Sentiment Analysis Using Dialogue Data: A Taxonomy of Literature

Laura E. M. Äyräväinen
University of Bath
lema21@bath.ac.uk

Joanne Hinds
University of Bath
jh945@bath.ac.uk

Brittany I. Davidson
University of Bath
bid23@bath.ac.uk

Abstract

In a rapidly advancing world of artificial intelligence and natural language processing, the need to understand human dialogue with information systems is ever more pressing. This paper presents a taxonomy of sentiment analysis using dialogue data, developed via a scoping literature review and employment of a widely recognized taxonomy method. By synthesizing the diverse approaches across 18 papers (comprising 22 dimensions and 328 categories), we present a framework that highlights the components underpinning current work, including application domains, data characteristics, sentiment analysis pipelines, methods and dialogue-specific information utilized. By offering a detailed, method-focused view of the existing research, our taxonomy aims to guide future studies that seek to integrate sentiment analysis into dialogue systems. We discuss issues in the current state of the literature and conclude by providing directions for future research.

Keywords: sentiment analysis, dialogue, taxonomy, natural language processing.

1. Introduction

In today's digitized world, the possibility of analyzing dialogue is becoming increasingly prevalent across a variety of information systems applications. For instance, chatbots (systems that interact with users via text or voice) are widely used in customer service (Li et al., 2016), banking (Mogaji et al., 2021) and health applications (Chin et al., 2023). Additionally, social media platforms provide access to public reactions and opinions on global events (e.g., Abd-Alrazaq et al., 2020). These applications can serve critical societal functions that range from providing organizations with valuable insights into dialogue dynamics surrounding products and services, to providing companionship to the elderly experiencing loneliness (Chen & Nakamura, 2021).

Sentiment analysis can perform a powerful role in dialogue systems because it can be used to gain understanding of users' attitudes, emotions, and feelings in real time and at scale (Gunasekaran, 2023). Insights from these analyses can be used to inform decision-making and policy, or to improve dialogue systems' responses and customized communication. The need for sentiment analysis in dialogue systems is becoming increasingly pertinent due to the rapid development and growing use of artificial intelligence (AI) to provide tailored, personalized, and intelligent services (El-Ansari & Beni-Hssane, 2023; Krugmann & Hartmann, 2024).

Currently, research on sentiment analysis in dialogue systems is diverse and it remains difficult to find a comprehensive synthesis of the literature. Consequently, we do not know what approaches are taken to analyze sentiment in dialogues, or which characteristics of the dialogue are most important for predicting sentiment. We attempt to fill this gap by answering the overarching question: What are the current approaches to dialogue sentiment analysis? To answer this question, we summarize current literature regarding (1) the research context (aims, theoretical motivation and application domains), (2) data characteristics (e.g., data source, dataset size, language, annotation) and (3) sentiment analysis methods (e.g., sentiment analysis pipeline, types of sentiment analysis systems, dialogue-specific information used).

To achieve this, we perform a scoping review of 18 papers that employ dialogue-specific approaches to analyze sentiment in textual, dyadic dialogues. We then rigorously analyze these papers to develop a taxonomy depicting key aspects of the dialogue sentiment analysis research comprising of 22 distinct dimensions and 328 categories following available guidelines (Nickerson et al., 2013; Usman et al., 2017). Taxonomies are widely adopted in information systems research to synthesize and

URI: https://hdl.handle.net/10125/109123
978-0-9981331-8-8

categorize research into comprehensible frameworks (Nickerson et al., 2013). By developing a taxonomy on sentiment analysis and dialogue data, we provide a resource that enables easy navigation of current literature by focusing on several key aspects of dialogue sentiment analysis, including dataset characteristics and methodological components. We aim to make complex computational materials more accessible, and to facilitate communication among researchers and non-technical users who are interested in sentiment analysis and dialogue data. The taxonomy is designed to help researchers identify gaps in existing literature and to locate suitable methodological approaches when designing and developing future work. Finally, we aim to encourage interdisciplinary research for the theoretically grounded development of sentiment analysis datasets and systems. We advocate for open science practices in reporting and sharing sentiment analysis pipelines and methods, as well as for carefully justified and thorough evaluation of sentiment analysis systems. As such, we conclude by providing recommendations for future research.

The remainder of this paper is structured as follows: first we introduce dialogue sentiment analysis, emphasizing its unique aspects and relevance in current research. Second, we outline our method for literature search and taxonomy development. Third, we present our findings on the current state of the literature, based on the developed taxonomy. Finally, we discuss the challenges identified in the dialogue sentiment research, consider the limitations of our work, and conclude with recommendations for future directions.

