ELMo_k}^{\text{task}}]$ to feed into the task RNN.

In a concatenation of representations from the forward and backward LSTMs, ELMo does not distinguish the interactions between the left and right contexts. The subsequent models like GPT (Generative Pre-Training) [Radford et al., 2018], and GPT2 [Radford et al., 2019] as standard AR models, use a left-to-right decoder; therefore, they only see the context to the left of a given token. These models do not perform optimally on sentence-level tasks such as Named Entity Recognition (NER) and Sentiment Analysis (SA). BERT [Devlin et al., 2019] is developed as a needed corrective toward learning context from both left and right directions. Devlin et al. [2019] aim to predict the randomly masked tokens of an input sequence depending on the context. They apply a Transformer encoder to gain insight into bi-directional contexts. BERT also incorporates a Next Sentence Prediction (NSP) mechanism that predicts whether the next incoming sentence is semantically aligned with the current one. This feature particularly aims to improve reasoning tasks such as Questions Answering (QA) and Language Inference (LI). Borrowing an intuition from Radford et al. [2018], similarly, in BERT, the first token of every sequence is always a special classification token ([CLS]), and the final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks. BERT is designed primarily for transfer learning, *i.e.*, fine-tuning on task-specific datasets. It outperforms previous state-of-the-art models in the eleven NLP tasks in the GLUE benchmark

[Wang et al., 2018] by a significant margin (score: %80.5.). Their results suggest that BERT is able to "learn" structural and contextual information of language. In the following, we discuss post-BERT methods with improved performance and architecture.

Liu et al. [2019b] propose RoBERTa (A Robustly Optimized BERT Pre-training Approach) with a few alterations to BERT, which achieves substantial improvements. The alterations include: (1) More extensive training with larger batches and more data; (2) Elimination of the next sentence prediction (NSP) mechanism; (3) Training on longer sequences; (4) Dynamic alteration of the masked positions during pre-training.

Lan et al. [2019] present two parameter-reduction techniques, factorized embedding parameterization and cross-layer parameter sharing, to lower memory consumption and increase BERT's training speed. They posit that the main reason for the ineffectiveness of the next sequence mechanism (NSP) is its lack of difficulty, as the negative examples are created by pairing segments from different documents; this mixes topic prediction and coherence prediction into a single task. ALBERT instead uses a sentence-order prediction (SOP) objective which focuses on inter-sentence coherence and is designed to address the ineffectiveness of the next sentence prediction (NSP) loss in BERT. SOP obtains positive examples by taking out two consecutive segments and negative examples by reversing the order of two consecutive segments from the same document, and as negative examples the same two consecutive segments but with their order swapped.

Yang et al. [2019] present XLNet model that serves to rectify two of the main

issues of BERT regarding independence assumption and input noise. (1) BERT is based on denoising auto-encoding. Specifically, for a text sequence x, BERT first constructs a corrupted version $\hat{x}$ by randomly setting a portion (*e.g.*, 15%) of tokens in x to a special symbol `[MASK]`. Let the masked tokens be $\overline{x}$. BERT assumes conditional independence of corrupted tokens and factorizes the joint conditional probability $p(\overline{x} \mid \hat{x})$ based on an independence assumption that all masked tokens $\overline{x}$ are separately reconstructed. (2) The symbols such as `[MASK]` are introduced by BERT during pre-training, yet they never occur in real data, resulting in a discrepancy between pre-training and fine-tuning. XLNet proposes a new auto-regressive (AR) method based on Permutation Language Modeling (PLM) [Uria et al., 2016] without introducing any new symbols.

More recent notable development of Transformer-based Language Models focusing on devising new training methods, noise infusion techniques, and multi-task learning approaches include **Mass** [Song et al., 2019], **UniLM** [Dong et al., 2019], **Electra** [Clark et al., 2019], **Bart** [Lewis et al., 2019], and **T5** [Raffel et al., 2020]. Due to the limited scope of the current investigation we do not survey their proposed methods with more details.

4.3.2 Supervised methods

While many modern NLP systems rely on word embeddings, previously trained in an unsupervised manner on large corpora, as base features, some studies in NLP have tried to explore data-intensive supervised transferable approaches. Inspired by the successful performance of supervised pre-training models in computer vi-

sion, McCann et al. [2017] propose a deep LSTM encoder from an attentional sequence-to-sequence model trained for machine translation (MT) to contextualize word vectors. They show that augmenting non-contextual word representations with context vectors (CoVe) improves performance over using only unsupervised word and character vectors on a wide variety of common NLP tasks: sentiment analysis, question classification, entailment, and question answering. Similarly, Conneau et al. [2017] present InferSent to obtain contextualized representations from a pre-trained natural language inference model on Stanford Natural Language Inference (SNLI) dataset. Their work shows the suitability of natural language inference for transfer learning to other NLP tasks.

4.4 Findings

While contextual embedding methods have an impressive performance on various natural language tasks, the inherent lack of interpretability of neural networks does not explain their behaviors. Prompted by such a concern, Tenney et al. [2019] use an edge probing task to look into how contextual word representations encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena. In a kindred effort, Liu et al. [2019a] investigate features of language captured or missed by contextualized vectors, transferability across different layers of the model, and the impact of pre-training on the linguistic knowledge and transferability. They find that (1) contextualized word embeddings do not capture fine-grained linguistic knowledge, (2) higher layers of RNN are

task-specific (with no such pattern for a transformer), and (3) pre-training on a closely similar task leads to a better performance comparing with language model pre-training. They report that contextual embedding techniques encode and learn important transferable features of language. In another study, Jawahar et al. [2019] provide a series of evidence and support through experimentation with BERT and conclude that BERT's intermediate layers encode a rich hierarchy of linguistic information, with surface features at the bottom, syntactic features in the middle, and semantic features at the top.

