I. Introduction

Sentiment analysis, a critical area of research in natural language processing (NLP), focuses on identifying and quantifying subjective information like attitudes and opinions expressed in text (Hutto & Gilbert, 2014). Its applications span customer feedback analysis, social media monitoring, market research, and political analysis, making it indispensable for understanding public opinion and aiding decision-making in businesses and organizations (Liu, 2020).

Early sentiment analysis methods relied on lexicon-based approaches, but they often failed to capture context and linguistic variations effectively (Tang et al., 2015). This led to the emergence of deep learning models, like recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, which have shown remarkable success in sentiment analysis tasks (Tang, Qin & Liu, 2015). RNNs, particularly long short-term memory (LSTM) networks, excel in capturing sequential dependencies. CNNs are adept at extracting local patterns from text (Chen et al., 2017). Transformers (Vaswani et al., 2017), popularized by models like BERT (Devlin et al., 2018), effectively capture contextual relationships and global dependencies.

Deep learning models have a significant advantage as they can automatically learn representations from large-scale datasets, eliminating the need for extensive feature engineering. They excel in capturing complex linguistic relationships and subtle sentiment nuances (Ma et al., 2017), which is especially achievable through unsupervised pre-training techniques like BERT.

However, deep learning approaches also present challenges. One major concern is their lack of interpretability (Lipton, 2018), making it challenging to understand their

decision-making process. Efforts are being made to develop interpretable attention mechanisms and explainability techniques to address this issue (Li et al., 2022; Minh et al., 2022). Additionally, deep learning models may struggle with generalization when faced with data from different domains or languages. Transfer learning techniques like domain adaptation and cross-lingual learning aim to improve performance in such scenarios (Chan et al., 2023). Biases in sentiment analysis models have also received significant attention (Kiritchenko & Mohammad, 2018). Addressing biases and promoting fairness in sentiment analysis models are essential to avoid perpetuating existing societal biases and ensure equal treatment across different groups.

This literature review aims to explore relevant research studies related to sentiment analysis with deep learning, discussing and evaluating the context of use, methodologies used and challenges faced. Section II enumerates and discusses key research studies identified; the section follows a structure and review format similar to selected sections of established literature review papers, specifically Sections 4.4 by Wankhade et al. (2022) and Section 4.3 by Xu et al. (2022). Section III explores the challenges and ethical implications of sentiment analysis. Section IV provides a summary and concludes the review.

II. Review of Deep Learning for Sentiment Analysis

Deep learning models have revolutionized sentiment analysis tasks, offering powerful capabilities for capturing sentiment patterns and extracting meaningful representations from text. This section provides an analysis of key papers that employ deep learning approaches for sentiment analysis. The research design, methodology, strengths, and limitations of each paper are critically evaluated. The discussion is

chronologically ordered in ascending order of the year of publication in order to provide a sense of the development of the field.

Tang et al. (2015) proposed a sentiment analysis model based on gated RNNs (GRNNs) to capture the sequential information present in text. The model employs a gate mechanism to control the flow of information and model long-term dependencies. One strength of this approach is that GRNNs effectively capture the sequential nature of text and consider the temporal dependencies in sentiment expression. The gate mechanism allows the model to control the flow of information, making it suitable for capturing long-term dependencies. Additionally, the model's document-level modelling approach provides a comprehensive representation of sentiment. However, GRNNs may suffer from the vanishing gradient problem, affecting their ability to capture long-term dependencies effectively. Furthermore, the model's performance may vary depending on the length and complexity of the input text. Additionally, the document-level modelling approach may overlook fine-grained sentiment variations at the sentence or phrase level (Tang, Qin & Liu, 2015).

The paper by Vaswani et al. (2017) introduced the Transformer model, which is a revolutionary architecture in natural language processing (NLP) that has profoundly impacted sentiment analysis. The work proposes a self-attention mechanism that allows the model to capture contextual dependencies in the input text, revolutionizing the way language sequences are processed. The Transformer introduces a novel architecture that does not rely on traditional recurrent or convolutional layers, addressing the limitations of previous models and achieving superior performance in various NLP tasks. Its self-attention mechanism enables the Transformer to capture long-range dependencies, crucial for understanding the sentiment expressed in longer text sequences. However, the paper lacks an in-depth analysis of the model's

interpretability. Understanding the inner workings of the self-attention mechanism is essential for gaining insights into the model's decision-making process and its potential biases in sentiment analysis. Moreover, the Transformer's large number of parameters and self-attention mechanism contribute to significant computational complexity, making it computationally intensive and challenging to deploy on resource-limited devices or in real-time applications. Nevertheless, this method is currently the best method of providing a level of explainability and interpretability in sentiment analysis models and deep learning models in general, which are notoriously opaque (Zanwar et al., 2023).