1.1. Background

Sentiment analysis has continued to grow in popularity since the early 2000s, enabling researchers and companies alike to generate sentiment data from various sources, including social media, emails and product reviews. (e.g., Mäntylä et al., 2018). Social media data has been used to gain insights into people and society across a variety of large-scale events, such as healthcare crises (Alamoodi et al., 2021) and electoral behavior (Skoric et al., 2020). A variety of sentiment analysis methods are established across fields. For instance, social sciences primarily use dictionary and word counting techniques, whereas computational sciences favor machine learning and AI approaches.

Traditionally, sentiment is analyzed at a sentence or document level, such that the unit of analysis is at most a single online review or a social media post. In other words, much of sentiment analysis research has focused on monologues or unidirectional text. A more complex scenario is that of a dialogue between two (or more) people, consisting of at least bidirectional utterances, where one discourse participant may influence the emotional tone of the following utterances from another discourse participant (Ma et al., 2023; Poria et al., 2019b). For example, if a friendly greeting from speaker A is met with a hostile response from speaker B, this likely affects how positive speaker A's next utterance will be. Tracking these types of sentiment dynamics in dialogues requires identification of additional information, such as the speaker, the order of utterances, and sentiment of previous utterances. This information can be used as additional context to predict sentiment of the unit of interest (e.g., the final utterance of a particular speaker). Apart from these aspects of dialogue data, sentiment analysis applied to dialogues offers an opportunity for more comprehensive linguistic analysis, such as consideration of pragmatic and discourse-level of texts, in addition to lexical level. Leveraging aspects of language beyond the lexical level holds great promise for improving nuanced comprehension and processing capabilities of sentiment analysis systems, thus steering them closer to genuine natural language understanding (Benamara et al., 2017).

Investigations into dialogue sentiment analysis have the potential to contribute to several fields, in theoretical, methodological, and practical ways. In theory development, this research can enrich our understanding of the linguistic mechanisms underpinning sentiment expression, perception and dynamics in interactive contexts, thereby contributing valuable insights to psycholinguistics and sociolinguistics. Methodologically, insights from conversational sentiment analysis may challenge and expand the existing frameworks of sentiment analysis, leading to the development of more sophisticated analytical tools that can navigate the complexities of human discourse. On a practical level, advancements in conversational sentiment analysis hold potential for improving human-computer interaction, from chatbots to virtual assistants.

The literature on sentiment analysis is extensive, with an accelerating pace of new publications (Cui et al., 2023; Mäntylä et al., 2018). Therefore, there is a growing need for resources that aid researchers and practitioners to navigate the field and its emerging sub-fields, such as sentiment analysis in dialogues. Systematic categorization of the literature, including data characteristics and adopted methodologies, is essential to inform further system development and to enable expansion to new domains and scholarly

inquiry in this area. To this end, we conduct a scoping review of dialogue sentiment analysis and develop a preliminary taxonomy of the key features of the literature.

2. Method

The protocol for the project was registered on the 13th February 2024 and is available on the OSF (https://osf.io/4ghj5). At this point, the search was already performed, and title-abstract screening had commenced, due to deadlines associated with the funding, which meant the searches needed to be tested and refined in advance. However, the original pre-registration was submitted without changes to the content. All supplementary materials for this study, including the final taxonomy and documentation, can be found on the OSF.

2.1. Literature search

The general inclusion criterion for articles was empirical primary studies written in English in which sentiment analysis is applied to textual dialogue data. More specifically, the inclusion criteria for study screening were as follows: (1) sentiment analysis must be applied to textual data (including audio transcriptions and original text form data. This also includes multimodal content where textual data is analyzed separately or can be separated from other modalities in the results). (2) The dialogue data must be dyadic (i.e., between two parties rather than multiparty conversations). (3) The sentiment analysis performed must be specific to dialogue content, such that the dialogue context is leveraged in determining the sentiment of utterances (e.g., information about the previous utterances, speaker characteristics, dialogue acts or other information specific to dialogues). In other words, studies where using dialogue data was trivial to the task were excluded (e.g., studies extracting sentiments from isolated utterances or the entire conversation as a single unit). (4) Outcome of sentiment analysis must be reported (primary studies were included, proof of concept papers or introductions to new systems without testing were excluded).

Four electronic databases were searched on 26th of January 2024: ACM Digital Library, IEEE Xplore, PsychInfo, and Scopus. With the aim of conducting a scoping review and developing a preliminary taxonomy of the literature, we kept the search string as inclusive and broad as possible, but only applied it to the titles of the studies. Searching for titles, abstracts and keywords returned a considerably large number of articles (over 1,000), many of which were irrelevant to our study. Thus, restricting the search to the titles of the papers ensured retrieving only highly relevant articles. The following search string was used: TITLE ("sentiment analysis" OR "sentiment detection" OR "sentiment classification" OR "opinion mining") AND TITLE (dialogue OR dyad OR conversation*).