4.5 Generative Language Models and Co-Creative Writing

The recent advances in generative AI, language model, and variational autoencoders (VAE), we discussed above, enable new kinds of user experiences around content creation. The unprecedented and dramatic increase in the potentiality of these models and the fidelity of created artifacts are fascinating yet a cause for concern on many fronts. These models are trained on massive amounts of uncurated data, which leads to encoded bias and radicalization along many lines. Despite the faithful mirroring of derogatory and often abusive language in their training data, there are environmental and financial costs attached to performance gains. [Strubell et al., 2019] estimate that a 0.1 increase in BLEU score results in an increase of $150,000 compute cost in addition to carbon emissions. Furthermore, the machine-generated content is often indistinguishable from human-generated con-

tent, which further highlights the greater societal, ethical, and cultural challenges the generative AI models are posing around ownership, privacy, authenticity, and security. However, despite their versatility and fluency, large language models are still in their infancy and show serious fatuity and weakness, which further underlines their importance as an object of study for the future of writing.

In what follows, I specifically focus on the utilization of generative language models in word processors and modern writing assistants and interrogate the space the exists between the human user and the writing tool. The unique positioning of this investigation at the intersection of HCI and HAI allows deepening our understanding of the human-machine co-creative configuration and exploring the opportunities and challenges of augmented user experiences.

We start by tracing back essential aspects of word processing as a technological invention that changed the course of writing. We then investigate contemporary issues related to co-creative generative writing using AI-powered word processors. In the scholarship of this kind, ethics, politics, and habits are brought to the forefront of our conversations about computational tools and their place in writing, a deeply anxious topic for those who analyze it closely.

Joining in this conversation, we argue that the writing teachers we discuss as early adopters of word processing software actually achieved a level of mastery of this software that has not been seen again, but indeed relates to debates circulating in the Digital Humanities, Human-centred AI, Human-Computer Interaction, and broad philosophies of technology about how much involvement a user should have in their software's creation and operations. Given the recent wave

of word processing software and related appendages that now utilize generative Transformer-based language models, word processing software is not only necessary to writing today but in fact, so concealed or black-boxed from its users that even those who build the interfaces may have little knowledge of the underlying systems that make the software possible. While most users have never explored the ethical nor philosophical dimensions that impact the very act of writing, the early adopters of word processing software that we discuss did, in fact, attempt to investigate all of these dimensions, displaying a "deep systems literacy" [Bridle, 2018] that illustrates the importance of human users being involved in making, operating, experimenting, investigating, and deeply caring for the software they use. We started by providing a brief historical account of word processors and writing software in Section 4.5.1.

4.5.1 Early Adopters of Word Processors and Writing Technologies

Through a practical and yet complex example of word processing software, we launch our inquiry by illustrating the traits that mark early adopters in writing classrooms. We first show that in the experimental days of this software, these instructors could be seen to enact what Albert Borgmann calls a "focal practice," which is a way of harmonizing oneself –through practice, investigation, and care– with the operations of the world in laboring over a "focal thing." Secondly, we describe that since that era, there has been an increase in word processing software operations and use, but a decrease in care and attention paid to it. Those

who call for harmony with machines, which includes the DH, HAI, and an ethics care arguments, the makers and doer debates, and some philosophers of technology more generally, echo the sentiments we see in the humanists who adopted word processing software in the 1980s. Attending to these pedagogues and the tremendous work they did is an attempt to argue for the continued importance of gathering around technology, like new software or otherwise, so that it may become a focal practice; we also offer new reasons to appreciate early adopters, who are often criticized as naive worshippers of new trends, when in many cases, they are more like the midwives of the newly created.

In the philosophy of technology, concealment or black-boxing of the machine is somewhat of a commonplace discussion point. Concealment of machines remains, however, only when they operate as they should. Mechanized devices–even electric ones– do often operate as they should: one could have a basic electric coffee maker that lasts fifteen years before it shows any problems. But for many reasons, computational machines are particularly precarious because of their reliance on code, and thus often, they suffer breakdowns that not only disrupt our use of them but actually pull the veil off their concealment, causing a rupture in our thought.

Far from concealment, the community of early adopters we studied circulated ideas that would become essential to the development of our current word processing tools and even to more general notions in AI. Their backgrounds in writing and rhetoric made them aware that in order for a machine to truly assist in writing beyond the model of a digital typewriter, it would require heuristic abilities and contain abilities to dialogue with the writer and learn from text. Experimentation

and competing uses of this software led to a proliferation of scholarship, most of which attempted to investigate many epistemological and ontological shifts happening in writing through computational word processing. Qualitative studies filled the journals, but they were rarely conclusive about whether or not word processing software had made their students better writers, more excited writers, or produced writing containing fewer errors.

4.5.2 Harmony & Rupture

Early adopters, who sometimes get the brunt of the critique from their colleagues for being naïve enough to enthuse over the newest technology, exhibited a kind of "deep systems literacy," to use a concept from James et al. [1890]. He writes that such a literacy "consists of much more than simple understanding, and might be understood and practiced in multiple ways. It goes beyond a system's functional use to comprehend its context and consequences. It refuses to see the application of any one systems a cure-all". Systems literacy is decidedly not the act of learning to code, a commonplace notion in today's rhetoric about computer literacy, which advocates computational thinking (solutionism, functionalism, technicism). We do not actually see the early adopters of word processing software in the classroom advocating for word processing as a cure-all for issues related to typewritten and handwritten text, but more of an attempt to, collectively, come up with a holistic way of understanding and making word processing software to serve the practice of writing and the teaching of writing.

