Unlocking the Pragmatics of Emoji: Evaluation of the Integration of Pragmatic Markers for Sarcasm Detection

Niamh Farnham

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

August 2022

Supervisor: Marina Iantorno

# Acknowledgements

The completion of this work could not have been possible without the staff of CCT College Dublin. Our programme coordinator Dr. David McQuaid who has delivered engaging classes and has gone above and beyond to ensure our learning was an engaging and enjoyable experience. My supervisor Marina Iantorno provided guidance and reassurance to me throughout the course of this work also. Each of the teaching faculty, library staff also had an invaluable role to play in the learning and development which is reflected in this work. Their contributions are sincerely appreciated and gratefully acknowledged.

I would like to extend thanks to each of my classmates who have contributed to ensure classes were engaging, provided an environment where we could each support each other during our learning and development and sharing in the achievements and challenges faced along the way. You have all undoubtedly helped me reach this point, for which I am very grateful.

Finally, I would like to acknowledge my family, friends and boyfriend who have each supported me in this process. You were all very patient when the only topic of conversation I was interested in was emojis.

Table Of Contents

[Acknowledgements i](#_Toc145513872)

[1 Abstract xi](#_Toc145513873)

[2 Introduction xii](#_Toc145513874)

[3 Research Problem and Statement of Objectives 1](#_Toc145513875)

[3.1 Problem Statement 1](#_Toc145513876)

[3.2 Problem Definition 1](#_Toc145513877)

[3.3 Research Objectives 2](#_Toc145513878)

[3.4 Delimitation 2](#_Toc145513879)

[4 Literature Review 3](#_Toc145513880)

[4.1 Theories of Emotion Classification 3](#_Toc145513881)

[4.1.1 Basic Theory 3](#_Toc145513882)

[4.1.2 Dimensional theory 4](#_Toc145513883)

[4.2 Application of Emotion Theory in Data Analytics 5](#_Toc145513884)

[4.2.1 Basic Theory 5](#_Toc145513885)

[4.2.2 Dimensional Theory 6](#_Toc145513886)

[4.2.3 Comparison of Theories 6](#_Toc145513887)

[4.3 State-Of-The-Art Text-Based Sentiment Analysis 7](#_Toc145513888)

[4.4 Sarcasm Detection 8](#_Toc145513889)

[4.4.1 Text-Based Sarcasm Detection 9](#_Toc145513890)

[4.4.2 Survey of Strategies for Sarcasm Detection 10](#_Toc145513891)

[4.5 Use of Emojis 14](#_Toc145513892)

[4.5.1 Emoji Sentiment Classification – Algorithmic Annotation 14](#_Toc145513893)

[4.5.2 Emoji Sentiment Classification – Manual Annotation 15](#_Toc145513894)

[4.6 Viability of Emojis to Improve Sarcasm Detection 16](#_Toc145513895)

[4.7 Chapter Conclusion 16](#_Toc145513896)

[5 Lexical Study of Emojis 18](#_Toc145513897)

[5.1 Dataset Selection 18](#_Toc145513898)

[5.1.1 Basic Emotional Theory 18](#_Toc145513899)

[5.1.2 Dimensional Theory 18](#_Toc145513900)

[5.2 Definition of Sentiment Parameters 18](#_Toc145513901)

[5.2.1 Basic Emotional Theory 18](#_Toc145513902)

[5.2.2 Dimensional Theory 19](#_Toc145513903)

[5.3 Emoji Feature Extraction 21](#_Toc145513904)

[5.3.1 Use of Emotion Vocabulary Embeddings 21](#_Toc145513905)

[5.3.2 Word Vectors of Basic Emotion Words 22](#_Toc145513906)

[5.3.3 Binary Word Association Lexicon 23](#_Toc145513907)

[5.3.4 Word-Emotion Intensity Lexicon 23](#_Toc145513908)

[5.3.5 Results 24](#_Toc145513909)

[5.4 Sentiment-Aware Vector Space Modification 24](#_Toc145513910)

[5.4.1 Evaluation Parameters 25](#_Toc145513911)

[5.4.2 Results 26](#_Toc145513912)

[5.5 Emoji Sentiment Prediction – Basic Theory 29](#_Toc145513913)

[5.5.1 Data Preparation 29](#_Toc145513914)

[5.5.2 Model Selection 29](#_Toc145513915)

[5.5.3 Model Evaluation 30](#_Toc145513916)

[5.6 Emoji Sentiment Prediction – Dimensional Theory 32](#_Toc145513917)

[5.6.1 Results and Methodology Evaluation 33](#_Toc145513918)

[5.7 Chapter Conclusion 33](#_Toc145513919)

[6 Pragmatics of Sarcasm 34](#_Toc145513920)

[6.1 Evaluation of Annotation Strategies 34](#_Toc145513921)

[6.2 Primary Research Methodology 35](#_Toc145513922)

[6.3 Survey Outcomes 35](#_Toc145513923)

[6.3.1 Demographics 35](#_Toc145513924)

[6.3.2 Annotations 36](#_Toc145513925)

[6.4 Statistical Analysis of Pragmatics in Sarcastic Content 39](#_Toc145513926)

[6.4.1 Structural and Sentiment Features 39](#_Toc145513927)

[6.4.2 Context-Based Features 41](#_Toc145513928)

[6.5 Quality Evaluation of Sarcasm-Annotated Datasets 44](#_Toc145513929)

[6.5.1 Survey Methodology 44](#_Toc145513930)

[6.5.2 Hashtag Annotation Strategy 45](#_Toc145513931)

[6.6 Chapter Conclusion 47](#_Toc145513932)

[7 Integration of Emoji Pragmatics into Sarcasm Detection 48](#_Toc145513933)

[7.1 Proposed Novel Methodology 48](#_Toc145513934)

[7.2 Proposed Architecture 49](#_Toc145513935)

[7.2.1 Model Selection 49](#_Toc145513936)

[7.2.2 Sentiment-Aware Attention Mechanism 49](#_Toc145513937)

[7.3 Data Preparation 50](#_Toc145513938)

[7.3.1 Expansion of Current Training Dataset 50](#_Toc145513939)

[7.3.1.1 Statistical Evaluation to Identify Aligned Features 50](#_Toc145513940)

[7.3.1.2 Results 53](#_Toc145513941)

[7.3.2 Data Cleaning 54](#_Toc145513942)

[7.4 Baseline Model Evaluation 55](#_Toc145513943)

[7.4.1 Neural Network Selection 55](#_Toc145513944)

[7.4.2 Evaluation Metrics 55](#_Toc145513945)

[7.4.3 Model Performance 56](#_Toc145513946)

[7.5 Attention Mechanism 57](#_Toc145513947)

[7.5.1 Architecture 57](#_Toc145513948)

[7.5.2 Model Performance 60](#_Toc145513949)

[7.6 Impact of Emojis on Model Performance 61](#_Toc145513950)

[7.7 Chapter Conclusion 61](#_Toc145513951)

[8 Evaluation 61](#_Toc145513952)

[8.1 Solution Strengths 61](#_Toc145513953)

[8.2 Solution Limitations 61](#_Toc145513954)

[8.3 Applications 61](#_Toc145513955)

[9 Conclusion 61](#_Toc145513956)

[10 References 61](#_Toc145513957)

[11 Appendix 61](#_Toc145513958)

[11.1 Word Vector Selection 61](#_Toc145513959)

[11.1.1 Consideration of Training Data 61](#_Toc145513960)

[11.1.2 Word Vector Bias 62](#_Toc145513961)

[11.2 Word Vector Optimisation Process 62](#_Toc145513962)

[11.2.1 Optimiser Selection 62](#_Toc145513963)

[11.2.2 Model Optimisation 63](#_Toc145513964)

[11.2.3 Normalisation 64](#_Toc145513965)

[11.3 Basic Theory Regression 65](#_Toc145513966)

[11.3.1 Feature Scaling 65](#_Toc145513967)

[11.3.2 Mitigation of Skew in Target Variables 65](#_Toc145513968)

[11.3.3 Assessment of Components for Model Selection for Basic Theory Regression 66](#_Toc145513969)

[11.4 Model Development for Basic Theory Regression 66](#_Toc145513970)

[11.4.1 Data Split 67](#_Toc145513971)

[11.4.2 Models Evaluated 67](#_Toc145513972)

[11.4.3 Target Variable Format Selection 68](#_Toc145513973)

[11.4.4 Model Optimisation 68](#_Toc145513974)

[11.4.5 Hyperparameter Tuning 68](#_Toc145513975)

[11.4.6 Cross-Validation 69](#_Toc145513976)

[11.4.7 Dimensionality Reduction 69](#_Toc145513977)

[11.4.8 Model Evaluation 69](#_Toc145513978)

[11.4.9 Neural Network Evaluation 70](#_Toc145513979)

[11.4.9.1 Model Selection 70](#_Toc145513980)

[11.4.9.2 Performance Evaluation 73](#_Toc145513981)

[11.5 Dimensional Theory Regression 73](#_Toc145513982)

[11.5.1 Selection of Pre Trained Models 73](#_Toc145513983)

[11.5.2 Evaluation of Architectures in Literature for Sentiment Regression 74](#_Toc145513984)

[11.6 Survey Components 77](#_Toc145513985)

[11.6.1 Harm and Risk Assessment 77](#_Toc145513986)

[11.6.2 Participant Sampling Strategy 79](#_Toc145513987)

[11.6.3 Sourcing Participants 80](#_Toc145513988)

[11.6.4 Survey Questions 80](#_Toc145513989)

[11.6.5 Survey Design 85](#_Toc145513990)

[11.7 Statistical Tests Used 87](#_Toc145513991)

[11.7.1 Selection Methodology 87](#_Toc145513992)

[11.7.2 Statistical Tests Used 88](#_Toc145513993)

[11.8 Pragmatic Features Evaluated 88](#_Toc145513994)

[11.9 Topic Modelling Process 90](#_Toc145513995)

[11.9.1 Data Preparation 90](#_Toc145513996)

[11.9.2 Models Assessed 92](#_Toc145513997)

[11.9.3 Model Selection and Evaluation 94](#_Toc145513998)

[11.9.4 Results 96](#_Toc145513999)

[11.10 Sarcasm Detection Model 97](#_Toc145514000)

[11.10.1 Data Preparation – Twitter Specific Text Features 97](#_Toc145514001)

[11.10.2 Data Preparation – Emoji Features 98](#_Toc145514002)

[11.10.3 Data Preparation – Grammatical Features 98](#_Toc145514003)

[11.10.4 Model Selection for Sarcasm Detection 99](#_Toc145514004)

List of Figures

[Figure 1 Tweet located using #fun in a sarcastic context. Methodology of (Saravia et al., 2021) labels this as content expressing joy. 5](#_Toc145512166)

[Figure 2 Fine-Grained Classification methodology of the recursive neural tensor network proposed by Socher et al. (2013). 8](#_Toc145512167)

[Figure 3 Sample of result of emoji sentiment classification using dimensional emotional theory. Coloured bars are proportional to p+, p0 and p-. Markers (black) represent . Grey markers represent the limits of the 95% confidence intervals for the . Note weary face (left) has high confidence that a negative classifier is appropriate, however ghost (right) cannot necessarily be classified as positive with 95% confidence. 21](#_Toc145512168)

[Figure 4 Cosine Similarity. 22](#_Toc145512169)

[Figure 5 Process to select model for basic theory regression task. 30](#_Toc145512170)

[Figure 6 Survey responses deconstructed by demographics. 36](#_Toc145512171)

[Figure 7 Age distribution of Twitter users. 36](#_Toc145512172)

[Figure 8 Classification results breakdown by gender and age. 37](#_Toc145512173)

[Figure 9 Distribution of labels assignments for each survey response. Where ‘I don’t know’ was selected 15.84% of the time (approximately 1-2 times per survey). Considering only sarcastic and non-sarcastic labels and controlling for unassigned text 68% of respondents answered within 2 responses of a 1:1 ratio, aligning closely with speculative labels. 37](#_Toc145512174)

[Figure 10 Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Feature. 52](#_Toc145512175)

[Figure 11 Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Subset. 53](#_Toc145512176)

[Figure 12 Data cleaning steps for neural network model training. 54](#_Toc145512177)

[Figure 13 Optimal GRU Model for Sarcasm Detection. 56](#_Toc145512178)

[Figure 14 Training Profile of Optimal GRU model. 56](#_Toc145512179)

[Figure 15 Optimal GRU Model for Sarcasm Detection with Attention Layer. 60](#_Toc145512180)

[Figure 16 Architecture and mechanism of transformer model attention layers. 77](#_Toc145512181)

[Figure 17 Outline of sampling for emoji usage survey questions. 85](#_Toc145512182)

[Figure 18 Single select radio buttons to collect discrete data. In built logic ensures that responses yield valid output using prompts to ensure only one option can be selected. Text fields may also contain conditional formatting to ensure instructions are followed. 86](#_Toc145512183)

[Figure 19 Free text field for emoji use questions is populated with a default value of the text string the participant has previously classified as sarcastic or non-sarcastic to improve response quality. This enables the participant to insert emojis with ease into the text prompt. The figure displays the desktop version of the survey which features a drop-down menu containing emojis to make responses easier where an emoji keyboard is not available. 86](#_Toc145512184)

[Figure 20 Flow chart to determine is data under evaluation is parametric or non-parametric. 87](#_Toc145512185)

List of Tables

Table 1 Research methodology summaries for sarcasm detection on twitter. 11

Table 2 Pearson Correlation scores for all considered prediction methods. Bolded Scores represent the greatest correlation (positive or inverse) observed for the emotion in the respective column, excluding human agreement scores. Average values are reported with respect to absolute correlation. 24

Table 3 Pearson Correlation scores for all considered prediction methods following vector-space modification- method 1. 27

Table 4 Pearson Correlation scores for all considered prediction methods following vector-space modification- method 2. 29

Table 5 Results of model development process for basic theory regression displaying the best outcomes for each model, correct to three significant figures. The selected optimal model is in bold for each basic emotion. Note that reported metrics are with reference to validation performance. 31

Table 6 Assessment of model accuracy considering the mean values of each basic emotion. 32

Table 7 Top 10 Most Used Emojis in Survey Results. 39

Table 8 Topics Identified using the Optimal Model. 41

Table 9 Statistical Evaluation of Emoji-Based Markers of Sarcasm. 42

Table 10 Statistical Evaluation of Emoji-Based Sentiment Skew in Tweets. 42

Table 11 Statistical Evaluation of Sentiment Congruence Between Emojis and Text. 43

Table 12 Statistical Evaluation of Text-Based Markers of Sarcasm. 43

Table 13 Statistical Evaluation of Context-Based Markers of Sarcasm – Topic Modelling. 44

Table 14 Comparison of Sarcastic Text Features between datasets. 46

Table 15 Baseline for Evaluation of Tweets for Organic Sarcasm Features. 51

Table 16 Performance Metrics of Optimal GRU Model for Sarcasm Detection. Each fold is detailed for the training and validation data to display the instability of the model. 57

Table 17 Performance Metrics of Optimal GRU Model for Sarcasm Detection. 60

Table 18 Summary of hyperparameters assessed. 63

Table 19 Strategies evaluated to mitigate skew for each basic emotion. 65

Table 20 Results of skewness transformation. 65

Table 21 Neural Network performance evaluation results. 73

Table 22 Survey questions to verify participant eligibility. 82

Table 23 Survey questions to establish demographics of participants. 83

Table 24 Survey questions about emoji usage. 83

Table 25 Selection of Statistical Tests. 87

Table 26 Summary of results from topic modelling for best models. 96

# 1 Abstract

# 2 Introduction

# 3 Research Problem and Statement of Objectives

## 3.1 Problem Statement

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 is a widely utilised and powerful tool in data analytics, contextual understanding, and ambiguity in text still present challenges to these models. Sarcasm is one such source of ambiguity which has been suggested to account for significant limitations in these models, largely due to limited semantic clues due to the short-form nature of online text; the models cannot know that the implied meaning of the text opposes the literal. Additionally, word vectors are largely the sole source of data considered for the purposes of the classification, which limits the scope for pattern recognition in prevalent model architectures such as neural networks. Emojis are widely used for the purpose of establishing tone and sentiment to the reader. This is intuitive to anyone who consumes digital content, however in the domain of natural language processing, there are few instances of research into the potential for consideration of emojis as a pragmatic indicator in text processing, however literature suggests that this approach has potential. The research problem can be summarised by the following question:

*Can the consideration of pragmatic cues from emojis in text improve outcomes for sarcasm detection?*

## 3.2 Problem Definition

The research problem proposed for this work can be clarified using the problem definition model as follows:

*Problem Identification:* Figurative literary devices such as sarcasm present significant challenges to modern natural language processing models as their identification relies heavily on tone, and context including interlocutor relationships which are not captured in the models. Emojis are often used to convey tone or sentiment within text, however literature exploring their contribution to sarcasm detection is sparse.

*Problem Clarification:* Lexical study of emojis in sarcastic text may be a key avenue for exploration to improve upon present state-of-the-art sarcasm detection for natural language processing.

*Problem Formulation:* Evaluation of emoji pragmatics in sarcastic content compared to literal counterparts may provide additional insight into the tone or context of online content. Integration of these patterns may increase the capacity of natural language processing models to consider these complex linguistic devices for more accurate sentiment analysis outcomes.

## 3.3 Research Objectives

To address the question, the following objectives are proposed:

*Objective 1:* Determine optimal emotion classification methodologies with respect to psychological principles and their implementation within a data analytics context for text and emoji classification. Propose feature extraction strategies for emojis which supplement limitations of pragmatic information in text-based analysis.

*Objective 2:* Establish patterns between text and emoji features which characterise sarcasm in online text. Propose a novel methodology for sarcasm detection which utilises emojis to improve pragmatic understanding of text.

*Objective 3:* Evaluate the impact of integrating emoji pragmatic information into sarcasm detection models. Propose implementations of this novel methodology within real-world natural language processing applications and critically assess its impact.

## 3.4 Delimitation

An intuitive delimitation to this work relates to the problem definition which dictates that the conclusions drawn may not apply where emojis are not used. The scope of the data used in this work was limited to tweets generated by native or near-native English speakers. While sarcasm is a universal concept worldwide, conclusions drawn on how it is conveyed within text can only apply to the English-speaking population as this work does not consider the possibility of dynamic semiotics of emojis across languages or cultures. The dataset also limits the scope of any conclusions drawn to twitter data as this work does not assess the consistency in pragmatic patterns across platforms, which may be relevant due to varying limitations and conventions relating to content length, structure and formality of language used.