All documents with at least one sentiment analysis-related search term and one dialogue-related search term in their titles were retrieved from the databases. The screening process is depicted as a PRISMA flowchart on the OSF. The literature search resulted in retrieval of 94 articles. After deduplication of the articles in Zotero, the final sample of documents to screen was 70 articles.

Titles and abstracts of the full sample of 70 studies were independently screened by two researchers (LA and JH). The inter-rater percent agreement for the screening was 74% (Cohen's κ = 0.29, 95% CI [0.00, 0.57]). All disagreements were resolved via discussion, and 58 papers were retained for full text screening. Considering the full text of the 58 papers, LA screened the full sample, and JH screened 50% of these. The inter-rater percent agreement for the screening was 79% (Cohen's κ = 0.60, 95% CI [0.32, 0.89]). The remaining six disagreements were resolved via discussion, and 18 papers were retained for the taxonomy.

2.2. Taxonomy development

Following Nickerson et al.'s (2013) guidelines, we first identified the intended use of the taxonomy: a resource for easy navigation of sentiment analysis literature, based on specific characteristics of interest, such as sentiment analysis methods used. The terminology used in describing the taxonomy is as follows: a dimension refers to the highest-level categories in the taxonomy, such as dataset size. Each dimension consists of categories (referred to as characteristics in Nickerson et al., 2013) with increasing level of granularity.

The classification structure for our taxonomy is best described as a faceted analysis (Usman et al., 2017), where the topic of interest (sentiment analysis study) is classified based on different aspects of the work (e.g., dataset, methods). Following the guidelines of Nickerson et al. (2013), both deductive (conceptual-to-empirical) and inductive (empirical-to-conceptual) approaches were employed in different iterations of the taxonomy development.

Given our background knowledge in the domain, conceptual-to-empirical approach was adopted for the

first iteration of the taxonomy development, whereby the initial dimensions of the taxonomy were constructed without examining the objects (sentiment analysis research papers) to be classified. The first five articles were categorized in the first iteration. The second iteration involved empirical-to-conceptual procedure, where the remaining sample of 13 studies were classified by inspecting their characteristics in each of the initial dimensions of the taxonomy. During this process, the dimensions and categories of the taxonomy may be modified, either by introducing new dimensions or categories or by removing or merging existing dimensions or categories.

3. Results

In this section, we summarize the findings from our scoping review, outlining the key dimensions and categories of the taxonomy (Figure 1). These are grouped into taxonomy dimensions relating to research context (e.g., application domains and theories), data characteristics (e.g., datasets, data sources, classification labels) and methods (e.g., sentiment analysis pipelines, types of systems, dialogue specific information used). More detailed information can be found on the OSF.

