Current challenges in detecting complex emotions from texts

Vinh Truong, RMIT University, vinh.truongnguyenxuan@rmit.edu.vn

Abstract

Textual emotion detection is a critical area of study with significant applications in business, education, and healthcare. Despite substantial theoretical advancements over the years, there are notable gaps in the practical implementation of these methods in the aforementioned fields. The techniques currently available do not yet seem ready for real-world application. This study offers a comprehensive review of existing approaches, datasets, and models used in textual emotion detection. Its primary objective is to identify the challenges faced in both current literature and practical applications. The findings reveal that textual datasets annotated with emotional markers are scarce, making it difficult to develop robust supervised classification models for this task. There is also a pressing need for improved models that can accurately categorize a wider range of emotional states distinctly. Finally, there is a demand for techniques capable of dimensionally detecting valence, arousal, and dominance scores from emotional experiences. These challenges stem not only from the models and applications themselves but also from the readiness of current approaches and datasets in the rapidly evolving fields of machine learning and affective computing.

1. Introduction

The rapid expansion of the Internet has fundamentally transformed how individuals communicate their feelings and viewpoints. Social networking platforms have emerged as essential channels for expressing emotions to a global audience. Through textual content, pictures, audio, and video, people share their experiences, opinions, and emotions, contributing to a vast and ever-growing pool of unstructured data. Text communication, in particular, has become predominant in web-based social media, generating an overwhelming amount of data (Kondo et al., 2022).

To make sense of this massive influx of information and understand human psychology, it is crucial to process the datasets as swiftly as it is produced. One effective method for achieving this is sentiment analysis, which identifies the polarity in texts, determining whether the author’s attitude towards an object, service, person, or place is positive, negative, or neutral. However, sentiment analysis alone is sometimes insufficient for capturing the full spectrum of human emotions. This is where emotion detection comes into play, offering a more precise understanding of an individual’s emotional or mental state (Ligthart et al., 2021).

Emotions are an inseparable component of human life, deeply influencing our decision-making processes and enhancing our ability to communicate effectively with the world. The recognition and understanding of these emotions, known as emotion detection or emotion recognition, involves identifying various feelings such as joy, sadness, or fury. Over the past few years, researchers have been diligently working to automate emotion recognition, striving to develop systems that can accurately interpret human emotions from diverse inputs (Pashchenko et al., 2022).

In practice, emotion detection encompasses a wide range of physiological and behavioral indicators. The advancement of emotion detection technologies requires robust models, diverse datasets, and ethical considerations. As researchers continue to refine these systems, the potential applications in areas such as mental health, customer service, and social media analysis remain vast and promising (Alvarez-Gonzalez et al., 2021).

One of the key applications of emotion detection is on social media data. The ability to accurately measure and analyze the emotional content of online content has become increasingly valuable as the volume of user-generated text has grown exponentially (Seyeditabari et al., 2018). Emotion detection can provide insights into consumer perceptions, public sentiment, and even mental health trends. Another promising application of emotion detection is in the domain of music recommendation and entertainment. By detecting the emotions evoked by a user's listening patterns, music players can make personalized recommendations that better match the user's emotional state and preferences (Adamov, 2017).

Textual emotion detection, a subset of emotion detection, aims to identify and extract detailed emotions from written texts. It is a rapidly growing field of research within the broader domain of affective computing and natural language processing (NLP). Textual emotion detection aims to identify and classify the emotional states expressed in textual data, which can have various applications ranging from sentiment analysis of social media content to personalized music recommendations (Kusal et al., 2022).

Researchers have developed various models to facilitate emotion detection, drawing from psychological theories and computational advancements. These models aim to capture the complexity of human emotions and provide a structured framework for analysis. The implementation of such models involves sophisticated algorithms and machine learning techniques capable of processing large volumes of textual data and identifying subtle emotional cues (Yusifov & Sineva, 2022).

However, detecting emotions from text presents complex challenges. Unlike physiological indicators, textual data lacks direct physical cues and relies heavily on linguistic nuances. The inherent ambiguities in language, the constant evolution of slang, and the introduction of new terminologies make textual emotion detection particularly difficult. Words and phrases can carry multiple meanings, and context plays a crucial role in interpreting the intended emotional tone distinctly (Tesfagergish et al., 2022).

Therefore, textual emotion detection is not limited to distinctly identifying primary psychological conditions such as happiness, sadness, and anger. It extends to more refined categorizations, often utilizing scales that encompass a broader spectrum of emotional states (Pashchenko et al., 2022). The fact is that the current textual emotion detection models excel in identifying basic emotions. They utilize advanced algorithms and deep learning techniques to analyze text data and categorize it into these fundamental emotional groups (Jain et al., 2024).

However, while they are effective at recognizing basic emotions, they face significant challenges in detecting more nuanced and complex emotional states (Acheampong et al., 2020). Emotions such as envy, guilt, pride, or admiration are often subtle and context-dependent, making accurate detection difficult for existing models. Translating text to represent a large number of distinct emotions remains a significant challenge, as most contemporary models have not yet achieved an acceptable level of accuracy for this task (Kamath et al., 2022b).

Furthermore, recent psychological advancements view emotions not as discrete entities but as continuous scales, adding another layer of complexity to an already challenging problem. Current models, therefore, is struggling with translating textual data into valence, arousal, and dominance (VAD) scores (Mohammad, 2021). Valence indicates the positivity or negativity of an emotion, arousal reflects the level of excitement or calmness, and dominance measures the degree of control or submissiveness. Accurately scoring text along these dimensions requires a deep understanding of context and nuances that current models often lack. This challenge is further exacerbated by the ambiguity and context-sensitivity of textual cues, leading to potential misinterpretations and inaccuracies in VAD scoring (Park et al., 2019).

Another major hurdle is the need for comprehensive and diverse datasets that accurately represent the wide array of human emotions. Many existing datasets are limited in scope, focusing predominantly on a few basic emotions and neglecting the richness and diversity of emotional experiences. Additionally, emotion detection systems need to be adaptable to different languages and cultural contexts, as emotional expressions can vary significantly across regions and cultures (Kajava et al., 2020).

To overcome these shortcomings, it is crucial to develop more advanced emotion detection models. Future research should aim to enhance the granularity and precision of emotion classification, incorporate a broader range of emotional categories, and improve the accuracy of VAD score translation. This may involve leveraging larger, more diverse datasets, developing more sophisticated algorithms, and integrating multi-modal data sources such as audio and video. By advancing these aspects, textual emotion detection can become more robust and reliable, better reflecting the complexities of human emotions in varied contexts.

This study, therefore, will contribute with the following:

- An examination of the various techniques and methods used in detecting complex emotions
- An overview of different publicly available complex emotion datasets
- An empirical evaluation of the fine-tuned models in detecting complex emotions
- An analysis of the challenges associated with approaches, dataset,ts, models and directions for future research in detecting complex emotions

2. Related Works

Comprehensive literature reviews in the field of textual emotion detection are notably scarce (Nandwani & Verma, 2021). Since emotion detection has evolved from sentiment analysis, it is logical to include systematic reviews on sentiment analysis as part of this survey. Sentiment analysis, which focuses on identifying positive, negative, or neutral sentiments within the text, has laid the foundational groundwork for more nuanced emotion detection. By examining systematic reviews in sentiment analysis, researchers can gain valuable insights into the methodologies, challenges, and advancements that have shaped the current landscape of emotion detection (Moher et al., 2010).

Ligthart et al. (2021) conducted a tertiary analysis of sentiment analysis, offering a thorough overview of various approaches, features, methodologies, and datasets. They also highlighted challenges and unaddressed issues, helping to identify areas needing further research. It presents a tertiary analysis of sentiment analysis, focusing exclusively on secondary studies. Similarly, Dang et al. (2020) reviewed deep learning methods used to address sentiment analysis problems, including sentiment polarity and comparative analyses of various deep learning techniques. It reviews the deep learning methods used to address sentiment analysis challenges, such as polarity detection and comparative analysis. However, the surveys were limited to sentiment analysis and lacked substantial information on emotion detection approaches and datasets.

Alswaidan and Menai (2020) discussed explicit and implicit emotion detection from text, presenting state-of-the-art approaches, their merits and demerits, and various corpora and lexicons available for emotion detection. Their review emphasized the significance of NLP tasks, the performance of different approaches, and highlighted open issues. It detailed state-of-the-art methods for textual emotion detection, including their features, advantages, and disadvantages, as well as the available corpora and lexicons. Similarly, Goyal and Tiwari (2017) reviewed various approaches, classifiers, application domains, and issues related to text emotion detection. Their review covered different methods, classifiers, and application areas, as well as the challenges in the field. However, the studies primarily focused on basic emotions like happiness, sadness, anger, and fear, potentially neglecting more nuanced and complex emotional states such as envy, guilt, pride, or admiration as this study will do.

Acheampong et al. (2020) surveyed the concept and main approaches used in textual emotion detection systems, discussing future methodologies, datasets, and outcomes along with their pros and cons. They provided annotated datasets suitable for emotion detection and highlighted open problems and future research directions. It surveys the concepts, primary methodologies, innovative approaches, utilized datasets, performance metrics, strengths, and weaknesses within the field of textual emotion detection systems. Similarly, Ain et al. (2017) presented a classification of sentiment analysis, surveying emotion theories, methods for polarity classification related to emotion mining, and valuable resources including datasets and lexicons. The survey covers emotion theories, polarity classification methods related to emotion mining, and essential resources such as lexicons and datasets. However, the study has emphasized theoretical advancements without sufficiently addressing the practical application and integration of emotion detection systems in real-world scenarios.

Kumar and Garg (2020) conducted a systematic literature review to explore current research in sentiment analysis focusing on context, identifying research gaps, and suggesting future directions. The review investigates and assesses current efforts in sentiment analysis within contextual frameworks, discussing limitations and proposing future research directions. In the meantime, Oberländer and Klinger (2018) centered on multimodal data fusion techniques, integrating audio, visual, and textual information. It surveyed datasets, compared emotion corpora, and standardized them into a common file format with a unified annotation schema. The study did not emphasize the challenges related to categorical and dimensional emotions. Additionally, the evaluation metrics used to assess model performance were not comprehensive enough to capture the full spectrum of emotional nuances. Standard metrics may not fully reflect the effectiveness of models in real-world applications.