Ma et al. (2017) proposed a sentiment analysis model based on a hybrid attention mechanism that combines self-attention and interactive attention. The authors aimed to improve interpretability by allowing the model to attend not only to important words but also to the interaction between different words in a sentence. The interactive attention mechanism captures the dependencies among words, providing insights into the contextual relationships and sentiment expression, as well as enhancing interpretability by highlighting the dependencies that contribute to sentiment predictions. The model's performance heavily relies on the availability of annotated datasets with aspect-level sentiment labels which are not readily available and are time-consuming and expensive to collect. Additionally, the interpretability of the interactive attention mechanism may not align with human intuition (Song et al., 2018), as it requires users to understand the complex interactions between words. The attention mechanism also introduces significant additional computational overhead and complexity.

Baziotis et al. (2017) proposed a deep learning model based on LSTM networks with attention mechanisms for message-level and topic-based sentiment analysis. The model employs hierarchical attention to capture relevant information at different levels of granularity. This approach has the advantage of focusing on the most informative parts of the text for sentiment analysis through the hierarchical attention mechanism. It achieves competitive performance in message-level and topic-based sentiment analysis tasks. Additionally, attention mechanisms enhance the model's interpretability by highlighting salient words and phrases. However, the hierarchical attention mechanism may introduce additional complexity and computational overhead. As in previous studies, this model's performance heavily relies on the availability and quality of annotated datasets for training. The interpretability aspect may be limited to the attention scores and may not provide a comprehensive understanding of sentiment analysis decisions (Baziotis, Pelekis & Doulkeridis, 2017).

Chen et al. (2017) proposed a sentiment analysis model that combines bidirectional LSTM (BiLSTM) with conditional random fields (CRFs) and a CNN. The BiLSTM-CRF component captures contextual information and utilizes the CRF layer for sequence labelling, while the CNN captures local patterns in the text. This approach has the advantage of effectively capturing contextual information and modelling the dependencies between words. The CRF layer ensures consistent sentiment predictions by considering the entire sequence. Moreover, the CNN component captures local patterns and strengthens the model's ability to capture sentiment nuances and local patterns. However, the model's performance may be affected by the complexity of the CRF layer, requiring more computational resources. Additionally, the sentence-level analysis may overlook the sentiment variations present within sentences. The combination of multiple components, i.e. LSTM with CRFs and CNNs,

introduces additional complexity and challenges in model training and optimization (Chen et al., 2017).

The work by Devlin et al. (2018) on BERT (Bidirectional Encoder Representations from Transformers) is a very prominent study which introduced a novel model, as well as pre-training and fine-tuning approach for sentiment analysis. BERT leverages a deep bidirectional transformer architecture and unsupervised pre-training on a large corpus to learn contextualized word representations. Fine-tuning on sentiment analysis tasks enables the model to capture subtle sentiment patterns effectively. The key strength of BERT lies in its ability to capture complex linguistic relationships and contextual dependencies, leading to highly accurate sentiment predictions. However, a drawback of the approach is its requirement for substantial computational resources for training and fine-tuning, limiting its applicability in resource-constrained environments. Additionally, the model's large size and complex architecture provide a significant challenge to interpretability as compared to simpler models (Devlin et al., 2018).

Wang and Cao (2020) proposed an aspect-specific sentiment analysis model that focuses on capturing sentiment variations for different aspects of a product or service. The model leverages a hierarchical architecture that incorporates aspect information to make fine-grained sentiment predictions. This approach has the advantage of effectively capturing aspect-level sentiment variations, providing detailed insights into different aspects of the target entity. Fine-grained sentiment predictions enhance the model's applicability in practical scenarios. However, as with previously explained models, this model's performance heavily depends on the availability and quality of aspect-labelled datasets. Moreover, fine-grained sentiment analysis introduces additional complexity and potential challenges in model training and evaluation (Wang & Cao, 2020).

III. Challenges and Ethical Concerns of Sentiment Analysis with Deep Learning

This section critically evaluates the ethical implications associated with sentiment analysis and deep learning, highlighting the potential challenges and risks involved.