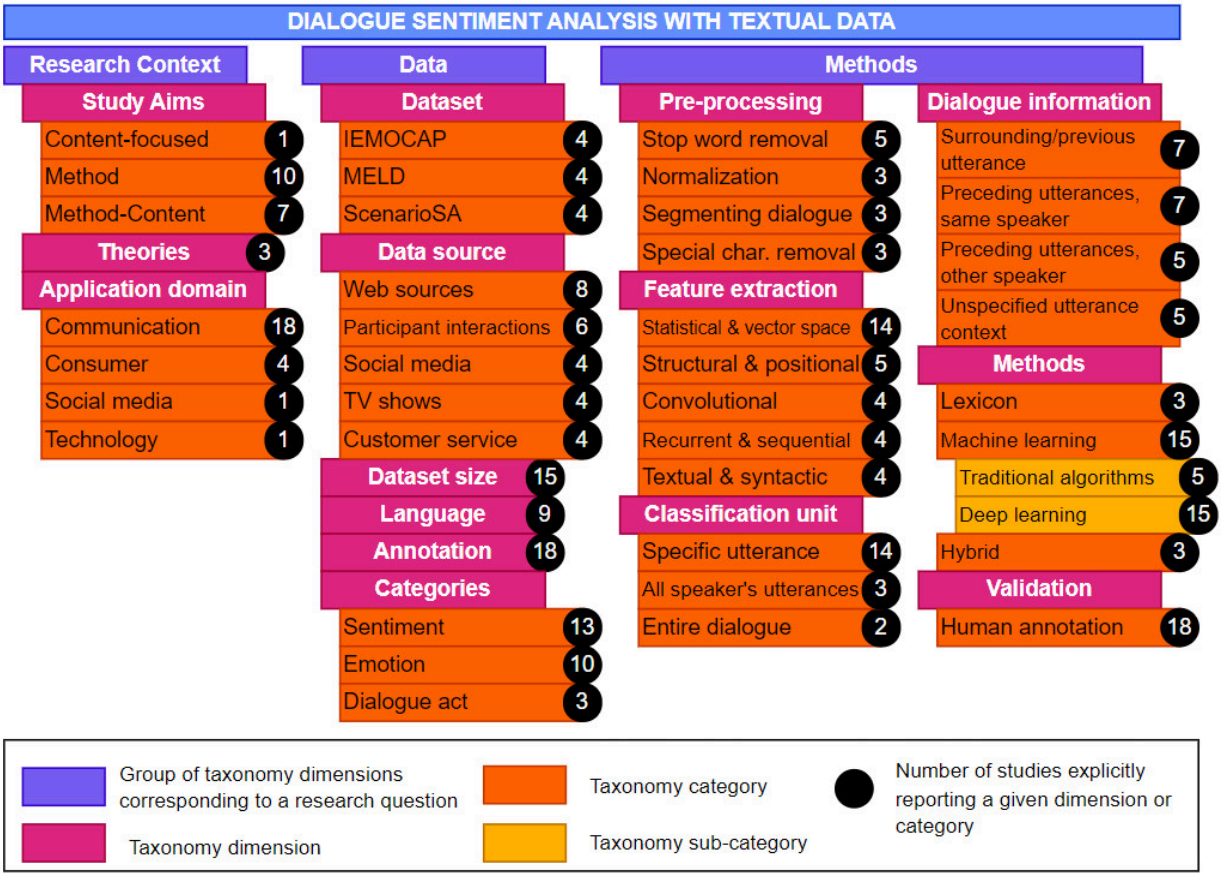

Figure 1. A simplified taxonomy of dialogue sentiment analysis literature (18 studies), depicting the key dimensions and categories within them.

3.1. Research context

The primary aims of the studies fell into three categories, which we refer to as Method studies, where sentiment analysis methods were evaluated or compared (10 studies), Content-focused studies, which aimed to answer a specific research question in the field (one study) and Method-Content studies, which consisted of both method evaluation and field-specific investigations (seven studies). The research questions in the field revolved around dialogue sentiment analysis, such as investigations of the importance of context in dialogue sentiment analysis (four studies) or the usefulness of joint topic modelling and sentiment analysis in dialogue sentiment analysis (two studies).

The domain of all studies was communication, as the studies revolved around sentiment analysis method evaluation or investigations in dialogue processing. Additional application domains were consumer-related (four studies), social media and technology (one study each). The consumer-related studies used customer support dialogues with interest in improving the customer support systems' language understanding capabilities.

The motivation for the proposed systems and investigations mainly came from previous work in computer science and related fields, without explicit mention of theories. By contrast, only three studies drew from psychological, sociological or linguistic theories or principles, such as Plutchik's wheel of emotions (Plutchik, 2001) or the principle of compositionality (Partee, 1995).

3.2. Data characteristics

Existing datasets were used in 10 studies, while the authors constructed their own datasets in eight studies. All 18 studies included either a description of the dataset or a citation to the original dataset. The most popular datasets were IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a) and ScenarioSA (Zhang et al., 2020), each used in four studies. These were followed by EmoContext dataset for task 3 in SemEval-2019 (Chatterjee et al., 2019) used in three studies and DailyDialoge (Li et al., 2017) and Mastodon (Cerisara et al., 2018) datasets, each used in two studies. The source of data used in the studies was most often from the web, such as websites that support online communication or dialogues between user and conversational agent (eight studies), participant interactions specifically for data collection purposes (six studies), social media, such as Twitter and Mastodon (four studies), TV-show dialogues (four studies) and communications between customers and service agents from different companies (four studies).

The size of the datasets used for sentiment analysis were typically reported as the number of dialogues, utterances and/or as the average number of turns per dialogue. We thus tracked measures of dataset size where possible, including all available information reported in each study. For studies with multiple datasets, the information about the largest dataset was recorded. The number of dialogues was reported in 15 studies, the number of utterances in 12 studies, (both were reported in 12 studies), and the average number of turns per dialogue was reported in four studies. Additionally, the unit of data was not clear in two studies, and no information regarding the dataset size was provided in one study. The total/maximum number of dialogues reported ranged from 5 to 32,915, with an average of 5,245.07 (*SD* = 8,703.65) and median of 2,214 across studies. The total/maximum number of utterances ranged from 508 to 102,979, with an average of 24,915.58 (*SD* = 30,533.09) and median of 13,708 across studies. The average turns per dialogue ranged from 5.32 to 7.

The language of the dialogues was reported explicitly in nine of the studies, and it was inferrable in the remaining nine studies based. The languages sentiment analysis was conducted in were English (12 studies), Chinese (four studies), Japanese, and Portuguese (one study each). The data modalities available in the datasets used were text only in 12 studies, audio, and text in one study, and audio, text and visual (image or video) in five studies.