To consider this more deeply, we turn to Albert Borgmann's concepts of the

device paradigm and its opposite, the focal thing/focal practice, as a helpful theory for analyzing how we might view this period in word processing history and also how we might be more generous to early adopters more broadly. In Borgmann's famous adaptation of Heidegger's philosophy of technology, he presents the device paradigm, in which difficult or laborious tasks are replaced with mechanized processes that help make things easier. Devices are problematic for a variety of reasons, but importantly, they are made and operated in such a way that they conceal their own conditions and invention from users. They are "glamorous in their appeal" [Strong and Higgs, 2010], making "no demands on our skills, strength, or attention, and it is less demanding the less it makes its presence felt" [Borgmann, 1987]. Since the device does not require that we understand its functions, tend to it, labor over it, and so forth, it has a "tendency to become concealed or to shrink". Such concealment or shrinking naturalizes it as part of our world, often helping it become totalizing or even necessary to daily life but invisibly operating.

In contrast, there are things in the world that do call for our attention, need tending, and require labor and skill to practice. Borgmann names these *focal things*. In this way, we see the use of Heidegger's concept of the *thing*, which "means a gathering, and specifically, a gathering to deliberate on a matter under discussion, a contested matter $\cdots$ They denote anything that in any way bears upon men, concerns them, and that accordingly is a matter for discourse" [Heidegger, 1971]. Focal things draw on this use of thing, calling us to gather and offering us bits of the good life: "Things, through their centering powers, unify means and ends, achievement and enjoyment, competence and consummation, mind and

body, body and world, individual and community, present and tradition, culture and nature" [Strong and Higgs, 2010]. We see that focal things insist on a kind of ethics attunement with the world around us, while devices slip by, more easily distracting us from ethics and attunement. Bennett draws upon the affinity between recognition of the force of things and ecological thinking and hopes that it brings an increased "deliberateness or intentionality" or in today's AI terminology, "explainability" that may lead to less thoughtless waste. A realization of our entanglement with these things is "compatible with a 'wise use' orientation to consumption".

4.6 Closing Thoughts

In the treatment of modern writing, one mediates the realization that technological things such as AI-powered word processors continue to be a chief modality in our practices and scholarship. In a time during which we can never divorce ourselves from the non-human world of technologies, grounding ourselves in focal practices with these technologies helps us resist a representationalist or instrumentalist understanding and uses of such things. Initiating focal practices further helps us revisit relationships between technology, ethics, and ecology. Rather than regarding things as ephemeral or secondary, they come into view as worthy of our time and primary to our analysis. We believe that, in particular, the most mundane pieces of our daily practice as scholars and pedagogues, or whatever, can illustrate our ethical commitments.

While we linger on practice more than waste in this part of the chapter, the concepts we discuss are inspired by the worry we have over our own practical ethics in the Anthropocene. Focal practices is one framework for a practical ethics of care for things that circulate in the fields of Science and Technology Studies and Digital Humanities. In this context, care necessitates an "ethical commitment and a sustained engagement with the well-being of things" [Jackson] and actuates an earnest culture of mutual responsibility and maintenance. A careful attentiveness evokes a realization of the materiality of electronics and machines, and the multiple forms of waste– from chemical pollution and material waste that issues from microchips and discarded computers to the obsolete software [Gabrys, 2011] bundled into them. Neglecting material care would render us mere users who contribute to the decay and sedimentation of objects we have created ourselves. Latour advises us to invest as much care into the stewardship of our technologies as we do into their creation (Latour, cited in Nowviskie [2013]). This conception of systemic management of technologies certainly does not advocate any notion of total mastery or dominance since, as masters, we would be exempted from any attending or concern toward the unexpected consequences of engagements. We think a focal practice entails more than just embracing and incorporating technology, it demands what Latour calls a "compositionist" notion of modernity that views "the process of human development as neither liberation from Nature nor as a fall from it, but rather as a process of becoming ever-more attached to, and intimate with, a panoply of non-human natures" [Latour, 2012].

To even our surprise, the writing teachers who committed themselves to early

adoption of word processing software and its new writing pedagogy might be the only practitioners of care, patience, and commitment who ever attended to many of the material and intangible changes that did happen and would happen in writing with word processing software. Through acts of maintenance, repair, and disassembly, they not only refused to subscribe to dominant narratives of progress– and perhaps unwittingly upgrade and uphold such narratives– they also helped expose the contingency and transience of the world of commodities.

Chapter 5

Ethics, Human-Centered AI, and Pedagogical Possibilities In CS Education

Chapter Overview[1]

Computer science holds the unfortunate distinction, at least historically, as a pragmatic and often non-critical domain of practice – in terms of both theory and praxis. Accordingly, in some circumstances, our participation in computational enterprises cannot be avoided, such as teaching Computer Science introductory

[1]The content presented in this chapter is in part from the conference paper:

H. Yaghoobian, and E. Kuecker. "Ethics of Critical Engagement in Positivist Enterprises", SEPES 2021, February 19-20 2021, Virtual, USA.

courses. Historically, the curricula for computer science and AI classes emphasize pure coding and solutionism. Nonetheless, we are witnessing a growing allegiance to *AI and data ethics* in both academia and industry, and purportedly, 200 university curricula claim to have complied with tech ethics [Fiesler, 2018]. This proliferation is a promising step forward and can be a useful corrective against discriminatory consequences concerning algorithmic systems, particularly in technical fields like CS and AI, where ethical issues are not discursively and normatively foregrounded. However, these efforts are more coordinated with conventional business ethics than more critical traditions of social justice Greene et al. [2019] expected to prevail in educational settings [O'neil, 2016]. It is required of educational systems to enable students to improve their technological sensibilities and skill sets to generate alternative possibilities that would challenge technocratic ideologies that are uncritical and limit innovation to the measures of efficiency and marketability.

Nudged by these concerns and energized by the creative possibilities of a classroom, we realize the integration of the critical and social dimensions of AI and computing technologies into CS courses as a key area of attention and propose a pedagogical approach largely through concepts that have their legacies in explainable AI (XAI), intersectional feminist activism, collective organizing [D'Ignazio and Klein, 2020], and critical theory [Kellner, 2003]. In this chapter, I will discuss the need for a shift in CS and AI education from a technology-oriented application to humanity-oriented applications that focus on cultivating students' critical and cognitive thinking. I argue that there are ways for ethical and critical engagement,

with enough room for creative manipulation. As reflected in the results from the surveys, shown in Figure 5.1, the proposed approach has resonated with %95.4 of the studied group of students.