Sentiment analysis often requires access to large amounts of user-generated content, such as social media posts, reviews, and comments. The collection and use of such data raise concerns regarding privacy and data protection. It is crucial to ensure that user consent is obtained, and data is handled in accordance with relevant regulations (Hemphill, Schöpke-Gonzalez & Panda, 2022). Additionally, the anonymization and aggregation of data should be implemented to mitigate privacy risks (Majeed & Lee, 2020). The importance of this is highlighted in high-profile incidents such as the Cambridge Analytica incident (Rosenberg, Confessore & Cadwalladr, 2018) and Unroll.Me indicent (Bowles, 2017) incident, among others. Transparency and clear communication are crucial to ensure that individuals are aware of how their data is being used for sentiment analysis purposes (Majeed & Lee, 2020).

The bias and fairness of deep learning models has also recently gained scrutiny. Deep learning models are susceptible to biases present in the training data, which can lead to biased sentiment analysis results. Biases can emerge from imbalanced training data, underrepresentation of certain groups, or biased annotations in labelled datasets (Kiritchenko & Mohammad, 2018). Such biases may have consequences and disproportionately impact vulnerable communities. Biased or inaccurate sentiment analysis results can perpetuate harmful stereotypes, discrimination, or unfair treatment of certain groups (Sweeney, 2013). It is essential to critically evaluate the

training data and implement bias detection and mitigation techniques (Dwork et al., 2012).

Deep learning-based sentiment analysis models have significant potential to be manipulated or misused. Adversarial attacks can be launched to manipulate sentiment predictions and influence public opinion (Wallace et al., 2019) such as the Cambridge Analytica (Rosenberg, Confessore & Cadwalladr, 2018). There is a need to develop and implement safeguards to prevent malicious manipulation and misuse of sentiment analysis systems.

The interpretability of deep learning models was a major theme in the previous sections. Deep learning models still largely lack interpretability, making it challenging to understand how they arrive at their predictions. The interpretability of deep learning models is a largely unsolved problem, despite the fact that it is crucial in domains where transparency, accountability, and explainability are required (Lipton, 2018). Research efforts are being made to develop interpretable models and explainability techniques to shed light on the decision-making process of deep learning models (Adak, Pradhan & Shukla, 2022). However, achieving a balance between interpretability and performance remains a challenge.

Hand-in-hand with interpretability is algorithmic transparency and accountability. The opaque nature of deep learning models poses challenges in ensuring algorithmic transparency and accountability. The functioning and decision-making mechanisms of complex deep sentiment analysis models are still largely opaque, and the accountability of these models is still unresolved (Diakopoulos et al., 2017).

IV. Summary and Conclusions

Sentiment analysis is a vital field in natural language processing (NLP) that identifies attitudes and opinions expressed in text. It finds applications in customer feedback analysis, social media monitoring, market research, and political analysis. Deep learning models like RNNs, CNNs, and transformers have been successful in sentiment analysis tasks. RNNs capture sequential dependencies, CNNs extract relevant features, and transformers handle contextual relationships and global dependencies.

The strength of deep learning models lies in automatically learning representations from large datasets, significantly reducing or even eliminating the need for feature engineering. They excel in capturing complex linguistic relationships and subtle sentiment nuances through techniques like BERT pre-training and have consistently achieved state-of-the-art predictive performance.

However, challenges persist, such as the lack of interpretability in deep learning models, hindering their understanding and accountability. The complexity of these models, coupled with the need for significant computational resources and high-quality annotated data, are additional challenges faced.

Furthermore, biases in sentiment analysis models are a concern. They can perpetuate societal biases, requiring efforts to promote fairness and equal treatment.

All in all, sentiment analysis using deep learning shows significant promise and has numerous potentially beneficial applications, but interpretability, generalization, and bias mitigation are crucial considerations for responsible and beneficial use.

V. References

Adak, A., Pradhan, B. & Shukla, N. (2022) Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review. *Foods*. 11 (10), 1500–1516. https://www.mdpi.com/2304-8158/11/10/1500.

Baziotis, C., Pelekis, N. & Doulkeridis, C. (2017) DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In: *Proceedings of the 11th international workshop on semantic evaluation*. 2017 pp. 747–754. https://aclanthology.org/S17-2126/.

