**Literature review**

Emojis have undoubtedly become a ubiquitous part of our online lives, first appearing in 1999 and their use becoming increasingly commonplace throughout digital communication. Their usage has resulted in an unprecedented evolution in how digital communication occurs, seamlessly integrating images into typography. While natural language processing (NLP) is a widely utilised and powerful tool in data analytics, contextual understanding, and ambiguity in text still present challenges to NLP models. Sarcasm is one such source of ambiguity which has been suggested to account for significant challenges to NLP models, for example, Maynard and Greenwood (2014) suggested that sarcasm accounted for up to a 50% drop in accuracy in their sentiment analysis model. While their dataset contained a disproportionate prevalence of sarcastic content and a small training dataset, the work demonstrates the need for accurate detection tools to aid in NLP tasks. Emojis are widely used with the purpose of establishing tone and clues towards sentiment to the reader. This is intuitive to anyone who consumes digital content in the modern age, however in the domain of data analysis and NLP, there are few instances of research into the potential of consideration of emojis towards the processing of text. This review of present relevant literature aims to establish the viability of consideration of emojis to aid in sarcasm detection through the critical evaluation of state-of-the-art methodologies in sentiment analysis and sarcasm detection.

1. *Theories of Emotion Classification:*

The categorisation of emotions presents a challenging question within the field of psychology, with debates on the state-of-the-art found throughout literature. The methodologies employed to achieve such categorisations fall primarily into two classes: basic emotion theory and dimensional theory (Thamm, 2006; Fontaine, 2013) which are fundamentally contradictory. The former initially introduced by Darwin (1872) proposed humans have a limited number of basic emotions which are present to enable fundamental tasks such as survival. These emotions can co-exist to form compound emotions and the degree of arousal for such compound emotions can further add granularity to secondary and tertiary categorisations (Anderson and Adolphs, 2014). While there is no consensus on the quantity or categories of the basic emotions (LeDoux, 1995; Damasio, 1997), broadly speaking all modern research has the same aim; to determine a discrete set of fundamental emotions, each possessing a distinctive neurophysiological signature (Stein and Oatley, 2008; Keltner *et al.*, 2019).

* 1. *Basic Theory*

Robert Plutchik, an initial prominent figure within the field found eight categories; anger, fear, disgust, surprise, anticipation, trust and joy (Plutchik, 1962; Imbir, 2017). His methodology surrounded the identification of emotions which arise from a behaviour with high survival value- for example fear stimulating the fight-or-flight response (LeDoux, 2012; Kravitz and Fernandez, 2015). While this adhered to the basic definition of the model, manual selection of emotions may have resulted in bias- an observation which could be concluded from his own work which stated that the nature of the individual has a significant impact on their experience of compound emotions and thus it appears pertinent to question whether the experiences of Plutchik contributed to his conclusions within this work. Later research developed this theory to increase objectivity in classifications, adding consideration of physiological manifestations of such emotions which must have unique and explainable characteristics (Keltner *et al.*, 2019). Ekman (1992) proposed seven emotions using this method however, later other works argue overlap in many of their categorisations. Jack, Garrod and Schyns (2014) found that disgust and anger share a wrinkled nose and fear and surprise share raised eyebrows and thus the proposed emotions were reduced to six. Additionally, the theory postulates that the emotions must be the result of survival instinct. Research in the field debate the inclusion of anger and disgust in addition to fear and surprise as separate entities respectively as these emotions are thought to have branched due to social functions therefore any works identifying such emotions separately may be invalid (Mansourian *et al.*, 2016). The most widely utilized groupings of basic emotions are the six primary emotions proposed by Ekman: anger, sadness, joy, disgust, fear and surprise and an alternative consisting of four: fear, anger, joy and sadness (Izard, 2007; Gu *et al.*, 2015; Mansourian *et al.*, 2016; Wang and Pereira, 2016), first presented by Izard who stated “people need the category label *fear* to explain flight to one another for safety, *anger* to explain the frustration of blocked goal responses, *joy* to explain the pride of achievement and *sadness* to explain the experience of life-changing loss”.