# 4 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.

## 4.1 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).

## 4.1.1 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”.

## 4.1.2 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.

## 4.2 Application of Emotion Theory in Data Analytics

## 4.2.1 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 seem 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.

## 4.2.2 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 identified 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.

## 4.2.3 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.

## 4.3 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).

## 4.4 Sarcasm Detection

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.

## 4.4.1 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.

## 4.4.2 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.

## 4.5 Use of Emojis

## 4.5.1 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).

## 4.5.2 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.

## 4.6 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.

## 4.7 Chapter 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.

# 5 Lexical Study of Emojis

This chapter aims to evaluate relevance of parameters for inclusion in an emoji lexicon with respect to the most optimal outcomes for later use to detect sarcasm. The most prevalent methodologies in literature are largely unfit for this purpose as sentiment is derived from the corresponding text, however, incongruency in text and emoji sentiment is widely cited as a marker of sarcasm thus there are validity concerns with regards to any works which deploy this strategy.

## 5.1 Dataset Selection

## 5.1.1 Basic Emotional Theory

By the nature of the problem domain, there is a necessity for human-annotated data, where emojis are presented independent of any text which may influence classifications. The EmoTag1200 dataset contains information regarding volunteers’ association of emojis to basic emotions, based on the Plutchik model. The presentation of emojis without any textual prompts makes this dataset uniquely suitable for this task, and methodologies for its generation are robust; multiple annotators were utilised, and agreement was monitored via Pairwise Pearson correlation and Krippendorff’s α. Steps were taken to mitigate fatigue bias through randomisation and reported results represent averages of the reported values.

## 5.1.2 Dimensional Theory

The dataset used in this section was the Emoji Sentiment Ranking dataset, containing rankings of the 751 most popular emojis annotated by 83 annotators as positive, negative, or neutral. Annotators were presented with the emoji alongside the text, however, were instructed to rank the emoji sentiment only. While bias associated with the textual component of the content cannot be totally ruled out in this case, no datasets presently exist which are created where emojis were annotated independently of text.

## 5.2 Definition of Sentiment Parameters

The following section aims to evaluate optimal approaches to define sentiment of emoji with regards to the two emotional theory models. Successful methodologies should generate parameters can fingerprint sentiment in terms of a wide array of features to facilitate highly accurate machine learning models to be constructed in subsequent work.

## 5.2.1 Basic Emotional Theory

The EmoTag1200 dataset contains information regarding the affinities of emoji to 8 basic emotions, which enable comparison and acknowledge that emotions are experienced synchronously with varying strengths based on the stimulus in question. Sentiment labels *c*, consist of continuous values representing affinity of the emoji to each emotion within set *e*:

where

Labels are most clearly equated to relative affinities to emotions; however, it is also reasonable to link these values to probabilities that a given emotion is the primary emotion associated with the emoji. Such a definition implies that:

where

The primary basic emotion is , the secondary is and the trend continues with decreasing affinity to the emoji. The former definition aligns more literally with the intention of original authors and provides increasing granularity which may be of value when more nuanced fingerprinting between emoji is necessary. The latter yields reduced granularity however greater accuracy of models generated for prediction of unseen emoji may be possible, and thus the result may be data of overall greater value.

## 5.2.2 Dimensional Theory

The Emoji Sentiment Ranking dataset contains information regarding the frequency to which emojis are classified as positive, negative, or neutral. These classifications were represented by discrete values which consider logical ordering of categories of equal distance apart to facilitate quantitative analysis:

Using these values, a discrete probability distribution for sentiment label *c*, can be determined where an assumption that the sum of the three respective probability distributions is equal to one is made to normalise data across emojis which appear at varying frequencies:

Where *c* is the sentiment label, and the following abbreviations were used to denote negativity, neutrality, and positivity:

(Negativity)

(Neutrality)

(Positivity)

Due to the low occurrence of annotation (*N* ≥ 5 is true for the lower quartile) presented in the dataset, relative frequency is not an ideal metric for approximation of probabilistic sentiment scores in many cases as will equal zero in events not observed in the data; such situations would result in bias due to the implication that certain events are impossible, whereas a low probability is more likely. For machine learning, contributes to overfitting as data is sparser. Where non-zero probabilities are assigned to unseen events, some information about their potential occurrence is preserved and thus more robust patterns can be captured. Additionally, where *N* is small averages are increasingly subject to skew. Probability distributions were thus determined using a Laplace estimate:

Where *k* is the cardinality of the class, in this case |c| = 3. The assumption of uniform distribution for the estimator is true for the dataset in question. The smoothing model was selected for its ability to address the issue of zero-probabilities and mitigate the impacts of small *N* where relevant. At larger *N*, the output approaches prior to any transformation. Such a feature of the estimator focuses its effects on cases where it is most necessary. Finally, defining as the mean of the probability distributions weighted against their discrete labels, a sentiment score can be determined:

This approach aims to acknowledge the varying perceptions of emoji sentiment, dependent upon an individuals’ personal usage, while retaining ability to evaluate the sentiment with regards to the consensus classification. An overall sentiment score for sentiment using discrete categorisations often employs a majority decision methodology, however in this context an alternative methodology which acknowledges a degree of subjectivity is preferable. The standard error of the mean is one such methodology which enables a more nuanced classification methodology:

The result of such a methodology enables the identification of a most probable classification and inclusion or exclusion of secondary potential classifications at a 95% confidence.

A diagram of emojis

Description automatically generated

Figure 3 Sample of result of emoji sentiment classification using dimensional emotional theory. Coloured bars are proportional to p+, p0 and p-. Markers (black) represent . Grey markers represent the limits of the 95% confidence intervals for the . Note weary face (left) has high confidence that a negative classifier is appropriate, however ghost (right) cannot necessarily be classified as positive with 95% confidence.

## 5.3 Emoji Feature Extraction

Previous works cite challenges regarding emojis lacking characteristics to enable feature extraction. The following work aims to evaluate several strategies to extract information regarding the sentiment of emoji for the purposes of improving outcomes of sarcasm detection.

## 5.3.1 Use of Emotion Vocabulary Embeddings

Word embeddings are representations of words as fixed-size vectors in multi-dimensional space, where each dimension captures some information about the word such as semantic and pragmatic relationships. Word2Vec is one such method to generate embeddings which has been observed to be particularly effective as capturing relationships between words (Fontaine, 2013). For example, given the male/female relationship is known:

*King – Man + Woman = Queen*

Relationships as above are effectively demonstrated where the results of the above transformation on the ‘King’ vector is very close to that of the ‘Queen’ vector. Given this observation, it is appropriate to use these vectors for comparative purposes, and by extension feature extraction. Emoji2Vec is an equivalent embedding which may supplements word-based corpus with emoji embeddings (Eisner et al., 2016). Authors cite greatest performance of this embedding with the Google News embedding for Twitter sentiment analysis tasks, thus this is the combination of embeddings which will be used. See appendix X for further discussion on the vectors selected.

Cosine similarity is a measure of similarity between two vectors, defined as follows:

Where:

And:

Which in the context of natural language processing provides an interpretable metric appropriate for the use-case it measures directional similarity, without consideration to magnitude and is robust to high dimensionality. As word frequency influences vector magnitude, this strategy eliminates related noise while conserving semantic and pragmatic relationships between words.

A diagram of a straight line

Description automatically generated

Figure 4 Cosine Similarity.

## 5.3.2 Word Vectors of Basic Emotion Words

This strategy aims to compare emojis () to each basic emotion () and defines emotion scores based on the cosine similarity between respective vectors ():

This approach eliminates noise related to distinctions for related vocabulary however this may be too strict where the goal is correlation to human-annotated data given the varied nuance that individuals assign to emotions. To address the strict parameters bounds of this strategy, several alternatives are proposed which employ less stringent boundaries on vocabulary for comparison.

## 5.3.3 Binary Word Association Lexicon

This method draws upon the EmoLex lexicon which assigns binary labels to convey association to basic emotions for each word. Vocabulary tagged for association with a basic emotion (), association to the basic emotion is ranked using cosine similarity and words (, where top- words) with greatest association are determined. Finally, the emotion score is determined as the average of resulting cosine similarities:

The method addressed the nuance related to individuals’ understanding of basic emotions, however words not contained within the EmoLex vocabulary are not considered, potentially disregarding relevant vocabulary. Additionally, this model filters for inclusion based upon cosine similarity to the emoji which fails to account for the degree of association between the basic emotion and the words in the subset. This likely results in negative skew in distribution of results for words uncorrelated to a given emotion.

## 5.3.4 Word-Emotion Intensity Lexicon

This method proposes an alternative approach which addresses the limitations associated with the use of a binary word lexicon, through the utilisation of the NRC-EIL lexicon which tags vocabulary with an intensity score for affinity to a given basic emotion. This model operates on a similar principle, however, selects the top-k words based upon intensity score, increasing the relevance of the considered vocabulary.

## 5.3.5 Results

Table 2 Pearson Correlation scores for all considered prediction methods. Bolded Scores represent the greatest correlation (positive or inverse) observed for the emotion in the respective column, excluding human agreement scores. Average values are reported with respect to absolute correlation.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source | Variant | Anger | Anticip | Disgust | Fear | Joy | Sadness | Surprise | Trust | Average |
|  | ***Basic Emotion Word Similarities*** | | | | | | | | | |
| Word2Vec |  | **0.36** | -0.01 | **0.30** | 0.10 | **0.44** | **0.36** | **0.09** | **0.15** | 0.23 |
|  | ***Emotion Lexicon Corpus Word Similarities*** | | | | | | | | | |
| EmoLex | k=5 | 0.35 | -0.16 | 0.25 | **0.39** | 0.36 | 0.30 | -0.04 | 0.14 | 0.25 |
| EmoLex | k=10 | 0.31 | -0.13 | 0.24 | 0.37 | 0.39 | 0.29 | **-0.09** | **0.15** | 0.25 |
| EmoLex | k=50 | 0.26 | **-0.18** | 0.28 | 0.34 | 0.40 | 0.29 | -0.02 | **0.15** | 0.24 |
| EmoLex | k=100 | 0.26 | -0.17 | **0.30** | 0.30 | 0.41 | 0.31 | 0.01 | 0.14 | 0.24 |
| NRC-EIL | k=10 | 0.26 | -0.16 | 0.13 | 0.20 | 0.30 | 0.32 | 0.05 | -0.01 | 0.18 |
| NRC-EIL | k=100 | 0.19 | -0.12 | 0.24 | 0.24 | 0.33 | 0.33 | -0.01 | 0.01 | 0.18 |
| NRC-EIL | k=300 | 0.18 | -0.13 | 0.23 | 0.25 | 0.31 | 0.33 | 0.02 | -0.02 | 0.18 |
|  | ***Human Annotation*** | | | | | | | | | |
|  |  | 0.67 | 0.74 | 0.65 | 0.59 | 0.78 | 0.69 | 0.64 | 0.72 | 0.69 |

Table \*\*\* compares the mean human-annotated emotion ratings against predicted scores for the proposed methods. Direct comparison to basic emotion word vectors yielded the overall highest performance. Notably the basic emotion was the most significant factor in performance overall, with some emotions achieving some level of correlation and others displaying zero, or weak negative correlation to their predicted values. The impact of context window size was varied across all subsets tested.

These results may be explainable by the nature of the relationships described by the vectors used. For example, *happiness* and *sadness* have a cosine similarity score of 0.42 using Google News Word2Vec model. This is capturing an intuitive relationship between the pair; however, within the context of this methodology, the implication that happiness is a core component of *sadness* is invalid.

## 5.4 Sentiment-Aware Vector Space Modification

This section aims to modify Googles News Word2Vec Embeddings to emphasise characteristics relevant to sentiment analysis while mitigating influence of contradictory semantic relationships. Two proposed strategies which leverage word-pair polarities or correlations between basic emotions to modify vectors were evaluated. The strategies implement a stochastic gradient descent model and Adam optimiser with two possible loss functions:

Where:

The goal of the above function is to maximize the similarity score of synonymous pairs of words ( where ), while minimizing the score for antonymous pairs () by considering their respective word vectors (). Following tuning, obtaining a min value ≈-7500 was achieved which approaches a theoretical minimum ≈-8000. Given the loss function which cannot account for degree of similarity, but rather optimises for polarity, several functions alternative loss values were considered.

The alternative loss function which directly considers Pearsons’ correlation () between basic emotions for relevant datasets:

The goal of this strategy is to optimise vector placement based upon Pearsons Correlation for all permutations () of Plutchiks basic emotions () using the human annotated data contained in the EmoTag1200 and NRL-EIL datasets respectively. Optimisations were weighted with respect to the correlation values to avoid shortcomings highlighted for the alternative strategy and thus loss functions could be minimised rather than considering varying degrees of loss. A depth discussion of the vector optimisation models is described in section 11.2.

## 5.4.1 Evaluation Parameters

The quality of outcomes was quantified by means of considering the Pearson’s correlation between the approximated values for each basic emotion, using methods outlined in section 5.3 against the human-annotated values. Pearson’s correlation measures the strength of association between variables:

Where:

As both total similarity and opposites provide valuable information for the purposes of a regression task, the best outcome is defined as the maximum absolute value for correlation, rather than the greatest correlation. While ideally an approximation method would yield a linear relationship to the human annotated data which can be captured by Pearson’s correlation, it is possible that more complex relationships, which are useful for regression modelling but not evident using this metric may arise. In this case, suitable results could be obtained used this metric only however it may be of value in expansions upon this work to consider alternative relationships between approximations and human-annotations to improve outcomes.

## 5.4.2 Results

Tables \*\*\* summarise the results of all vector space modification tests using previous estimators and drawing correlation to human-annotations for each emoji in the EmoTag1200 dataset. In the case of every basic emotion except for *anger*, modified vectors displayed improved suitability for the purpose of sentiment analysis using the Plutchik model. Results indicate the Emoji2Vec vectors were not limited by contradictory semantic relationships to the same extent as the word vectors, evident by the improved outcomes using modified word vectors alongside the original Emoji2Vec vectors. Given the disproportionate representation of emoji relative to words as a manner to convey sentiment rather than information, this outcome is reasonable. The quality of outcomes across the emotions seems to be related to the strength of ties between the basic emotion and sentiment polarity. For example, *joy* and *sadness* are intuitively linked to *positive* and *negative* sentiments respectively and yielded reasonably favourable results in the analysis. Conversely *surprise* is more ambiguous to classify based on polarity and yielded lower correlation. Such an observation highlights the limitations of the proposed objective functions; they rely on the polarity of the emotions and cannot necessarily acknowledge the core feeling that they represent. While this is a relevant component of the emotions, additional emphasis on the pragmatic component may improve outcomes in future improvements upon the proposed models. Improved correlation was achieved to human-annotated values in the EmoTag1200 dataset which will improve outcomes for the regression model discussed in section X. Additionally and importantly, this result leverages the emoji vector only to obtain results, thus enabling the extrapolation of data to all emojis in the Emoji2Vec vocabulary. If additional parameters were considered, improved outcomes were possible however this would reduce the quantity of emojis that could be considered for the regression task to those which the additional parameters were available.