Data annotation by human raters was reported in all 18 studies, with additional labels provided by the customers involved in the dialogues used in one study. The number of annotators was most often three (nine studies), followed by two, five and seven annotators (three studies each). However, full details about the annotation process varied between the datasets, including how disagreements between annotators were handled to generate the final annotation, information about the annotators, the instructions given to the annotators, and the details and measures of inter-rater reliability.

The dialogue data were categorized into a variety of classification categories, which were most commonly sentiment (13 studies), emotion (10 studies) or dialogue act (three studies) categories. The sentiment categories ranged from 2-class systems (e.g., negative, non-negative or positive, negative) to 5-class systems (e.g., very negative/positive, negative/positive, neutral). The emotion categories ranged from 3-class to 9-cass systems, including different variations of basic emotions such as anger, happiness, sadness,

excitement, frustration, fear, surprise, and others. The dialogue act categories ranged from 4-class systems (with labels inform, questions, directives, commissive) to 28-class systems (e.g., acknowledge, agree, appreciation). Overall, 23 unique classification category sets were used across the 18 studies.

3.3. Sentiment analysis methods

Pre-processing of the dialogues was reported in eight studies, while all 18 studies included some description of feature extraction (see Figure 1). However, the features and feature extraction methods used in the studies were very diverse and often complex, such as different deep learning models or new approaches for extracting the necessary dialogue information.

The unit of text classified by different sentiment analysis systems was most commonly a specific utterance in the dialogue (14 studies), followed by all utterances of a given speaker (three studies) and the entire dialogue (two studies).

Different types of information were utilized in determining the sentiment of the unit of interest. As several classifiers were compared in most studies, this information also varied by a classifier used within a single study. The most common information considered were surrounding utterances without specifying the speaker and preceding utterances from the same speaker (seven studies each). This was followed by preceding utterances from the other speaker and unspecified utterance context (five studies each). Overall, while the distance of the utterance context considered varied across studies, most approaches in each of these categories tended to utilize maximal amount of context, i.e., all the preceding or surrounding utterances. The context information used remained unspecified in five studies. By contrast, the speaker information used was not specified in 12 studies, for all the approaches used within a study (this excludes studies in which the baseline methods were not described in detail). Finally, the benefit of considering the topic of the dialogues in sentiment classification was considered in two studies.

Most studies employed traditional sentiment analysis (14 studies), i.e., sentiment detection without considering the target of the sentiment. The topic of the conversation or utterance was tracked in three studies and aspect-based sentiment analysis was applied in one study.

All except one study included evaluation of a new sentiment analysis system, framework, or method for dialogue sentiment analysis. Most studies compared the performance of their proposed sentiment analysis method to several baseline methods, and multiple classifiers were compared in 12 out of 18 studies. The number of different sentiment analysis classifiers evaluated in the studies ranged from one (four studies) to 19 (one study). A total of 120 classifiers were evaluated across the 18 studies, out of which 84 were unique methods (including a handful of different versions of the same method).

The sentiment analysis methods used in the studies were classified into the following broad categories: Lexicon-based, Machine learning, and Hybrid methods. The specific lexicon-based methods used were SentiStrength (two studies), Bag of Words and SenticNet (one study each). Machine learning methods covered 'traditional machine learning' methods such as SVMs, ensemble methods (e.g., Random Forest, AdaBoost), through to deep learning models (e.g., LSTMs, BERT and its variants, Conversational Memory Networks). Out of 15 studies that reported using at least one machine learning approach, traditional algorithms were used in five studies and deep learning methods in 15 studies. The hybrid category includes sentiment analysis systems that leverage a combination of machine learning, rule-based and/or lexical resources. These methods were used in three studies. Validation in all 18 studies was based on human annotation.

4. Discussion

Our taxonomy aims to provide an in-depth overview of the new and complex application of sentiment analysis to dialogue, which will only expand into other types of natural language processing (NLP) tasks as we continue to develop more tools interacting with people in real-time. The overview of the literature provided by our taxonomy work reveals several interesting topics to consider. Below, we discuss these findings, particularly regarding challenges we identified in the current state of research in dialogue sentiment analysis. We emphasize the methodological quality of research as a crucial component for advancing the field.

4.1. Research context

Considering the theoretical background of sentiment analysis approaches, a critical aspect often overlooked in our sample of studies was the conceptual validity of the classification labels used in annotation. Different subjective states, such as emotion, sentiment and opinion, differ in important characteristics, including the level of consciousness, duration of time and cultural influence associated

with these subjective states (Munezero et al., 2014). Due to these differences, subjective states are also expressed differently in natural language. It is therefore not trivial how the subjective state of interest is defined and subsequently annotated in language data (e.g., sentiment or emotion categories? Which categories are included?). In our sample, authors frequently borrowed categories from previous computer science literature without explicitly considering their psychological validity or relevance to real-world sentiment expressions. This was also seen in very few studies drawing from psychological or linguistic theories. Neglecting the psychological validity of classification labels can lead to the inclusion of arbitrary or poorly defined categories that may not accurately capture the nuances of human sentiment. As a result, sentiment analysis systems trained and evaluated on such datasets may produce skewed or inaccurate results.