Seyeditabari et al. (2018) examined textual emotion detection techniques, methodologies, and models, highlighting deficiencies in many approaches developed for detecting emotions in text. It argues that numerous techniques, methodologies, and models in this field are inadequate for various reasons. According to Nandwani and Verma (2021), the vast availability of online data on consumer and citizen decisions, ratings, and opinions seems virtually limitless. Consequently, various methods and tools have been developed to process online texts and extract as much information as possible, including the analysis of emotions and sentiments. However, the studies did not adequately address challenge of generalizability across different languages and cultural contexts. Emotional expressions can vary widely across cultures, and models trained on specific datasets may not perform consistently across different linguistic and cultural settings.

Minaee et al. (2021) offered an overview of deep learning-based text classification models, discussing their methodological advancements, commonalities, and limitations. It provided a comprehensive review of deep learning models for text classification, emphasizing their technical contributions and weaknesses. Poria et al. (2017) explored affective computing and multimodal affect analysis frameworks, focusing on the integration of text, visual, and audio data. It examined fusion techniques for multimodal data and discussed their applications in affective computing and multimodal affect analysis. The study did not review the approaches, datasets, or models used in emotion detection as in the proposed study.

Kowsari et al. (2019) presented a comprehensive overview of text classification algorithms, feature extraction methods, dimensionality reduction techniques, and evaluation methodologies, emphasizing both their limitations and real-world applications. It covered a wide range of topics including text classification algorithms, diverse approaches to feature extraction, existing algorithms, dimensionality reduction methods, and various evaluation techniques. Kusal et al. (2022) discussed recent advances in emotion detection research, including emotion models, datasets, detection algorithms, characteristics, limitations, and potential future approaches in both text and speech-based emotions. Despite these surveys, a comprehensive study addressing datasets, technical approaches, application domains, and future directions in textual emotion detection is still lacking.

The surveys presented so far in the field of textual emotion detection often lack a comprehensive examination of datasets, technical methodologies, application domains, and future directions in the context of categorial and dimensional emotions (aka complex emotions). This review identifies these gaps and sets the stage for future research endeavors in this evolving area. By focusing on a thorough survey encompassing diverse datasets, innovative techniques, application domains, and emerging trends, this review aims to provide a comprehensive understanding of the current landscape of textual emotion detection. It seeks to pave the way for further advancements by identifying areas where more robust methodologies and interdisciplinary approaches can enhance the accuracy and applicability of emotion detection systems in real-world contexts.

3. Research Methodology

A Systematic Literature Review (SLR) is a rigorous research method that follows a defined process to systematically discover, analyze, and understand existing information relevant to a specific research issue in a fair and reproducible manner (Kitchenham et al., 2010). This SLR follows the PRISMA guideline, a set of recommendations for reporting and structuring systematic reviews and meta-analyses of data, as published by Moher et al. (2010). The authors' research methodology is organized into three main sections: selection criteria, inclusion/exclusion criteria and selection results.

3.1. Selection Criteria

This study primarily utilized IEEE, Science Direct, Scopus, and Web of Science databases to search for documents related to textual emotion detection. It devised a specialized query to retrieve relevant articles across these databases and then employed a filtering process to refine results aligned with their research objectives. This process involved removing duplicates, applying inclusion and exclusion criteria, filtering based on title and abstract relevance, and conducting full-text screening.

To gather relevant data for their research, this study employed a strategic use of search keywords and queries. It utilized a combination of primary and secondary keywords, enhancing their search with Boolean operators such as AND and OR. This method allowed them to refine their searches across various databases. The specific keywords they used included terms like "emotion detection", "text", "artificial intelligence", "deep learning", "machine learning", "sentiment analysis", "emotion recognition", "dataset", and "fine-tuned model". These keywords were applied in searches focusing on both the abstract and the title of the manuscripts.

By carefully selecting these keywords, this study aimed to compile a comprehensive collection of manuscripts related to the field of textual emotion detection. Their focus was on studies that leveraged artificial intelligence techniques. This targeted approach ensured that the gathered data was highly relevant and specific to their research objectives. Using keywords related to AI and emotion detection, the team could identify studies that utilized advanced computational methods to analyze and interpret human emotions in text.

The strategic application of Boolean operators further refined their search results, enabling the researchers to combine various concepts and exclude irrelevant studies effectively. For instance, using "AND" helped narrow down the results to manuscripts that included multiple key concepts simultaneously, while "OR" expanded the search to include studies that covered at least one of the specified keywords. This meticulous approach to keyword selection and search query formulation was crucial in assembling a robust dataset for their research on textual emotion detection using artificial intelligence.

3.2 Inclusion/Exclusion Criteria

The author developed a rigorous set of inclusion criteria to ensure that only the most pertinent research articles were selected for their study. These criteria required that articles be published between 2015 and the present, ensuring the relevance and currency of the data. Additionally, the inclusion criteria specified that the articles must be either journal papers or book chapters, guaranteeing the quality and credibility of the sources. The research needed to span various domains, including computer science, engineering, psychology, decision sciences, and social sciences, to provide a multidisciplinary perspective on the topic.

Once the relevant articles were identified based on the inclusion criteria, the author applied exclusion criteria to further refine the selection. Articles that were not related to textual emotion detection or artificial intelligence were eliminated to maintain focus on the core subject. Additionally, articles not published in English were excluded to ensure the author could accurately interpret and analyze the findings. This careful screening process helped the author build a robust and relevant dataset for their research.

By establishing these criteria, the author ensured a systematic and unbiased selection process. The inclusion criteria allowed the author to gather a comprehensive and up-to-date collection of high-quality research articles, while the exclusion criteria helped eliminate irrelevant studies. This meticulous approach was essential for conducting a thorough and reliable analysis of textual emotion detection using artificial intelligence, providing a solid foundation for the author's research findings and conclusions.

3.3 Selection Results

The initial search query for textual emotion detection yielded a substantial total of 706 articles. By applying the predefined inclusion and exclusion criteria, this number was reduced to 232 articles. Further refinement based on the relevance of keywords, titles, and abstracts, as well as the removal of duplicate entries, resulted in a more manageable 156 articles. After a thorough full-text screening and a rigorous quality assessment, the selection was finally narrowed down to 65 articles deemed suitable for the final manuscript reading.

The comprehensive literature review conducted on these 65 articles will be summarized into four key areas: emotion theories, approaches to textual emotion detection, datasets, and models. Each area will be meticulously analyzed to extract valuable insights and to ensure a structured and coherent presentation of findings. This categorization will facilitate a detailed exploration of the current state of research in textual emotion detection, highlighting both the advancements and the methodologies employed in this field.

In each of these areas, the review will identify gaps related to the detection of complex emotions from text. By pinpointing these gaps, the study aims to shed light on the challenges and limitations that current approaches face. This critical analysis will not only provide a comprehensive overview of existing research but also pave the way for future studies to address these identified shortcomings, ultimately advancing the field of textual emotion detection using artificial intelligence techniques.

4. Emotion Theories

Emotion models fundamentally delineate how one emotion differs from another, providing a structured framework to understand the complex landscape of human feelings. Generally, these models represent emotions in two primary forms: categorial and dimensional.

Categorial Emotions

Categorial representations break down emotions into finite groups. Tomkins and Robert McCarter firstly developed an influential model of emotions that emphasizes the fundamental role emotions play in human behavior and experience. Tomkins, a pioneering psychologist, proposed that emotions are primary motivational systems that drive human behavior more powerfully than other drives such as hunger or sex. His work, complemented by McCarter, helped form a comprehensive theory of emotions (Tomkins & McCarter, 1964).

Paul Ekman went further to categorize emotions into six basic categories: happiness, sadness, anger, disgust, surprise, and fear based on cross-cultural studies (Ekman & Oster, 1979). Ekman's primary methodology involved developing the Facial Action Coding System (FACS), a tool that categorizes facial expressions into specific action units (Ekman, 1993). The universality of Ekman's findings has been both a strength and a point of contention in the current literature. While his work laid the foundation for emotion recognition and understanding, critics argue that cultural nuances and individual differences might influence the interpretation of facial expressions by just a limited set of universal emotions (Plaza-del-Arco et al., 2022).

Robert Plutchik's model (Plutchik, 1984) builds on some of Ekman's concepts. Plutchik acknowledged the existence of primary emotions and their combinations that lead to complex emotions. However, he expanded the number of primary emotions to eight, organized into opposing pairs: joy versus sadness, trust versus disgust, anger versus fear, and surprise versus anticipation. Plutchik's model emphasizes the interaction and oppositional nature of these primary emotions, providing a nuanced understanding of emotional dynamics.

Robert Plutchik's contribution expands the emotional landscape beyond basic (facial) emotions, introducing the "Wheel of Emotions." Plutchik identified eight primary emotions arranged in a circular diagram, emphasizing the dynamic nature of emotional experiences. His model allows for the combination and intensification of emotions, offering a more nuanced understanding of human feelings (Plutchik, 2001). The Wheel of Emotions has been praised for capturing the complexity of emotional experiences. However, critics argue that the model might still oversimplify the intricate nature of emotions and interactions when the number of emotions identified is still too limited (Alkaabi et al., 2022; Cortis, 2021).

Both Ekman and Plutchik made significant contributions to the understanding of emotions and how they are expressed and experienced. Ekman's work is often associated with facial expressions linked to basic emotions, while Plutchik's model provides a more nuanced view of the complex interplay between emotions. Compared with Ekman, Plutchick offered a more systematic way of organizing emotions, especially through its one-dimensional pairs (Tesfagergish et al., 2022).

In contrast, Ortony, Clore, and Collins (OCC) challenged the concept of “basic emotions” proposed by Ekman and Plutchik (Ortony et al., 2022). They contended that emotions stem from individuals' interpretations of events and can vary in intensity. The OCC model expands the discrete representation to include 22 emotions, incorporating categories such as relief, envy, reproach, self-reproach, appreciation, shame, pity, disappointment, admiration, hope, fears-confirmed, grief, gratification, gloating, liking, and disliking. This comprehensive approach offers a broader spectrum of emotional experiences, highlighting the complexity and variability of human emotions.

The Hourglass of Emotions revisited model by Susanto et al. (2020) enhances the original Hourglass of Emotions model proposed by Cambria et al. (2012). This revision addresses several shortcomings identified in the original model, including inappropriate color associations, the inclusion of neutral and ambiguous emotions, and the absence of polar emotions such as calmness and eagerness. Moreover, the revisited model corrects associations of antithetic emotions and incorporates self-conscious or moral emotions, which were lacking in the original.

The trend in categorial representations of emotions has evolved from Ekman’s model, which emphasized emotions as primary motivational systems, to more nuanced frameworks like those proposed in the Hourglass of Emotions. This progression reflects a growing recognition of the complexity and variability of human emotions beyond simple categorizations.