5.1 Introduction

Historically, curricula for CS classes emphasize only the technical aspects of computing. However, the ever-increasing role of computing in our lives poses tremendous challenges to educators in CS to rethink the pedagogical choices and processes to deploy tools and methods in creative and ethical ways. Toward developing new literacies and communication frameworks, this chapter describes the integration of various critical elements borrowed from the emerging area of human-centered AI and feminist theory, and critical theory, into CS introductory course in the CS department at the University of Georgia. We discuss the generative potential of relating theory to practice to enhance students' ability to critique the underlying assumptions about the neutrality of computing technologies. While this approach might be an atypical practice in most CS classes, it allowed me to help get CS students considering the ethical implications of their work and non-CS students participating through their disciplinary knowledge or everyday life experience.

5.2 Foundations of Computational Disciplines

Heavily quantitative disciplines, like Computer Science (CS) and Artificial Intelligence (AI), emphasize objectivity, validity, bias, and reality, which are all de-

fined in relation to statistical inferences from data and models. The foundations for this logic incorporate a theory of knowledge in which theories, hypotheses, models, and data represent the world and a theory of truth in which knowledge emerges if substantive theories, hypotheses, models, and data speak to and support the conception of the world. This epistemology relies on a view of 'reality' and is committed to a singular representable reality that is the object of study; and also the ability for science practitioners to represent this reality and test its representations.

While these foundations help navigate the terrains of practice, they also create impasses that have no clear solution for the scientists who uphold them. For example, some social justice scholars have recently begun criticizing algorithms because of their ability to oppress people. Safiya Noble's work suggests that search engine algorithms are helping structure society's visions of what it means to be a "black girl" through simple searches that yield far from "objective" results [Noble, 2018]. Eubanks [2018] has shown how algorithms decide who gets healthcare, who gets insurance, and more. But who is responsible for creating these algorithms? Members of the computer science community. The foundations of the field offer nothing about the origins of CS' epistemology and ontology, nor how they relate to practices and ethics in the communities that utilize them. Too often, the term "science" sociologically acts to buffer the field from foundational critiques within it.

Triumphant accounts of technological progress remove AI from its social and cultural context and remain tethered to the limited calculus of efficiency and

the reductive logic of technological fixes. Accordingly, computational traditions stand reluctant to discuss ethics in their scholarly communities or embed it into the curricula because doing so would force them to address that it is, in fact, impossible to be objective, and therefore, they have not simply studied reality, but actually created it. This is particularly true in CS, as David Berry mentions, given that "software, computation, and code define our contemporary situation, becoming part of the metaphors by which it is even possible to think today" [Berry, 2015].

CS itself would wish to hold onto its meanings of data, its conceptions of reality, and its deterministic correspondences so that its work can remain settled and incontestable. In sciences, the impasse, or perhaps the vicious cycle, is that a representation or a tool used to test correspondence is only comparable to pre-existing representations with pre-existing tools, underscoring what John Dewey refers to as "stumbling block of empiricists in trying to account for science on an empirical basis" [Dewey, 1930]. Ironically, through this type of positivism, the scientist becomes an active participant in the production of what they abstractly call 'reality.' To a non-expert, this conception of reality to the material of everyday life is hidden. Worse, to a student who is being initiated into the CS field, almost no insight is provided into the practical implications of choosing among various, incommensurable representations or tools. There are no ethics classes, no alternative methods presented, no treatises on how to trouble the tools we are to use. The enormous ethical implications of the kind of work a computer scientist might eventually do are materially and discursively shunted to the side.

5.3 Toward a Human-centered AI Educational Model

The existing framework of computer sciences classes emphasize only the technical dimension of concepts in computing and information science. In Topics for Computing course topics include data representation, network protocols, encryption and security, mathematical modeling, and more. We were aware that while computing machines have eliminated some kinds of errors, a whole horizon of new issues and errors has appeared, but this existing curriculum creates a narrative that suggested each of these computational topics were solutions, all part of the myth of progress where technological progress is the only kind of progress, and progress is the only direction of movement. The ethical movement had no place in the existing curriculum.

One of the issues that has come into being is that computing machines have become so complex that we do not know how much trust we can place in them. In fact, many of them are "black-boxed" from being understood even at the hardware level, not allowing even someone with a PhD. in computer science/computer engineering to repair their own machine. This forecloses the opportunity for a person to do what Sequeira [2018] has described as "unmaking." Unmaking is the ability to tinker with something in order to "experience" it before "comprehending it." The history of computer science is full of anecdotes of homemade experimentation with devices, though the future histories of computer science may have very few, given the purposeful black-boxing and obfuscation of both hardware and software.

Indeed, consider the linguistic norm of "computer user," indicating that one is not a maker, creator, editor, or even knower. Such a passive position is very difficult to defy. Ontologically speaking, we are humans navigating a built world, and this built world was actually constructed by humans, but such construction is often veiled from our view. While this is true of many things, computation is a particularly potent example.

AI on the human condition or human-centered AI (HAI)[2], is a type of AI that demands explainable and interpretable computation and decision-making processes and is contingent upon continuous (re)adjustments of AI algorithms through human context and societal and cultural phenomena to enhance human and environmental welfare. Toward a sustainable AI, we need to be cognizant of the greater impacts of AI on humans, and by extension, nature and this shift become actualized with the climate of "classroom" as a site of participation, engagement, and interaction.

As I worked on the introductory CS class in the Computer Science department at the University of Georgia, I wanted to create a curriculum that emphasized HAI points, while also offering windows into subverting it. In CS, the core of pedagogical practices often remains limited to the traditional repertoires of science that do not allow for inherently transdisciplinary and critical endeavors. One is to learn what is offered and not question it, largely because of a fixation on efficiency. When problems in computation are discussed, the conversations often ignore the ways in which those problems cannot actually be solved with technical solutions.