Table 3 Pearson Correlation scores for all considered prediction methods following vector-space modification- method 1.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source | Variant | Anger | Anticip | Disgust | Fear | Joy | Sadness | Surprise | Trust | Average |
|  |  |
| Word2Vec |  | -0.01 | 0.10 | 0.04 | 0.13 | 0.21 | 0.10 | -0.02 | 0.05 | 0.08 |
|  |  |
| EmoLex | k=5 | -0.10 | 0.10 | 0.00 | 0.05 | 0.16 | 0.06 | -0.01 | -0.01 | 0.06 |
| EmoLex | k=10 | 0.00 | 0.04 | -0.04 | 0.05 | 0.17 | 0.07 | -0.09 | 0.00 | 0.06 |
| EmoLex | k=50 | 0.01 | 0.16 | -0.02 | 0.09 | 0.19 | 0.00 | **-0.10** | 0.04 | 0.08 |
| EmoLex | k=100 | 0.01 | 0.16 | -0.03 | 0.12 | 0.15 | 0.02 | -0.09 | 0.02 | 0.08 |
| NRC-EIL | k=10 | 0.04 | 0.08 | 0.12 | 0.00 | 0.20 | -0.01 | 0.03 | 0.02 | 0.06 |
| NRC-EIL | k=100 | -0.01 | 0.14 | 0.00 | 0.00 | 0.17 | 0.08 | -0.03 | 0.00 | 0.05 |
| NRC-EIL | k=300 | 0.00 | 0.19 | -0.04 | 0.04 | 0.17 | 0.05 | -0.05 | 0.01 | 0.07 |
|  |  |
| Word2Vec |  | 0.01 | 0.07 | -0.06 | -0.04 | 0.04 | 0.10 | 0.00 | -0.06 | 0.05 |
|  |  |
| EmoLex | k=5 | -0.04 | 0.17 | 0.03 | -0.03 | 0.00 | 0.03 | 0.04 | -0.04 | 0.05 |
| EmoLex | k=10 | -0.03 | 0.15 | 0.03 | -0.02 | 0.05 | 0.02 | 0.03 | -0.02 | 0.04 |
| EmoLex | k=50 | -0.05 | 0.15 | 0.00 | -0.03 | 0.02 | 0.01 | 0.02 | 0.03 | 0.04 |
| EmoLex | k=100 | -0.05 | 0.14 | -0.01 | -0.03 | -0.01 | 0.01 | 0.02 | 0.03 | 0.04 |
| NRC-EIL | k=10 | -0.08 | 0.17 | 0.00 | -0.04 | 0.05 | -0.04 | -0.02 | 0.05 | 0.06 |
| NRC-EIL | k=100 | -0.06 | 0.15 | -0.01 | -0.03 | 0.00 | 0.02 | 0.02 | 0.03 | 0.04 |
| NRC-EIL | k=300 | -0.06 | 0.14 | -0.01 | -0.04 | 0.01 | 0.02 | 0.03 | 0.05 | 0.05 |
|  |  |
| Word2Vec |  | 0.04 | -0.23 | 0.11 | 0.06 | **-0.25** | 0.12 | 0.04 | -0.13 | 0.12 |
|  |  |
| EmoLex | k=5 | 0.07 | -0.24 | 0.05 | 0.06 | -0.17 | 0.12 | 0.01 | -0.14 | 0.11 |
| EmoLex | k=10 | 0.06 | -0.23 | 0.04 | 0.08 | -0.04 | 0.12 | 0.00 | -0.15 | 0.09 |
| EmoLex | k=50 | 0.06 | -0.23 | 0.09 | 0.11 | -0.12 | 0.14 | 0.03 | -0.16 | 0.12 |
| EmoLex | k=100 | 0.08 | -0.22 | 0.12 | 0.14 | -0.11 | 0.13 | 0.03 | -0.17 | 0.13 |
| NRC-EIL | k=10 | 0.06 | **-0.25** | 0.16 | 0.16 | -0.10 | 0.07 | -0.01 | -0.15 | 0.12 |
| NRC-EIL | k=100 | 0.10 | -0.21 | 0.16 | 0.15 | -0.16 | 0.10 | 0.03 | -0.17 | 0.14 |
| NRC-EIL | k=300 | 0.11 | -0.21 | 0.17 | 0.16 | -0.15 | 0.09 | 0.03 | **-0.18** | 0.14 |
|  |  |
| Word2Vec |  | **0.16** | 0.08 | 0.10 | 0.19 | 0.15 | 0.02 | 0.01 | 0.16 | 0.11 |
|  |  |
| EmoLex | k=5 | 0.03 | -0.08 | 0.12 | 0.08 | 0.07 | 0.06 | 0.05 | -0.08 | 0.07 |
| EmoLex | k=10 | 0.03 | -0.08 | 0.11 | 0.10 | 0.06 | 0.07 | 0.05 | -0.10 | 0.08 |
| EmoLex | k=50 | 0.01 | -0.09 | 0.09 | 0.10 | 0.06 | 0.07 | 0.06 | -0.14 | 0.08 |
| EmoLex | k=100 | 0.01 | -0.08 | 0.09 | 0.10 | 0.05 | 0.08 | 0.06 | -0.15 | 0.08 |
| NRC-EIL | k=10 | 0.03 | -0.09 | 0.04 | 0.10 | 0.06 | 0.08 | 0.06 | -0.16 | 0.08 |
| NRC-EIL | k=100 | 0.00 | -0.08 | 0.07 | 0.10 | 0.07 | 0.09 | 0.05 | -0.15 | 0.08 |
| NRC-EIL | k=300 | 0.00 | -0.08 | 0.08 | 0.09 | 0.06 | 0.08 | 0.05 | -0.16 | 0.08 |
|  |  |
| Word2Vec |  | **0.16** | 0.08 | 0.10 | **0.19** | 0.15 | 0.02 | 0.01 | 0.16 | 0.11 |
|  |  |
| EmoLex | k=5 | 0.06 | -0.13 | 0.11 | 0.16 | 0.06 | **0.22** | -0.08 | -0.08 | 0.11 |
| EmoLex | k=10 | 0.06 | -0.13 | 0.14 | 0.14 | 0.07 | 0.21 | -0.06 | -0.11 | 0.12 |
| EmoLex | k=50 | 0.04 | -0.14 | **0.18** | 0.13 | 0.09 | 0.20 | -0.02 | -0.15 | 0.12 |
| EmoLex | k=100 | 0.05 | -0.15 | 0.17 | 0.12 | 0.08 | 0.19 | -0.02 | -0.16 | 0.12 |
| NRC-EIL | k=10 | 0.03 | -0.14 | 0.09 | 0.15 | 0.13 | 0.20 | -0.01 | -0.11 | 0.11 |
| NRC-EIL | k=100 | 0.03 | -0.15 | 0.14 | 0.12 | 0.09 | 0.21 | -0.05 | -0.16 | 0.12 |
| NRC-EIL | k=300 | 0.04 | -0.15 | 0.14 | 0.12 | 0.08 | 0.21 | -0.04 | -0.17 | 0.12 |
|  |  |
|  |  | 0.67 | 0.74 | 0.65 | 0.59 | 0.78 | 0.69 | 0.64 | 0.72 | 0.69 |

Table 4 Pearson Correlation scores for all considered prediction methods following vector-space modification- method 1. Original emoji vectors and transformed word vectors used.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source | Variant | Anger | Anticip | Disgust | Fear | Joy | Sadness | Surprise | Trust | Average |
|  | ***Basic Emotion Word Similarities (Loss ≈-7500)*** | | | | | | | | | |
| Word2Vec |  | 0.13 | 0.00 | 0.31 | 0.10 | 0.41 | 0.22 | 0.11 | 0.01 | 0.16 |
|  | ***Emotion Lexicon Corpus Word Similarities (Loss ≈-7500)*** | | | | | | | | | |
| EmoLex | k=5 | 0.29 | -0.03 | 0.23 | 0.35 | 0.41 | 0.20 | -0.15 | -0.14 | 0.23 |
| EmoLex | k=10 | 0.28 | -0.03 | **0.34** | 0.37 | **0.50** | 0.27 | -0.11 | 0.04 | 0.24 |
| EmoLex | k=50 | 0.31 | -0.15 | 0.32 | 0.32 | **0.50** | 0.33 | -0.07 | 0.20 | 0.28 |
| EmoLex | k=100 | 0.32 | -0.17 | **0.34** | 0.36 | 0.44 | 0.36 | 0.09 | 0.11 | 0.27 |
| NRC-EIL | k=10 | 0.28 | -0.18 | 0.32 | 0.29 | 0.35 | 0.29 | **0.20** | -0.10 | 0.25 |
| NRC-EIL | k=100 | **0.35** | -0.04 | 0.28 | 0.39 | 0.43 | 0.40 | -0.07 | 0.17 | 0.27 |
| NRC-EIL | k=300 | 0.33 | -0.03 | 0.30 | **0.42** | 0.48 | **0.44** | 0.03 | 0.20 | 0.28 |
|  | ***Basic Emotion Word Similarities (Loss ≈-5000)*** | | | | | | | | | |
| Word2Vec |  | 0.30 | 0.07 | 0.02 | 0.36 | 0.16 | 0.03 | -0.08 | -0.01 | 0.13 |
|  | ***Emotion Lexicon Corpus Word Similarities (Loss ≈-5000)*** | | | | | | | | | |
| EmoLex | k=5 | 0.12 | -0.18 | 0.19 | 0.28 | 0.30 | 0.24 | -0.11 | -0.09 | 0.19 |
| EmoLex | k=10 | 0.17 | -0.17 | 0.23 | 0.26 | 0.27 | 0.27 | -0.07 | -0.07 | 0.19 |
| EmoLex | k=50 | 0.11 | -0.17 | 0.21 | 0.17 | 0.20 | 0.24 | 0.00 | -0.13 | 0.15 |
| EmoLex | k=100 | 0.10 | -0.18 | 0.21 | 0.15 | 0.20 | 0.21 | 0.04 | -0.12 | 0.15 |
| NRC-EIL | k=10 | 0.09 | -0.12 | 0.15 | 0.11 | 0.30 | 0.16 | 0.03 | -0.06 | 0.13 |
| NRC-EIL | k=100 | 0.09 | -0.14 | 0.18 | 0.16 | 0.24 | 0.20 | 0.02 | -0.09 | 0.14 |
| NRC-EIL | k=300 | 0.07 | -0.14 | 0.18 | 0.15 | 0.21 | 0.21 | 0.03 | -0.10 | 0.14 |
|  | ***Basic Emotion Word Similarities (Loss ≈-3000)*** | | | | | | | | | |
| Word2Vec |  | -0.17 | -0.10 | -0.02 | -0.11 | 0.04 | -0.09 | -0.16 | -0.17 | 0.11 |
|  | ***Emotion Lexicon Corpus Word Similarities (Loss ≈-3000)*** | | | | | | | | | |
| EmoLex | k=5 | -0.12 | -0.08 | 0.09 | -0.03 | 0.11 | -0.09 | 0.06 | -0.21 | 0.10 |
| EmoLex | k=10 | -0.09 | -0.08 | 0.09 | -0.03 | 0.09 | -0.06 | 0.04 | -0.19 | 0.08 |
| EmoLex | k=50 | -0.10 | -0.12 | 0.06 | 0.00 | 0.08 | -0.04 | 0.03 | -0.19 | 0.08 |
| EmoLex | k=100 | -0.11 | -0.12 | 0.04 | 0.00 | 0.09 | -0.03 | 0.04 | -0.16 | 0.07 |
| NRC-EIL | k=10 | -0.18 | -0.15 | 0.01 | -0.02 | 0.19 | -0.07 | 0.05 | -0.13 | 0.10 |
| NRC-EIL | k=100 | -0.13 | -0.13 | 0.05 | -0.03 | 0.09 | -0.04 | 0.06 | -0.17 | 0.09 |
| NRC-EIL | k=300 | -0.11 | -0.12 | 0.06 | -0.03 | 0.09 | -0.04 | 0.06 | -0.17 | 0.09 |
|  | ***Basic Emotion Word Similarities (Loss ≈-1000)*** | | | | | | | | | |
| Word2Vec |  | 0.31 | -0.04 | 0.21 | 0.29 | 0.19 | 0.09 | -0.03 | 0.11 | 0.16 |
|  | ***Emotion Lexicon Corpus Word Similarities (Loss ≈-1000)*** | | | | | | | | | |
| EmoLex | k=5 | 0.09 | -0.15 | 0.17 | 0.20 | 0.13 | 0.14 | -0.03 | -0.20 | 0.14 |
| EmoLex | k=10 | 0.09 | -0.15 | 0.17 | 0.18 | 0.14 | 0.15 | -0.02 | -0.17 | 0.13 |
| EmoLex | k=50 | 0.06 | -0.17 | 0.17 | 0.12 | 0.14 | 0.15 | 0.01 | -0.16 | 0.12 |
| EmoLex | k=100 | 0.06 | -0.16 | 0.18 | 0.11 | 0.14 | 0.15 | 0.03 | -0.17 | 0.13 |
| NRC-EIL | k=10 | 0.10 | -0.16 | 0.07 | 0.08 | 0.17 | 0.16 | 0.04 | -0.13 | 0.11 |
| NRC-EIL | k=100 | 0.04 | -0.15 | 0.15 | 0.12 | 0.16 | 0.16 | 0.02 | -0.15 | 0.12 |
| NRC-EIL | k=300 | 0.04 | -0.15 | 0.15 | 0.11 | 0.15 | 0.15 | 0.02 | -0.16 | 0.12 |
|  | ***Basic Emotion Word Similarities (Loss ≈-500)*** | | | | | | | | | |
| Word2Vec |  | 0.30 | 0.03 | 0.28 | 0.28 | 0.39 | 0.11 | 0.05 | **0.22** | 0.21 |
|  | ***Emotion Lexicon Corpus Word Similarities (Loss ≈-500)*** | | | | | | | | | |
| EmoLex | k=5 | 0.09 | **-0.20** | 0.18 | 0.24 | 0.19 | 0.19 | -0.08 | -0.16 | 0.17 |
| EmoLex | k=10 | 0.11 | -0.19 | 0.25 | 0.21 | 0.18 | 0.22 | -0.06 | -0.14 | 0.17 |
| EmoLex | k=50 | 0.08 | -0.18 | 0.22 | 0.16 | 0.17 | 0.19 | 0.00 | -0.13 | 0.14 |
| EmoLex | k=100 | 0.08 | -0.17 | 0.23 | 0.14 | 0.17 | 0.19 | 0.02 | -0.14 | 0.14 |
| NRC-EIL | k=10 | 0.06 | -0.16 | 0.10 | 0.18 | 0.25 | 0.21 | 0.04 | -0.08 | 0.14 |
| NRC-EIL | k=100 | 0.07 | -0.15 | 0.17 | 0.14 | 0.21 | 0.21 | 0.00 | -0.11 | 0.13 |
| NRC-EIL | k=300 | 0.07 | -0.15 | 0.16 | 0.14 | 0.18 | 0.21 | 0.02 | -0.13 | 0.13 |
|  | ***Human Annotation*** | | | | | | | | | |
|  |  | 0.67 | 0.74 | 0.65 | 0.59 | 0.78 | 0.69 | 0.64 | 0.72 | 0.69 |

Table 5 Pearson Correlation scores for all considered prediction methods following vector-space modification- method 2.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Source | Variant | Anger | Anticip | Disgust | Fear | Joy | Sadness | Surprise | Trust | Average |
|  | ***Basic Emotion Word Similarities (Both Modified Vectors)*** | | | | | | | | | |
| Word2Vec |  | 0.19 | -0.13 | 0.21 | 0.16 | 0.20 | 0.11 | -0.08 | -0.12 | 0.15 |
|  | ***Emotion Lexicon Corpus Word Similarities (Both Modified Vectors)*** | | | | | | | | | |
| EmoLex | k=5 | 0.10 | -0.13 | 0.08 | 0.19 | 0.16 | 0.14 | -0.07 | -0.07 | 0.12 |
| EmoLex | k=10 | 0.09 | -0.11 | 0.10 | 0.16 | 0.16 | 0.16 | 0.02 | -0.04 | 0.11 |
| EmoLex | k=50 | 0.06 | -0.12 | 0.11 | 0.10 | 0.17 | 0.13 | 0.02 | -0.03 | 0.09 |
| EmoLex | k=100 | 0.06 | 0.11 | 0.10 | 0.09 | 0.16 | 0.13 | 0.03 | -0.04 | 0.09 |
| NRC-EIL | k=10 | 0.01 | -0.09 | 0.07 | 0.05 | 0.21 | 0.09 | -0.02 | 0.00 | 0.07 |
| NRC-EIL | k=100 | 0.02 | -0.07 | 0.11 | 0.08 | 0.16 | 0.11 | 0.01 | 0.02 | 0.07 |
| NRC-EIL | k=300 | 0.01 | -0.08 | 0.09 | 0.08 | 0.15 | 0.12 | 0.03 | 0.00 | 0.07 |
|  | ***Basic Emotion Word Similarities (Modified Word Vectors Only)*** | | | | | | | | | |
| Word2Vec |  | **0.36** | -0.11 | **0.38** | 0.41 | **0.44** | 0.18 | -0.11 | **-0.24** | 0.28 |
|  | ***Emotion Lexicon Corpus Word Similarities (Modified Word Vectors Only)*** | | | | | | | | | |
| EmoLex | k=5 | 0.24 | -0.17 | 0.18 | **0.42** | 0.34 | 0.26 | **-0.16** | -0.11 | 0.24 |
| EmoLex | k=10 | 0.22 | -0.16 | 0.25 | 0.33 | 0.30 | 0.28 | -0.06 | -0.05 | 0.21 |
| EmoLex | k=50 | 0.19 | **-0.20** | 0.27 | 0.25 | 0.31 | 0.27 | -0.02 | -0.09 | 0.20 |
| EmoLex | k=100 | 0.20 | -0.19 | 0.29 | 0.24 | 0.30 | 0.28 | 0.00 | -0.12 | 0.20 |
| NRC-EIL | k=10 | 0.12 | -0.14 | 0.14 | 0.23 | 0.40 | 0.16 | -0.04 | -0.04 | 0.16 |
| NRC-EIL | k=100 | 0.14 | -0.11 | 0.22 | 0.22 | 0.31 | 0.27 | -0.01 | -0.06 | 0.17 |
| NRC-EIL | k=300 | 0.15 | -0.13 | 0.21 | 0.22 | 0.27 | **0.29** | 0.03 | -0.08 | 0.17 |
|  | ***Human Annotation*** | | | | | | | | | |
|  |  | 0.67 | 0.74 | 0.65 | 0.59 | 0.78 | 0.69 | 0.64 | 0.72 | 0.69 |

## 5.5 Emoji Sentiment Prediction – Basic Theory

## 5.5.1 Data Preparation

Prior to the implementation of any models, preparation and evaluation of the data is essential to make an informed decision with regards to the model selected for the task in addition to optimising model performance. Data preparation and an assessment of the features of the task for this purpose have been detailed in section 11.4.

## 5.5.2 Model Selection

Model selection was carried out using the following approach:

A diagram of a company

Description automatically generated

Figure 5 Process to select model for basic theory regression task.

The goal of the approach was to evaluate each considered model in its most robust, optimised configuration to ensure the most accurate model could be identified. Each step in this process is discussed in section 11.4.

## 5.5.3 Model Evaluation

Table \*\*\* reports the results of model evaluation for each basic emotion. The goal was to identify the model which yielded an outcome which minimised mean absolute error and mean squared error, while maximising R2 score. Where the optimal metrics for a basic emotion were distributed across multiple models the determination of the optimal model considered relative trade-offs between the metrics in each case. For example, where a model had a slightly lower mean absolute error than its alternative, but a much higher R2 score, a cost-benefit analysis would favour this model. Metrics for training, validation and test were all considered in the context of one another, where close alignment was viewed with preference, as this signals good generalisation.

As the models have varying means for their respective values it was important to consider the mean when assessing model performance in each case. Table \*\*\* uses the mean values to provide an indication of the accuracy of the models. While this cannot consider the distribution of error across the set, the low mean squared errors achieved in each optimal model indicates that error is broadly distributed in proportion to the value of the label. Most models had moderate abilities to explain the underlying patterns in the training data, evidence by their respective R2 scores. However, this was not the case for *anticipation* and *surprise*, which had poor performance. This consolidates previous hypothesis relating to the limitations of the proposed objective functions to modify the word vectors as these emotions are distinct from the others in that they can apply in both positive and negative contexts and are distinguished by additional features.

Accuracy scores averaging 90% were achieved for the basic emotion regression (range 86-92%). A likely limitation of the performance was the limited data for training. An expansion of the annotated data may improve the outcomes of these models or enable the use of models capable of capturing more complex patterns in future works. Supplementary features gleaned from other sources may have improved outcomes for this task, however the use of word vectors only to for feature extraction makes expansion upon this dataset more accessible as the emoji corpus expands, which is a primary strength of this approach.