Dimensional Emotions

Dimensional models, in contrast, position emotions within a uni- or multi-dimensional space, illustrating the relationships between emotions and their relative occurrences. The spatial arrangement in these models helps depict the intensity and interaction of different emotions.

Russell's Valence, Arousal, and Dominance (VAD) model is a dimensional approach to understanding and categorizing emotions. Proposed by James A. Russell in the 1980s, this model posits that emotional experiences can be described along three continuous dimensions: valence, arousal, and dominance (Russell, 1980; Russell & Mehrabian, 1977).

Valence refers to the positivity or negativity of an emotion. It measures how pleasant or unpleasant an emotional experience is, ranging from highly negative (e.g., sadness, anger) to highly positive (e.g., happiness, joy). This dimension captures the intrinsic attractiveness or averseness of an emotion. Arousal indicates the intensity of the emotional experience, from low to high. Low arousal emotions include states like calmness or boredom, while high arousal emotions encompass feelings like excitement or anxiety. This dimension reflects the level of activation or energy associated with an emotion. Dominance describes the degree of control or influence an individual feels over an emotion, ranging from feelings of submission or passivity to feelings of control or empowerment. For instance, fear might be associated with low dominance (feeling overwhelmed or controlled by the emotion), while anger might be associated with high dominance (feeling in control or powerful).

The VAD model offers a comprehensive and flexible framework for understanding emotions, accommodating the vast array of human emotional experiences without limiting them to discrete categories. It allows for the mapping of emotions onto a three-dimensional space, where different emotions can be plotted based on their valence, arousal, and dominance levels. This dimensional approach helps capture the complexity of emotions and their interrelationships, providing a more nuanced understanding of how emotions are experienced and expressed. Its ability to account for the subtle gradations and overlaps between different emotions makes it a valuable tool for researchers and practitioners aiming to better understand and interact with human emotions (Posner et al., 2005).

Several dimensional emotion models have later emerged, enhancing and expanding the original framework to capture more nuanced aspects of emotional experiences. These models build on the foundational VAD dimensions to address various limitations and incorporate new insights from ongoing research in psychology, neuroscience, and affective computing.

The Circumplex Model of Affect is one of the most notable extensions of Russell's VAD model (Russell, 1980). In this model, emotions are plotted in a circular arrangement around two primary dimensions: valence and arousal. This circumplex structure allows for the visualization of emotional states in a way that highlights their relationships and transitions. For instance, emotions like excitement (high arousal, positive valence) and calmness (low arousal, positive valence) can be easily compared and contrasted. The Circumplex Model underscores the continuous and dynamic nature of emotional experiences.

Proposed by Bakker et al. (2014), the PAD model stands for Pleasure (similar to valence), Arousal, and Dominance. This model has been particularly influential in environmental psychology, where it is used to assess people's emotional responses to different environments. The PAD model's inclusion of the dominance dimension provides a more comprehensive understanding of how people perceive control and influence in various situations (Bakker et al., 2014).

The Geneva Emotion Wheel (GEW) is another dimensional emotion model (Scherer et al., 2013). It is designed to facilitate the assessment and communication of emotions. It arranges emotions in a circular format, similar to the Circumplex Model, but also includes additional dimensions and emotional categories. The GEW allows users to pinpoint their emotional state on a wheel that combines valence and arousal with other nuanced emotional distinctions. This model is particularly useful for self-reporting and research involving emotional granularity (Scherer et al., 2013).

The Affective Slider is a more recent tool developed for capturing emotional responses in a user-friendly and visually intuitive way (Betella & Verschure, 2016). It provides sliders for valence, arousal, and dominance, allowing users to indicate their emotional state along these dimensions easily. This tool is beneficial in digital and interactive environments where quick and efficient emotional assessment is required. The Affective Slider builds on the VAD model by incorporating modern design principles to enhance usability and accuracy in emotion measurement.

Table 1: Emotion models

Model	**Type of model**	**Emotional states**	**No. of states**
Tomkins model (Tomkins & McCarter, 1964)	Categorial	Disgust, surprise-Startle, anger-rage, anxiety, fear-terror, contempt, joy, shame, interest, excitement	9
Ekman model (Ekman & Oster, 1979)	Categorial	Anger, disgust, fear, joy, sadness, surprise	6
Plutchik Wheel of Emotions (Plutchik, 1980)	Categorial	Joy, sadness, trust, disgust, fear, anger, surprise, anticipation	8
Valence, Arousal and Dominance (Russell, 1980)	Dimensional		–
Shaver model (Shaver et al., 1987)	Categorial	Sadness, joy, anger, fear, love, surprise	6
OCC (Ortony et al., 2022)	Categorial	Joy, distress, hope, fear, satisfaction, fears-confirmed, relief, disappointment, pride, shame, admiration, reproach, gratification, remorse, gratitude, anger, love, hate, liking, disliking	22
The Hourglass of Emotions (Cambria et al., 2012)	Categorial	Pleasantness, ecstasy, joy, serenity, contentment, anticipation, vigilance, interest, anticipation, optimism, anger, rage, anger, annoyance, irritation, disgust, loathing, disgust, boredom, dislike, sadness, grief, sadness, pensiveness, dissatisfaction, surprise, amazement, surprise, distraction, shock	24
Pleasure, Arousal, and Dominance (Bakker et al., 2014)	Dimensional		-
Geneva Emotion Wheel (Scherer et al., 2013)	Dimensional		-
The Hourglass Model revisited (Susanto et al., 2020)	Categorial	Admiration, adoration, aesthetic appreciation, amusement, anxiety, approval, awe, awkwardness, boredom, calmness, confusion, craving, disappointment, disgust, distress, doubt, ecstasy, embarrassment, empathic pain, enthusiasm, entrancement, envy, excitement, fear, gratitude, guilt, horror, interest, joy, longing, nostalgia, pain, pride, realization, relief, remorse, sadness, satisfaction, shame, surprise, sympathy, triumph, warmth, wonder, love, hate, trust, confusion	48

Shaver's model of emotions focuses on the hierarchical organization of emotions and their relationships with each other (Shaver et al., 1987). Izard developed the Differential Emotions

Theory (DET), which posits that emotions are discrete, fundamental, and biologically rooted in humans (Izard, 1992). Lövheim's model maps emotions onto a cube defined by three neurochemical axes: dopamine: Associated with pleasure and reward, influencing emotions like joy and interest, noradrenaline: linked to arousal and alertness, affecting emotions such as fear and anger, serotonin: related to mood regulation, impacting emotions like sadness and contentment (Lövheim, 2012).

Emotions according to the dimensional models are described along continuous dimensions, capturing the complexity and variability of emotional experiences.

As summarized in Table 1, categorical and dimensional models of emotion offer essential frameworks for comprehending and analyzing human emotions. Categorical models categorize emotions into distinct groups, such as happiness, sadness, anger, and fear. These models are instrumental in simplifying the complex spectrum of human emotions into manageable and recognizable categories, making it easier for researchers and practitioners to study emotional responses and their implications in various contexts.

Dimensional models, on the other hand, provide a spatial framework for emotions, illustrating the relationships and intensity of different emotional states. These models, such as Russell's Circumplex Model of Affect, plot emotions on axes like arousal and valence, offering a more nuanced understanding of how emotions can vary and intersect. The spatial representation allows for the visualization of subtle differences and overlaps between emotions, capturing the complexity of human emotional experiences more effectively than categorical models.

Advances in these models, like the Hourglass of Emotions, have further enhanced our ability to categorize and understand a broader range of emotions. These advanced dimensional models expand the number of emotional states that humans can experience and illustrate how emotions can overlap and influence one another. By providing a more comprehensive and detailed map of emotional experiences, these models enable a deeper exploration of human affective states, improving our ability to analyze and interpret the rich tapestry of human emotions in both research and practical applications.

This study, therefore, concludes that:

Finding 1: *There are two recent developments in emotion theories. The first is the expansion of the number of categorical emotions. The second is the scaling of emotional experiences into continuous dimensions of valence, arousal, and dominance.*

5. Emotion Detection

Emotion detection approaches are diverse methodologies employed to identify and analyze human emotions from various data sources. These approaches leverage advancements in artificial intelligence, machine learning, and natural language processing to interpret emotional cues from text, speech, facial expressions, and physiological signals.

5.1. Approaches

The primary approaches for identifying text-based emotions include keyword-based, rule-based, machine learning-based, and deep learning-based methods.

- Keyword-based Approach: This method centers on identifying keyword occurrences within a text and comparing them against annotations in the dataset. Initially, an emotion keyword list is generated from lexical resources such as WordNet or WordNet-Affect. The dataset undergoes preprocessing, followed by matching keywords from a predefined list with emotion words found in the text. The intensity of the emotion keyword is then evaluated, negation cues are considered, and finally, the emotion tag is assigned. Studies such as Perikos and Hatzilygeroudis (2013), and Shivhare et al. (2015) are based on this approach.
- Rule-based Approach: This method utilizes linguistic rules to detect emotions in text. Initially, the dataset undergoes text preparation, which includes data cleaning, tokenization, and part-of-speech tagging (POS tagging). Rules for extracting emotions are then developed using statistical and linguistic principles, with each word's probabilistic association recorded. The most effective rules are chosen for the dataset to identify emotion labels. Studies by Udochukwu and He (2015), and Liu and Cocea (2017) are examples of this approach.
- Machine Learning-based Approach: This approach enables systems to learn and enhance their performance based on experience. Machine learning algorithms classify text into different emotion categories, often employing supervised learning techniques. Initial steps include text preprocessing, such as tokenization, part-of-speech tagging (POS tagging), and lemmatization. Relevant features are extracted, prioritizing those with the greatest information value. The algorithm is subsequently trained using these features and emotion labels, enabling the system to predict emotions accurately from new data. Notable studies include De Bruyne et al. (2018), Suhasini and Srinivasu (2020), Singh et al. (2021), and Allouch et al. (2018).
- Deep Learning-based Approach: This method employs neural network layers to learn from unlabeled or unstructured data, a subset of AI-based machine learning. Initially, the dataset undergoes preprocessing steps such as tokenization, stop words removal, and lemmatization. Embeddings are then created to represent tokens as numerical vectors. These vectors are input into deep neural network layers associated with emotion labels, where patterns are identified and utilized to predict labels accurately. Key studies in this area include Baziotis et al. (2018), Can et al. (2018), Basile et al. (2019), Shrivastava et al. (2019), (Xiao et al., 2019), Rathnayaka et al. (2019), Bian et al. (2019), Mohammad (2021), Xia et al. (2018), Diogo et al. (2021), and Polignano et al. (2019)

The review has shown that there are several methods for detecting emotions in text, each representing a distinct approach in textual emotion detection. These methods encompass different tasks and processes, emphasizing the importance of aligning specific tasks with each method. Commonly utilized techniques include keyword-based, rule-based, machine learning-based, and deep learning-based approaches. Each technique follows a distinct process to identify emotions within textual data. Among these, deep learning and machine learning techniques are predominantly favored by researchers due to their advanced capabilities in accurately classifying and interpreting emotional content.