4.2. Data characteristics

Although human annotation was used as ground truth in each included study, the rigor in generating the labels and reporting the annotation process varied. For instance, while practically all datasets were annotated by at least two annotators, the procedures for handling disagreements differed, from using the judgement of only one of the annotators, to majority voting and adjudication by an "expert".

These types of variability across datasets make evaluation of sentiment analysis methods challenging. It is difficult to assess sentiment analysis methods with several datasets when the reliability of the annotations across datasets is not comparable. Findings based on unreliable ground truth can lead to misguided sentiment analysis system development, and potentially waste research efforts. As another example, the large number of classification categories across datasets also complicates evaluation of sentiment analysis systems. The number of categories directly influences the complexity of the classification task, where a higher number of categories can make the task more challenging for classifiers (Hartmann et al., 2023). Sentiment analysis systems trained and evaluated on datasets with different numbers of categories may exhibit varying levels of performance, making it difficult to assess their true capabilities across different contexts. Each set of categories effectively represents a distinct classification task, with different levels of complexity and granularity. To address this challenge, it is essential to evaluate sentiment analysis systems on multiple datasets with the same number of classification categories. This approach allows for a more meaningful comparison of performance across different contexts and ensures that success is attributed to the sentiment classification task itself rather than the specific dataset characteristics.

4.1. Sentiment analysis methods

In our sample, additional sources of information were explored to augment dialogue sentiment analysis, such as simultaneous dialogue act recognition or topic modelling. We also note interest in leveraging multimodal information (e.g., audio or visual cues), investigated in several studies (which were excluded from our scoping review due to our focus on textual data). This raises a broader question regarding the type of information that is most beneficial for sentiment analysis systems. While the usefulness of different types of information are explored and demonstrations of performance improvements are provided (at least with some datasets), these performance improvements are often modest. Based on our sample of studies, it appears that the priority of investigations is not in identifying and understanding the core features of dialogue that are most predictive of sentiment via systematic investigations (cf., Carvalho et al., 2023). Instead, the current research efforts are dispersed across various novel, but potentially less impactful approaches. In order to steer the research efforts into maximally beneficial areas, we need to ask what kind of performance improvement is sufficient to justify increased system complexity?

The vast majority of methods proposed for dialogue sentiment analysis were based on deep learning, often leveraging a combination of several deep learning-based feature extraction methods and classifiers. Deep learning methods, such as complex neural networks (e.g., CNNs and RNNs) are often considered black box methods due to the difficulty in interpreting how these models produce their output (Choudhary et al., 2022). Furthermore, as the complexity of the systems increases, it becomes more challenging to report all the relevant information for reproducibility, or to keep track of what information should be reported. Unfortunately, no models were shared in our sample of studies, however, one would hope that the process to train the model(s) (including sharing the training data) would be clear and well-documented to allow for reproduction and to align with open science principles.

In our sample, well justified comparisons of performance across sentiment analysis methods were rare (reported in only one study), which seems to reflect the general situation in NLP research (Dror et al., 2018). More broadly, attention should be paid on

appropriate and well thought through method evaluation, including baseline model selection, suitable for the specific task. Although some ablation studies and method comparisons aiming to uncover the important characteristics of dialogue sentiment analysis systems were present in individual studies, the literature in dialogue sentiment analysis seems to suffer from the same fragmentation as the sentiment analysis literature more generally (e.g., Äyräväinen et al., 2023), that is, new systems are introduced at a fast pace, without comprehensive and standardized evaluation of these systems. With such diverse approaches to dialogue sentiment analysis, it is increasingly difficult to identify the specific aspects of different methods that are particularly beneficial in dialogue sentiment analysis. This hinders building on promising approaches and focusing on areas of improvement, and as such, prevents efficient evolution of reliable and successful dialogue sentiment analysis systems.

5. Limitations and Future directions

Comprehensive knowledge of previous work is necessary for efficient advancements in the field of dialogue sentiment analysis. Our work is limited by the streamlined literature search, which increases the likelihood of missing relevant studies, thus leaving room for a more exhaustive literature review in the future. Additionally, due to the complexity of the proposed systems in the sampled studies, our categorisation of these is best seen as one potential solution for the method classification. However, our preliminary taxonomy serves as a starting point for mapping the literature, and our documentation should facilitate further development of this taxonomy, with new dimensions or categories added as seen relevant.