Table 6 Results of model development process for basic theory regression displaying the best outcomes for each model, correct to three significant figures. The selected optimal model is in bold for each basic emotion. Note that reported metrics are with reference to validation performance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Performance Metric** | | |
| **Mean Absolute Error** | **Mean Squared Error** | **R2 Score** |
|  | **Anger** | | |
| Linear Regression | 0.0112 | 0.000230 | 0.148 |
| Random Forest | 0.0136 | 0.000290 | -1.41 |
| XGBoost Regression | 0.0127 | 0.000283 | -1.35 |
| Support Vector Regression | 0.0318 | 0.00112 | -2.63 |
| **Gaussian Process Regression** | 0.00915 | 0.000171 | 0.365 |
| KNN Regression | 0.0110 | 0.000292 | -0.0858 |
|  | **Anticipation** | | |
| Linear Regression | 0.0118 | 0.000217 | -0.238 |
| Random Forest | 0.0122 | 0.000201 | -0.143 |
| XGBoost Regression | 0.0142 | 0.000251 | -0.432 |
| Support Vector Regression | 0.0136 | 0.000241 | -0.373 |
| Gaussian Process Regression | 0.0117 | 0.000185 | -0.0557 |
| **KNN Regression** | 0.0114 | 0.000189 | -0.0735 |
|  | **Disgust** | | |
| **Linear Regression** | 0.0121 | 0.000270 | 0.310 |
| Random Forest | 0.0145 | 0.000419 | -0.0697 |
| XGBoost Regression | 0.0147 | 0.000456 | -0.164 |
| Support Vector Regression | 0.0206 | 0.000543 | -0.385 |
| Gaussian Process Regression | 0.0126 | 0.000273 | 0.303 |
| KNN Regression | 0.0159 | 0.000419 | -0.0691 |
|  | **Fear** | | |
| Linear Regression | 0.0119 | 0.000319 | 0.0607 |
| Random Forest | 0.0162 | 0.000351 | 0.0578 |
| XGBoost Regression | 0.0183 | 0.000477 | -0.282 |
| Support Vector Regression | 0.0263 | 0.000840 | -1.25 |
| **Gaussian Process Regression** | 0.0113 | 0.000170 | 0.543 |
| KNN Regression | 0.0158 | 0.000397 | -0.0670 |
|  | **Joy** | | |
| Linear Regression | 0.0236 | 0.00114 | -0.0148 |
| Random Forest | 0.0313 | 0.00147 | -0.312 |
| XGBoost Regression | 0.0350 | 0.00186 | -0.663 |
| Support Vector Regression | 0.0295 | 0.00115 | -0.0254 |
| **Gaussian Process Regression** | 0.0199 | 0.000614 | 0.451 |
| KNN Regression | 0.0287 | 0.00113 | -0.00802 |
|  | **Sadness** | | |
| **Linear Regression** | 0.0152 | 0.000381 | 0.417 |
| Random Forest | 0.0166 | 0.000669 | -0.0241 |
| XGBoost Regression | 0.0159 | 0.000462 | 0.292 |
| Support Vector Regression | 0.0303 | 0.00106 | -0.619 |
| Gaussian Process Regression | 0.0144 | 0.000408 | 0.375 |
| KNN Regression | 0.0189 | 0.000721 | -0.104 |
|  | **Surprise** | | |
| **Linear Regression** | 0.0103 | 0.000156 | -0.0299 |
| Random Forest | 0.0106 | 0.000178 | -0.0484 |
| XGBoost Regression | 0.0131 | 0.000246 | -1.05 |
| Support Vector Regression | 0.0141 | 0.000270 | -1.25 |
| Gaussian Process Regression | 0.00973 | 0.000141 | -0.173 |
| KNN Regression | 0.0111 | 0.000199 | -0.659 |
|  | **Trust** | | |
| **Linear Regression** | 0.0128 | 0.000303 | 0.202 |
| Random Forest | 0.0169 | 0.000495 | -0.300 |
| XGBoost Regression | 0.0173 | 0.000530 | -0.393 |
| Support Vector Regression | 0.0176 | 0.000410 | -0.0776 |
| Gaussian Process Regression | 0.153 | 0.000519 | -0.411 |
| KNN Regression | 0.0171 | 0.000484 | -0.272 |

Table 7 Assessment of model accuracy considering the mean values of each basic emotion.

|  |  |  |  |
| --- | --- | --- | --- |
| **Basic Emotion** | **Mean Value for Affinity to Emoji** | **Mean Absolute Error** | **% Accuracy Approximation** |
| Anger | 0.122 | 0.00915 | 92.50 |
| Anticipation | 0.137 | 0.0114 | 91.68 |
| Disgust | 0.102 | 0.0121 | 88.14 |
| Fear | 0.120 | 0.0113 | 90.58 |
| Joy | 0.138 | 0.0199 | 85.58 |
| Sadness | 0.118 | 0.0152 | 87.12 |
| Surprise | 0.122 | 0.0103 | 92.20 |
| Trust | 0.131 | 0.0128 | 90.23 |

## 5.6 Emoji Sentiment Prediction – Dimensional Theory

The nature of dimensional theory-based sentiment metrics does not facilitate solutions similar to that which was implemented for basic theory parameters as word vectors associated with *positive* and *negative* vocabulary would necessitate the consideration of context and topic to a much greater extent making this a more complex and nuanced problem set. No human-annotated datasets were identified which provide equivalent data to the EmoLex of NRL-EIC lexicons thus the proposed adaptation of the previous methodology (a) compared emoji vectors directly to the vectors of the respective labels, and (b) compared label vectors to the vocabulary of the iSarcasm dataset. Such vocabulary was selected as it represents the vernacular of the population which may use sarcasm in online content thus the words highly associated with the emojis based on their vectors were likely to be relevant choices.

Results obtained using both options performed poorly with respect to generating predictions which correlated to the human-annotated data, and consequently was not useful in predicting information for emojis unseen in the dataset. This approach was unable to obtain accuracies >50% and thus alternative strategies were explored.

Given the limitations in data quantities available to train vectors and the subsequent regression models, pre-trained models were postulated to generate improved outcomes as they benefit from increased volume of training data and computational power available during their training process. Such factors yield a model capable of capturing more complex relationships thus results which mirror human annotations more closely. Several pre-trained models were assessed with regards to their abilities to predict dimensional theory data which correlated to human annotations assessed via Pearson’s correlation. Models selected for assessment based on suitability of architecture and use in similar problem sets within literature (detailed in section 11.5).

## 5.6.1 Results and Methodology Evaluation

The optimal model identified was the latest Twitter RoBERTa Base Sentiment which achieved a Pearson’s correlation of 0.83 to the human annotated sentiment scores. Fine-tuning the model using the human-annotated data did not result in improvement to this score, likely due to the small quantity of data available for the purpose therefore the model was deployed for the regression task without optimisation.

While the results obtained yielded good correlation to the human annotated data, the input of a single emoji in each case to obtain the results does not necessarily align with the intended use for the model; the attention mechanisms cannot provide contextual information from surrounding information and thus outputs may be limited to some degree by ambiguity. Alternative methodologies would likely involve the input of emoji-containing strings of text, however in the context of the problem set this also has limitations; the sentiment may be skewed by the content of the text. This concept also applies to some extent even where the emoji is inputted in isolation due to the training data. While the approach has limitations, the solution would involve the obtaining of a larger training dataset with annotations. Such a task is outside the scope of this work however may be an avenue for future consideration.

## 5.7 Chapter Conclusion

The aim of this chapter was to establish viable methods for feature extraction of emojis, with focus on extraction of sentiment markers. Intuitive and interpretable definitions for features were proposed which enabled the comparison of human-annotated labels to predictions generated during feature extraction. Previous literature highlighted limitations in feature extraction as a primary constraint upon the use of emoji for natural language processing tasks. Sentiment focused feature extraction was performed solely using the emoji vectors, as this ensures the strategy can be implemented for any emojis which are released in the future. The vectors were additionally trained on information from the Unicode Consortium, which is released alongside every new emoji further consolidating this aspect of the strategy. These vectors were modified to improve their sentiment awareness which achieved improved outcomes in the proposed pipelines for prediction of emoji sentiment metrics for basic emotions. Areas for future improvements in this pipeline largely surround the lack of annotated data to be used in training which limited the complexity of models viable for use and by extension the complexity of patterns which could be learned. The objective functions proposed during the vector space modification additionally may be improved by adding to the information they capture; strategies which consider aspects unrelated to polarity of sentiment or relative association have significant potential to improve outcomes of this work. The work carried out in this chapter has established sentiment-based feature parameters for approximately 1660 emojis available presently from the Unicode Consortium, significantly expanding upon the data available from human-annotated datasets available presently with comparable data. The pipeline is viable for future expansions upon this work as is does not rely on any parameters external to the vectors for the emoji.

# 6 Pragmatics of Sarcasm

## 6.1 Evaluation of Annotation Strategies

The literature review conducted highlighted validity concerns with regards to annotation and sampling. Sarcastic tweets were primarily sourced across the surveyed literature through identification of key hashtags such as ‘#sarcasm’. This strategy was shown to result in an unrepresentative sample of sarcastic tweets. This is likely to reduce performance on sarcasm detection models and thus true performance of such models is difficult to assess. One study found that only 15% of tweets labelled as sarcastic using this methodology were true labels- highlighting significant shortcomings to this strategy.

Datasets with a more robust sampling methodology exist, however there is little data available in these datasets which contain emojis. This works proposes a survey which aims create a more representative sample of sarcastic online content containing emojis. Previous work has been conducted to improve upon this poor annotation strategy however the majority of the tweets collected do not contain emojis; the iSarcasm dataset collected self-reported sarcastic and non-sarcastic tweets from participants, alongside a rephrase of the tweet in a more literal style. The work was subject to quality control by a linguistics professional therefore it is likely that the results are more representative of organic sarcastic content than the previously discussed sampling strategy.

## 6.2 Primary Research Methodology

This primary research aims to adapt this dataset to evaluate emoji use patterns in sarcastic and non-sarcastic tweets. The adaptation will consist of the addition of emojis to known sarcastic and non-sarcastic text by survey participants. The methodology will collect quantitative data regarding demographics of the sampled population in addition to quantitative data regarding emoji use in sarcastic and control content. The components of the survey including sampling strategies, question selection and format and design optimisation are discussed in section 11.6.

The goal of this work is to generate a dataset of verified sarcastic and control data which is richer in emojis. There are some limitations to this strategy; these are tweets where the author did not originally use emojis. While the use of emojis is known to be systematic in nature, it is a reasonable assumption that two people create a tweet with matching sentiment and pragmatic intent, where one uses emojis and the other does not. A more optimal approach would ask participants to self-report a sample of their tweets containing emojis as sarcastic and non-sarcastic to generate a dataset rather than providing prompts due to the possibility of irrelevant text to the participant reducing the quantity of usable results for each participant. The issue of relevance is addressed by enabling participants to not assign classifications to text which is not relevant or understood by them. The alternative of submission of sample text containing emojis with classifications of sarcastic and non-sarcastic would likely yield more relevant results to the participant however this approach significantly increases the effort required from participants which would decrease response rate and possibly reduce the likelihood of the task being completed as instructed. Raw results for the survey can be found in section \*\*\*.

## 6.3 Survey Outcomes

## 6.3.1 Demographics

The survey yielded 87 responses which can be deconstructed as follows:

A screenshot of a video game

Description automatically generated

Figure 6 Survey responses deconstructed by demographics.

Responses were not distributed evenly across subsets of the population, with gender skewing towards female and younger individuals. The shape of the age-related data is logical given that emojis are used disproportionately by younger people. No individuals were identified for participation over 65 years of age, which is likely attributable to the same observation. The age distribution of the survey does however seem to align reasonably well with the distribution of Twitter users. Given the age bins are not aligned it is plausible to conclude that the survey respondents follow a very similar age distribution to organic Twitter users, a result which may be desirable in this context. While a positive result has been observed for age distribution, gender alignment to Twitters user-base is not closely aligned with a global gender distribution skewing 70.4% towards male users. English speaking countries with reportable data cite a greater proportion of female users (averaging 41.5%). Given the contradictory skew in gender representation within the survey data, evaluation must be carried out to identify any differences which correlate to gender.

A blue graph with black border

Description automatically generated

Figure 7 Age distribution of Twitter users.

## 6.3.2 Annotations

Given the 1:1 split of speculative labels for each survey, the distribution of sarcastic and non-sarcastic assignments is encouraging for the overall quality of the data returned by the participant population due to one noteworthy implication; although participants were explicitly informed of the surveys focus on their assessment of sarcasm within the text prompts, this does not appear to have impacted the classifications. This point addresses a consideration with regards to survey validity; the comprehension of the underlying nature of a study could conceivably introduce variance in the subjects’ responses. Hence, the potential for such an influence warrants consideration in the evaluation of the implications of any conclusions drawn.

The distribution of data within figure \*\*\* additionally lends credibility to the data with respect to its incongruent skew in gender distribution compared to the Twitter user-base; this parameter is uninfluenced by gender where usable classifications are applied. Both gender and age were found to have minimal impact on the rates of sarcasm reported in responses.

A blue and orange bars

Description automatically generated A colorful graph with black border

Description automatically generated

Figure 8 Classification results breakdown by gender and age.

A screenshot of a graph

Description automatically generated

Figure 9 Distribution of labels assignments for each survey response. Where ‘I don’t know’ was selected 15.84% of the time (approximately 1-2 times per survey). Considering only sarcastic and non-sarcastic labels and controlling for unassigned text 68% of respondents answered within 2 responses of a 1:1 ratio, aligning closely with speculative labels.

Inter-annotator patterns may be evaluated through the consideration of the quality control questions. Where text prompt is controlled, comparative work between participants is more intuitive. Agreement was not achieved in many cases, which is expected where context or the individuals’ personal beliefs are of greater relevance to the perception of the text prompt. Consider the following prompt where the beliefs of the participant were likely to have an impact on their response:

*Vaccine dose 1. Thank you, science.*

Responses labelled sarcastic:

*Vaccine dose 1. Thank you, science.* 💀

*Vaccine dose 1. Thank you, science.* 😜

Responses labelled non-sarcastic:

*Vaccine dose 1. Thank you, science.*👍

*Vaccine dose 1. Thank you, science.*👍

*Vaccine dose 1. Thank you, science.* 💛

However, in cases where the text prompt does not represent content that has ties to personal beliefs agreement between annotators was high:

Text prompt:

*was not aware that Crocs were appropriate business casual attire.*

Responses were universally labelled sarcastic:

*was not aware that Crocs were appropriate business casual attire.*

*was not aware that Crocs were appropriate business casual attire.* 😒

*was not aware that Crocs were appropriate business casual attire.* 😅

*was not aware that Crocs were appropriate business casual attire.*😂

*was not aware that Crocs were appropriate business casual attire.* 🙄

*was not aware that Crocs were appropriate business casual attire.* 🙄

*was not aware that Crocs were appropriate business casual attire.*🤔

Such an observation indicates that successful sarcasm detection models must implement highly sophisticated models which can weigh the degree to which topics are tied to beliefs which are polarised across the population. This adds additional complexity to the necessary context awareness, making this a challenging problem set to overcome.

## 6.4 Statistical Analysis of Pragmatics in Sarcastic Content

## 6.4.1 Structural and Sentiment Features

The following work aims to establish which structural, or sentiment parameters have statistical significance when identifying sarcasm in short form text prompts. Section 11.7 discusses the selection methodology for statistical tests selected in each case. Results indicate that emojis were more frequently added to tweets considered to be sarcastic by the annotator, indicating that consideration of emojis is of particular benefit in the detection of sarcasm. The emojis in sarcastic and non-sarcastic content did overlap to some extent, however the context which they are used displayed incongruence. Sarcastic content evaluated largely used emojis with literal implications of positive affect alongside negative events conveyed within the text. This result consolidates observations throughout literature that emojis are used to clarify tone, where the intended sentiment by the author is incongruent with that of the literal text meaning.

Table 8 Top 10 Most Used Emojis in Survey Results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sarcastic Tweets** | | **Non-Sarcastic Tweets** | |
| **Emoji** | **Occurrences** | **Emoji** | **Occurrences** |
| 😂 | 42 | 😂 | 17 |
| 😅 | 19 | 😭 | 17 |
| 😒 | 18 | 😡 | 13 |
| 😏 | 17 | 👍 | 12 |
| 🙈 | 13 | 😁 | 11 |
| 😭 | 13 | 😬 | 10 |
| 😡 | 10 | 😒 | 9 |
| 😉 | 10 | 😊 | 8 |
| 💀 | 10 | 💛 | 7 |
| 😁 | 9 | 😍 | 6 |

Sarcastic tweets containing 😂 emoji:

*Love living in a capitalist society where I look forward to getting a surgery where I’m legit GETTING AN INTERNAL BODY PART REMOVED bc it means I’ll get a few days off from work xoxo* 😂

*The only thing I got from college is a caffeine addiction* 😂

*"2 things I love: 1. Being woken up by construction work an hour before my alarm. 2. Sarcasm."*😂

*Gonna go cry now some no face told me to lose weight x* 😂

Non-Sarcastic tweets containing 😂 emoji:

*I never thought I'd say this, but I have become one of those people who like bounty bars.* 😅😂

*Quarantine Day 256: Dear Ancient Greeks, re: the blanket-wearing I get it. #QuarantineLife* 😂

*@sadgirlkali Make your husband agree to let you die first. FUN date topic for discussion*😂

*"If I could have changed anything about my childhood, I would have never watched SpongeBob -no one"*😂

Evaluation of sentiment and structural parameters relating directly to emojis and the text prompt itself was carried out to determine parameters which may be of importance for sarcasm detection. Such parameters were defined as any parameters which had a statistically significant difference between sarcastic and non-sarcastic subsets within the survey results. The emojis in sarcastic content were found to have greater neutrality than those in the non-sarcastic subset, and the skew towards the use of positive sentiment was found to be greater. These results may be indicative of the use of emojis to subdue perceived sentiment.

Several structural markers were additionally evaluated, based on postulation that they may be indicative of sarcasm. Interjections were found to occur at greater frequencies in sarcastic content. User mentions were additionally found to be more frequent in non-sarcastic content. The capitalisation style of content was found to differ across the subsets with mid-word capitalisation and capitalised words occurring more frequently in non-sarcastic content. These observations lend support to evidence that users who generate sarcastic content do so in a more colloquial idiom, with less regard for legibility of the content; by extension the rigidity of conformance to punctuation and grammar may be an additional parameter for exploration as a distinguishing factor between the subsets.

Each parameter was assessed via its distribution across the entire sample set or subset initially. This methodology may be limited by destructive interference between overall positive and negative sentiment content. To evaluate this further, additional statistical evaluation was carried out, where text prompts were divided into subset based on their polarities. These tests confirmed hypothesis that opposing polarities were adding noise to the reported results, obtained from the overall dataset, and highlighted additional features of importance for later sarcasm detection work. This observation displays differing distinctions between sarcastic and non-sarcastic content, where polarity is controlled. This indicates that successful identification of sarcasm must be carried out with respect to the sentiment of the text.

## 6.4.2 Context-Based Features

Given the consensus across literature of the importance of context in identification of sarcastic content, evaluation to determine potential markers must consider this. Topic modelling was carried out to identify relevant topics for the overall dataset and compared to that of the sarcastic subset. The topic modelling process is documented in section 11.9.