- Machine learning: Machine learning approaches address emotion detection by classifying texts into distinct emotion categories using various algorithms. These techniques can be categorized as supervised or unsupervised, with supervised methods being more common in textual emotion detection. Classical machine learning algorithms remain widely utilized for emotion detection tasks, often performing comparably or even better than deep learning approaches, particularly on smaller datasets. Popular supervised machine learning algorithms include Naive Bayes (NB), Support Vector Machines (SVM), Random Forest (RF), Neural Networks (NN), k-Nearest Neighbor (k-NN), Logistic Regression (LR), and Decision Trees (DT). They fall into three main categories: supervised, unsupervised, and reinforcement learning. Supervised classifiers utilize inputs with known output labels, while unsupervised classifiers handle unlabeled data. Reinforcement learning classifiers learn through trial and error, using a limited set of annotated data for labeling and a larger amount of unlabeled data (Ariely et al., 2023).
- Deep Learning: Deep learning models utilize intricate structures composed of multiple neural network layers to extract complex information from input data. In textual emotion detection, common models include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and transformer-based pre-trained models. CNNs, typically used in image processing, have recently been applied to Natural Language Processing (NLP) to identify data patterns through convolutional filters. RNNs are popular in NLP due to their ability to store and process sequential information. Attention networks focus on significant data parts, and transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) and its variants (RoBERTa, DistilBERT, SpanBERT), Generative Pre-Training (GPT), and XLNet utilize attention mechanisms with encoders and decoders. These models improve relational context extraction, eliminate long-term sequence dependency issues, and allow parallel input sequence processing. These deep learning algorithms significantly enhance the performance of textual emotion detection tasks. Deep learning classifiers fall under the category of unsupervised learning, but they can also be supervised or semi-supervised (Ain et al., 2017).

Between machine learning and deep learning, deep learning is superior for detecting emotions as it can more effectively capture the abstract meanings of texts, particularly with the advent of Transformer models (Nandwani & Verma, 2021). Transformers utilize attention mechanisms to understand the context and relationships within the text, allowing them to grasp nuanced emotional cues and subtleties that traditional machine learning models often miss. This advanced capability enables deep learning models to deliver more accurate and sophisticated emotion classification, enhancing applications in areas such as sentiment analysis, customer feedback interpretation, and human-computer interaction.

5.2. Transformers-based approaches

The advent of large pre-trained transformer models such as BERT, GPT, and T5 has transformed the field of NLP and emotion detection. These models can be fine-tuned on specific tasks using relatively small datasets, leveraging the extensive knowledge they have acquired from vast amounts of data (Zanwar et al., 2022).

- ### BERT

The Bidirectional Encoder Representations from Transformers (BERT), introduced by Devlin et al. (2018), revolutionized natural language processing (NLP) by harnessing transformer architecture to enhance various NLP tasks such as sentiment analysis (SA), question answering (QA), and text summarization (TS). BERT's architecture and training methodology have established new benchmarks in the field, driving significant advancements in understanding and generating human language.

BERT utilizes the encoder component of the transformer model as its sub-structure. Its training process comprises two distinct phases: pre-training and fine-tuning. During pre-training, BERT employs Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) mechanisms. MLM involves masking some words in a sentence and training the model to predict these masked words based on contextual cues from the surrounding text. This enables BERT to grasp bi-directional contexts, meaning it can infer a word's meaning based on both its preceding and following words (Zanwar et al., 2022).

During pre-training, BERT was trained on an extensive dataset comprising 16GB of text sourced from the BooksCorpus dataset and English Wikipedia. Following pre-training, BERT undergoes supervised fine-tuning on specific NLP tasks, e.g. emotion detection (Tesfagergish et al., 2022). During this phase, the output layer of the BERT model is substituted with task-specific layers, and the model is further trained on labeled datasets relevant to the task at hand. This two-stage training process enables BERT to efficiently apply its deep understanding of language across a wide array of applications including detecting emontions in text.

Despite its numerous advantages, BERT has some limitations. One significant drawback is its primary support for monolingual classification tasks, which restricts its application in multilingual environments. This constraint can be a major limitation for global applications that require processing and understanding multiple languages. Additionally, BERT has a fixed input size, typically up to 512 tokens, which can be restrictive when dealing with longer texts. This fixed input size can lead to context fragmentation, making it challenging to process and understand longer documents or conversations fully (Wagner, 2021).

BERT also struggles with pragmatic inference, where understanding the implied meaning or intent behind a text is necessary. This shortcoming becomes evident in tasks that require deep semantic comprehension beyond the literal interpretation of words. For example, detecting sarcasm, irony, or nuanced emotional expressions can be problematic for BERT. This limitation affects its performance in various natural language understanding tasks that depend on grasping the underlying context and intent, rather than just the surface meaning (Hoyos, 2021).

- ## XLNet

The development of advanced language models has revolutionized natural language processing (NLP), with XLNet emerging as a significant breakthrough. Introduced by Yang et al. (2019), XLNet is an auto-regressive language model that integrates the Permutation Language Model (PLM) and the Transformer-XL architecture. As a variant of the BERT model, XLNet introduces a novel training objective and advanced mechanisms to overcome the limitations of previous models, achieving state-of-the-art (SOTA) results in various NLP tasks.

The fundamental difference between BERT and XLNet lies in their training objectives. BERT employs a Masked Language Model (MLM) approach, where it masks certain words in a sentence and then predicts these masked words using a bi-directional context. This approach enables BERT to understand the context from both directions, enhancing its language understanding capabilities. However, BERT's method also introduces a fixed-length limitation, constraining its ability to handle long sequences effectively (Yang et al., 2019).

In contrast, XLNet uses a permutation objective. This approach allows the model to learn bi-directional context by considering all possible permutations of words in a sentence. By training on these permutations, XLNet can capture a more comprehensive understanding of the contextual relationships between words, enabling it to learn from a diverse set of word orders. This process enhances the model's ability to generalize and understand complex contextual relationships, leading to improved performance in various NLP tasks (Nandwani & Verma, 2021).

Furthermore, XLNet incorporates the Transformer-XL model, which addresses the fixed-length limitation of BERT. Transformer-XL introduces recurrence mechanisms and positional encoding, allowing the model to capture long-term dependencies across sequences. This capability is crucial for processing lengthy texts and maintaining context over extended passages (Dai et al., 2019).

Another significant architectural innovation in XLNet is the implementation of two-stream self-attention. This mechanism provides target-aware representations by separately computing content and query streams. The content stream processes the context, while the query stream focuses on the specific target position. This dual-stream approach allows XLNet to better capture the relationships between different parts of the input sequence, further enhancing its contextual understanding (Alvarez-Gonzalez et al., 2021).

Despite its numerous advantages, XLNet's sophisticated architecture and extensive training requirements introduce significant computational complexity. The permutation-based training objective and two-stream self-attention mechanisms demand substantial computational resources, making the model computationally intensive. This complexity can pose challenges for deployment, especially in resource-constrained environments (Kusal et al., 2022).

- ROBERTA

The Robustly Optimized BERT Pre-training Approach (RoBERTa), introduced by Liu (2019), marks a substantial advancement in natural language processing (NLP) through its optimization of the original BERT model. RoBERTa's modifications and refinements highlight the critical importance of hyper-parameters, pre-training datasets, and text encoding techniques in achieving superior NLP performance.

RoBERTa was designed to address and enhance several methodological aspects of BERT, focusing on optimizing hyper-parameters and investigating their influence on model performance. Key areas of improvement included the use of larger pre-training datasets, the transition from static to dynamic Masked Language Modeling (MLM), adjustments in batch sizes, and the elimination of the Next Sentence Prediction (NSP) task. These changes aimed to refine the pre-training process and boost the model's overall efficacy (Liu, 2019).

During the pre-training phase, RoBERTa utilized an extensive dataset comprising 160GB of uncompressed text data sourced from five English language datasets. These datasets encompassed the original BERT data sources, CC-News data, Stories dataset by Trinh and Le (2018), and Open Web data, with respective sizes of 16GB, 75GB, 31GB, and 38GB of sentences. This extensive data corpus allowed RoBERTa to learn from a diverse range of textual information, enhancing its ability to generalize across various NLP tasks.

One of the critical enhancements in RoBERTa was the shift from static to dynamic masking during MLM. Unlike static masking, where the same mask is applied to each training instance for every epoch, dynamic masking generates a new mask for each instance in each epoch. This approach prevents overfitting to specific masked positions and allows the model to learn more robustly. Although dynamic masking offered only slight improvements in multitask performance, it contributed to a more generalized model (Acheampong et al., 2020).

RoBERTa's training involved experimenting with different batch sizes, ultimately determining that a batch size of 2,000 was optimal. Larger batch sizes helped stabilize the training process and improve the model's performance. Additionally, the model utilized Byte-Pair Encoding (BPE) over traditional word or character-level encoding. BPE's ability to capture subword information proved more effective, leading to better representation of the input text and improved performance in downstream tasks (Bostrom & Durrett, 2020).

RoBERTa demonstrated substantial improvements over its predecessor BERT and other contemporary models like XLNet. The extensive pre-training data and methodological refinements enabled RoBERTa to achieve state-of-the-art performance across various NLP benchmarks. As shown in comparative studies, e.g. Adoma et al. (2020), RoBERTa consistently outperformed BERT in tasks such as sentiment analysis, question answering, and text summarization.

The primary advantages of RoBERTa stem from its enhanced pre-training approach and the use of larger datasets, which result in better performance across a wide range of tasks. Its ability to outperform both BERT and XLNet underscores its effectiveness in capturing and leveraging contextual information.

However, similar to BERT and XLNet, RoBERTa primarily supports monolingual tasks, which can restrict its effectiveness in multilingual environments. Emotion detection often requires understanding subtle linguistic nuances across different languages, a capability that RoBERTa's training may not fully capture without extensive adaptation or multilingual pre-training (Suhasini & Srinivasu, 2020).