[2]https://hai.stanford.edu/

Heath [2021], one of the first professional programmers in the US, has a concept she dubs "technochauvinism," which stipulates that all problems can be solved through technical solutions. But the major fallacy of this is that all computational problems are entangled in social and natural contexts. The dominant framework of CS, however, is inherently impervious to ethics and renders such concerns as a marginal subject. Carl Sagan posits "we live in a society exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology." In an introductory class like Topics in Computing (TiC), the student audience is actually ideal for testing out a curriculum that centers on this kind of everyday ethics and critical considerations as they intersect with CS Education. The students in the class could be from any major, and many of them would not have their minds made up about CS yet, unlike a graduate audience, for example.

5.4 Ethics-oriented Critical Engagement in Computer Science Education

My proposed tactic for an ethics-oriented critical engagement in CS education had to first do away with the instrumental view of technology. Instead, I decided to operate with a literacy model called "systems literacy," a theory that is born more from Anthropocene studies than science itself. "True literacy in systems consists of much more than a simple understanding and might be understood and practiced in multiple ways. It goes beyond a system's functional use to comprehend its context and consequences. It refuses to see the application of any one system as

a cure-all" [Bridle, 2018]. In other words, deep systems literacy keeps the original goals of the course intact but sees them only as the surface. With each topic, the goal is to figure out its context and consequences, which automatically delivers CS into transdisciplinary discussions.

As an example of how those conversations are taken deeper, in the case of the topic of data structures and algorithms used to analyze data, the curriculum was enhanced to include issues with the categorical representation of data, which is related to algorithmic discrimination. Or similarly, after getting familiar with the fundamentals of hardware and software, issues related to sustainability in the forms of both electronic waste and digital rubbish are discussed. I continued to infuse the technical contents of the course with elements borrowed from debates in AI and sustainability, data feminism; automation, and the future of the workforce, which are all becoming part of the continual debates under the umbrella of HAI and HCI.

Another important issue in CS education and critical engagement centers not on the curriculum precisely, but issues surrounding pedagogy and the field at large. The underrepresentation of women studying the discipline, for example, is often discussed as merely a recruitment problem, and solutions are proposed about how to attract women to the field. But there seems to be a fundamental misunderstanding of the precise problem, given that at least in this class, there were many students who identified as women and showed interest in CS. They were there, in the room, ready to engage in CS. Given that they are such a marginal part of the CS community both historically and currently, it was useful to gather

their input on what interests they had in the field.

In this way, a critical issue in CS was apparent in the very fabric of the class itself. I experimented with the use of "exit tickets" as a pedagogical practice to complement the deep systems literacy curriculum focus. After each class, students filled out an assessment about the class topics and pedagogic delivery so that instructional feedback could be a constant flow throughout the semester. Of particular note, this critical issue in CS greatly showed itself, with students reporting a great interest in having discussions and readings centered on feminism and women in CS, desiring a space to talk about it. Purposely, the class was structured with open days that could be filled with whatever was of interest to the students as long as it related to the class topics– this allowed the class to naturally flow in a feminist direction guided by curious students, thus feminist issues became one key part of the class, though it would have never been in regular CS curriculum.

Ultimately, this experiment in syllabus redesign and shifting the priorities of a basic CS class revealed that there are ways to engage with computational disciplines in order to improve them. On the one hand, it was proved that issues in computation are part of everyone's everyday lives, which is precisely why students, when asked how to fill the gaps in the schedule, immediately had the topics of feminist issues and AI on their minds– these are things that intersect with their computational lives. This further troubles the problem of "users" versus "creators," empowering learners to elevate their status from user to creator in the course of the class.

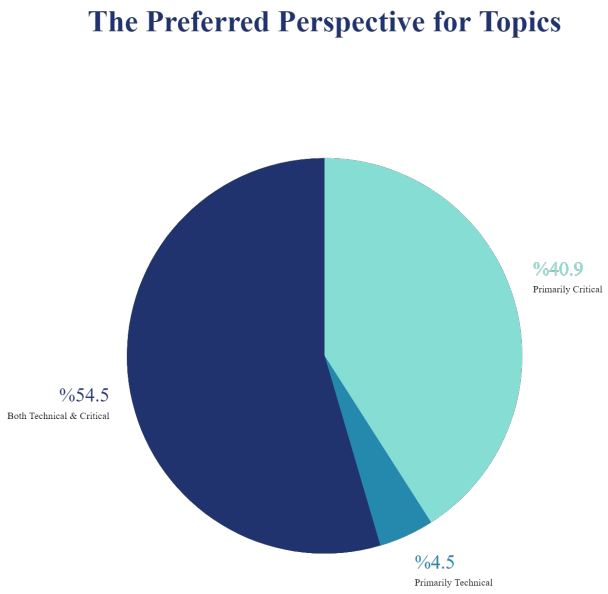

Figure 5.1: Students' Preferred Perspective for Analysis of Topics

The other thing that we found was that basic introductory courses offer something that we need to keep. While we could entirely deny computational methods and technological logic, we instead enjoyed dwelling with an inherited, old-fashioned syllabus. It forced my thought toward actionable changes, at least as they might impact 30 or so young people—many of these essential, introductory curriculum elements needed to remain. I did not wish to do away with the parts where students learned the basics of computer operation, for example, and instead replace that time on critical readings. Knowing the practical operations of something, however dreary, is actually an essential element of XAI and systems literacy. The impact of the introductory class truly relies on both, in alignment, so that experience can aid comprehension, as Sequeira said, though, for us, com-

prehension of a critical CS class would ideally appear when students, years after taking the class, still think about the various yet-to-be-solved issues that were brought up during the class discussion. The ability to notice these problems as they come up over and over in their daily lives, tell others about them, and perhaps change their relationships to computation in light of such grievances– that would be what it means to comprehend the scope of this kind of topic.