Table 9 Topics Identified using the Optimal Model.

|  |  |
| --- | --- |
| **Overall Dataset Topic Interpretation** | **Sarcastic Subset Topic Interpretation** |
| Expression of Preference | Leisure Activities |
| Reflection or Contemplation of Past Events | Opinions of Others (In the Media) |
| Personal Experiences OR Expectations | Appearance and Clothing |
| Routine Life | Opinions and Thoughts |

The topics gleaned from the respective training sets display distinctions, where those relating to the overall dataset seem to be introspective and the converse can be observed in the sarcastic topics. This aligns with what is known intuitively about sarcasm; when published online, it is often intended to glean greater attention from the reader compared to literal text conveying the same message.

The relative affinities of each topic learned from the whole dataset was compared for sarcastic and non-sarcastic text prompts. Results indicate that average affinities are lesser in topics 3 and 4, which are the most introspective topics. It should be noted that such an observation is largely subjective and while it may provide indications to aid in later sentiment analysis, further results of a more quantitative nature would be necessary to make conclusions in this regard.

Table 10 Statistical Evaluation of Emoji-Based Markers of Sarcasm.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sarcastic Content** | | | | | **Non-Sarcastic Content** | | | | | **Significant difference between Sarcastic and Non-Sarcastic Content** | | |
| **Overall** | **Positive**  **Subset** | | **Negative**  **Subset** | | **Overall** | **Positive**  **Subset** | | **Negative**  **Subset** | | **Overall** | **Positive**  **Subset** | **Negative**  **Subset** |
| **Frequency of Emoji Use** | | | | | | | | | | | | |
| 0.980 | | 0.979 | | 0.980 | 0.607 | | 0.602 | | 0.611 | Yes | Yes | Yes |
| **Position of Emojis in Text** | | | | | | | | | | | | |
| 0.947 | | 0.943 | | 0.954 | 0.972 | | 0.961 | | 0.980 | No | No | No |
| **Sentiment Score of Emojis Used** | | | | | | | | | | | | |
| 0.101 | | 0.312 | | -0.0430 | 0.130 | | 0.496 | | -0.181 | No | Yes | Yes |
| **Degree of Positivity of Emojis Used** | | | | | | | | | | | | |
| 0.337 | | 0.453 | | 0.258 | 0.372 | | 0.572 | | 0.203 | No | Yes | Yes |
| **Degree of Negativity of Emojis Used** | | | | | | | | | | | | |
| 0.236 | | 0.141 | | 0.301 | 0.243 | | 0.0754 | | 0.384 | No | Yes | Yes |
| **Degree of Neutrality of Emojis Used** | | | | | | | | | | | | |
| 0.406 | | 0.406 | | 0.353 | 0.385 | | 0.353 | | 0.413 | Yes | Yes | Yes |
| **Degree of Anger Expressed through Emoji** | | | | | | | | | | | | |
| 0.0699 | | 0.0709 | | 0.0692 | 0.0711 | | 0.0713 | | 0.0709 | No | No | No |
| **Degree of Anticipation Expressed through Emoji** | | | | | | | | | | | | |
| 0.0788 | | 0.0801 | | 0.0779 | 0.0802 | | 0.0799 | | 0.0804 | No | No | No |
| **Degree of Disgust Expressed through Emoji** | | | | | | | | | | | | |
| 0.0582 | | 0.0596 | | 0.0573 | 0.0593 | | 0.0595 | | 0.0591 | No | No | No |
| **Degree of Fear Expressed through Emoji** | | | | | | | | | | | | |
| 0.0689 | | 0.0699 | | 0.0681 | 0.0696 | | 0.0690 | | 0.0701 | No | No | No |
| **Degree of Joy Expressed through Emoji** | | | | | | | | | | | | |
| 0.0810 | | 0.0826 | | 0.0709 | 0.0816 | | 0.0809 | | 0.0822 | No | No | No |
| **Degree of Sadness Expressed through Emoji** | | | | | | | | | | | | |
| 0.0687 | | 0.0689 | | 0.0686 | 0.0693 | | 0.0675 | | 0.0708 | No | No | No |
| **Degree of Surprise Expressed through Emoji** | | | | | | | | | | | | |
| 0.0708 | | 0.0720 | | 0.0699 | 0.0718 | | 0.0712 | | 0.0723 | No | No | No |
| **Degree of Trust Expressed through Emoji** | | | | | | | | | | | | |
| 0.0809 | | 0.0826 | | 0.0798 | 0.0822 | | 0.0820 | | 0.0823 | No | No | No |

Table 11 Statistical Evaluation of Emoji-Based Sentiment Skew in Tweets.

|  |  |  |
| --- | --- | --- |
| **Parameter 1** | **Parameter 2** | **Skew Observed** |
| **Positive/Negative Sentiment in Sarcastic Tweets** | | |
| 0.337 | 0.236 | Yes |
| **Positive/Negative Sentiment in Non-Sarcastic Tweets** | | |
| 0.372 | 0.242 | Yes |

Table 12 Statistical Evaluation of Sentiment Congruence Between Emojis and Text.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sarcastic Content** | | | | | **Non-Sarcastic Content** | | | | | **Significant difference between Sarcastic and Non-Sarcastic Content** | | |
| **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | **Negative Subset** |
| **Average Text-Emoji Sentiment Incongruence – Positivity Score** | | | | | | | | | | | | |
| 0.278 | | 0.355 | | 0.225 | 0.271 | | 0.362 | | 0.191 | No | No | Yes |
| **Average Text-Emoji Sentiment Incongruence – Negativity Score** | | | | | | | | | | | | |
| 0.266 | | 0.169 | | 0.334 | 0.241 | | 0.150 | | 0.321 | No | No | No |
| **Average Text-Emoji Sentiment Incongruence – Neutrality Score** | | | | | | | | | | | | |
| 0.197 | | 0.224 | | 0.179 | 0.236 | | 0.280 | | 0.198 | Yes | Yes | Yes |

Table 13 Statistical Evaluation of Text-Based Markers of Sarcasm.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sarcastic Content** | | | | | **Non-Sarcastic Content** | | | | | **Significant difference between Sarcastic and Non-Sarcastic Content** | | | | |
| **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | | **Negative Subset** | |
| **Average Text Sentiment Score** | | | | | | | | | | | | | | |
| -0.0324 | | 0.594 | | -0.451 | 0.0184 | | 0.687 | | -0.557 | No | | Yes | | Yes |
| **Average Text Positivity Score** | | | | | | | | | | | | | | |
| 0.319 | | 0.654 | | 0.0933 | 0.361 | | 0.712 | | 0.0584 | No | | Yes | | Yes |
| **Average Text Neutrality Score** | | | | | | | | | | | | | | |
| 0.330 | | 0.286 | | 0.362 | 0.297 | | 0.262 | | 0.326 | Yes | | No | | No |
| **Average Text Negativity Score** | | | | | | | | | | | | | | |
| 0.351 | | 0.0600 | | 0.544 | 0.342 | | 0.0249 | | 0.615 | No | | Yes | | Yes |
| **Average Tweet Length** | | | | | | | | | | | | | | |
| 19.0 | | 18.7 | | 19.2 | 18.5 | | 17.0 | | 19.8 | No | | No | | No |
| **Average Number of Hashtags** | | | | | | | | | | | | | | |
| 0.104 | | 0.164 | | 0.0634 | 0.148 | | 0.235 | | 0.0725 | No | | Yes | | No |
| **Average Number of Laughter Indicators** | | | | | | | | | | | | | | |
| 0.0232 | | 0.0357 | | 0.0146 | 0.00836 | | 0.00602 | | 0.0104 | No | | Yes | | No |
| **Average Number of Capitalised Words** | | | | | | | | | | | | | | |
| 1.89 | | 1.79 | | 1.97 | 2.31 | | 2.46 | | 2.17 | Yes | | Yes | | Yes |
| **Average Number of User Mentions** | | | | | | | | | | | | | | |
| 0.168 | | 0.157 | | 0.176 | 0.343 | | 0.464 | | 0.238 | Yes | | Yes | | Yes |
| **Average Instances of Pragmatically Relevant Punctuation** | | | | | | | | | | | | | | |
| 0.406 | | 0.371 | | 0.429 | 0.460 | | 0.211 | | 0.674 | No | | Yes | | Yes |
| **Average Instances of Affirmatives** | | | | | | | | | | | | | | |
| 0.574 | | 0.650 | | 0.522 | 0.549 | | 0.440 | | 0.642 | No | | Yes | | Yes |
| **Average Instances of Negations** | | | | | | | | | | | | | | |
| 0.516 | | 0.329 | | 0.644 | 0.518 | | 0.247 | | 0.751 | No | | Yes | | Yes |
| **Average Instances of Intensifiers** | | | | | | | | | | | | | | |
| 0.423 | | 0.329 | | 0.488 | 0.354 | | 0.343 | | 0.363 | No | | No | | Yes |
| **Average Instances of Interjections** | | | | | | | | | | | | | | |
| 1.55 | | 1.37 | | 1.67 | 0.864 | | 0.717 | | 0.990 | Yes | | Yes | | Yes |
| **Average Instances of Mid-Word Capitalisation** | | | | | | | | | | | | | | |
| 0.490 | | 0.493 | | 0.488 | 0.816 | | 1.04 | | 0.621 | Yes | | Yes | | Yes |

Table 14 Statistical Evaluation of Context-Based Markers of Sarcasm – Topic Modelling.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sarcastic Content** | | | | | **Non-Sarcastic Content** | | | | | **Significant difference between Sarcastic and Non-Sarcastic Content** | | | | |
| **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | | **Negative Subset** | | **Overall** | **Positive Subset** | | **Negative Subset** | |
| **Topic 1 – Overall Topics** | | | | | | | | | | | | | | |
| 0.344 | | 0.352 | | 0.339 | 0.348 | | 0.342 | | 0.354 | No | | No | | No |
| **Topic 2 – Overall Topics** | | | | | | | | | | | | | | |
| 0.244 | | 0.260 | | 0.233 | 0.207 | | 0.215 | | 0.201 | No | | No | | No |
| **Topic 3 –Overall Topics** | | | | | | | | | | | | | | |
| 0.187 | | 0.215 | | 0.168 | 0.218 | | 0.217 | | 0.219 | Yes | | No | | Yes |
| **Topic 4– Overall Topics** | | | | | | | | | | | | | | |
| 0.226 | | 0.173 | | 0.262 | 0.226 | | 0.225 | | 0.227 | No | | Yes | | No |

## 6.5 Quality Evaluation of Sarcasm-Annotated Datasets

The literature review highlighted validity concerns regarding previous convention with regards to obtaining data labelled as sarcastic and non-sarcastic. While it is known that the use of tags like #sarcasm and #irony to obtain positively labelled data for classification yields an unrepresentative training dataset of sarcastic content, however no work could be identified which evaluated this impact. Given the iSarcasm dataset presents the most robust methodology to obtain sarcastic labelled data which is as representative as possible of organic sarcastic content, all comparative work has been carried out using this dataset as a benchmark for comparison. This section has two aims:

* Compare quantitative features of the iSarcasm dataset sarcastic text to the survey results. Critically evaluate the results to assess the survey methodology.
* Assess the impact, if any, a poor annotation strategy has on the features of the sarcastic labelled data.

## 6.5.1 Survey Methodology

Of the features evaluated, 66% were found to have no deviation between the iSarcasm dataset and the survey results. Given the sampling of the survey text prompts originated from a subset of the iSarcasm dataset which contained no emojis, this may explain some of the observed incongruence. Features which signify tone were the primary sources of significant difference between the two sets. Hashtags, laughter indicators, affirmatives, negations, intensifiers, interjections, and mid-word capitalisations were all found in varying frequencies. This may indicate that the presence of emoji for the original author of the tweets’ felt tone was sufficiently evident based on these other indicators to convey their intentions. Analysis which compared a selection of these features between sarcastic labelled tweets with and without emoji did not show significant differences between the sets, so it is unlikely that this is a validity concern with regards to the survey methodology.

## 6.5.2 Hashtag Annotation Strategy

A dataset which utilised #sarcasm and #irony to label a series of tweets for sarcasm detection was sourced for the purposes of this comparison. 28% of evaluated features were found to have no significant difference to the baseline data, with differences being universal across all feature categories. This result provides strong indication that any sarcasm detection models trained using datasets labelled in this manner yield inaccurate outcomes, confirming postulation of previous validity concerns. For a more robust assessment of the impact of this annotation strategy, additional data from alternative datasets may be assessed for alignment with features in the hashtag annotated dataset and the iSarcasm dataset respectively. This observation similarly applies to any future human-annotated data which is collected using the iSarcasm state-of-the-art method for sarcastic/non-sarcastic data annotation.

Table 15 Comparison of Sarcastic Text Features between datasets.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Result** | | | | | | | | | | | **Statistically Significant Difference** | | | | | | | | | | | |
| **iSarcasm** | | | | | | **Dataset annotated using hashtags** | | | | | **iSarcasm** | | | | | | | **Dataset annotated using hashtags** | | | | |
| **Overall** | **Positive**  **Subset** | | | **Negative**  **Subset** | | **Overall** | | **Positive**  **Subset** | | **Negative**  **Subset** | **Overall** | **Positive**  **Subset** | | | **Negative**  **Subset** | | | **Overall** | | **Positive**  **Subset** | | **Negative**  **Subset** |
| **Emoji Use Frequency1** | | | | | | | | | | | | | | | | | | | | | | |
| 1.35 | 1.47 | | | 1.21 | | 0.102 | | 0.217 | | 0.0449 | Yes | | Yes | | Yes | | | Yes | | Yes | | Yes |
| **Position of Emojis in Text** | | | | | | | | | | | | | | | | | | | | | | |
| 0.895 | 0.898 | | | 0.893 | | 0.976 | | 0.943 | | 0.992 | No | No | | | No | | | Yes | | Yes | | Yes |
| **Sentiment Score of Emojis Used** | | | | | | | | | | | | | | | | | | | | | | |
| 0.160 | 0.307 | | | -0.0691 | | 0.269 | | 0.303 | | 0.189 | Yes | No | | | No | | | Yes | | No | | Yes |
| **Degree of Positivity of Emojis Used** | | | | | | | | | | | | | | | | | | | | | | |
| 0.361 | 0.447 | | | 0.227 | | 0.405 | | 0.429 | | 0.346 | No | No | | | Yes | | | Yes | | No | | Yes |
| **Degree of Negativity of Emojis Used** | | | | | | | | | | | | | | | | | | | | | | |
| 0.201 | 0.140 | | | 0.296 | | 0.135 | | 0.127 | | 0.156 | No | No | | | No | | | Yes | | No | | Yes |
| **Degree of Neutrality of Emojis Used** | | | | | | | | | | | | | | | | | | | | | | |
| 0.379 | 0.368 | | | 0.396 | | 0.401 | | 0.361 | | 0.498 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Anger Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0666 | 0.0687 | | | 0.0634 | | 0.0684 | | 0.0668 | | 0.0724 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Anticipation Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0757 | 0.0771 | | | 0.0736 | | 0.0772 | | 0.0750 | | 0.0824 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Disgust Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0551 | 0.0563 | | | 0.0531 | | 0.0566 | | 0.0544 | | 0.0619 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Fear Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0655 | 0.0672 | | | 0.0629 | | 0.0670 | | 0.0650 | | 0.0717 | No | No | | | No | | | No | | No | | No |
| **Degree of Joy Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0795 | 0.0795 | | | 0.0796 | | 0.0787 | | 0.0764 | | 0.0841 | No | No | | | Yes | | | No | | Yes | | Yes |
| **Degree of Sadness Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0656 | 0.0675 | | | 0.0627 | | 0.0666 | | 0.0648 | | 0.0711 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Surprise Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0683 | 0.0694 | | | 0.0665 | | 0.0686 | | 0.0666 | | 0.0734 | No | No | | | No | | | No | | No | | Yes |
| **Degree of Trust Expressed through Emoji** | | | | | | | | | | | | | | | | | | | | | | |
| 0.0782 | 0.0790 | | | 0.0771 | | 0.0791 | | 0.0763 | | 0.0859 | No | No | | | No | | | No | | No | | Yes |
| **Average Text-Emoji Sentiment Incongruence – Positivity Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.299 | 0.371 | | | 0.205 | | 0.264 | | 0.360 | | 0.216 | No | No | | | No | | | No | | No | | No |
| **Average Text-Emoji Sentiment Incongruence – Negativity Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.257 | 0.153 | | | 0.391 | | 0.251 | | 0.208 | | 0.273 | No | No | | | No | | | No | | Yes | | Yes |
| **Average Text-Emoji Sentiment Incongruence – Neutrality Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.236 | 0.241 | | | 0.229 | | 0.190 | | 0.251 | | 0.160 | Yes | No | | | Yes | | | Yes | | No | | Yes |
| **Average Text Sentiment Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.159 | 0.719 | | | -0.558 | | -0.0655 | | 0.588 | | -0.390 | Yes | Yes | | | No | | | Yes | | Yes | | Yes |
| **Average Text Positivity Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.462 | 0.766 | | | 0.0740 | | 0.270 | | 0.624 | | 0.0940 | Yes | No | | | No | | | Yes | | Yes | | Yes |
| **Average Text Neutrality Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.235 | 0.188 | | | 0.294 | | 0.395 | | 0.341 | | 0.421 | Yes | Yes | | | No | | | Yes | | Yes | | Yes |
| **Average Text Negativity Score** | | | | | | | | | | | | | | | | | | | | | | |
| 0.303 | 0.0463 | | | 0.632 | | 0.335 | | 0.0353 | | 0.485 | No | Yes | | | Yes | | | No | | Yes | | Yes |
| **Average Tweet Length** | | | | | | | | | | | | | | | | | | | | | | |
| 19.2 | 18.2 | | | 20.4 | | 15.4 | | 14.3 | | 15.9 | No | No | | | No | | | Yes | | Yes | | Yes |
| **Average Number of Hashtags2** | | | | | | | | | | | | | | | | | | | | | | |
| 0.189 | 0.228 | | | 0.139 | | 2.19  1.19 | | 2.57  1.57 | | 2.00  1.10 | Yes | Yes | | | Yes | | | Yes  Yes | | Yes  Yes | | Yes  Yes |
| **Average Number of Laughter Indicators** | | | | | | | | | | | | | | | | | | | | | | |
| 0.00610 | 0.0109 | | | 0.000 | | 0.0120 | | 0.0120 | | 0.0120 | Yes | Yes | | | Yes | | | Yes | | Yes | | Yes |
| **Average Number of Capitalised Words** | | | | | | | | | | | | | | | | | | | | | | |
| 2.02 | 1.82 | | | 2.28 | | 2.44 | | 2.21 | | 2.55 | No | No | | | No | | | Yes | | Yes | | No |
| **Average Number of User Mentions** | | | | | | | | | | | | | | | | | | | | | | |
| 0.360 | 0.348 | | | 0.375 | | 0.500 | | 0.386 | | 0.557 | Yes | Yes | | | Yes | | | Yes | | Yes | | Yes |
| **Average Instances of Pragmatically Relevant Punctuation** | | | | | | | | | | | | | | | | | | | | | | |
| 0.372 | | 0.337 | 0.417 | | 0.448 | | 0.301 | | 0.521 | | No | | | No | | No | Yes | | No | | Yes | |
| **Average Instances of Affirmatives** | | | | | | | | | | | | | | | | | | | | | | |
| 0.524 | | 0.500 | 0.556 | | 0.332 | | 0.283 | | 0.356 | | No | | | Yes | | No | Yes | | Yes | | Yes | |
| **Average Instances of Negations** | | | | | | | | | | | | | | | | | | | | | | |
| 0.659 | | 0.500 | 0.861 | | 0.508 | | 0.301 | | 0.611 | | No | | | Yes | | Yes | Yes | | Yes | | Yes | |
| **Average Instances of Intensifiers** | | | | | | | | | | | | | | | | | | | | | | |
| 0.238 | | 0.196 | 0.292 | | 0.164 | | 0.133 | | 0.180 | | Yes | | | Yes | | Yes | Yes | | Yes | | Yes | |
| **Average Instances of Interjections** | | | | | | | | | | | | | | | | | | | | | | |
| 1.122 | | 0.870 | 0.144 | | 0.788 | | 0.867 | | 0.749 | | Yes | | | Yes | | Yes | Yes | | Yes | | Yes | |
| **Average Instances of Mid-Word Capitalisation** | | | | | | | | | | | | | | | | | | | | | | |
| 0.707 | | 0.576 | 0.875 | | 1.77 | | 1.63 | | 1.84 | | Yes | | | No | | Yes | Yes | | Yes | | Yes | |

1Given the survey questions were sampled from a subset of the iSarcasm dataset (tweets with no emojis), the converse subset was used for this evaluation to avoid overlap of features. Reported results for iSarcasm emoji use frequency is therefore skewed and unrepresentative of a realistic value. For this reason, these reported values were excluded from the calculation of aligned features for the two datasets.