* 1. *Dimensional theory*

The alternative framework for categorisation of emotions, dimensional theory places emotions at unique points within a two-dimensional plane classifying them based on valence and arousal (Russell, 1980; Russell and Barrett, 1999; Barrett and Russel, 2015). This model presents emotions in a more fluid manner, representing transitory neuropsychological states converse to the discrete categorizations in basic theory, and addresses a limitation with regards to overlapping characteristics between emotions. Dimensional theory is preferable in cases where greater flexibility and consideration to the nuance of emotional experience between individuals adds value, however a majority of literature using this theory relies on self-classification of emotions on scales representing valence and arousal making classification of ‘like’ emotions across a population difficult due to bias of subjects, although this is an intuitive limitation when considered in the context of the question itself. The more rigid definitions of emotions in basic theory may be advantageous for comparison across a population as they are more objective, even when self-reporting is utilized.

* 1. *Application of Emotion Theory in Data Analytics:*
     1. *Basic Theory*

In the domain of data analytics, fine-grained emotion classification is a widely utilized tool in natural language processing, however the principles of either emotion theory seems to be under-utilized. Discrete emotion labels representing primary, secondary, and tertiary emotions are common, however the effects of the subjectivity of such categories are evident with poor generalisation and lack of agreement in label assignment where multiple subjects assign labels to the same content. The core goal of basic emotional theory is to address such a limitation and with clearly differentiated discrete labels thus addressing these limitations directly. The largest dataset for fine-grained emotion classification ‘GoEmotions’ is a human-annotated dataset created from Reddit content, with 27 emotion categories (Alon and Ko, 2020). Subjectivity of classifications is evident in the discussion of outcomes when assigned labels were compared across subjects, where related emotions were relatively evenly distributed across the labels. Google argues this as evidence of efficacy of the model, however in 6% of cases no inter-annotator agreement was found (which was defined in this case as two or more of the annotators) even when multiple classifications were permissible to be assigned, which indicates limitations in the label classes, most likely due to the high dimensionality, as this shortcoming was not observed for datasets where lesser dimensionality was used. It was also noted that the work has poor ability to generalise- with a notable reduction in performance when predictions are made for text sourced outside of Reddit. Saravia et al. (2021) use the 8 emotions outlined by Plutchik, while the reduced dimensionality showed improvements in generalisation due to reduced subjectivity for assignments, it should be noted that the assignment of classifications was based on vocabulary found in hashtags within tweets, which essentially condenses more subjective assignments into the more objective classifications made by the Plutchik model and fails to address the nature of hashtag use- hashtags are not always used to consolidate emotion found in the text, conversely they are often employed to convey sarcasm amongst other figurative devices (Kunneman *et al.*, 2015; Sykora, Elayan and Jackson, 2020).

Graphical user interface, text, application

Description automatically generated

*Figure 1* Tweet located using #fun in a sarcastic context. Methodology of (Saravia et al., 2021) labels this as content expressing joy.

Works utilizing the four and six basic emotions cite fewer limitations due to subjectivity of their categories regardless of the methodologies used to assign labels (Zheng, Mountstephens and Teo, 2020; Pandey, 2021), consolidating evidence that the present state-of-the-art basic emotion models are successful in their aim of creating discrete classifications of emotions which in the context of data analytics is evidenced to improve outcomes in fine-grained emotion classification tasks.

* + 1. *Dimensional Theory*

The dimensional theory of emotion classification is less prominent in natural language processing, however observations by Grgić, Podobnik and Carvalho (2022) provide proof of concept and results which indicate the methodology generalises well between different text sources thus this method may be preferable where generalization is a concern which applies in some cases utilizing discrete emotion categories (Alon and Ko, 2020). This work presents a three-dimensional plane, where the final dimension representing ‘expectancy’. This final dimension could not be located in any literature in the domain of psychology researching dimensional theory, however there is evidence suggesting the degree of anticipation may have an impact on emotion, so its inclusion is not necessarily illogical (Castelfranchi and Miceli, 2018). Rao (2019) highlights the limitations of a dimensional theory approach for sentiment analysis using a simplified one-dimensional scale, finding natural language processing models have a bias towards classifying at the extremes of the scale.