Due to the recent interest in generative AI, we note that these systems may become more relevant in the future of dialogue sentiment analysis (this method category is already in our taxonomy, where one study evaluated generative models). However, being designed for dialogue generation, the current performance of these systems in detecting unidirectional sentiment falls short from that of deep learning models specifically trained for sentiment analysis (Anas et al., 2024). Future impact of generative AI in dialogue sentiment analysis thus likely relies on fine-tuning the systems for this specific task. Yet, the opacity of generative AI models, like other deep learning techniques, raises concerns about their interpretability. Further research in this area should thus focus on enhancing the explainability and accountability of these models.

Based on our findings of the current state of dialogue sentiment research, we provide recommendations, outlined in the remainder of this section. The field of dialogue sentiment analysis is clearly in its infancy, and it would greatly benefit from more principled exploration of dialogue sentiment, via systematic inquiry of the core characteristics that best predict sentiment in conversational contexts. This could be achieved by shifting from competitive to comprehensive testing, where the aim is to increase in-depth understanding of what is required to efficiently analyze sentiment in dialogues, rather than limited comparisons of performance metrics achieved by systems that differ from each other in several ways. These efforts would also benefit from more interdisciplinary research, leveraging knowledge from linguistics and psychology in both system development and dataset construction. For instance, Benamara et al. (2017) outline how more linguistically motivated work, including discourse information and other contextual phenomena, should be utilized in future sentiment analysis approaches. Ultimately, these efforts should also help with moving towards more standardized datasets, when the relevant dialogue characteristics are established, thus also facilitating more informative and standardized system comparisons.

When using existing dialogue datasets, researchers should consider whether the classification categories and other annotated information (e.g., dialogue acts, speaker intent) are suitable for their specific purposes. Critically assessing the reliability of these annotations (e.g. inter-annotator agreement, disagreement handling) is also needed. The choice of dialogue information used for the task at hand should ideally be justified, via theory or empirical evidence. At present, the empirical evidence remains scarce, but, for instance, deciding on the utterance context used and whether to track speaker information may be informed by the few studies addressing these questions (e.g., Carvalho et al., 2023).

Commitment to open science practices is paramount for efficient advancements in the field. This includes both meticulous reporting of the full sentiment analysis pipeline and sharing datasets (Gebru et al., 2021), code and models (Mitchell et al., 2019). For instance, to advance efficient exploration of the core characteristics of dialogues for sentiment prediction, researchers should provide a clear explanation of the information, such as type of context and speaker information employed in their systems. This may be local context (such as adjacent, intra-speaker or inter-speaker utterances) or global context (entire conversation threads or inter-speaker dynamics), and description of how exactly this

context is integrated into the sentiment analysis process. Clarifying the role of context and speaker is crucial for multiple reasons: (1) it enhances the system's transparency, thus facilitating a deeper understanding of its operational mechanics; (2) it aids in the system's further refinement and development by pinpointing areas of strengths and weaknesses; and (3) it empowers users to make informed assessments regarding the system's credibility and reliability.

6. Conclusion

Based on our scoping review, we developed a taxonomy considering the application of various sentiment analysis approaches to textual dialogue data. This was to provide a framework for navigating complex literature, to facilitate understanding of the currently used datasets, methods, and other attributes of the studies, and to offer an overview of the work conducted so far. Despite considerable and increasing interest in sentiment analysis more generally (e.g., Cui et al., 2023), the specific application to dialogue data is an emerging area of research, with diverse approaches for leveraging dialogue-specific information in sentiment detection. It is important for this work to continue, but we urge the research community to focus on more multidisciplinary and theoretically driven approaches, ensuring well-documented techniques and modelling, committing to open science practices and to ensure appropriate testing and evaluation of proposed systems. This work has a number of applications, such as improved dialogue systems in customer service, education and mental health services. Thorough understanding of sentiment in dialogues is thus fundamentally important.

7. Acknowledgements

This work has been funded by the UK Government grant awarded to BID and JH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

8. References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. Journal of medical Internet research, 22(4), e19016. https://doi.org/10.2196/19016

Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert systems with applications, 167, 114155. https://doi.org/10.1016/j.eswa.2020.114155

Anas, M., Saiyeda, A., Sohail, S. S., Cambria, E., & Hussain, A. (2024). Can Generative AI Models Extract Deeper Sentiments as Compared to Traditional Deep Learning Algorithms?. IEEE Intelligent Systems, 39(2), 5-10. https://doi.org/10.1109/mis.2024.3374582

Äyräväinen, L. E., Hinds, J., & Davidson, B. I. (2023). Sentiment Analysis in Digital Spaces: An Overview of Reviews. arXiv preprint https://doi.org/10.48550/arXiv.2310.19687

Benamara, F., Taboada, M., & Mathieu, Y. (2017). Evaluative language beyond bags of words: Linguistic insights and computational applications. Computational Linguistics, 43(1), 201-264. https://doi.org/10.1162/coli_a_00278