2Given the sampling of the text for the dataset relies on the presence of #sarcasm or #irony for inclusion, the dataset reports values disproportionately high compared to natural hashtag use rates. In every instance either hashtag appeared once only therefore the values used for statistical evaluation considered comparison both with and without these labels. When assessing the feature for a statistically significant difference.

## 6.6 Chapter Conclusion

This chapter established features which are significantly different between sarcastic and non-sarcastic tweets using data obtained from a survey. The survey methodology proposed implemented a method to integrate emojis into sarcastic and non-sarcastic content which aimed to capture data more representative of organically occurring sarcasm on Twitter than previous convention of annotation using keyword hashtags and was less labour intensive for participants than previous works which use manual identification of sarcasm in Twitter users previously published tweets, based on author classification.

Statistical analysis found survey responses to be sufficiently representative of the relevant population on Twitter and evaluated a range of structural, sentiment and contextual features in the survey results to identify features with significant differences in sarcastic and non-sarcastic content. The identified differences were consistent with intuitive outcomes and indicate that such markers may provide value in subsequent sarcasm detection. This work additionally identified the importance of context on the presentation of sarcasm; where its presentation is notably different in positive and negative content and more likely to be expressed in topics which are polarising.

Where good alignment was found in patterns in data between content in the survey results and the iSarcasm data, the evaluated dataset which implemented a keyword hashtag strategy to identify sarcastic content showed significantly different features across all categories assessed, providing strong indicators that this method does not yield content which is representative of sarcasm on Twitter.

This chapter has established an understanding of trends in features which may aid in the identification of sarcasm in addition to insights into the impacts of differing annotation practices. Such work provides a framework to collect additional data for training and provides insights relevant to feature extraction for sarcasm detection.

# 7 Integration of Emoji Pragmatics into Sarcasm Detection

## 7.1 Proposed Novel Methodology

Work carried out to date has identified several features which may serve to improve outcomes for sarcasm detection models. Chapter 6 has concluded that collection methodologies for annotated data for training has significant implications for the data obtained in terms of the structural and sentimental features. Based upon this assessment, expansions upon the current training dataset must be carried out mindfully to ensure that the characteristics of organic sarcastic data are retained. Additional observations regarding features which have a statistically significant difference in sarcastic and non-sarcastic text were identified; this will form the basis for feature extraction prior to model training. Additional observations which will be subject to evaluation for consideration in the proposed model include:

* Both semantics and sentiment have been identified as features with distinct markers in sarcasm. Word vectors are limited in their capacity to effectively convey both semantic relationships and sentiment concurrently; such an effect is fundamental to the nature of language; where sentiment of two words is opposing, the words may be semantically linked. Consider *good* and *bad* which are semantically adjacent however the sentiment is opposite.
* Subjectivity of sarcasm identification is increased where the topic relates to personal beliefs rather than more generalised humour. An observation relating to this effect is that the belief held by the majority seems to be classified as non-sarcastic at higher rates and the converse is true where the belief is regarded as more controversial. There were notable differences in the use of emojis in these cases- sentiment congruence was high between emojis and text in the non-sarcastic content and the opposite was true for the sarcastic content.
* The presentation of sarcasm is different in varying contexts. Negative sarcastic tweets were observed to use emojis disproportionately to reduce perceived negativity. The converse was true to some extent for positive sarcastic content; however, the effect was less universal.

## 7.2 Proposed Architecture

## 7.2.1 Model Selection

Given the complexity of the task and prevalence of similar architectures identified in literature, neural networks will be the model class evaluated for the task. Selection Criteria will consider the architectural features with respect to the task in addition to observations from literature on their suitability.

## 7.2.2 Sentiment-Aware Attention Mechanism

This section aims to evaluate an approach to ensure both semantic and sentiment information can be considered for the purpose of sarcasm detection as previous evaluation has determined that both features are likely to play a role in understanding the underlying patterns in sarcastic content. The consideration of two sets of word vectors, optimised for semantic and sentiment information respectively is a possible solution for exploration, which may enable a more nuanced representation of the words by the model. This option may improve outcomes for highly specific tasks such as sarcasm detection. However, there are limitations to this approach; this would significantly increase complexity and increase data requirements for the task. Additionally, where the vectors contain overlapping information, there may be interference, reducing model performance necessitating increasingly robust measures against noise which would likely impede the learning of subtle nuances which characterise sarcasm. This evaluation leads to a conclusion that a better approach would consider these features in a manner which does not include additional word vectors.

The fundamental purpose of attention mechanisms is to mirror cognitive attention within text. Attention mechanisms are not generally utilised for the purpose of enhancing the sentiment awareness of a model, however in this context this poses potential to ensure both sentiment and semantic information can be considered in the model architecture. To achieve this, the traditional attention mechanism will be modified to weight word importance with regards to their polar sentiment.

## 7.3 Data Preparation

## 7.3.1 Expansion of Current Training Dataset

Evaluation to this point has established that sarcasm detection is a highly complex process, and thus is likely to require a large dataset for training. Previous work in chapter 6 has established that annotation strategies which rely on weak labelling using hashtags yield text which represents sarcasm significantly differently than human-annotation, however this is not to say that all data contained in these datasets are unrepresentative of organic sarcasm. This work aims to evaluate a series of datasets annotated for sarcasm detection collected in this manner and identify any text which is aligned with observations about the human-annotated data which was shown to be characteristic of sarcastic text previously. The annotation strategies for the evaluated datasets are as follows:

1. Annotation based on the presence of #sarcasm, #sarcastic or #irony in the text.
2. Annotation of based upon the presence of #sarcastictweet and #not in addition to a set of offensive vocabulary words.
3. Annotation strategy based on the presence of key words and hashtags in addition to semi-supervised learning techniques.
4. Annotation strategy based on the presence of key words and hashtags in addition to semi-supervised learning techniques.
5. Annotation based on the idea that sarcastic content consists of text containing both positive and negative verb phrases in the same document (tweet) in addition to a set of key words.

## 7.3.1.1 Statistical Evaluation to Identify Aligned Features

The goal of this work was to establish a method for removal of data which was unrepresentative of organic sarcasm present in the datasets due to the limitations of hashtag-based annotation strategies. Such a task was carried out by establishing a range for features previously observed to have a statistically significant difference between sarcastic and non-sarcastic text occur in the validated data at 95% confidence. Using this analysis, data can be selectively omitted from the dataset to leave only data which is the most aligned with the data found using validated methodologies. Standard calculations for a 95% confidence interval were implemented with the following adjustments to fit the context of the data:

* Where measuring instances of a feature, the minimum value for the lower limit of the range at 95% confidence was defined as zero to align with what is possible.
* Rounding of the results to up the nearest integer was utilised in cases where metrics represent the counting of features, as this must result in an integer value. Justification for this action is clear considering the example of the hashtags per tweet parameter; the range at 95% confidence is defined as 0.00 < x < 0.967 which results in the exclusion of all instances where a hashtag occurs, which is not necessarily valid. The adjustment of the obtained result in cases like this are reasonable as the dataset which establishes these boundaries is relatively small and the range must be viewed in the context of the limitations associated with small datasets.

Table 16 Baseline for Evaluation of Tweets for Organic Sarcasm Features.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | | **Standard Deviation** | | **Range at 95% confidence** | |
|  | **Positive Subset** | **Negative Subset** | **Positive Subset** | **Negative Subset** | **Positive Subset** | **Negative Subset** |
|  | **Emoji-Based Features** | | | | | |
| **Emojis Per Tweet** | 0.569 | 0.507 | 0.803 | 0.680 | UL: 2.17  LL: 0.00 | UL: 1.87  LL: 0.00 |
| **Sentiment Score** | 0.323 | -0.0514 | 0.412 | 0.413 | UL: 1.15  LL: -0.500 | UL: 0.775  LL: -0.878 |
| **Degree of Positivity** | 0.466 | 0.261 | 0.255 | 0.200 | UL: 0.977  LL: 0.00 | UL: 0.661  LL: 0.00 |
| **Degree of Negativity** | 0.143 | 0.312 | 0.178 | 0.229 | UL: 0.498  LL: 0.00 | UL: 0.770  LL: 0.00 |
| **Degree of Neutrality** | 0.391 | 0.427 | 0.155 | 0.118 | UL: 0.701  LL: 0.0800 | UL: 0.662  LL: 0.191 |
|  | **Text-Based Features** | | | | | |
| **Degree of Positivity** | 0.667 | 0.0707 | 0.282 | 0.0698 | UL: 1.23  LL: 0.102 | UL: 0.210  LL: 0.00 |
| **Degree of Negativity** | 0.0520 | 0.589 | 0.0695 | 0.246 | UL: 0.191  LL: 0.00 | UL: 1.08  LL: 0.00 |
| **Degree of Neutrality** | 0.281 | 0.340 | 0.247 | 0.205 | UL: 0.775  LL: 0.00 | UL: 0.751  LL: 0.00 |
| **Hashtags per Tweet** | 0.164 | 0.104 | 0.637 | 0.431 | UL: 1.44  LL 0.00 | UL: 0.967  LL: 0.00 |
| **Laughter Markers per Tweet** | 0.0336 | 0.0116 | 0.211 | 0.107 | UL: 0.455  LL: 0.00 | UL: 0.226  LL: 0.00 |
| **Affirmatives per Tweet** | 0.502 | 0.533 | 0.844 | 0.828 | UL: 2.19  LL: 0.00 | UL: 2.19  LL: 0.00 |
| **Negations per Tweet** | 0.379 | 0.651 | 0.779 | 1.06 | UL:1.94  LL: 0.00 | UL: 2.78  LL: 0.00 |
| **Intensifiers per Tweet** | 0.247 | 0.320 | 0.483 | 0.678 | UL: 1.21  LL: 0.00 | UL: 1.67  LL: 0.00 |
| **Interjections per Tweet** | 1.08 | 1.70 | 1.74 | 1.70 | UL: 4.57  LL: 0.00 | UL: 4.78  LL: 0.00 |
| **Relevant Punctuation per Tweet** | 0.423 | 0.528 | 1.03 | 1.29 | UL: 2.59  LL: 0.00 | UL: 3.11  LL: 0.00 |
| **User Mentions per Tweet** | 0.227 | 0.258 | 0.584 | 0.754 | UL: 1.40  LL: 0.00 | UL: 1.77  LL: 0.00 |
| **Capitalised Words per Tweet** | 2.00 | 2.13 | 2.19 | 2.28 | UL: 6.38  LL: 0.00 | UL: 6.69  LL: 0.00 |
| **Mid-Word Capitalisations per Tweet** | 0.551 | 0.603 | 1.35 | 1.69 | UL: 3.25  LL: 0.00 | UL: 3.94  LL: 0.00 |

Previous chapters establish that sarcasm presents differently where there is positive and negative sentiment in the text. For this reason, evaluation of each dataset has been broken down into these subsets for a more granular approach to the task. Figures X and X show results of the analysis for each dataset.

A group of blue and white bars

Description automatically generated

Figure 10 Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Feature.

A screenshot of a graph

Description automatically generated

Figure 11 Atypical Feature Identification in Datasets Annotated using Hashtag Strategy by Subset.

## 7.3.1.2 Results

In every case a significant proportion of text prompts were found to contain atypical presentations of sarcasm based on the established baseline from the results contained in the datasets which utilised human-annotation methods. Given the more aligned results observed for datasets 3 and 4, the addition of machine learning improves outcomes for training data collection. Details with regards to the techniques were not provided within the associated literature therefore greater depth of evaluation is difficult. Across the datasets available, 72% of the tweets were found to display incongruent characteristics compared to the present gold-standard collection strategy used in the iSarcasm dataset. Given the assessed datasets are the most prominently referenced across sarcasm detection literature, validity concerns are evident with regards to the true capabilities of resulting models to identify organic sarcasm online.

For the purposes of expansion upon the present training dataset, a cautious approach was implemented which includes only data which has no features which present atypically compared to the defined baseline. This yielded an additional 38593 sarcastic tweets for consideration during model training.

This strategy ensures that the bias associated with poor annotation strategy is mitigated, however the results are limited to align with the characteristics of the previously validated data. In any instance using this strategy, the scope of what is represented in the final dataset is limited to what aligns with the baseline. Where some valid instances of organic sarcasm are unrepresented in the baseline, they will be excluded from the collected data also. However, this serves to exclude data which is unrepresentative of how sarcasm usually presents; it is not typical for sarcastic content to contain hashtags like #sarcasm therefore the training data should reflect this. Expansion upon the availability of human-annotated data in future would serve to improve outcomes for representation of organic sarcasm within the dataset, leading to a more comprehensive understanding of patterns which are present in sarcastic content.

## 7.3.2 Data Cleaning

To ensure patterns learned for sarcastic content are as robust as possible, it is important to ensure that the data is processed appropriately prior to model training. The process for cleaning was dependent on the type of model to be trained. The processes have been detailed in section 11.10. Following the data cleaning process, a control subset was generated which omitted all emojis from the text.

A diagram of a process

Description automatically generated with medium confidence

Figure 12 Data cleaning steps for neural network model training.

## 7.4 Baseline Model Evaluation

## 7.4.1 Neural Network Selection

Neural networks are a primary area of potential identified in literature for sarcasm detection tasks, with the primary models used being CNN, LSTM and GRU models. Suitability based on their respective architectures are evaluated in section 11.10.4. Each identified model was hyperparameter tuned using a similar methodology to that which was described previously in section 11.4.5, with the goal of identifying the most optimal model with regards to accuracy and generalisation capabilities.

Following hyperparameter tuning, three models with the greatest performance metrics were re-trained using 5-fold cross validation to evaluate their robustness to varied data. The results were used to identify the optimal model.

## 7.4.2 Evaluation Metrics

The assessment metrics were adjusted to account for the nature of the task, binary classification. During hyperparameter tuning accuracy was the primary metric used to evaluate model performance. The metric reports the proportion of correct predictions made by the model. This was selected based on the ease of interpretation and its suitability given the balanced nature of the dataset.

Following initial tuning, during the assessment of models with varying data input, F1 score was additionally considered which provides insight into the precision and recall of the model.

Initial evaluation during the hyperparameter tuning process identified the GRU model as achieving maximum accuracy while minimizing loss. To balance computational cost with assessment of as many iterations of the model as possible during hyperparameter tuning cross validation was performed on the three models identified with the greatest performance and the final optimal results was selected from this pool.

## 7.4.3 Model Performance

Initial evaluation during the hyperparameter tuning process identified the GRU models as the greatest performing, which is aligned with expectations with respect to their architectural features and literature assessment discussed previously.

A screenshot of a computer

Description automatically generated

Figure 13 Optimal GRU Model for Sarcasm Detection.

Although this was the optimal outcome for the assessed models, evaluation of results display overfitting. Over the duration of training, loss increases and the accuracy trends upwards, but the curve is unstable. Across folds there is notable variance in performance, indicating the model is not capable of learning robust features.

A graph of performance metrics

Description automatically generated

A graph of a graph showing the difference between epidermis and epidermis

Description automatically generated

Figure 14 Training Profile of Optimal GRU model.

Test accuracy result of ≈63% in the context of binary classification indicates a model that can to some extent discriminate between sarcastic and non-sarcastic content, given the accuracy consistently exceeds 50%, the expected accuracy of a model which predicts labels at random. However, given the variance between tests, and high loss scores, the models performance on unseen data cannot be stated with certainty, thus is an undesirable solution. Given this approach utilised the word vectors as the sole source of information to train the model, this result may be related to the over-reliance of the model on semantic information, where sentiment is also relevant. Further adaptations to the model which enable the consideration of sentiment to a greater extent may be of value for the task. Section 7.5 discusses the implementation of a strategy to provide enhanced sentiment awareness to the model.

Table 17 Performance Metrics of Optimal GRU Model for Sarcasm Detection on unseen data.

|  |  |
| --- | --- |
| **Accuracy** | 0.6255 |
| **F1 score** | 0.6829 |

## 7.5 Attention Mechanism

## 7.5.1 Architecture

To integrate sentiment awareness into the model in the most computationally efficient manner, which minimising the addition of noise which may arise from vectors with overlapping information, an attention mechanism which weights the semantic and contextual information gleaned from the word vectors alongside sentiment embeddings.