Additionally, RoBERTa's ability to infer pragmatic and nuanced meanings, such as sarcasm or implicit emotions, may be limited. Emotion detection often involves understanding not just explicit emotional expressions but also subtle cues and contextual clues that indicate underlying emotions. RoBERTa's training may not inherently capture these deeper semantic layers required for accurate emotion classification across diverse contexts (Zanwar et al., 2022).

• GPT

Unlabeled data, abundant on the internet, holds immense potential for various ML tasks. The advent of Web 2.0 has introduced an enormous volume of such data that presents significant potential if effectively utilized (Puri & Catanzaro, 2019). Properly structured and labeled, this data can greatly enhance the accuracy and efficiency of ML models. However, while this unstructured data, when labeled and structured, can lead to remarkable outcomes, the manual efforts required for labeling are both labor-intensive and time-consuming (Wang et al., 2018).

The field of machine learning (ML) has consistently demonstrated that the use of structurally labeled data tends to yield superior results compared to using unlabeled data. This principle has positioned supervised learning at the forefront of many successful ML applications. Supervised learning, which relies on labeled data, often outperforms unsupervised learning because the labels provide clear guidance for the model (Tesfagergish et al., 2022). This creates a bottleneck in fully exploiting the wealth of available data.

To address these challenges, the Generative Pre-trained Transformer (GPT) models leverage semi-supervised learning approaches, combining the strengths of both supervised and unsupervised learning (Radford et al., 2019). To do so, the models adopt a semi-supervised learning approach. This methodology combines the benefits of both supervised and unsupervised learning. Initially, GPT models are pre-trained on large unlabeled datasets, allowing them to learn general language patterns. They are then fine-tuned on smaller, labeled datasets specific to particular tasks. This two-step process enables GPT models to leverage the vast amount of available unlabeled data while still benefiting from the accuracy of supervised learning (Peres et al., 2023).

The GPT models are also built using Transformer decoders. The architecture of the original GPT includes 12 transformer layers and 12 attention heads, forming a robust transformer decoder capable of processing large volumes of text data. The primary task of the GPT model is to predict the next token in a sequence. This is achieved by passing input texts through the model's multi-head attention layers and feed-forward layers, ultimately producing a probability distribution of the next possible token via a softmax layer (Floridi & Chiriatti, 2020).

The GPT-2 model advanced the concept of semi-supervised learning by significantly increasing the architecture and data size. It included normalization layers and was designed to predict the next sentence in a sequence, demonstrating that language models could learn tasks without direct supervision (Radford et al., 2019). GPT-2 was available in four different sizes, with the largest model containing 1.5 billion parameters and trained on 40 gigabytes of text data (Jain et al., 2024).

The GPT-3 model, introduced by Brown et al. (2020), further scaled up the architecture to 175 billion trainable parameters. While retaining the fundamental structure of GPT-2, GPT-3 incorporated alternating dense and locally banded sparse attention patterns. This extensive scaling enabled GPT-3 to excel in a wide range of NLP tasks, such as question answering, text classification, and semantic similarity assessments, without the need for fine-tuning.

Table 2: Transformer models

Model	Advantages	Disadvantages
BERT	- Strong performance on a variety of NLP tasks - Bidirectional context understanding - Widely adopted and supported with pre-trained models - Good at capturing word dependencies	- Computationally expensive and resource-intensive - Large model size - Slower inference time - Requires substantial fine-tuning
DistilBERT	- More efficient and faster than BERT - Smaller model size and reduced memory usage - Retains most of BERT's performance with fewer parameters	- Slightly lower performance compared to full-sized BERT models - May miss some nuances captured by larger models
XLNet	- Superior performance on many NLP benchmarks - Permutation-based training captures bidirectional context without masking - Handles longer context better due to autoregressive pre-training	- Even more computationally intensive than BERT - Complex architecture - Slower training times
RoBERTa	- Improved training approach over BERT, leading to better performance - Robust and versatile for various NLP tasks - No Next Sentence Prediction (NSP), simplifying pre-training	- Very large model size - High computational costs - Requires a lot of data for training
GPT-4	- Excels at generating coherent and contextually relevant text - Versatile across a wide range of conversational and creative tasks - Fine-tuning allows for customization in various applications - Strong few-shot learning capabilities	- Prone to generating incorrect or nonsensical information - May produce biased or inappropriate responses - Lack of factual grounding (may not provide accurate or up-to-date information) - Resource-intensive, especially for large deployments

As shown in Table 2, the advancement of GPT-4 marks a significant leap in artificial intelligence, showcasing remarkable improvements in natural language processing capabilities. Compared to its predecessors, GPT-4 exhibits a deeper understanding of context, enabling it to generate more coherent and contextually appropriate responses (Achiam et al., 2023). This iteration incorporates extensive training on diverse datasets, enhancing its ability to handle a broader range of topics with greater accuracy.

GPT-4 proves highly suitable for sentiment analysis due to its advanced natural language processing capabilities and nuanced understanding of context. Leveraging an extensive training dataset, GPT-4 can accurately discern the subtleties in text that convey emotions, attitudes, and opinions. Its ability to process and interpret complex language patterns allows it to identify and categorize sentiments with high precision, whether the text is overtly emotional or subtly suggestive (Katz et al., 2024).

In conclusion, the use of structurally labeled data continues to yield the best results in machine learning, but the vast amount of unlabeled data available today presents an opportunity that cannot be ignored. The GPT models, through their semi-supervised learning approach, effectively bridge the gap between these two types of data, leveraging the strengths of both supervised and unsupervised learning. As the models have evolved from GPT to GPT-4 now, they have demonstrated significant improvements in performance and versatility, albeit with increased resource demands. These advancements highlight the ongoing potential for semi-supervised learning approaches in the future of machine learning and natural language processing.

This study, therefore, concludes that:

***Finding 2**: GPT, especially GPT-4 represents a remarkable advancement in the field of NLP, leveraging innovative training objectives and architectural features to overcome the limitations of previous models like BERT. It should be employed in the field of emotion detection, especially when the number of emotions as conceptualized keeps increasing and emotion detection is moving from a multi-level classification task to a zero-shot one.*

6. Datasets

In textual emotion detection, researchers can either create their own datasets or utilize publicly available ones. This section reviews publicly available and useful datasets that feature reliable labeling or annotation processes. These datasets are widely employed by researchers in textual emotion detection and are accordingly described. They are grouped into two groups of basic emotion and complex emotion ones.

6.1. Basic emotion datasets

As one of the most popular emotion datasets, the EmotionLines dataset represents a significant effort in the field of natural language processing and emotion recognition, providing researchers with a rich resource for studying emotions as expressed in dialogue. This dataset comprises

29,245 labeled utterances extracted from 2,000 dialogues, offering a diverse collection of emotional expressions for analysis and modeling (Chen et al., 2018).

EmotionLines categorizes each utterance into one of eight emotion labels: anger, disgust, fear, happiness, sadness, surprise, neutral, and non-neutral. These labels encompass the spectrum of emotions as defined by Ekman's six basic emotions (anger, disgust, fear, happiness, sadness, surprise) along with a neutral category and a non-neutral category (Ekman, 1993).

One of the notable challenges of the EmotionLines dataset is the imbalance in class distributions among the emotional categories. This imbalance can affect the performance of machine learning models trained on the dataset, as they may become biased towards predicting the majority classes (such as neutral or happiness) more accurately than the minority classes (such as anger or disgust) (Poria et al., 2018).

Similarly, the EmotionPush dataset represents a pioneering effort in the realm of natural language processing and emotion analysis, focusing specifically on private dialogues in social spoken-language interactions. Developed as a repository of instant message logs and corresponding read event logs from real conversations on Facebook Messenger, EmotionPush comprises a total of 162,031 message logs. Key to its creation is the emphasis on ensuring data privacy and utility through innovative masking techniques and novel task proposals (Huang & Ku, 2018).

Privacy is paramount in datasets derived from private conversations, and EmotionPush addresses this concern by meticulously masking all named entities. Each entity is anonymized using a code composed of its type and a unique identifier, thus safeguarding the identities of individuals while maintaining the integrity of the conversational data for analysis. Furthermore, to balance data utility with privacy, the dataset is released partially in its original textual form and partially in the form of word embeddings, ensuring that researchers can explore both semantic and syntactic aspects of the conversations (Huang & Ku, 2018).

The availability of EmotionPush has significant implications for advancing research in natural language processing, particularly in emotion-aware computing and dialogue systems. Researchers can leverage this dataset to train and evaluate machine learning models that understand and respond to emotional cues in real-time conversations. By analyzing patterns in emotional expressions and response dynamics, advancements can be made in sentiment analysis, affective computing, and the development of empathetic AI-driven interfaces (Luo & Wang, 2019).

Researchers are also using the EmoryNLP dataset. The dataset stands as a comprehensive resource for studying emotional expressions within narrative contexts, offering a rich tapestry of annotated utterances across various scenes and episodes. Developed to capture the nuanced spectrum of human emotions as portrayed in fictional narratives, EmoryNLP comprises 97 episodes, 897 scenes, and a total of 12,606 annotated utterances (Zahiri & Choi, 2018). Each utterance within this dataset is meticulously labeled with one of seven emotions, drawn from the primary emotions: sad, mad, scared, powerful, peaceful, and joyful, alongside a default category of neutral.

The structure of EmoryNLP facilitates a detailed exploration of emotional dynamics within narrative discourse. Episodes and scenes provide contextual frameworks within which utterances are analyzed, capturing the interplay between characters, events, and emotional states throughout a narrative arc. The annotation process ensures that each utterance is classified according to its predominant emotional content, enabling researchers to delve into the distribution and portrayal of emotions across different narrative contexts (Chatterjee et al., 2019).

Willcox's feeling wheel, a conceptual framework for categorizing emotions, serves as the basis for EmoryNLP's emotion annotation (Willcox, 1982). This approach aligns with established psychological theories of emotion, providing a structured yet nuanced representation of emotional states ranging from core emotions like sadness, anger, and fear to more nuanced states like powerfulness, peacefulness, and joyfulness. By incorporating these categories, EmoryNLP enriches the dataset with a diverse palette of emotional expressions that reflect the complexity of human affective experiences (Zhong et al., 2019).

SemEval, short for Semantic Evaluation, is a series of international workshops on semantic evaluation of natural language processing (NLP) systems. (Baziotis et al., 2018). It is organized under the auspices of the Association for Computational Linguistics (ACL). The main objective of SemEval is to provide a standardized platform for evaluating and comparing the performance of various NLP systems on a variety of semantic tasks, including datasets for sentiment classification. (De Bruyne et al., 2018).