Chapter 6

Conclusions and Future Work

The main focus in this dissertation is placed upon exploring the interconnectivities of various overlapping areas in Computer Science and addressing the discontinuities. The body of work presented in this dissertation attempts to fill important gaps in computational and theoretical approaches to emotion and subjectivity analysis through the lens of language. Through the manuscript, I address four research questions, which are discussed below. I now briefly return to some of the findings and concepts explored and discussed in previous chapters.

How is language a gate to emotionality and subjectivity? Are there linguistic constructs beyond lexical and syntactical structures that contribute to the intensity of emotions and feelings?

Chapter 2 investigates this question in the context of memory narratives. Our proposed structure-aware approach to emotion in memory narratives of the studied

population demonstrates a willingness to search for a new paradigmatic framework for the interpretation of affective memory and its function from an empirically social and linguistic perspective. We identified a linguistic construct, which we dubbed "discontinuity" that was capable of recruiting empirical and discursive support. Discontinuity is the frequency with which shifts in verb tenses occur in a memory narrative. In this study, we find that the salience of emotions and subjectively experienced feelings are reflected in language use both within and beyond the morphological and syntactic structures. Therefore, there is a sensitive and interactive realm of experience that precedes the linguistic expressions and legitimately counts as cognition. In this sense, quantitative models used in this exploration allowed us to look into other ways of grappling with the mystery of the human past. We find that verb tense shifts function discursively, and the affective dimensions of subjectivity are evidently manifested in autobiographical memory narratives. We discover a distinct link between the level of discontinuity and how the surveyed population defined themselves politically.

Can the shape of discourse be utilized toward more effective implicit intentionality and sarcasm? Can we de-bias the currently existing models for sarcasm detection by leveraging the hierarchical structures in language?

The study presented in Chapter 3.2 is an exploration of discourse-aware detection of sarcasm, with eyes toward rectifying the user bias in approaches propose hitherto. We devise a hierarchical re-current attention network architecture that

leverages the hierarchical information embedded in discourse trees as context toward improving the performance of the model. Because of the ever-evolving and figurative nature of sarcastic utterances, its identifications always remains a challenging task for both humans and machines.

How do computers affect the writing process? Will research in AI and generative language models change the way we think about communicating with each other and with computers? What are the possible risks associated with this technology?

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. In Chapter 3.1.2, we offer an empirical investigation of BERT and Post-BERT models and examine their underlying structures, promises, and pitfalls. We think understanding the limitations of LMs and putting their success in context is important, particularly at these early stages of their adoption and development. Therefore, we offer a deeply analytical and philosophical account of the interactive space between the human and the emerging technological writing tools. Grounding this investigation in human-centered AI allows us to critique the potential follow-on risks, the inherent black-boxing, and the lack of interpretability of these models.

How can the research in Explainable AI and Data Ethics inform and transform the pedagogical practices in CS Education?

The unifying theme of this dissertation has been an empirical and theoretical investigation of an ethics-oriented, interpretable, and context-dependent computation. These efforts have molded, informed, and ultimately transformed the requirements for a more contemporary model for CS educations. Most of the previously imposed requirements were formed in an era filled with entirely different problems surrounding computation. Today, they mostly fall short in preparing a CS student for the grand societal challenges. Chapter 5 is an explanatory account of the reasons I think introductory CS classes should foster the cultivation of critical and computational thinking skills. I believe infusing technical and computational concepts with critical perspectives enriches students' technological imagination, broaches limits and generates new possibilities for pushing beyond the typical use for both hardware and software. As reflected in student evaluations, this approach has resonated with students at both undergraduate and graduate levels at the University of Georgia, Computer Science Department.

6.1 Path Forward

Studies presented in this dissertation not only reach across various disciplines but also across methodologies, trying to bring multiple modes of inquiry and interpretation from sciences and humanities together. I do not view computational models and frameworks as the primary ends of my work. Specifically, I view the

computational domain as a site of convergence where people from across fields and hierarchies come together to make inquiries about phenomena that can not be easily pigeonholed into the confines of a single discipline. This dissertation tries to shift the epistemologies that define the possibilities of quantitative research to examine timely topics of broad interest to multiple scholarly communities. These studies will move research in both AI and HCI forward while demonstrating the need for interdisciplinarity in both fields.

While my past and ongoing efforts have led to a series of work that essentially interrogates the interactive spaces between human and computing technologies, I plan to pursue to integrate computational methodologies with humanistic evidence toward powerful effects.

Bibliography

A. Agrawal, A. An, and M. Papagelis. Leveraging transitions of emotions for sarcasm detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1505–1508, 2020.

L. M. Alcoff. *Visible identities: Race, gender, and the self*. Oxford University Press, 2005.

D. Bamman and N. A. Smith. Contextualized sarcasm detection on twitter. In *Ninth international AAAI conference on web and social media*. Citeseer, 2015.

F. Barbieri, H. Saggion, and F. Ronzano. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, 2014.

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

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

D. M. Berry. *Critical theory and the digital*. Bloomsbury Publishing USA, 2015.

S. K. Bharti, K. S. Babu, and S. K. Jena. Parsing-based sarcasm sentiment recognition in twitter data. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1373–1380. IEEE, 2015.

B. A. Biesecker. Remembering world war ii: The rhetoric and politics of national commemoration at the turn of the 21st century. *Quarterly Journal of Speech*, 88(4):393–409, 2002.

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

A. Borgmann. *Technology and the character of contemporary life: A philosophical inquiry*. University of Chicago Press, 1987.

J. Bridle. *New dark age: Technology and the end of the future*. Verso Books, 2018.

R. Brown and J. Kulik. Flashbulb memories. *Cognition*, 5(1):73–99, 1977.

S. D. Brown and P. Reavey. Contextualising autobiographical remembering: An expanded view of memory. *Collaborative Remembering: Theories, Research, and Applications*, 2017.