Sentiment metrics selected for use in this layer were limited to degree of positivity and negativity. Such a selection directly addresses the limitations of adjacent semantics and opposing sentiment, in addition to providing means to learn patterns relating to sentiment incongruency, differing presentation of sarcasm in positive and negative contexts which have each been observed to contribute to what makes sarcasm. This work was limited to the evaluation of the effects of dimensional theory centric embeddings as these parameters were shown to have statistically significant differences in sarcastic text compared to non-sarcastic text. The evidence for parameters generated from basic theory data was more limited. Given this observation, the trade off between increasing computational complexity and likelihood of noise against the increased granularity of the data which enhance learning of subtle patterns indicated that the results were not likely to yield favourable results from their inclusion. The values previously generated in chapter X were limited to emojis only; no dataset could be identified which contains predictions for the up-to-date corpus therefore this resource is of significant value for such a task. Word values were obtained from the SentiWordNet dictionary which provides polar information for degree of positivity and negativity for a large set of words. The sentiment data was converted to embeddings and scaled to integer values for ease of interpretation:

Where and represent the positive and negative sentiment intensity of the words respectively where large values indicate greater intensity towards the given polarity and the converse for lesser values. Out of vocabulary words were represented by 0, indicating an objective word. Given an input string:

Sentiment embeddings (,) containing and respectively can be constructed for string *t*:

The embeddings are introduced into the attention layer of the optimised GRU model and the weighted contribution of the sentiment of each word is determined via a matrix multiplication between the sentiment values and the weight matrix:

Which is subsequently adjusted by a bias term:

Where:

Where *o* represents the output of the GRU layer and and represent positive and negative sentiment intensity embeddings respectively. Each respective transformation is calculated using independent weight and bias to enable varying importance to be assigned to each data type, providing a more flexible framework for optimisation during model training. The summation of these weights is obtained, and a hyperbolic tangent function is applied:

Where:

This introduces non-linearity to the model, increasing its capacity to learn complex patterns which cannot be entirely represented by linear relationships. Without the use of an activation *n* layers in a model can be reduced to a single linear layer, which mirrors linear regression thus is only capable of learning linear functions within the feature space.

Given a definition of linear regression as follows:

Two linear layers can be reduced to a single linear layer as follows:

Where and . Such a definition can be extended to *n* layers by induction. The hyperbolic tangent activation specifically was selected for the introduction of non-linearity due to its mapping of values to a range of -1 to 1 which mitigates instability and exploding gradients, which is not achieved by alternatives such as the sigmoid function. The function was additionally widely documented within literature for similar implementations within neural network and attention architectures.

A softmax function is applied to the output of the tanh function to convert unnormalized attention scores into interpretable probability distributions, which can be added to a total of 1 where greater probability is assigned to higher scores:

Where attention weights, represent the contribution of each token in the string to the final text representation. Finally, the attention weights are applied to the input word vectors:

This approach allows the model to learn the most relevant semantic information from the vectors in addition to considering the sentiment information for its prediction. The use of sentiment determined externally to the model enables the model to learn based off prior knowledge learned from word tokens previously, while also considering the most computationally efficient method to introduce sentiment information.

## 7.5.2 Model Performance

Evaluation metrics indicate the introduction of the attention layer to the model was effective in improving the models’ performance for the purposes of sarcasm detection. On an identical training dataset, the increased sentiment information and context weighting afforded by the attention layer yielded an improvement in accuracy ≈15.8%. The training profile of the model indicated increased robustness in pattern recognition and decreased overfitting, implying the sentiment information and sentiment-context weighting was a key limitation of methodologies which rely solely on semantic information within word vectors.

**A screenshot of a computer program

Description automatically generated**

Figure 15 Optimal GRU Model for Sarcasm Detection with Attention Layer.

Improvements are similarly observed in the F1 score; from 0.6829 to 0.7539 consolidating observations of increased performance and highlighting the enhancement of generalisation and overall performance of the adapted model. Loss values across training displayed a smoother profile, showing increased stability of the model which further lends confidence to the result. Results using this iteration of the model can be viewed with increased confidence in comparison to the previous model thus results would be increasingly predictable in comparison to the previous model, where deployed for use on unseen data.

Table 18 Performance Metrics of Optimal GRU Model for Sarcasm Detection with Sentiment-Aware Attention Layer on unseen data.

|  |  |
| --- | --- |
| **Accuracy** | 0.7246 |
| **F1 score** | 0.7539 |

## 7.6 Impact of Emojis on Model Performance

Given emojis are used disproportionately to convey tone within online content, it is possible that sentiment incongruence may be disproportionately relevant where their contribution to an input is considered. Presently, this information is under-utilised for sarcasm detection tasks in online content, despite this observation. Removal of the consideration of emoji from the training data yielded slight decrease in performance metrics; 1.9% and 1.4% decrease in accuracy and F1 scores respectively. Validation loss however displayed a 13% increase where emojis were omitted, indicating that information obtained from emoji data is of importance in the generation of a robust model. Where emojis are not used, it is possible that accuracy on unseen data is similar based on the performance metrics, however the outcome on unseen data is much less predictable.

Table 19 Performance Metrics of Optimal GRU Model for Sarcasm Detection with Sentiment-Aware Attention Layer on unseen data.

|  |  |
| --- | --- |
| **Accuracy** | 0.7110 |
| **F1 score** | 0.7436 |

## 7.7 Chapter Conclusion

This chapter initially presented broad characteristics of sarcasm which contribute to its complexity for detection in Twitter content based upon work carried out in previous chapters. Characteristics of greatest importance were postulated to include sentiment features underrepresented in present models and context. Such features were used to construct a novel sentiment-aware attention mechanism which enabled tokens within the text string to be weighted based upon their polar sentiment in addition to their contextual importance. A GRU neural network was selected for the task based upon prevalence in literature for similar tasks, architectural suitability, and performance in initial evaluation. The addition of the attention mechanism into the model yielded an improvement in accuracy ≈15.8% and improved stability. Such an outcome consolidated hypothesis that sentiment may be an under-utilised component of sarcasm detection where standard word vector methodologies are implemented, which are prevalent in present literature. The impact can be attributed at least partially to the sentiment weights within the attention layer as a control test using only context vectors did not yield improved results from the model without attention.

The chapter concludes by addressing one core research question; Can emojis be used to improve outcomes for sarcasm detection? The omission of information relating to emojis from the dataset provides a slight

# 8 Evaluation

## 8.1 Solution Strengths

## 8.2 Solution Limitations

## 8.3 Applications

# 9 Conclusion

# 10 References

# 11 Appendix

## 11.1 Word Vector Selection

## 11.1.1 Consideration of Training Data

The availability of emoji vectors is sparse, however a majority train using short form online content such as tweets or Instagram captions. This methodology may have a significant limitation in this context; incongruence in sentiment between emoji and the text may be hard to identify in sarcastic content. Where this is already accounted for within the pragmatic information of the vector itself this incongruence may be less visible. It is unlikely that this would omit incongruence totally as there are no emojis which are used universally to convey sarcasm however emojis which are most frequently used in sarcastic content would be disproportionately impacted. The emoji2vec vectors were trained using official emoji descriptions, which largely omits any noise introduced by emojis being used in contexts contrary to their literal sentiment. However, these descriptions are highly literal and may not capture the nuances of certain use-cases. In the context of use to extract emotion data, this is not necessarily important, however vectors will likely require fine-tuning to provide more useful contextual information when deployed in a sarcasm detection model.

The Google News Word2Vec vectors, trained on news articles are unlikely to be impacted to a significant extent by sarcasm thus by similar logic incongruence is likely to be more evident using this selection. The developers of emoji2vec describe this as the most appropriate set of word vectors to be used alongside their vectors and thus this is a logical choice for use in the regression task. However, there are similar limitations with regards to their use in later sarcasm detection for twitter content given the significant differences in vernacular used between the two sets of text.

## 11.1.2 Word Vector Bias

Word vectors by nature, carry the biases of the dataset for which they were trained. The impacts of this in the case of the emoji vectors are likely to be minimal given the descriptions which they are trained are highly objective, with no identified use of vocabulary that may introduce bias such as adjectives or adverbs. The converse is the case however in the case of the Google News vectors which have been shown to contain significant bias in terms of gender, socioeconomic status, and race (Bolukbasi et al., 2016). This bias is important to retain within the data as it serves to add context, which is highly valuable to the detection of sarcasm, and sentiment more broadly. However, consideration must be given to whether biases within the media are also representative of that which would be found in the considered tweets. A comparative study relevant to this use-case is a potential area of expansion of this work in future, however the work of Curto et al., 2022 comparing bias in Google News Word2Vec and Twitter GloVe Vectors can provide an indication of the impacts of bias in the word vectors. The work found that the variance in bias was minimal for many topics however the Google News vector displayed greater bias based on socio-economic status and the GloVe vectors displayed greater skew towards negative sentiment in the context of discrimination. This will be a consideration while carrying out hyperparameter tuning.

## 11.2 Word Vector Optimisation Process

## 11.2.1 Optimiser Selection

Given the complex nature in the relationships between basic emotions, and context playing a key role in pragmatics captured in such vectors, determining a true optimum presents challenge both in defining loss functions and optimiser selection. Two proposed loss functions outlined in section X, aim to exploit known relationships between vocabulary which are postulated to increase sentiment awareness. The complexity of the problem set and high dimensionality of the vectors imply a loss landscape containing many local minima which may hinder optimisation where a function which cannot escape local minima to locate a global minimum is selected. Two optimisation functions were assessed for this task which are robust in escaping local minima:

*Stochastic Gradient Descent:* Gradient descent approaches operate by means of an iterative descent down a slope to locate a minimum for the loss function, defined as a slope equal to zero. To avoid convergence at local minima, the scale of each adjustment at each step is defined as:

Repeated until convergence

Where such a method reduces the step size as the slope approaches zero. Considering this function in the context of a three-dimensional plane, it is evident that there is no mechanism which avoids convergence within a local minimum or saddle point. Hence, an adjustment of this function which leverages a single randomly sampled loss gradient in each step is more appropriate:

Repeated until convergence

This adjustment avoids convergence at saddle points or local minima as the random sample may point away from a local minimum as it may not lie around this particular minimum in the loss contour, allowing the model to escape these points, where the summation of all results would not.

*Adam Optimisation:* Adam is a more sophisticated alternative to the stochastic gradient descent model, which introduces a variable learning rate during training. This method adapts learning rates ()using an exponentially decaying average of past squared gradients and implements an exponentially decaying average of past gradients to update vector direction ():

Where and denote Hadamard (element-wise) product and division respectively and is a smoothing term to prevent division by zero. Given , the update direction has momentum, which pushes the loss away from local minima to locate the global. The adaptive learning rate, is scaled by such that larger gradients result in smaller learning rates. The consideration the two moving averages of the gradients smooths noise, which is likely to feature prominently in the data under consideration. Such smoothing is particularly effective around saddle points, where gradients approach zero in many dimensions.

## 11.2.2 Model Optimisation

To select the optimal model for the task, hyperparameter tuning was carried out in both cases to monitor loss over time and best outcome for the loss functions. Hyperparameters assessed in each case were as follows:

Table 20 Summary of hyperparameters assessed.

|  |  |  |
| --- | --- | --- |
| **Model** | **Hyperparameter** | **Values Considered** |
| Stochastic Gradient Descent | Learning rate | 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 |
| Adam | Learning rate | 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 |
| Batch sizes | 32, 64, 128 |
| Betas | (0.9, 0.999), (0.8, 0.9), (0.7, 0.8) |
| Use regularisations | True, False |

A grid search model was employed to iterate through hyperparameters as the task was deemed not to have excessive computational cost to necessitate skipping permutations of hyperparameters. Relatively large learning rates were assessed to explore the solution space more comprehensively and increase convergence rate. Generally larger learning rates are undesirable due to their unstable gradients and tendency to overshoot optimal points, however given the adaptive learning rate in the Adam model, this effect was offset and smooth descent to optimal points were observed. A larger range considering smaller learning rates was considered for the gradient descent given the constant learning rate. The Adam optimiser additionally considered varying and which control moving average and second moment decay rate respectively. Varying aims to find balance between slow adaptation to mitigate the impacts of noise (smaller value) and faster adaptation which aids the optimiser where large gradients are present (larger value). Similarly, is varied to consider the outcomes where outlier values have reduced impact (smaller value) and greater response to changes in magnitude (larger value). Batch size is considered for the purposes of considering the balance between mitigating noise and reaching the true minimum point rather than navigating around it. The approach to batch size ideally would additionally consider over and underfitting more directly, however in this case fit was assessed through the consideration of the data at varying degrees of loss against human-annotated data.

The Adam optimiser yielded better results with smooth loss curves, contrasting the gradient descent model, indicating that the assumption of complexity in the loss landscape was accurate. The Adam model enabled the initial use of a greater learning rate to navigate this, with mechanisms to escape local minima which was not possible in the gradient descent model due to its constant learning rate.

## 11.2.3 Normalisation

To ensure training would not be prematurely halted due to the direction of vector shifts, and reduce impact vector length related noise, after each iteration vectors were normalised to unit length by dividing each vector by its Euclidean norm. The result is each step moves vectors across a spherical plane of radius 1, rather than leaving the movement plane undefined. This approach preserves cosine similarity values which are not impacted by magnitude, however when considering emotions, it is possible that magnitude may play a role in their identity; for example, it intuitive that *mad* indicates a less intense emotion than *furious* which should be reflected in their respective vector magnitudes. This information is lost using this approach. However, it is unclear the degree to which this information may be of value in the context of the comparisons in question.

## 11.3 Basic Theory Regression

## 11.3.1 Feature Scaling

The available data for the task falls into several classes: basic emotions (8 features), dimensional theory polar emotions (2 features), degree of neutrality and sentiment score. This presents a challenge with regards to feature scaling. It is unlikely that a simplistic strategy of scaling evenly across all features will appropriately represent the patterns present thus an alternative strategy was implemented. The aforementioned four parameter classes were scaled relative to each other using a simplistic min-max scaler. A secondary scaling was applied within the two classes containing multiple features to scale them proportionate to their class. This approach considers that not every emoji evokes equal magnitude of emotion and prevents the disproportionate consideration of one emotional model based on the quantity of features it contains.

## 11.3.2 Mitigation of Skew in Target Variables

Four approaches were evaluated for each basic emotion to address the skew in the data:

Table 21 Strategies evaluated to mitigate skew for each basic emotion.

|  |  |
| --- | --- |
| **Transformation** | **Formula** |
| Log transform |  |
| Square root transformation |  |
| Cube root transformation |  |
| Reciprocal transformation |  |

And the degree of improvement was defined by the skewness metric:

Where E is the expectation operator and X is a random variable and the best identified outcome is that which is closest to zero.

Table 22 Results of skewness transformation.

|  |  |  |
| --- | --- | --- |
| **Basic Emotion** | **Best transformer** | **Best skew value obtained** |
| Anger | Reciprocal transformation | -1.176 |
| Anticipation | Cube root transformation | 0.001 |
| Disgust | Reciprocal transformation | -0.706 |
| Fear | Reciprocal transformation | -0.095 |
| Joy | No transformation | 0.040 |
| Sadness | Reciprocal transformation | -1.171 |
| Surprise | Reciprocal transformation | -0.031 |
| Trust | Log transformation | 0.109 |

Each optimal transformation was performed on the respective data and stored in separate columns. This action was performed as the performance of certain regression models is adversely affected by skew in data. Where these parameters are used to yield outcomes for regression tasks a reverse operation can be performed on predictions to revert the data to its original form.

## 11.3.3 Assessment of Components for Model Selection for Basic Theory Regression

*Data volume:* The dataset contains 150 rows for use in the regression task, a notably small volume of data which will likely limit the model selection to more simplistic options, which perform better on smaller volumes of data.

*Parameters for training:* A significant volume of approximations have been generated prior to and following vector adjustments. Given parameters which display greater linear correlation to the target variable will be of greater value to the model, three subsets of appropriations with the greatest similarity to the target variable will be selected in each case to aid in model training. Based on the same logic, dimensionality reduction during optimisation will also be performed with consideration to similarity.

*Target variables:* Both target variables with skew mitigated and without transformation will be evaluated and the option which yields the best results will be used. Previous data evaluation which focused on correlation data highlighted that basic emotions sometimes display correlation or negative correlation to the others which indicates that a multivariate regression approach may be viable. Such an approach would account for both the relationship between the target and input parameters in addition to the relationships between each target variable. Traditional univariate regression may provide a more accurate result due to the small data volume limiting the bandwidth to increase complexity thus both options will be considered.

## 11.4 Model Development for Basic Theory Regression

The overall goal of the model development process was to identify the best possible model for the regression task with regards to its accuracy and robustness. This section discusses in depth the process outlined in section X.

## 11.4.1 Data Split

The data available was initially split into two subsets of train and test data at a ratio of 80:20. This represents a high-typical split for similar use cases, which was deemed appropriate given the small quantity of data available as it retains as much data as possible for the training set, while ensuring the test subset is of sufficient size to give reliable indications of the models’ performance. The purpose of the test data is to withhold some data from the training process to understand how the model performs on unseen data. Within the training process, a secondary data split is performed during cross-validation. These processes were implemented to ensure that recorded performance is attributable to the models’ predictive capabilities, rather than the selection of train and test data for a particular instance.

## 11.4.2 Models Evaluated

The model selection process was created with the purpose of identifying the most optimised and robust model possible for the regression task. Given the limitations in data volume a focus on models with more simplistic architectures was appropriate. The following models were assessed:

*Linear Regression:* Models the relationship between a dependent variable and one or more independent variables by fitting the parameters to a linear equation. This model is simplistic therefore performs well on smaller volumes of training data. The Pearson’s correlation metrics evaluated previously show moderate to high-moderate correlation between parameters and the target variables which indicates that this model may be appropriate for the task.

*Random Forest:* Combines multiple decision tree regressors to predict the target variable. This model is similarly simplistic thus appropriate for the small dataset. If generalisation is a challenge, the aggregation of predictions from multiple decision trees can mitigate the impacts of overfitting. Overfitting has been identified as a potential problem given the predictions not displaying total correlation due to the limitations associated with the vector information and approximation methodologies highlighted previously.