Bowles, N. (2017) Unroll.Me Service Sifts Through Users' Email, Sold Data to Uber. *The New York Times*. https://www.nytimes.com/2017/04/24/technology/personal-data-firm-slice-unroll-me-backlash-uber.html.

Chan, J.Y. Le, Bea, K.T., Leow, S.M.H., Phoong, S.W. & Cheng, W.K. (2023) State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*. 56 (1), 749–780. doi:10.1007/S10462-022-10183-8.

Chen, T., Xu, R., He, Y. & Wang, X. (2017) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*. 72, 221–230. doi:10.1016/j.eswa.2016.10.065.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*. 11 October 2018

Association for Computational Linguistics (ACL). pp. 4171–4186. https://arxiv.org/abs/1810.04805v2.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H.V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S. & Wilson, C. (2017) Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. (2012) Fairness through awareness. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012 pp. 214–226. doi:10.1145/2090236.2090255.

Hemphill, L., Schöpke-Gonzalez, A. & Panda, A. (2022) Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*. 9 (1), 643–657. https://www.nature.com/articles/s41597-022-01773-w.

Hutto, C. & Gilbert, E. (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*. 2014 pp. 216–225.

Kiritchenko, S. & Mohammad, S.M. (2018) Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In: *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*. 2018 Association for Computational Linguistics (ACL). pp. 43–53. doi:10.18653/v1/s18-2005.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J. & Dou, D. (2022) Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*. 64 (12), 3197–3234. doi:10.1007/S10115-022-01756-8.

Lipton, Z.C. (2018) The Mythos of Model Interpretability. *Queue*. 16 (3), 31–57. doi:10.1145/3236386.3241340.

Liu, B. (2020) *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Ma, D., Li, S., Zhang, X. & Wang, H. (2017) Interactive Attention Networks for Aspect-Level Sentiment Classification. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2017 pp. 4068–4074. http://alt.qcri.org/semeval2014/task4/.

Majeed, A. & Lee, S. (2020) Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access*. 9, 8512–8545. https://ieeexplore.ieee.org/abstract/document/9298747/.

Minh, D., Wang, H.X., Li, Y.F. & Nguyen, T.N. (2022) Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*. 55 (5), 3503–3568. doi:10.1007/S10462-021-10088-Y.

Rosenberg, M., Confessore, N. & Cadwalladr, C. (2018) How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

Song, H., You, J., Chung, J. & Park, J.C. (2018) Feature attention network: interpretable depression detection from social media. In: *Proceedings of the 32nd Pacific Asia conference on language, information and computation*. 2018 pp. 613–622. https://aclanthology.org/Y18-1070.pdf.

Sweeney, L. (2013) Discrimination in online Ad delivery. *Communications of the ACM*. 56 (5), 44–54. doi:10.1145/2447976.2447990.

Tang, D., Qin, B. & Liu, T. (2015) Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 conference on empirical methods in natural*. 2015 Association for Computational Linguistics. pp. 1422–1432. https://aclanthology.org/D15-1167.pdf.

Tang, D., Wei, F., Qin, B., Yang, N., Liu, T. & Zhou, M. (2015) Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*. 28 (2), 496–509.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is All you Need. In: *Conference on Advances in Neural Information Processing Systems (NIPS)*. 2017 pp. 6000–6010.

Wallace, E., Feng, S., Kandpal, N., Gardner, M. & Singh, S. (2019) Universal Adversarial Triggers for Attacking and Analyzing NLP. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019 Association for Computational Linguistics. pp. 2153–2162. https://github.com/Eric-Wallace/universal-triggers.

Wang, Z. & Cao, J. (2020) Multi-Task Learning Network for Document-level and Multi-aspect Sentiment Classification. In: *IEEE Fifth International Conference on Data Science*. 2020 pp. 171–177. https://ieeexplore.ieee.org/abstract/document/9172423/.

Wankhade, M., Rao, A.C.S. & Kulkarni, C. (2022) A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. 55 (7), 5731–5780.

Xu, Q.A., Chang, V. & Jayne, C. (2022) A systematic review of social media-based sentiment analysis: Emerging trends and challenges. *Decision Analytics Journal*. 3, 100073.

Zanwar, S., Li, X., Wiechmann, D., Qiao, Y. & Kerz, E. (2023) What to Fuse and How to Fuse: Exploring Emotion and Personality Fusion Strategies for Explainable Mental Disorder Detection. *Findings of the Association for Computational Linguistics (ACL)*. 8926–8940. https://aclanthology.org/2023.findings-acl.568/.