* + 1. *Comparison of Theories*

No work comparing the two models on a ‘like’ problem set could be identified for use in natural language processing, however it is likely that the preferable model would be context-based. Optimal practice during use of discrete emotion categories involves the use of one of the emotion groupings proposed in basic theory to facilitate valid annotation of training data as subjectivity of categories is minimized. For use of continuous scales from dimensional theory best practice is less evident due to the limited relevant work available however, the annotation of training data would be greatly influential on the results. Random sampling during annotation would ensure preceding data does not impact the labels assigned on average where a continuous scale to rank emotions is used due to the primacy and recency effects (Fang, van Kleef and Sauter, 2018; Cowen *et al.*, 2019), which would have a greater impact here than in cases with more objective discrete categories. Manual annotation of training data is widely used across both models and thus results are subject to human perceptions and bias (Ekman, 2020; Feldborg *et al.*, 2021). While bias is a limitation with regards to validity management, the discussion may be more nuanced in this context as the bias mirrors that which exemplifies the human experience of emotion therefore may not have detrimental effects on results. Mitigation of inaccurate labelling to ensure relative homogeneity across annotators would involve all subjects being briefed to ensure an understanding of the definition of each class. Labels should be applied by multiple subjects, monitoring inter-iterator agreement (Mesevage, 2021). Text where no consensus is achieved and annotators who consistently output results contrary to the majority may be unreliable and removal of such information from the dataset may be necessary. Random sampling during annotation would ensure preceding data does not impact the labels assigned on average where a continuous methodology of assignment is used.

1. *State-Of-The-Art Text-Based Sentiment Analysis*

State-of-the-art text-based sentiment analysis generally aims to classify sentiment of text with increasing granularity. Context extraction from text was cited as a key area to improve upon current NLP models capacity to achieve this (Acheampong, Wenyu and Nunoo-Mensah, 2020). Yusof, Mohamed and Abdul-Rahman (2019) evaluated the effects of the consideration of both contextual and semantic information which resulted in improved outcomes in more ambiguous cases, compared to a control which considered keyword sentiment only. Katz, Ofek and Shapira (2015) consolidated these observations, expanding upon research into the scope of the improvement. Their findings suggest that the most significant improvements were observed in cases of noisy and unstructured text, but improvements were observed even in more structured text samples.

Present literature classifying text into emotion-based sentiment categories deploy a range of strategies. Lee, Teh and Pak (2021) defined 12 emotions across a scale with respect to their polarity. The study was successful in correlating word usage to emotions but limitations for this work noted that text in isolation was not the most effective method to achieve this and greater accuracy may be achieved with the inclusion of additional parameters. Authors acknowledged that their approach could not account for the presence of sarcasm, which is common in the context of the data collected. Nandwani and Verma (2021) cited similar limitations of their work, classifying text into 6-8 emotion categories using text only. Of the models proposed, the most successful achieved approximately 94% accuracy, with sarcasm and irony being cited as limitations of their model also. Schuff et al. (2017) showed similar limitations but proposed an interesting strategy to towards improving reliability of labels used in model training; a chance-corrected measure of inter-annotator agreement was used to assess the reliability of the labels used in the training dataset, reducing the impact of inaccurate or labels representing outlier opinions.

The first identified instance of fine-grained sentiment analysis in this review carried out by Socher et al. (2013) trained a recursive neural tensor network, obtaining unremarkable accuracy; 46%. This was improved upon by Peters et al. (2018) which obtained 55% accuracy using a Bi-Attentive Classification Network on the same dataset. This model is the first to attempt to use semantics to improve accuracy using EMLo embeddings and the degree of improvement upon previous state-of-the-art indicates that semantics play a significant role in accuracy for such a task. Improvement upon semantic embeddings largely dictate developments in fine-grained sentiment analysis at present. While neural language models have made significant improvements, deep learning models have greater capacity to capture semantic context from text and thus achieve greater accuracy for fine-grained sentiment analysis. The BERT (Bidirectional Encoder Representations from Transformers) model is one such example of state-of-the-art, with its success attributable to the vast and varied training corpus enabling good generalisation abilities and its bidirectional transformer architecture allowing the capture of dependencies in the forward and reverse direction increasing contextual understanding. Masked language modelling was used during training; a sample of tokens in a sequence were masked and the model was trained to predict them based on the surrounding tokens- this encouraged a deeper understanding of language syntax and semantics. Throughout the analysis of such state-of-the-art works, the limitations associated with figurative linguistic devices were regularly cited.

A picture containing screenshot, text, diagram, line

Description automatically generated

*Figure 2* Fine-Grained Classification methodology of the recursive neural tensor network proposed by Socher et al. (2013).