For example, the SemEval-2019 Task 3 dataset represents a significant contribution to the field of natural language processing, specifically focusing on emotion recognition and classification within textual data. Developed by Chatterjee et al. (2019), this dataset comprises a total of 30,000 texts, consisting of 15,000 emotion-labeled texts and an additional 15,000 unlabeled texts. The labeled texts are categorized into three primary emotions: happy, sad, and angry, providing a structured foundation for studying emotional expressions in various linguistic contexts.

In addition to the labeled texts, the dataset includes a substantial set of 15,000 unlabeled texts. These texts serve a dual purpose: they provide a pool for researchers to explore semi-supervised learning approaches, where algorithms can leverage both labeled and unlabeled data to improve classification accuracy and robustness. This aspect of the dataset encourages the development of innovative methodologies that can harness large amounts of unannotated data to enhance the performance of emotion recognition systems. (Cortis, 2021).

While the SemEval-2019 Task 3 dataset provides a valuable resource for studying basic emotions like happiness, sadness, and anger, challenges such as ambiguity in emotional expressions and cultural variations in emotion perception remain pertinent. Moreover, future research directions may involve expanding the dataset to include additional emotional categories, exploring multimodal approaches that incorporate visual and auditory cues alongside textual data, and adapting models to handle nuanced emotional expressions beyond the core emotions defined in the current dataset. (Basile et al., 2019).

6.2. Complex emotion datasets

While there are many datasets of basic emotions available both in literature and in practice, there are not that many for complex emotions.

Only recently, the GoEmotions dataset emerged as a comprehensive resource for studying the spectrum of complex emotions expressed in online conversations. Curated from Reddit comments, this dataset comprises 58,000 meticulously labeled instances, each annotated across 27 distinct emotion categories along with a Neutral label. Developed to capture the diverse emotional nuances inherent in digital communication, GoEmotions offers a nuanced exploration of how individuals express and perceive emotions in an online context. (Demszky et al., 2020).

GoEmotions distinguishes itself through its expansive range of emotion categories, totaling 27 labels that encompass a broad spectrum of emotional states. These categories include admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, and surprise. Such granularity allows researchers to delve deeply into specific emotional nuances and variations that might otherwise be overlooked in broader sentiment analyses (Kamath et al., 2022b).

The availability of GoEmotions has significantly bolstered research in natural language processing (NLP) and sentiment analysis. Researchers and practitioners can utilize this dataset to train and evaluate machine learning models capable of accurately detecting and interpreting a wider array of emotions in text. Applications span sentiment analysis tools, emotion-aware chatbots, social media monitoring systems, and beyond, where understanding emotional nuances is crucial for enhancing user interaction and engagement.

Similarly, in the domain of dimensional emotions, only a few datasets available recently. Notably, developed by Buechel and Hahn (2022), EmoBank comprises 10,000 sentences carefully curated to encompass a diverse range of genres, ensuring broad applicability and relevance across different linguistic contexts. It is a comprehensive text corpus meticulously annotated with emotion using the psychological Valence-Arousal-Dominance (VAD) scheme.

One of the distinguishing features of EmoBank lies in its dual annotation approach, capturing not only the emotions expressed by writers but also the emotions perceived by readers. This dual perspective enriches the dataset by providing insights into how emotional content is conveyed and interpreted in textual communication. Such annotations offer a comprehensive view of emotional dynamics within language, facilitating deeper analyses into the alignment or divergence between intended and perceived emotional states. (Park et al., 2019).

Table 3: Datasets

Dataset	Data size	Sentiments/emotions	Range
Emotion Lines	29245 labeled utterances from 2000 dialogues	anger, disgust, fear, happiness, sadness, surprise, neutral, and non-neutral	8
Emotion Push	91,000 records	joy, sadness, anger, fear, disgust, and surprise	6

Emony NLP	60,000 records	anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.	8
SemEval Tasks	9613 reviews in SST-2	Positive and negative	2
	SemEval- 2014 (Task 4): 5936 reviews for training and testing	Positive, negative, and neutral	3
	SemEval- 2018 (Affect in dataset task): 1758 reviews for testing	Anger, Joy, sad and fear	4
GoEmotions	58,000 records	Admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise, neutral	28
EmoBank	10,548 records	Valence, Arousal Dominance model (VAD)	-

In addition to expressive and perceptive annotations, a subset of the EmoBank corpus has been annotated according to Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. This annotation schema, established by Paul Ekman, categorizes emotions based on universal facial expressions and physiological responses, providing a standardized framework for understanding emotional states across cultures and contexts. (Lee et al., 2022).

Table 3 shows that the most popular existing datasets in natural language processing categorize emotions into a limited range of 6 to 10 categories, which simplifies the complexity of human emotional expression but may not capture the full spectrum of nuanced emotions. An exception to this is the GoEmotions dataset, which stands out with its annotation of 28 distinct emotion categories. However, this richness comes with inherent balance limitations, as some emotions may be underrepresented compared to others, which can affect the performance of models trained on this data.

Despite the availability of datasets like GoEmotions that expand the number of emotion categories, there remains a scarcity of datasets that comprehensively address additional dimensions such as valence, arousal, and dominance like EmoBank. These dimensions are crucial in understanding the intensity, positivity or negativity, and control associated with emotions, which are essential for applications ranging from affective computing to mental health assessments. The scarcity of datasets covering these scales underscores a significant gap in emotion data resources, limiting the development and accuracy of models that aim to capture the multifaceted nature of human emotions in textual data.

***Finding 3**: Most available datasets categorize emotions into 6-10 basic categories. Limited datasets annotate complex emotions with more emotional categories or those scaled in continuous dimensions.*

7. Fine-tuned Models

Fine-tuned models have become a cornerstone in modern machine learning, especially in natural language processing (NLP) and computer vision. Fine-tuning involves taking a pre-trained model, already trained on a large dataset, and adapting it to a specific task or dataset. (Wagner, 2021). This approach leverages the pre-existing knowledge encoded in the model, enabling efficient training with less data and computational resources. Fine-tuning has proven highly effective in improving model performance across various tasks, including text classification, sentiment analysis, and image recognition. (Hernández-Álvarez, 2021).

One of the primary advantages of fine-tuning is its ability to transfer learning from a general domain to a specific one. For instance, a model pre-trained on a large corpus of general text can be fine-tuned to excel in a specialized domain like medical literature or legal documents. This transfer of learning significantly reduces the amount of labeled data required for training, as the model already understands fundamental language structures and patterns. Fine-tuned models also benefit from the robustness and generalization capabilities developed during the initial pre-training phase, which helps them perform well even with smaller, task-specific datasets. (Chowdhary & Chowdhary, 2020).

The process of fine-tuning involves several steps, starting with selecting a pre-trained model that closely aligns with the target task. The chosen model is then further trained on the target dataset, adjusting the model's weights through backpropagation. Hyperparameter tuning is often necessary to optimize the performance, involving adjustments to learning rates, batch sizes, and other training parameters. Fine-tuning also requires careful consideration of overfitting, as the model can easily become too specialized for the fine-tuning dataset, losing its ability to generalize. (Ramachandran et al., 2022).

Notably, the final models are often trained on domain-specific datasets like emotion data to achieve high performance on the target task and platform. This fine-tuning allows the models to capture the unique linguistic patterns and emotional expressions present in user-generated content across platforms. Next, the model is compared with baseline models using various parameters to quantify its performance. (Yin et al., 2019). Model evaluation metrics are essential for this purpose.

A confusion matrix is generated, which provides counts of correct and incorrect predictions based on known actual values. This matrix displays true values for data fitting according to emotional categories. Researchers evaluate their models using metrics such as accuracy, precision, recall, and the F1 score.

- Accuracy summarizes the overall performance of the model across all classes, particularly useful when all classes are of equal importance. It is calculated as the ratio of the number of correct predictions to the total number of predictions.
- Precision measures the model's accuracy in categorizing a sample as positive. It is determined by the ratio of correctly categorized positive samples to the total number of positive samples, whether correctly or incorrectly categorized.

- Recall evaluates the model's ability to identify positive samples. It is calculated by dividing the number of positive samples correctly identified by the total number of positive samples.
- F1 Score assesses the model's effectiveness in identifying positive samples, balancing precision and recall. It is calculated as the harmonic mean of precision and recall, providing a single metric that considers both false positives and false negatives.

There are many fine-tuned emotion detection models which can detect basic or complex emotions in text. For the basic emotions, most of the models can be detected with high accuracy, precision, recall, and f1 rates. However, for more complex emotions, the performance of these models remains a challenge (Gaind et al., 2019).

Until recently, fine-tuned emotion detection models with the use GoEmotions can classify 27 different emotions with quite low accuracy (Gaind et al., 2019). Some models can only achieve an accuracy rate of 50%. There is still a challenge to detect complex emotions in text, especially when the emotions are more nuanced or context-dependent (Wu et al., 2017). Detecting 27 emotions is still a challenge, and more research is needed to improve the performance of these models, especially in handling cultural differences and individual variations in the expression of emotions (Mehta & Pandit, 2018).

ROBERTA can detect emotions better than BERT. But even these models struggle with more subtle or mixed emotions, and their performance can vary across different languages and cultural contexts. There are also challenges in collecting and annotating high-quality datasets for emotion detection, which is crucial for training robust models (Kamath et al., 2022a). Not many fine-tuned models used GPT, especially GPT-4 for emotion detection.

Table 4: Fine-tuned model evaluation results

Table 4 summarizes the evaluation results of the most used fine-tuned emotion detection models from Hugging Face. Hugging Face is a leading company and open-source community focused on advancing natural language understanding and AI technology. They are renowned for their contributions to developing state-of-the-art transformer models like BERT and GPT, which have revolutionized various NLP tasks such as language translation, sentiment analysis, and text generation. Beyond its innovations in model architecture, Hugging Face is widely recognized for its platform, which provides tools and resources for researchers and developers to easily access, train, and deploy pre-trained models through their Hugging Face Hub. This hub serves as a centralized repository for sharing and discovering NLP models, datasets, and training scripts, fostering collaboration and accelerating advancements in the field of artificial intelligence (Jain, 2022).