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, 335-359. https://doi.org/10.1007/s10579-008-9076-6

Carvalho, I., Silva, C., & Oliveira, H. G. (2023). The importance of context for sentiment analysis in dialogues. IEEE Access. https://doi.org/10.1109/access.2023.3304633

Cerisara, C., Jafaritazehjani, S., Oluokun, A., & Le, H. T. (2018, August). Multi-task dialog act and sentiment recognition on Mastodon. In COLING. https://hal.science/hal-01838323/

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019, June). SemEval-2019 task 3: EmoContext contextual emotion detection in text. In Proceedings of the 13th international workshop on semantic evaluation (pp. 39-48). https://doi.org/10.18653/v1/s19-2005

Chen, S., & Nakamura, M. (2021, November). Generating personalized dialogues based on conversation log summarization and sentiment analysis. In The 23rd International Conference on Information Integration and Web Intelligence (pp.217-222). https://doi.org/10.1145/3487664.3487695

Chin, H., Lima, G., Shin, M., Zhunis, A., Cha, C., Choi, J., & Cha, M. (2023). User-chatbot conversations during the COVID-19 pandemic: study based on topic modeling and sentiment analysis. Journal of medical Internet research, 25, e40922. https://doi.org/10.2196/40922

Choudhary, S., Chatterjee, N., & Saha, S. K. (2022). Interpretation of black box nlp models: A survey. arXiv preprint https://arxiv.org/pdf/2203.17081

Cui, J., Wang, Z., Ho, S. B., & Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. Artificial Intelligence Review, 1-42. https://doi.org/10.1007/s10462-022-10386-z

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018, July). The hitchhiker's guide to testing statistical significance in natural language processing. In

Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 1383-1392). https://doi.org/10.18653/v1/p18-1128

El-Ansari, A., & Beni-Hssane, A. (2023). Sentiment analysis for personalized chatbots in e-commerce applications. Wireless Personal Communications, 129(3), 1623-1644. https://doi.org/10.1007/s11277-023-10199-5

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Gunasekaran, K. P. (2023). Exploring sentiment analysis techniques in natural language processing: A Comprehensive Review. arXiv preprint https://arxiv.org/pdf/2305.14842

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75-87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Krugmann, J. O., & Hartmann, J. (2024). Sentiment Analysis in the Age of Generative AI. Customer Needs and Solutions, 11(1), 1-19. https://doi.org/10.1007/s40547-024-00143-4

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017, November). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 986-995). https://aclanthology.org/I17-1099

Li, P., Yan, Y., Wang, C., Ren, Z., Cong, P., Wang, H., & Feng, J. (2016, July). Customer voice sensor: A comprehensive opinion mining system for call center conversation. In 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 324-329). https://doi.org/10.1109/icccbda.2016.7529578

Ma, J., Rong, L., Zhang, Y., & Tiwari, P. (2023). Moving From Narrative to Interactive Multi-Modal Sentiment Analysis: A Survey. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3610288

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229). https://doi.org/10.1145/3287560.3287596

Mogaji, E., Balakrishnan, J., Nwoba, A. C., & Nguyen, N. P. (2021). Emerging-market consumers' interactions with banking chatbots. Telematics and Informatics, 65, 101711. https://doi.org/10.1016/j.tele.2021.101711

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. IEEE transactions on affective computing, 5(2), 101-111. https://doi.org/10.1109/taffc.2014.2317187

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. European Journal of Information Systems, 22(3), 336-359. https://doi.org/10.1057/ejis.2012.26

Partee, B. (1995). Lexical semantics and compositionality. An invitation to cognitive science: Language, 1, 311-360. https://www.cs.brandeis.edu/~jamesp/classes/cs216-2009/readings2009/ParteeSemanticsAndCompositionality.pdf

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist, 89(4), 344-350. https://doi.org/10.1511/2001.28.344

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019a). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1050

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019b). Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 7, 100943-100953. https://doi.org/10.1109/access.2019.2929050

Skoric, M. M., Liu, J., & Jaidka, K. (2020). Electoral and public opinion forecasts with social media data: A meta-analysis. Information, 11(4), 187. https://doi.org/10.3390/info11040187

Usman, M., Britto, R., Börstler, J., & Mendes, E. (2017). Taxonomies in software engineering: A systematic mapping study and a revised taxonomy development method. Information and Software Technology, 85, 43-59. https://doi.org/10.1016/j.infsof.2017.01.006

Zhang, Y., Zhao, Z., Wang, P., Li, X., Rong, L., & Song, D. (2020). ScenarioSA: a dyadic conversational database for interactive sentiment analysis. IEEE Access, 8, 90652-90664. https://doi.org/10.1109/access.2020.2994147