1. *Conversational agents*

Chatbots are conversational agents which are capable of simulating conversation using natural language. The applications of such technologies are vast including customer service, pedagogy and healthcare and thus significant investments are being made to improve such tools (Forbes, 2022). State-of-the-art chatbots such as ChatGPT and Bard utilise a transformer architecture to process language input. ChatGPT utilises a generative pre-training transformer trained on a vast dataset of online content up to 2021 (Gozalo-Brizuela and Garrido-Merchán, 2023). Bard uses Googles LaMBDA (Language Model for Dialogue Applications) and is trained on the ‘Infiniset’ and is capable of real-time web searching thus has no limitations with regards to recency of data (Aydin and Karaarslan, 2023). Neither model relies on pre-defined outputs to queries which is common in more rudimentary chatbot architectures (Gozalo-Brizuela and Garrido-Merchán, 2023), a distinction which places them at present state-of-the-art, however the way outputs are achieved differs; the latter extracts intent from the users’ query to inform its response while the former utilises statistical patterns to generate output. While no literature could be identified assessing the chatbots’ relative performance with regards to generation of sarcastic output or identification of sarcastic input, based on assessment of similar mechanisms in other contexts, it is likely that both approaches would present limitations with regards to sarcasm. A basic text was performed using ChatGPT, which itself highlighted this as a limitation (see *figure 3*). Statistical approaches often fail to capture subtle contextual nuance which characterises sarcasm and intent-extraction relies to some degree on sentiment within text which is fundamentally incongruent with the intent of the message where sarcasm is utilised. Neither chatbot utilises emojis in any capacity at present, however this may be an area of future opportunity to improve their abilities to understand and react appropriately to user intent. Ethical concerns have been highlighted with regards to this, with Véliz (2023) arguing that any integration of emojis would further enhance the ability of these tools to ‘manipulate’ and illicit emotions from their users. While these concerns should be viewed in the broad context of conversational agents as these models develop increasingly natural feeling conversational abilities the consideration of emojis is not an objectively poor ethical decision- rather if used to enhance semantic understanding in more nuanced situations, their integration would simply enable the creation of more accurate responses from these models.

A screenshot of a chat

Description automatically generated

*Figure 3* ChatGPT self-assessment of sarcasm detection capabilities.

Sarcasm in chatbots is a sparse topic in literature to date; Ili´c, Nakano and Hajnal (2019) created a chatbot trained to generate sarcastic responses to input queries using an LSTM encoder-decoder model. While the model could generate sarcastic responses, it was inconsistent in the quality of delivery and responses did not always adhere to logic or fact. The authors highlighted challenges with regards to objective metrics for evaluation of model performance. Noever and McKee (2023) expanded upon this field in the creation of a ‘factual sarcastic chatbot’ known as Marv. While these works highlight the developing capabilities of such tools to generate sarcastic output, no work could be identified which applied this in a more organic manner; these models necessitate every response be sarcastic, however when used organically, sarcasm is only a natural response in certain contexts. There are significant development opportunities with regards to organic integration of sarcasm into conversational agents with Missy Cummings, an Associate Professor at MIT stating that a sarcastic robot is a ‘holy grail’, however she also discussed related challenges due to the lack of tone (Zawacki, 2015). Additional tone cues may be a key element of development in this field.

1. *Sarcasm Identification*

Figurative linguistic devices refer to any text which possesses a sub-context beyond its literal meaning aiming to add nuance, context, or tone Peters (2022). Examples of such devices include sarcasm, irony, hyperbole, and metaphor. As these devices rely heavily on implied meaning opposing the literal, detection using natural language processing models presents a challenging problem set.

Sarcasm is typically associated with the expression of frustration, anger, or annoyance and its delivery relies heavily on situational awareness or context (Airaksinen, 2020). Structural markers of sarcasm vary significantly across languages both in terms of importance and the structure itself, with languages containing highly structured, codified systems of formality in speech such as Japanese and Korean being more readily identified through structure in addition to context as sarcasm is often conveyed through the exploitation of such structures (Wetzel, 2004). Sarcasm has more methodical usage in these languages, where deliberate misuse of verbs relevant to the interlocutor and situation are used making structural identifiers more useful for identification purposes (Brown, 2013; Iseli, 2023). Conversely, English does not distinguish the relative status of those involved in an exchange and thus honorific sarcasm is not observed. There is little research which has been found to identify sarcasm based on structural markers of text in English, with the split interrogative structure being only marker identified during this review which does not make up a significant proportion of sarcasm found in the English language. The structure of the device is composed of a question which is followed by a tag representing a possible answer, for example: *Who do you think you are, my mother?* (Michaelis and Feng, 2015; Botteri, 2016). While this structure is disproportionately represented in sarcastic text, it is only indicative of sarcasm where the tag is an unexpected response to the question posed. This structure also has a sincere use case where it portrays ongoing performance or proposing future outcomes of current efforts (Clark and Fox Tree, 2002; Michaelis and Feng, 2015). External to the sparse structural markers identified, sarcasm is overwhelmingly identified through ques in body language and tone, which are absent in text analysis (Persicke *et al.*, 2013; Hiremath and Patil, 2021), making its identification within text a challenge referred to as the “Achilles’ heel” of sentiment analysis and other natural language processing tasks in literature (Majumdar *et al.*, 2021).