J. Bruner. *Acts of meaning.* Harvard university press, 1990.

K. Buschmeier, P. Cimiano, and R. Klinger. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, 2014.

E. Camp. Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs*, 46 (4):587–634, 2012.

J. D. Campbell and A. N. Katz. Are there necessary conditions for inducing a sense of sarcastic irony? *Discourse Processes*, 49(6):459–480, 2012.

P. Carvalho, L. Sarmento, M. J. Silva, and E. De Oliveira. Clues for detecting irony in user-generated contents: oh...!! it's" so easy";-. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56, 2009.

W. Chafe. Some things that narratives tell us about the mind. *Narrative thought and narrative language*, 79:98, 1990.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, 2017.

T. Dadu and K. Pant. Sarcasm detection using context separators in online discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 51–55, 2020.

A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015.

D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116, 2010.

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

J. Dewey. The quest for certainty: A study of the relation of knowledge and action. *The Journal of Philosophy*, 27(1):14–25, 1930.

C. D'Ignazio and L. F. Klein. *Data feminism.* MIT Press, 2020.

L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075, 2019.

D. Edwards. Discourse, cognition and social practices: The rich surface of language and social interaction. *Discourse Studies*, 8(1):41–49, 2006.

J. Eisterhold, S. Attardo, and D. Boxer. Reactions to irony in discourse: evidence for the least disruption principle. *Journal of Pragmatics*, 38(8):1239–1256, 2006.

K. Elkins and J. Chun. Can gpt-3 pass a writer's turing test? *Journal of Cultural Analytics*, 1(1):17212, 2020.

V. Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, 2018.

D. I. H. Farías, V. Patti, and P. Rosso. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, 16(3): 1–24, 2016.

W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.

E. Fersini, F. A. Pozzi, and E. Messina. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *2015 IEEE*

International Conference on Data Science and Advanced Analytics (DSAA), pages 1–8. IEEE, 2015.

C. Fiesler. Tech ethics curricula: A collection of syllabi. *https://cfiesler.medium.com/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18*, 2018.

E. Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 392–398, 2012.

S. Fleischman. Discourse functions of tense-aspect oppositions in narrative: Toward a theory of grounding. 1985.

L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

J. Gabrys. *Digital rubbish: A natural history of electronics*. University of Michigan Press, 2011.

A. Ghosh and T. Veale. Fracking sarcasm using neural network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169, 2016.

A. Ghosh and T. Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, 2017.

R. González-Ibánez, S. Muresan, and N. Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*, pages 581–586. Association for Computational Linguistics, 2011.

D. Greene, A. L. Hoffmann, and L. Stark. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

S. Greene and P. Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511, 2009.

M. Halbwachs. *On collective memory*. University of Chicago Press, 1992.

V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and social psychology Review*, 19(4):307–342, 2015.

D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea. Cascade: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, 2018.

M. Heath. Meredith broussard, artificial unintelligence: How computers misunderstand the world. *International Journal of Communication*, 15:3, 2021.

M. Heidegger. Poetry, language, thought. 1971.

M. Heidegger. *Being and time*. Suny Press, 2010.

I. Hernández-Farías, J.-M. Benedí, and P. Rosso. Applying basic features from sentiment analysis for automatic irony detection. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 337–344. Springer, 2015.

A. Hogenboom, F. Frasincar, F. De Jong, and U. Kaymak. Using rhetorical structure in sentiment analysis. *Commun. ACM*, 58(7):69–77, 2015.

S. Ilic, E. Marrese-Taylor, J. Balazs, and Y. Matsuo. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7, 2018.

S. L. Ivanko and P. M. Pexman. Context incongruity and irony processing. *Discourse Processes*, 35(3):241–279, 2003.

S. J. Jackson. Debates in the digital humanities 2019: Part v][chapter 38 debates in the digital humanities 2019 part v][chapter 38.

K. Jaidka, I. Singh, J. Lu, N. Chhaya, and L. Ungar. A report of the cl-aff offmychest shared task: Modeling supportiveness and disclosure. In *Proceedings of the AAAI-20 Workshop on Affective Content Analysis, New York, USA, AAAI*, 2020.

W. James, F. Burkhardt, F. Bowers, and I. K. Skrupskelis. *The principles of psychology*, volume 1. Macmillan London, 1890.

S. Javdan, B. Minaei-Bidgoli, et al. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 67–71, 2020.

G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, 2019.

Y. Ji and J. Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 13–24, 2014.

A. Joshi, V. Sharma, and P. Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, 2015.

A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, 2016.

A. Joshi, P. Bhattacharyya, and M. J. Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):73, 2017.

D. Kellner. Toward a critical theory of education. *Democracy & Nature*, 9(1): 51–64, 2003.

E. A. Kensinger. Remembering the details: Effects of emotion. *Emotion review*, 1(2):99–113, 2009.

M. Khodak, N. Saunshi, and K. Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

D. M. Korngiebel and S. D. Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *NPJ Digital Medicine*, 4(1):1–3, 2021.

M. Kraus and S. Feuerriegel. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79, 2019.

R. J. Kreuz and G. M. Caucci. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4. Association for Computational Linguistics, 2007.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

B. Latour. Love your monsters in love your monsters: Postenvironmentalism and the anthropocene. *Breakthrough Journal*, 2012.

C. J. Lee and A. N. Katz. The differential role of ridicule in sarcasm and irony. *Metaphor and symbol*, 13(1):1–15, 1998.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

M. Li, C. Lang, M. Yu, Y. Lu, C. Liu, J. Jiang, and W. Huang. Scx-sd: Semi-supervised method for contextual sarcasm detection. In *International Conference on Knowledge Science, Engineering and Management*, pages 288–299. Springer, 2020.

C. Liebrecht, F. Kunneman, and A. van den Bosch. The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Com-*

putational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 29–37, 2013.

B. Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019a.

P. Liu, W. Chen, G. Ou, T. Wang, D. Yang, and K. Lei. Sarcasm detection in social media based on imbalanced classification. In *International Conference on Web-Age Information Management*, pages 459–471. Springer, 2014.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

S. Lukin and M. Walker. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, 2013.

W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

E. Manning. *Always more than one: Individuation's dance*. Duke University Press, 2013.

J. Märkle-Huß, S. Feuerriegel, and H. Prendinger. Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 1142–1151. HICSS, 2017.

J. R. Martin and P. R. White. *The language of evaluation*, volume 2. Springer, 2003.

B. Massumi. *Parables for the virtual: Movement, affect, sensation*. Duke University Press, 2002.

D. G. Maynard and M. A. Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA, 2014.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

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

A. Mishra, D. Kanojia, S. Nagar, K. Dey, and P. Bhattacharyya. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, 2016.

S. Mukherjee and P. K. Bala. Sarcasm detection in microblogs using naïve bayes and fuzzy clustering. *Technology in Society*, 48:19–27, 2017.

U. Neisser. Memory: What are the important questions. *Memory observed: Remembering in natural contexts*, pages 3–19, 1982.

K. Nimala, R. Jebakumar, and M. Saravanan. Sentiment topic sarcasm mixture model to distinguish sarcasm prevalent topics based on the sentiment bearing words in the tweets. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2020.

S. U. Noble. *Algorithms of oppression*. New York University Press, 2018.

B. Nowviskie. Resistance in the materials. *URL: http://nowviskie.org/2013/resistance-in-thematerials/[25 August 2015]*, 2013.

D. Nozza, E. Fersini, and E. Messina. Unsupervised irony detection: a probabilistic model with word embeddings. In *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 68–76. SCITEPRESS, 2016.

J. K. Olick, V. Vinitzky-Seroussi, and D. Levy. *The collective memory reader*. Oxford University Press, 2011.

C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

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

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

D. B. Pillemer, A. B. Desrochers, and C. M. Ebanks. *Remembering the past in the present: Verb tense shifts in autobiographical memory narratives.* Lawrence Erlbaum Associates Publishers, 1998.

R. A. Potamias, G. Siolas, and A.-G. Stafylopatis. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, pages 1–12, 2020.

E. Probyn. Writing shame. In M. Gregg and G. J. Seigworth, editors, *The Affect Theory Reader*, chapter 3, pages 71–90. Duke University Press, 2010.

T. Ptáček, I. Habernal, and J. Hong. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, 2014.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

R. Rakov and A. Rosenberg. " sure, i did the right thing": a system for sarcasm detection in speech. In *Interspeech*, pages 842–846, 2013.

P. Ramachandran, P. J. Liu, and Q. Le. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, 2017.

A. Reyes and P. Rosso. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision support systems*, 53(4):754–760, 2012.

A. Reyes and P. Rosso. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614, 2014.

A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74: 1–12, 2012.

A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

M. Richardson. Writing trauma: Affected in the act. *New Writing*, 10(2):154–162, 2013.

E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, 2013.

B. Robinson and M. Kutner. Spinoza and the affective turn: A return to the philosophical origins of affect. *Qualitative Inquiry*, 25(2):111–117, 2019.

J. Sequeira. *Other Paradises: Poetic Approaches to Thinking in a Technological Age*. John Hunt Publishing, 2018.

T. Sharot, M. R. Delgado, and E. A. Phelps. How emotion enhances the feeling of remembering. *Nature neuroscience*, 7(12):1376–1380, 2004.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936, 2019.

D. P. Spence. *Narrative truth and historical truth: Meaning and interpretation in psychoanalysis*. WW Norton & Company, 1984.

D. Strong and E. Higgs. 1. borgmann's philosophy of technology. In *Technology and the Good Life?*, pages 17–37. University of Chicago Press, 2010.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

M. Taboada, K. Voll, and J. Brooke. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report*, 2008.

K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, 2015.

I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. Bowman, D. Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

J. Tepperman, D. Traum, and S. Narayanan. " yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth international conference on spoken language processing*, 2006.

O. Tsur, D. Davidov, and A. Rappoport. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle. Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17(1): 7184–7220, 2016.

C. Van Hee, E. Lefever, and V. Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, 2018a.

C. Van Hee, E. Lefever, and V. Hoste. We usually don't like going to the dentist: Using common sense to detect irony on twitter. *Computational Linguistics*, 44 (4):793–832, 2018b.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

T. Veale and Y. Hao. Detecting ironic intent in creative comparisons. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 765–770. IOS Press, 2010.

M. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International*

Conference on Language Resources and Evaluation (LREC'12), pages 812–817, 2012.

W. R. Walker, R. J. Vogl, and C. P. Thompson. Autobiographical memory: Unpleasantness fades faster than pleasantness over time. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 11(5):399–413, 1997.

W. R. Walker, J. Skowronski, J. Gibbons, R. Vogl, and C. Thompson. On the emotions that accompany autobiographical memories: Dysphoria disrupts the fading affect bias. *Cognition & Emotion*, 17(5):703–723, 2003.

B. C. Wallace, L. Kertz, E. Charniak, et al. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, 2014.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

S. I. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, 2012.

Z. Wang, Z. Wu, R. Wang, and Y. Ren. Twitter sarcasm detection exploiting a

context-based model. In *international conference on web information systems engineering*, pages 77–91. Springer, 2015.

D. Wilson. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10): 1722–1743, 2006.

H. Wittels. *Humblebrag: The art of false modesty*. Grand Central Publishing, 2012.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

E. Zerubavel. *Time maps: Collective memory and the social shape of the past*. University of Chicago Press, 2012.

X. Zhu, P. Sobihani, and H. Guo. Long short-term memory over recursive structures. In *International Conference on Machine Learning*, pages 1604–1612, 2015.