No	**Model name**	**Download**	**Dataset**	**No. of emotions**	**Accuracy**	**Precision**	**Recall**	**F1**
1	bhadresh-savani/bert-base-uncased-emotion dair-ai/emotion	71.8 k	dair-ai/emotion	6	0.926	0.890	0.879	0.884
2	j-hartmann/emotion-english-distilroberta-base	1.1 M	google-research-datasets/go_emotions	27	0.225	0.071	0.167	0.095
3	michellejieli/emotion_text_classifier	262 k	google-research-datasets/go_emotions	27	0.257	0.071	0.127	0.085
4	bhadresh-savani/bert-base-uncased-emotion	71.8 k	google-research-datasets/go_emotions	27	0.058	0.038	0.107	0.042
5	bhadresh-savani/distilbert-base-uncased-emotion	70.4 k	google-research-datasets/go_emotions	27	0.055	0.036	0.102	0.039
6	cardiffnlp/twitter-roberta-base-emotion	21 k	google-research-datasets/go_emotions	27	0.019	0.007	0.039	0.012
7	arpanghoshal/EmoRoBERTa	14.7 k	google-research-datasets/go_emotions	27	0.418	0.529	0.395	0.438
8	SamLowe/roberta-base-go_emotions-onx	16.1 k	google-research-datasets/go_emotions	27	0.474	0.575	0.396	0.450
9	SamLowe/roberta-base-go_emotions	2,8 M	google-research-datasets/go_emotions	27	0.474	0.575	0.396	0.450
10	monologg/bert-base-cased-goemotions-original	30.2k	google-research-datasets/go_emotions	27	0.484	0.538	0.410	0.460

Table 4 provides a detailed overview of the most downloaded fine-tuned models specifically designed for emotion detection. It illustrates that models like bhadresh-savani/bert-base-uncased-emotion and dair-ai/emotion can achieve impressive accuracy rates, such as 92.6%, when detecting the six basic emotions. This highlights the effectiveness of these models in accurately categorizing fundamental emotional states, which is crucial for many applications in natural language processing and sentiment analysis.

However, the challenge intensifies when extending these models to detect a broader spectrum of emotions, as evidenced by the significant drop in accuracy observed in the evaluation. For instance, the accuracy for detecting 27 emotions drops drastically to just 5.5% and 5.8% for the fine-tuned models mentioned. Even the highest-performing model achieves a modest accuracy of 48.4%, accompanied by an F1 score of 0.46, underscoring the difficulty in accurately capturing and categorizing a wider array of nuanced emotional expressions. These findings highlight the complexities and limitations in scaling emotion detection models beyond basic emotional categories, necessitating further research and advancements in model development to improve accuracy and performance across a broader range of emotions.

This study, therefore, concludes that:

Finding 4: *The evaluation metrics revealed that the accuracy and F1 score of models trained on the GoEmotions dataset are significantly lower than those of models trained on datasets with 6-8 emotion categories. The F1 score is only around 0.5. That could be the quality of the dataset or the approach is not appropriate to detect complex objects like emotions.*

8. Discussion

Textual Emotion Detection is a critical task within the field of Natural Language Processing (NLP), aiming to identify and classify emotions conveyed through text. While advancements have been made, Textual Emotion Detection continues to pose significant challenges for researchers and developers. These challenges stem from various factors, including data limitations, the inherent complexity of human emotions, and the dynamic nature of conversational contexts. This study explores these challenges in detail and discusses potential opportunities to overcome them.

8.1. Emotion Theories

The review has highlighted two significant recent developments in emotion theories. The first development is the expansion of the number of categorical emotions. Traditionally, emotion theories have focused on a limited set of basic emotions, typically ranging from six to ten categories. However, recent research has pushed the boundaries by identifying and categorizing a broader spectrum of emotions. This expansion allows for a more nuanced understanding of human emotional experiences, accommodating a wider variety of emotional states beyond the basic ones such as happiness, sadness, anger, and fear. By recognizing a greater number of distinct emotions, researchers can better capture the complexity and richness of human emotional life.

The second development is the scaling of emotional experiences into continuous scores of valence, arousal, and dominance. This approach moves beyond the discrete categorization of emotions to consider the dimensions along which emotions vary. Valence refers to the positivity or negativity of an emotion, arousal indicates the level of activation or intensity, and dominance reflects the degree of control or influence exerted by the emotion. By measuring emotions on these continuous scales, researchers can gain a more detailed and dynamic picture of emotional experiences. This dimensional approach is particularly useful in applications such as affective computing and human-computer interaction, where capturing the subtleties of emotional responses can enhance the user experience.

Both developments represent significant strides in the field of emotion research. The expansion of categorical emotions provides a richer framework for understanding the diversity of emotional experiences, while the continuous scaling of emotions offers a more precise and flexible tool for measuring and analyzing emotions. Together, these advancements enhance our ability to study and interpret human emotions, paving the way for more effective applications in areas ranging from mental health to artificial intelligence. By embracing both the categorical and dimensional aspects of emotions, researchers can develop more comprehensive models that reflect the full complexity of human emotional life.

8.2. Emotion Detection Approaches

GPT (Generative Pre-trained Transformer), especially GPT-4, stands out as a significant leap forward in natural language processing, introducing novel training objectives and architectural innovations that surpass the capabilities of earlier models such as BERT. Unlike BERT, which relies on masked language modeling tasks, GPT employs an autoregressive approach, predicting the next word in a sequence, thereby fostering a deeper understanding of context and coherence in language. This approach allows GPT to generate more fluent and contextually appropriate text, making it particularly suitable for applications requiring nuanced comprehension of emotional nuances and subtleties.

In the realm of emotion detection, GPT's capabilities become especially advantageous as the conceptualization of emotions expands, and the task evolves from traditional multi-label classification to more complex zero-shot learning scenarios. Zero-shot learning in emotion detection involves the model inferring the emotional content of text without prior training on specific emotion labels, relying instead on its broader linguistic understanding and contextual reasoning abilities. GPT's ability to generate coherent text and understand intricate relationships within language positions it well for such tasks, potentially enhancing accuracy and adaptability in capturing a diverse range of emotional expressions.

As the demand grows for emotion detection systems capable of handling increasingly nuanced emotional states, GPT's flexibility and generative prowess offer promising avenues for advancing the field. Its adaptability to zero-shot learning scenarios means that it can potentially generalize across diverse datasets and adapt to new emotional categorizations without the need for extensive retraining, thus paving the way for more sophisticated and context-aware applications in sentiment analysis, mental health assessment, and human-computer interaction.

8.3. Datasets

Most available datasets for emotion categorization focus on a limited range of 6-10 categories. This conventional approach captures basic emotions such as happiness, sadness, anger, and fear, which are sufficient for many applications but fall short of representing the full spectrum of human emotional experiences. These basic categories provide a foundational understanding but lack the nuance required for more sophisticated analyses, leaving a gap in comprehensive emotion research and application.

The GoEmotions dataset stands out as a notable exception, offering a more extensive range of 28 emotion categories. This broader categorization allows for a more detailed exploration of emotional states, capturing subtleties that are often missed by other datasets. However, despite its richness, the GoEmotions dataset has balance limitations, meaning that some emotions are underrepresented. This imbalance can skew the results of models trained on this dataset, potentially affecting their accuracy and reliability in real-world applications.

In addition to the categorical limitations, there is a significant scarcity of datasets addressing the valence, arousal, and dominance scales. These continuous measures of emotional intensity and quality are crucial for applications requiring a nuanced understanding of emotions, such as affective computing and human-computer interaction. The lack of such datasets hampers the development of models that can accurately interpret and respond to the full range of human emotional experiences. As the demand for sophisticated emotion data grows, the creation and refinement of datasets incorporating these dimensions will be essential for advancing the field and improving the performance of emotion-driven applications.

8.4. Fine-tuned Models

The evaluation metrics revealed a significant disparity in performance between models trained on the GoEmotions dataset and those trained on datasets with 6-8 emotion categories. Specifically, models utilizing the GoEmotions dataset exhibit notably lower accuracy and F1 scores. The F1 score, a crucial measure that balances precision and recall, hovers around 0.5 for these models. This contrasts sharply with the higher scores typically achieved by models dealing with fewer emotion categories, indicating that the expanded range of emotions in the GoEmotions dataset presents additional challenges for accurate classification.

This discrepancy could be attributed to several factors, one of which is the quality of the GoEmotions dataset itself. Despite its comprehensive categorization, the dataset suffers from balance limitations, with certain emotions being underrepresented. This imbalance can skew the training process, leading to a model that performs well on more common emotions but poorly on those that are less frequent. Consequently, the overall performance metrics, including the F1 score, are adversely affected, reflecting the model's struggle to generalize across the full spectrum of emotions included in the dataset.

Another possibility is that the approach used to detect and classify emotions may not be suitable for handling the complexity inherent in the GoEmotions dataset. Emotions are multifaceted and context-dependent, making them challenging to categorize accurately, especially when dealing with a larger set of categories. Traditional machine learning and even some advanced deep learning techniques may fall short in capturing the subtle nuances and overlapping nature of emotions. As such, new methodologies or enhancements to existing approaches may be necessary to improve the detection and classification of complex emotional states, ensuring that models can achieve higher accuracy and more reliable F1 scores when working with comprehensive emotion datasets like GoEmotions.

9. Conclusions

The surge in digital online media has significantly increased the demand for advanced analytical techniques to understand emotions conveyed through text. Textual emotion detection has become a pivotal area within this landscape, utilizing artificial intelligence (AI) to analyze emotional content in digital communications. This study provides a comprehensive review of the literature on textual emotion detection, examining methodologies, datasets, fine-tuned models, evaluation metrics, and the challenges faced in this growing field.

Artificial intelligence has been crucial in advancing methodologies for textual emotion detection. The literature highlights several AI approaches, including deep learning, machine learning, rule-based, and keyword-based techniques. Deep learning and machine learning are particularly prominent due to their ability to automate feature extraction and effectively handle large datasets. These approaches employ complex algorithms and neural networks to identify and classify emotions from textual data with high accuracy.

Feature extraction methods vary from traditional lexical features to sophisticated embeddings generated by deep learning models, capturing nuanced emotional undertones in digital texts. The systematic literature review conducted in this study reveals diverse methodologies used in textual emotion detection, often based on emotion models like Ekman's six basic emotions or Plutchik's wheel of emotions. Datasets for training and evaluation are critical, with various publicly available datasets sourced from social media, online reviews, and other digital communications providing rich repositories of emotional expressions.