There are strong indications that these challenges may be related to the difficulties faced in generating a strict definition for sarcasm; literature largely concludes that a single overarching definition of sarcasm cannot be determined. In an academic context, the definition often relates to instances where the quality maximum is violated in pragmatic theory (Grice, 1975). Nuances have been added to such definition in some works for example, Hadi (2013) define sarcasm as any utterance which ‘echoes’ an expectation that has been violated and Sperber and Wilson (1981) add to this with their explanation; the speaker of sarcasm is not performing genuine speech, but rather pretending to with an expectation that the audience will recognise this act. Further expansion of the definition adds an ‘allusional pretence’ which dictates that the sarcasm must not only be pragmatically insincere but additionally must imply a failed expectation or deviation from the norm (Kumon-Nakamura, Glucksberg and Brown, 1995). While these definitions do provide a baseline set of characteristics, there is an element of subjectivity in all definitions identified and thus definition-based sarcasm identification models are not optimal or commonly presented. Most modern research into sarcasm identification where manual annotation of sarcastic content is carried out relies on the individuals intuitive understanding of the concept; a methodology which is largely viewed as effective considering sarcasm is socially determined.

* 1. *Text-Based Sarcasm Detection*

In text, context is widely regarded as the most important factor identification of sarcasm for humans. Social and interpersonal factors are known to affect the use and interpretation of sarcasm (Gibbs, 1994, 2003; Giora, 2012). One such example of this observed by Kreuz (2018) found that with increasing familiarity between individuals the frequency of use of sarcasm increased. Additionally, when interlocutors share increased commonality, they can more readily identify sarcasm (Kreuz and Link, 2002). The two effects are tied together by the *principle of inferability* which is a widely accepted theory stating that speakers only employ sarcasm when there is evidence that receivers will correctly identify it (Kreuz and Roberts, 1995). While the nature of sarcasm use by humans is clearly explained by the conclusions in these works, they also highlight the nature of the limitations of sarcasm detection in an online textual context; machine learning often cannot know the context of the relationships between interlocutors as the receiver of information is often undetermined when the text is created meaning primary mechanisms for detection for humans are not viable. This effect is further amplified in shorter form text such as tweets or online reviews, where character constraints are tight this context is limited and tweets often exist in sequence where sarcasm may only be evident in the context of the overall thread, resulting in excluded context where tweets are considered as single entities. Filatova (2012) presents a ‘crowdsourced corpus’ of sarcastic and non-sarcastic pairs of Amazon reviews for products, sourced from volunteers. The paired reviews were re-classified by new annotators, where context of the product and star rating were provided, and by a separate population where context was not provided. With increasing availability of context there was a significant increase in agreement between participants. While this has a positive implication no assessment of the corpus on ability to classify sarcasm programmatically was carried out. The work concluded that the most significant limitation of sarcasm identification related to absence of context rather than differing definitions of sarcasm between participants; with such an effect being increasingly pronounced in shortform text. In an online context, short-form text and sarcasm are highly interlinked, presenting a dichotomy; context is increasingly limited but also more important.

* 1. *Survey of Strategies for Sarcasm Detection*

Regardless of the challenges presented, the topic of sarcasm detection in text is a vibrant area of research in the domain of data analysis. Table 1 summarizes recent approaches taken with regards to sarcasm detection in online text.

*Table 1* Research methodology summaries for sarcasm detection on twitter.

|  |  |  |  |
| --- | --- | --- | --- |
| **Author(s)** | **Year** | **Labelling approach** | **Algorithm** |
| Zhang, Zhang and Fu | 2016 | Sarcastic and non-sarcastic | Gated Recurrent Neural Network (GRNN) |
| Abercrombie and Hovy | 2016 | Sarcastic and non-sarcastic | Logistic Regression |
| Ghosh and Veale | 2016 | Sarcastic and non-sarcastic | CNN, Long Short-Term Memory (LSTM) Network, Deep Neural Network (DNN) |
| Tungthamthiti, Shirai and Mohd | 2016 | Sarcastic and non-sarcastic | SVM |
| Saha, Yadav and Ranjan | 2017 | Polarity: Positive, negative, and neutral | Naïve Bayes |
| Tay et al. | 2018 | True and false | Multi-Dimensional Intra-Attention Recurrent Network (MIARN) |
| Ren, Ji and Ren | 2018 | Polarity: Positive, negative | CNN |
| Parmar, Limbasiya and Dhamecha | 2018 | Polarity: Positive, negative | Feature-based Composite Approach (FBCA) |
| Parde and Nielsen | 2018 | Polarity: Positive, negative | Naïve Bayes |
| Bouazizi and Ohtsuki | 2019 | Polarity: Positive, negative, and neutral | Random Forest |
| Sharma et al. | 2019 | Sarcastic and non-sarcastic | LSTM |
| Kumar and Garg | 2019 | Sarcastic and non-sarcastic | Naïve Bayes, Gradient Boosting and Random Forest |
| Garg and Duhan | 2020 | Positive, negative Sarcastic and non-sarcastic | SVM |
| Khotijah, Tirtawangsa and Suryani | 2020 | Sarcastic and non-sarcastic | LSTM |
| Sykora, Elayan and Jackson | 2020 | Sarcastic and non-sarcastic | Random Forest |
| Dutta and Mehta | 2021 | Sarcastic and non-sarcastic | CNN-LSTM |
| Eke, Norman and Shuib | 2021 | Sarcastic and non-sarcastic | LSTM |
| Moores and Mago | 2022 | Sarcastic and non-sarcastic | DNN |
| Prasanna, Shaila and Vadivel | 2023 | Sentence structure, Polarity: Positive, negative | Decision tree |
| Pandey and Singh | 2023 | Sarcastic and non-sarcastic | LSTM |
| Sharma et al. | 2023 | Sarcastic and non-sarcastic | DNN |

Of the surveyed literature, there is no consensus drawn on the ideal classifier model for the task, however the majority of works have largely converged on the use of ‘sarcastic’ and ‘non-sarcastic’, or equivalent variants for training data annotation. Regardless of the popularity of such a labelling convention, there are significant limitations which are not addressed in any of these works; the datasets most prominently utilized for this research, where disclosed are heavily skewed towards data which is not representative of organic occurrences of sarcasm on the platform or in digital text as a whole- the tweets present in the datasets are sourced based on indicator hashtags such as #sarcasm, which results in a dataset which disproportionately represents tweets from celebrities and influencers who are anxious to avoid misinterpretation of their content by their audience (Tsur, Davidov and Rappoport, 2010; González-Ibáñez, Muresan and Wacholder, 2011; Bamman and Smith, 2015). Observations within these works also show that these hashtags are in some cases used where sarcasm is not utilised, thus resulting in noise in the dataset. Sarcasm by nature is aligned with negative and often offensive sentiment, which is contrary to intent of these tweets designed to engage audiences in a relatable and non-offensive manner. Additionally, use of such hashtags to label sarcasm is not the primary way sarcasm is presented on the platform, or more broadly in text and thus models trained using datasets with are likely to have an inherent bias towards a certain presentation of sarcasm. (Abercrombie and Hovy (2016) postulate that the nature of the sarcasm in these datasets is disproportionately subtle to avoid negative reception by the audience. The surveyed literature seldom discusses generalisation capabilities of their proposed models, however where this is a consideration outside of twitter data containing highly unambiguous hashtags as labels, the outlook is consistently poor. No evidence could be identified which supports any theory that sarcasm is expressed fundamentally differently on twitter, therefore there is a strong indication that the root cause of such an observation is related to the nature of the data annotation. Sykora, Elayan and Jackson (2020) evaluated one such dataset which sourced tweets in this manner and found just 15% of tweets annotated as sarcastic to have accurate labels and highlighted concerns regarding the validity of the model accuracies reported within papers utilising this method of annotation. Authors discuss the need for more critical evaluation of sentiment analysis results in future work in addition to new data preparation methodologies. Abercrombie and Hovy (2016) propose an improved data sampling strategy which aims to account for the principles of inferability through use of tweet threads and avoids use of tweets containing highly unambiguous hashtags thus addressing primary limitations of modern sarcasm detection research. This work developed a hand-annotated corpus of contextualised sarcastic and control (non-sarcastic) twitter exchanges which outperformed the control dataset, indicating potential to improve upon present conventions which have significant validity concerns with regards to dataset relevance. However, the work acknowledges that further efforts are necessary to assess the ability of the method to improve generalisation of sarcasm detection in text. A notable point discussed which is supported by others within both NLP (Filatova, 2012) and cognitive psychology (Utsumi, 2000; Gibbs and Colston, 2007) domains agree that additional context aids in the identification of sarcasm by humans. Such context can involve an understanding of the relationship between interlocutors or information about the behaviours of the individual outside of the scope of the specific interaction. Availability of such information increases agreement between annotators and accuracy in manual annotation- this may be essential to state-of-the-art annotation in this context and improve sarcasm detection outcomes overall.