Despite advancements, several challenges hinder the broader adoption and effectiveness of textual emotion detection. Accurately labeling data with distinct emotions is difficult, especially with more than ten emotions. There is also a scarcity of datasets for valence, arousal, and dominance scores, and issues with imbalanced datasets leading to biased models and low accuracy. The review identifies gaps in existing literature, suggesting the need for further research to improve model generalizability and develop robust methods for handling diverse linguistic and cultural expressions of emotion. These insights will guide the development of more accurate and effective textual emotion detection systems, enhancing our ability to connect and communicate in the digital age.

References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adamov, A. Z. A., Eshref (2017). Opinion mining and Sentiment Analysis for contextual online-advertisement. In.

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. 2020 17th International Computer

Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP),

Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, *8*(6).

Alkaabi, N., Zaki, N., Ismail, H., & Khan, M. (2022). Detecting Emotions behind the Screen. *AI*, *3*(4), 948-960.

Allouch, M., Azaria, A., Azoulay, R., Ben-Izchak, E., Zwilling, M., & Zachor, D. A. (2018). Automatic detection of insulting sentences in conversation. 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE),

Alswaidan, N., & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, *62*(8), 2937-2987.

Alvarez-Gonzalez, N., Kaltenbrunner, A., & Gómez, V. (2021). Uncovering the limits of text-based emotion detection. *arXiv preprint arXiv:2109.01900*.

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International journal of artificial intelligence in education*, *33*(1), 1-34.

Bakker, I., Van Der Voordt, T., Vink, P., & De Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology*, *33*, 405-421.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., Rosso, P., & Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. Proceedings of the 13th international workshop on semantic evaluation,

Baziotis, C., Athanasiou, N., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., & Potamianos, A. (2018). Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

Betella, A., & Verschure, P. F. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, *11*(2), e0148037.

Bian, C., Zhang, Y., Yang, F., Bi, W., & Lu, W. (2019). Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision*, *13*(3), 329-337.

Bostrom, K., & Durrett, G. (2020). Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Buechel, S., & Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.

Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, February 21-26, 2011, revised selected papers,

Can, E. F., Ezen-Can, A., & Can, F. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. *arXiv preprint arXiv:1806.04511*.

Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P. (2019). SemEval-2019 task 3: EmoContext contextual emotion detection in text. Proceedings of the 13th international workshop on semantic evaluation,

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., & Ku, L.-W. (2018). Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Chowdhary, K., & Chowdhary, K. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

Cortis, K., Davis, Brian (2021). Over a decade of social opinion mining: a systematic review. In (Vol. 54, pp. 4873-4965).

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

De Bruyne, L., De Clercq, O., & Hoste, V. (2018). LT3 at SemEval-2018 Task 1: A classifier chain to detect emotions in tweets. Proceedings of The 12th International Workshop on Semantic Evaluation, June 5–6, 2018, New Orleans, Louisiana,

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Diogo, P. M. J., Sousa, M. O. C. L. e., Rodrigues, J. R. G. d. V., Silva, T. A. d. A. M. d. A. e., & Santos, M. L. F. (2021). Emotional labor of nurses in the front line against the COVID-19 pandemic. *Revista Brasileira de Enfermagem*, *74*(Suppl 1), e20200660.

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, *48*(4), 384.

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual review of psychology*, *30*(1), 527-554.

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

Gaind, B., Syal, V., & Padgalwar, S. (2019). Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

Goyal, S., & Tiwari, N. (2017). Emotion recognition: a literature survey. *Int. J. Technol. Res. Eng*, *4*(9), 1502-1524.

Hernández-Álvarez, M. G., Sergio L. . (2021). Detection of Human Trafficking Ads in Twitter Using Natural Language Processing and Image Processing. In (Vol. 1213 AISC, pp. 77-83).

Hoyos, W., Aguilar, J., Toro, M. . (2021). Dengue models based on machine learning techniques: A systematic literature review. In (Vol. 119).

Huang, C.-Y., & Ku, L.-W. (2018). Emotionpush: Emotion and response time prediction towards human-like chatbots. 2018 IEEE Global Communications Conference (GLOBECOM),

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Jain, B., Goyal, G., & Sharma, M. (2024). Evaluating Emotional Detection & Classification Capabilities of GPT-2 & GPT-Neo Using Textual Data. 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence),

Jain, S. M. (2022). Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems* (pp. 51-67). Springer.

Kajava, K., Öhman, E., Hui, P., & Tiedemann, J. (2020). Emotion preservation in translation: Evaluating datasets for annotation projection. *Proceedings of Digital Humanities in Nordic Countries (DHN 2020)*.

Kamath, R., Ghoshal, A., Eswaran, S., & Honnavalli, P. (2022a). An enhanced context-based emotion detection model using roberta. 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT),

Kamath, R., Ghoshal, A., Eswaran, S., & Honnavalli, P. B. (2022b). Emoroberta: An enhanced emotion detection model using roberta. IEEE International Conference on Electronics, Computing and Communication Technologies,

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230254.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., & Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and software technology*, *52*(8), 792-805.

Kondo, Y., Asatani, K., & Sakata, I. (2022). Evaluating Emerging Technologies on the Gartner Hype Cycle by Network Analysis : A Display Technology Case Study. In.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150.

Kumar, A., & Garg, G. (2020). Systematic literature review on context-based sentiment analysis in social multimedia. *Multimedia tools and Applications*, *79*(21), 15349-15380.

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2022). A Review on Text-Based Emotion Detection--Techniques, Applications, Datasets, and Future Directions. *arXiv preprint arXiv:2205.03235*.

Lee, L.-H., Li, J.-H., & Yu, L.-C. (2022). Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(4), 1-18.

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, 1-57.

Liu, H., & Cocea, M. (2017). Fuzzy rule based systems for interpretable sentiment analysis. 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI),

Liu, Y., Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, Stoyanov, Veselin. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses*, *78*(2), 341-348.

Luo, L., & Wang, Y. (2019). Emotionx-hsu: Adopting pre-trained bert for emotion classification. *arXiv preprint arXiv:1907.09669*.

Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, *114*, 57-65.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, *54*(3), 1-40.

Mohammad, S. M. (2021). Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 323-379). Elsevier.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery*, *8*(5), 336-341.

Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, *11*(1), 81.

Oberländer, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. Proceedings of the 27th international conference on computational linguistics,

Ortony, A., Clore, G. L., & Collins, A. (2022). *The cognitive structure of emotions*. Cambridge university press.

Park, S., Kim, J., Ye, S., Jeon, J., Park, H. Y., & Oh, A. (2019). Dimensional emotion detection from categorical emotion. *arXiv preprint arXiv:1911.02499*.

Pashchenko, Y., Rahman, M. F., Hossain, M. S., Uddin, M. K., & Islam, T. (2022). Emotional and the normative aspects of customers' reviews. *Journal of Retailing and Consumer Services*, *68*, 103011.

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*.

Perikos, I., & Hatzilygeroudis, I. (2013). Recognizing emotion presence in natural language sentences. Engineering Applications of Neural Networks: 14th International Conference, EANN 2013, Halkidiki, Greece, September 13-16, 2013 Proceedings, Part II 14,

Plaza-del-Arco, F. M., Martín-Valdivia, M.-T., & Klinger, R. (2022). Natural language inference prompts for zero-shot emotion classification in text across corpora. *arXiv preprint arXiv:2209.06701*.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Elsevier.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, *1984*(197-219), 2-4.

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.

Polignano, M., Basile, P., de Gemmis, M., & Semeraro, G. (2019). A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention. Adjunct publication of the 27th conference on user modeling, adaptation and personalization,

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98-125.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715-734.

Puri, R., & Catanzaro, B. (2019). Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

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

Ramachandran, K., Mary, A. A. S., Hawladar, S., Asokk, D., Bhaskar, B., & Pitroda, J. (2022). Machine learning and role of artificial intelligence in optimizing work performance and employee behavior. *Materials Today: Proceedings*, *51*, 2327-2331.

Rathnayaka, P., Abeysinghe, S., Samarajeewa, C., Manchanayake, I., Walpola, M. J., Nawaratne, R., Bandaragoda, T., & Alahakoon, D. (2019). Gated recurrent neural network approach for multilabel emotion detection in microblogs. *arXiv preprint arXiv:1907.07653*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, *11*(3), 273-294.

Scherer, K., Shuman, V., Fontaine, J., & Soriano, C. (2013). The GRID meets the Wheel: Assessing emotional feeling via self-report. In *Components of emotional meaning: A sourcebook* (pp. 281-298). Oxford University Press.

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061.

Shivhare, S. N., Garg, S., & Mishra, A. (2015). EmotionFinder: Detecting emotion from blogs and textual documents. International Conference on Computing, Communication & Automation,

Shrivastava, A., Amudha, J., Gupta, D., & Sharma, K. (2019). Deep learning model for text recognition in images. 2019 10Th international conference on computing, communication and networking technologies (ICCCNT),

Singh, P., Srivastava, R., Rana, K., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, *229*, 107316.

Suhasini, M., & Srinivasu, B. (2020). Emotion detection framework for twitter data using supervised classifiers. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19* (pp. 565-576). Springer.

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, *35*(5), 96-102.

Tesfagergish, S. G., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences*, *12*(17), 8662.

Tomkins, S. S., & McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and motor skills*, *18*(1), 119-158.

Trinh, T. H., & Le, Q. V. (2018). A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Udochukwu, O., & He, Y. (2015). A rule-based approach to implicit emotion detection in text. Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20,

Wagner, M. (2021). Machine learning in a digital age: The future is now. In.

Wang, X., Ye, Y., & Gupta, A. (2018). Zero-shot recognition via semantic embeddings and knowledge graphs. Proceedings of the IEEE conference on computer vision and pattern recognition,

Willcox, G. (1982). The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, *12*(4), 274-276.

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21-40.

Xia, J., Zhang, J., Sun, W., Zhang, B., & Wang, Z. (2018). Finite-time adaptive fuzzy control for nonlinear systems with full state constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(7), 1541-1548.

Xiao, Z., Chen, Y., Dou, W., Tao, Z., & Chen, L. (2019). MES-P: An emotional tonal speech dataset in Mandarin with distal and proximal labels. *IEEE Transactions on Affective Computing*, *13*(1), 408-425.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Yusifov, E., & Sineva, I. (2022). An Intelligent System for Assessing the Emotional Connotation of Textual Statements. 2022 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF),

Zahiri, S. M., & Choi, J. D. (2018). Emotion detection on tv show transcripts with sequence-based convolutional neural networks. Workshops at the thirty-second aaai conference on artificial intelligence,

Zanwar, S., Wiechmann, D., Qiao, Y., & Kerz, E. (2022). Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. *arXiv preprint arXiv:2212.09465*.

Zhong, P., Wang, D., & Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.