Three text-based sarcasm detection frameworks were highlighted as state-of-the-art through this section of the review:

* Fracking Sarcasm using Neural Networks (FSNN) which utilises a CNN, LSTM and DNN in combination to detect sarcasm on the sentence level (Ghosh and Veale, 2016).
* Contextualized Affect Representations for Emotion Recognition (CARER) uses a multi-layer CNN architecture with a matrix of enriched patterns which preserve semantic relationships between word clusters to detect emotion (Saravia *et al.*, 2021).
* Role of Conversation Context for Sarcasm Detection (RCCSD) which uses conditional LSTM networks with sentence-level attention on conversational context and response (Ghosh, Fabbri and Muresan, 2017).

The survey of present methodologies indicates that current state-of-the-art sarcasm detection methods are converging on a single data annotation convention, which presents concerns regarding the accuracy of outcomes of the resulting models, and thus any reports of accuracies where this annotation methodology has been utilised should be viewed with increased scepticism with regards to their true ability to detect sarcasm. Outcomes indicate that sarcasm detection may be achievable using many methods, however the accuracy which can be achieved is not necessarily clear. Notably, punctuation, case-shifting or embedded pictographs such as emojis were universally removed during data pre-processing in all cases assessed in this section of the review. This is contrary to expectations as such features are frequently used to convey tone or add context and emotion cues (Briscoe, 1997; Boutet *et al.*, 2021; Hand *et al.*, 2022), which is the primary limitation of sarcasm detection cited in the surveyed literature. Works which add consideration to these features do exist however this research is sparse at present. The most successful text-based sarcasm detection approaches each present differing strategies to consider additional context in addition to the text itself- this provides a strong indication that additional context is the key to improvement of these tools in future works.

1. *Use of Emojis*
   1. *Emoji Sentiment Classification-Algorithmic Annotation*

Research into the impact of emojis on the perception of sentiment in text is an emerging research field at present both in the context of NLP and linguistics. The classification of the sentiment of the emoji itself is essential for such a task, with official Unicode documentation (Unicode Consortium, 2023) and manually built lexicons being primary sources for such information. The meaning of emojis is largely socially determined, the former does not align with semantic and usage and thus lexicons are preferred. (Ahanin and Ismail, 2022) uses a multi-label method utilising discrete classifications of 11 emotions. The work proposes a balanced weighted PMI algorithm to classify the emojis, validated against corresponding text polarity and the 11 emotion categories. This work does not consider incongruency in sentiment between the text and the emojis thus the lexicon may not be valid. Karthik, Nair and Anuradha (2018) detail the use of machine learning classifiers, artificial neural networks and convolutional neural networks to predict emoji sentiment programmatically. Poor performance (<50% accuracy) was observed for all approaches with the exception of the convolutional neural network which yielded accuracy of 97%. Following the assessment of each model a manual assignment of polarities to emojis was utilised for the remainder of the work. The improved accuracy observed for the convolutional neural network may not necessarily be opposing the hypothesis that incongruency in sentiments may reduce the performance of such classifiers as such neural networks are ‘fully connected’ and thus overfitting occurs disproportionately compared to other models assessed in this work (Pavlitskaya, Oswald and Zöllner, 2022). The work did not discuss reasons for the poor performance of the other models; however, the methodology aligns with that of Ahanin and Ismail (2022) and thus the same logic applies. Yoo and Rayz (2021) follow a similar methodology and cited emojis with double meaning as the primary validity concern relating to their work.

This limitation was identified in all assessed works where similarity scores to text are utilised- no works could be identified which address this limitation although in some cases it was acknowledged. The consideration of figurative use-cases for emojis is likely a key limitation as many emojis are used disproportionately in a figurative manner; for example, Emojipedia found that in only 7% of cases the 🍑 emoji referred to fruit (Emojipedia, 2023). No works could be identified which discuss quantitatively the use of individual emojis in sarcastic and non-sarcastic manners, however it is known that the population is used to convey figurative speech and tone regularly (Kaye, Wall and Malone, 2016).

* 1. *Emoji Sentiment Classification-Manual Annotation*

Hogenboom et al. (2013) built a lexicon utilising manual assignment of sentiment where annotators classified emojis based on a 5-point polarity scale. In the review of current methods in this work, manual assignment was the only methodology identified which avoided reliance on text comparison and was found to have high accuracy; in approximately 90% of cases all three annotators reached consensus with regards to the sentiment of the emoji and. Works in this field primarily aim to assign labels based on polarity rather than discrete emotion categories. This may be an area of opportunity as discrete emotion categories would reduce subjectivity and increase contextual clues for the text- no works evaluating this methodology could be identified during this review.

1. *Viability of Emojis to Improve Sarcasm Detection*

While computational emoji classification presents significant limitations due to the frequent use to convey figurative linguistic devices, the use of accurate corpus’ are essential to the validity of works which use emojis to gain insight into sentiment of text. Literature evaluating the topic in depth is sparse however references to future opportunities are relatively frequent. Riordan (2017) indicated that use of emojis has predictable and strategic use in text and thus patterns observed in the intent of their use may be a valid method to provide valuable context. Hand et al. (2022) evaluated the impact on perception of text sentiment using a positive, negative, and neutral face emojis. Their work focused on the congruency of the emoji and text sentiment determining an amplification of sentiment where there was an agreement in sentiment and an increase in ambiguity where incongruency was observed. The work postulates that this may be indicative of sarcasm however did not explore a causal link. Subramanian et al. (2021) evaluated the consideration of emojis in a deep neural network for sarcasm detection and reported some improvement upon control tests however this work made use of the aforementioned poor data annotation strategy and no context awareness in the analysis of the text thus reported results do not necessarily represent the true potential of such a strategy and have validity concerns. (Prasad *et al.*, 2017) proposes a work which integrates emoji and slang dictionaries to consider their respective contributions towards sarcasm detection in tweets. The methodology is robust with regards to annotation- utilising a human-annotated dataset, however consideration towards context is omitted and the classification algorithms employed are not in line with what is presently regarded as state-of-the-art in this context. This work shows an improvement where slang and emojis are considered compared to where they are not however their respective contributions to the noted improvement are not disclosed.

Conclusion:

There is a clear dichotomy between works considered evaluating emoji and text-based sentiment analysis respectively. Literature pertaining to text-based analysis highlight a lack of context as a primary limitation in cases where the tone is not evident from the words themselves but rather the subtext- where sarcasm is found. Emojis are regarded as valuable tools to enhance semantic understanding of text, while works aiming to consider emojis have evidence of potential, the sample sets evaluated are generally not of high quality and annotation strategies present validity concerns. The approaches with regards to sentiment analysis are not consistent across text and emoji-based work- with granularity of classifications being the primary reason for this discrepancy. In the case of text-based work, increasing granularity in emotion-based categories is the direction of development in the field- however evidence suggests that abiding by classifications outlined in basic emotion theory achieves better outcomes due to greater objectivity in groupings. In emoji-sentiment analysis the classifications are notably less granular- often operating in terms of polarity. Classifications in literature based on textual comparison do not address the impact of sarcasm on the outcomes and thus are not regarded as reliable. Valid human-annotated strategies are the present most viable present option for sentiment-labelled emoji corpus’ and similar is true for annotation of sarcastic text for model training. Based on the work reviewed, emojis present great potential in this problem set and avenues for development may involve exploration of the potential of sentiment analysis classifiers for emojis and text respectively in addition contextual clues between the two. Sentiment labels have not yet been considered in depth in this context- however many works have highlighted their potential.

The review has highlighted a key direction towards improvement upon current state-of-the-art sarcasm detection; as it has been shown to rely heavily on contextual clues, consideration of the embedded emojis within the text is a viable research avenue towards achieving improved outcomes. Current work in the field largely relies on annotation strategies which do not lend validity to reported results- future works should aim to improve upon this convention to glean results which are representative of real-world use cases of such tools.