*XGBoost Regression:* An adaptation of the Random Forest regressor which constructs an ensemble of decision trees sequentially, where each tree is modified to correct errors in its predecessor. Predictions are weighted to optimise the overall loss function which enables the capture of more complex patterns than Random Forest, however in this case the increased complexity may be of detriment due to the data volume.

*Support Vector Regression:* An adaptation of linear regression which defines a hyperplane that best fits the data while minimizing points which fall outside of these boundaries. The model reduces overfitting using a margin of error around the plane where the data can reside, and its simplicity enables the implementation with small datasets.

*Gaussian Process Regression:* A model which describes probability distributions over many functions to capture uncertainty and use a probabilistic framework for regression. The model works well for limited sample sized, while still capturing more complex, non-linear relationships and mitigating impacts of noise. Uncertainty metrics may additionally be of value in ensuring the final model is suitably robust.

*K-Nearest Neighbours Regression:* Predicts the value of a target variable by averaging the values of its k closest neighbours. The approach is simplistic and does not rely on large datasets to generate predictions. This is a potentially viable option as throughout the data preparation, an emphasis on generating approximations where similar emotions have similar vectors was implemented, which should return similarity scores which align to similar alternatives. This may not be sufficient given the correlations achieved previously, which may indicate that there is insufficient cohesion between similar emojis to obtain accurate results.

## 11.4.3 Target Variable Format Selection

The target variables were previously identified to not have a normal distribution and thus alternatives were generated which adjust the values to have a normal distribution. Both these parameters in addition the original labels were assessed to obtain the most optimal outcomes.

The task was identified to have potential for both univariate and multivariate regression. Both options were assessed to ensure comprehensive coverage of consideration to possible optimal models.

## 11.4.4 Model Optimisation

The optimisation process for the models had two goals. To determine the parameters which generate the most accurate predictions and was robust to altering the data. Hyperparameter tuning was deployed to determine the best possible outcome performance each model was capable of. Cross-validation was implemented to evaluate how robust the performance was.

## 11.4.5 Hyperparameter Tuning

The hyperparameter grids for each respective model were generated with the goal of assessing the impact of a broad range of hyperparameters on the outcomes of the models. Such hyperparameters were identified in the relevant documentation for each model. The values for each hyperparameter in each case aim to cover a broad range of possible options, spanning a range that covers typical values found in similar implementations in literature, adding a buffer above and below the range for comprehensive assessment. Where the hyperparameter has categorical options, these were selected based upon their potential suitability following assessment against the problem set and data available in cases where the number of options was too large for use.

Given the small data volume, the computational cost of the training processes is reasonably low, even where many hyperparameters are present in the grid. For this reason, the grid search method for tuning was deployed to provide the most comprehensive assessment of the selected hyperparameters, which considers every combination of hyperparameters within the defined grid space.

## 11.4.6 Cross-Validation

To ensure the model performance is not dependent on the specific combination of train and validation data used in a single instance, a cross validation model was implemented to ensure the optimal identified outcome was robust when predicting unseen data. A k-fold method was used which splits the data into five components and combines them to form k iterations of train and validation data. Five folds was the k-value deemed appropriate as this represents an 80:20 split of training and validation data, which was selected per the logic of the split prior to the training stage.

## 11.4.7 Dimensionality Reduction

The selection of three sets of prediction parameters was implemented with the goal of each subset being able to mitigate the impacts of their respective limitations. However, this strategy operates upon the assumption that the error is inconsistent across each set of approximations, which may not be the case and results in a large quantity of features being introduced during training. This strategy has a significant potential for underperformance without dimensionality reduction due to overfitting, multicollinearity and increasing complexity with a large quantity of features. Several subsets of data were generated by splitting the data into subsets for each approximation method in addition to selectively excluding features with low correlation to the target parameter, per the Pearson’s correlations determined previously.

## 11.4.8 Model Evaluation

The models were assessed in terms of their performance based on three metrics which work together to provide a broad picture of the performance:

*Mean absolute error:* Determines the mean absolute difference between predicted and actual values. This metric is selected to provide an easily interpretable metric to understand the error in the predicted values.

*Mean squared error:* Quantifies the average of the squared differences between predicted and actual values, which provides an indication of the overall magnitude of prediction errors with larger penalties applied to greater errors. Given the mean absolute error provides a mean value which does not provide much information regarding the distribution of the error across each individual prediction, the additional information provided by the mean squared error is of value to supplement this limitation.

*R2 score:* Measures the proportion of variance in the target variable which can be explained by the input parameters and can be considered a measure of the ‘goodness of fit’ of the model. This metric was included to provide a more comprehensive understanding of the models’ performance, such as the models ability to generalise which is also essential to understand when considering model performance.

## 11.4.9 Neural Network Evaluation

## 11.4.9.1 Model Selection

Outcomes from model selection using traditional machine learning models found that in multiple cases models capable of learning more complex patterns performed best. To ensure a comprehensive evaluation of models for the task, several neural networks were assessed for univariate regression. While these models generally require larger volumes of training data to learn patterns, they are also capable of modelling more complex patterns. The assessment was carried out to understand if this trade off yielded a favourable result in this case. The models assessed were as follows:

*Feedforward Neural Network:* A model which learns patterns by passing data through sequential layers, applying weighted transformations and activation functions.

*Convolutional Neural Network:* A model which uses convolutional layers to automatically extract relevant features from the input followed by fully connected layers to map the features to the output. It is more commonly used in image processing; however similar implementations have been identified in literature.

*Radial Basis Function Neural Network:* A model which uses a radial basis function as the activation function in its hidden layer. They excel at approximation and interpolation tasks, making them highly suitable for data which contains complex relationships. Additionally, they are effective using data which is unevenly distributed which may improve outcomes given the non-parametric nature of parameters available.

The neural network models are capable of modelling more complex patterns than the previous models, however with increased complexity there is a necessity for a greater volume of data, which is unavailable for this task. *Anger* was one such emotion which performed best using a more complex model: the Gaussian Process Regressor, therefore this emotion was selected for initial evaluation of the models. Due to increase computational complexity, initial evaluation was carried out for each model with some manual hyperparameter tuning and model with greatest potential (Feedforward Neural Network) was more extensively tuned. This enabled a more extensive tuning process to be carried out with a more comprehensive search of the hyperparameters with the resources available. The initial stage of the process aimed to obtain parameters which resulted in the model of greatest accuracy:

*Dense layers:* Varying the number of layers present in a neural network generally increases its capacity to learn complex patterns, however as complexity increases, the possibility of overfitting and the quantity of data necessary to obtain meaningful results increases. Models containing between 1 and 6 dense layers were evaluated to find a balance complexity and generalisation. The dense layer units were additionally varied to further optimise this effect. Unit values were varied from 8 to 4096 in steps of 128 units to comprehensively assess the optimal options for each layer in the model.

Dense layers calculate the dot product of the input and a weight matrix, which is transformed via an activation function. Several common activation functions were assessed in each layer; ReLU, SELU, elu and swish. The selection of activation functions centred around methods which restrict negative values as sparse representation can aid in generalisation, which is likely to be a challenge given the limited annotated data. These activations also simplify the optimisation landscape thus decreasing training times which enables a more comprehensive search of other hyperparameters with the available resources. ReLU is a very common activation function for many problem sets and returns positive values unchanged and converts negative values to zero:

ReLU:

The ability to output zero differs from alternatives which can only approximate zero. The result of this feature is a more simplistic model which is desirable in this case due to labelled data availability. However, if the output is consistently zero for all inputs, the neuron becomes inactive and stops contributing to learning. The ELU function operates on a similar principle, adding a constant to smooth negative value:

ELU:

This is a highly popular adaptation of ReLU which addresses the limitation of inactivity, and generally converges faster, however it is more computationally expensive due to its non-linearity. The SELU function addresses the same limitation using self-normalisation:

SELU:

The normalisation in this function, generally have more stable gradients than ELU activated models, however there is significantly less implementation in literature therefore their advantages and disadvantages may not be comprehensively understood. Finally, the swish activation introduces non-linearity for negative inputs using the sigmoid function:

Swish: where

Such an equation results a non-monotonic first derivative and smoothing. The function has been shown to outperform ReLU however is more computationally expensive due to increased linearity.

*Dropout and batch normalisation layers:* Initial evaluation of the neural networks displayed a discrepancy between train and test data performance even after convergence, which indicated poor generalisation. Such an observation highlighted the necessity for a robust mechanism to mitigate overfitting. Optional dropout layers were considered between dense layers to prevent overfitting. These layers randomly select a portion of the neurons to deactivate. This differs from the previously mentioned limitation of the ReLU function as the output layer is scaled in proportion to the dropout rate. Dense layers discourage neurons from becoming too specialised; thus, the neural network must learn more robust features. Dropout rates were varied to ensure optimal outcomes could be obtained from these layers. Batch normalisation was evaluated in a similar manner to the dense layers. During each iteration, these layers normalise the inputs by scaling them to have unit variance, which is performed within each batch of training data. This stabilises the distribution of inputs, encouraging more consistent updates to weights, leading to improved generalisation on unseen data.

*Learning rate:* Learning rate updates the degree to which weights are updated after each iteration of training. This is a highly influential parameter to model performance thus many potential values were assessed. Excessively large or small learning rates result in premature convergence and a suboptimal outcome or an extremely slow convergence respectively. No exploration of adaptive learning rates was implemented in this case to limit the computational cost of the tuning process; however, this may be an alternative approach where greater computing resources are available.

*Efficiency control:* Given the large range of parameters to be assessed and the relatively long training times associated with more complex models, measures were implemented to improve efficiency. While these do not directly impact the performance of a model, it allows the tuning process to explore a greater range of hyperparameters, which leads to a greater probability of identifying an optimal outcome. An early stopping callback was implemented to halt model training if the performance did not improve for 5 epochs based on the mean absolute error of the validation data. This is a relatively strict approach; however, this serves to reduce consideration of models with unstable gradients in addition to halting training where the model has stopped improving. The effect of this is that more combinations of hyperparameters (500 trials conducted) can be tested in a reasonable period, and more epochs (300) can be used so models which generate accurate predictions but are inefficient can also be considered. Batch normalisation and the selected activation functions additionally serve to improve convergence rates and by extension overall process efficiency by stabilising inputs.

From the hyperparameter tuning process, three models which displayed the best performance were progressed to cross validation to evaluate their generalisation capabilities further. A 5-fold cross validation was implemented per the logic of the traditional neural networks.

## 11.4.9.2 Performance Evaluation

The results of this evaluation are detailed in table X. Even where significant action was taken to ensure good generalisation, the result is an unstable model with poor alignment in metrics across the validation and test data. The error metrics show better performance in the test data, even where the R2 score is significantly worse. In both cases the R2 score indicates that the model cannot explain the patterns in the model thus the true performance should the model be deployed for prediction of unseen data cannot be understood reliably. For this reason, neural networks were excluded from consideration for the regression task.

Table 23 Neural Network performance evaluation results.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Validation Data Performance** | **Test Data Performance** |
| Mean Absolute Error | 0.0125 | 0.00670 |
| Mean Squared Error | 0.000278 | 0.000852 |
| R2 Score | 0.000393 | -0.155 |

## 11.5 Dimensional Theory Regression

## 11.5.1 Selection of Pre Trained Models

The purpose of this section is to narrow the scope of potential neural network architectures for assessment for a regression task to generate dimensional emotional theory parameters. This task has been broken down into two phases:

*Phase 1:* Identification of architectures commonly used in relevant literature and critical evaluation of their features with regards to the specific task.

*Phase 2:* Using the most suitable identified architecture from phase 1, assess several pre-trained models which fall into this category for the regression task.

## 11.5.2 Evaluation of Architectures in Literature for Sentiment Regression

Sentiment is fundamentally linked to context and thus models which successfully capture sentiment are likely to contain characteristics which enable the capture of long-term dependencies. Neural network architectures are widely utilised for such purposes. Their suitability is largely attributable to aspects of their architecture such as their:

*Ability to handle sequential data*: As language is sequential in nature, sentiment is often linked to word order. As neural networks handle data sequentially, their outcomes are often improved compared to more simplistic models.

*Robustness against noise:* Language is inherently noisy and variable due to variance in vernacular across a population in addition to relatively frequent errors in spelling or grammar. Such an effect is known to be more prominent in online content, where such features are sometimes used to add nuance to pragmatics. In the context of a problem set where there is potential for such features to add insight, this feature of such architectures may possess limitations if this information cannot be extracted in another manner.

Several models which would fall under this category are prominently found throughout literature evaluating sentiment analysis methodologies and specifically in the domain of sarcasm detection due to these features. The three primary models identified were as follows:

LSTM: The LSTM mechanism of selective memory is made up of a cell state, hidden state, and gates. The NN models sequential data by propagating over time through the connection of sequential events using the hidden state. This component captures dependencies by considering both the previous step and current output:

However, a feature of such a mechanism result in all previous steps being considered in the current step when implemented in isolation due to the chain rule:

Because:

With the ultimate result being either vanishing or exploding gradients and thus limitations on the abilities to capture long term dependencies in isolation. The cell state mitigates this effect through the filtering of less relevant information from the memory through the forget gate. This information, in addition to the weights from the input gate enable the model to learn which time steps contain important information, resulting in weights for each time step being represented in proportion to their understood importance to the model. Literature evaluating this architecture is not consistent with regards to its assessment of the efficacy of such a mechanism; with some studies citing the model as effective to capture long-term dependencies, and others postulating that the mechanism may ‘dilute’ important information over time. Similar contrasting observations are found in the sentiment across literature with regards to their robustness against noise, with the former asserting that noise information is filtered efficiently during training and the latter arguing the converse and observing amplification of noise. In this context, what is traditionally regarded as noise may provide pragmatic cues with respect to potential sarcasm as discussed above, these cues may be indicative that context of the problem set plays a part in the efficacy of the architecture to model valid patterns in the data, which explains contrasting observations within literature. No works could be identified which assessed this hypothesis, however this may be an area for future research.

GRU: GRU models address vanishing and exploding gradients using a more simplistic mechanism than LSTM models. These models omit the cell state and regulate memory using gates. These architectures utilise an update gate to dictate the information which is retained from the previous step in series with a reset gate which dictates the information which should be eliminated. Compared to LSTMs for natural language processing tasks, due to their more simplistic architecture GRU models seem to perform better where shorter data sequences are used, possibly providing greater potential in the context of the short form content as is used for the problem set in question.

Present state-of-the-art for sentiment analysis also includes significant volume of models which contain an alternative mechanism for memory. Transformer-based models leverage self-attention mechanisms to capture dependencies of long and short ranges.

These models are based upon encoder-decoder architectures, which are capable of processing data in parallel due to their attention mechanism which avoids processing data in parallel in favour of processing the sequences as a whole. The encoder consists of several layers, each containing two sublayers. The first sublayer generates self-attention and the second consists of a fully connected feed-forward network with two linear transformations and ReLU activation:

Where each layer uses its own weights and bias parameters. Given there is no inclusion of recurrence, there is no embedded manner to consider the relative position of words. To address this, positional encodings are added to the word embeddings. The decoder consists of several layers, which are each composed of three sublayers: the first decodes the previous input to extract positional information and apply attention. The attention in decoders is distinguished from that found in the encoder cells as they do not consider all words, but rather only words which have occurred before the current. The second layer contains a self-attention mechanism which receives information from the previous sublayers of decoders and the encoders output keys and values. The final decoder sublayer consists of a fully connected feed-forward neural network, like that of the second sublayer in the encoder cells.

The attention layers operate by passing each word in the sequence through the embedding and positional encoding layers to generate their respective vectors. The result is passed into the encoder where it is first processed by the attention module. The sequence is passed through three separate layers which each produce a matrix. These layers consist of query which defines the word for which the attention is to be calculated and key and value are compared to the query with regards to their relevance. These transformations are trainable operations which are adjusted to produce the desired output predictions over the course of training, quantified by the attention score, defined as the dot product between the query matrix and a transpose of the key matrix:

An intermediate matrix is produced, consisting of a multiplication between all combinations of the words in the respective matrices. A second dot product calculation is performed between the intermediate matrix and the value matrix to produce the attention score.

A diagram of a number

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A diagram of a graph

Description automatically generated

Figure 16 Architecture and mechanism of transformer model attention layers.

Given such a mechanism which learns based on similarities and differences in the input words, the architecture is aligned with the previous methodology for word vector transformation which yielded improved results compared to the original word vectors used in the basic theory regression task. Additionally, the lack of necessity for labelled data is advantageous adaptations of the strategy implemented for the basic theory regression could not yield acceptable levels of accuracy. The key characteristics of the training data which may have contributed to this result may be the small size of the training set, and the potential for complexity in the relationships between the emoji and labels, which could not be captured by less complex algorithms, more appropriate for the available training data. While traditional neural networks necessitate large, labelled datasets, the converse is true for transformer-based models which learn based on patterns between elements in the dataset, eliminating the need for such a resource intensive annotation process.

Each identified option presents advantages and disadvantages with regards to their architecture, and evaluation of each option may be the best approach to determining the best performance in this use case. However, considering the identified neural network options which necessitate annotated data, no models were identified which were trained using emoji and thus these options were omitted from consideration. Transformer-based models for dimensional theory sentiment analysis trained using emoji were identified, falling into the BERT category. Several iterations of BERT models tuned for various contexts to achieve outputs of dimensional-based sentiment scores were evaluated for correlation to the human-annotated dataset using Pearson’s correlation.

## 11.6 Survey Components

## 11.6.1 Harm and Risk Assessment

When conducting research involving human participation, it is important that the potential for societal benefit is balanced against the risks involved to the participants. The purpose of this assessment is to establish the magnitude and probability of any risks or discomforts which may be experienced by the participants such as harm and privacy so that adequate disclosures can be made to potential participants, ensuring their consent to participate is a fully informed decision.

By nature, sarcastic content disproportionately contains subject matter and inferences which may cause offence to some, thus psychological harm was deemed a primary potential cause of harm for this work and assessments were focused on mitigating this harm. The dataset for which the tweets are sourced for this survey carried out a screening process to identify and exclude content which may be considered excessively harmful. Given the definition of what was considered excessively harmful was not stated, a secondary screening was carried out on the content which would aim to remove content related to the following categories: violence, abuse of animals or humans, crime, terrorism, eating disorders, suicide, pornography, and exploitation of vulnerable populations. Given participants were not limited to Ireland, this list reflects the consensus of multiple English-speaking countries with regards to what is considered harmful content in the online space. This screening process did not highlight any content which may be classified as significantly harmful, indicating that the initial screening was effective in addition to the likelihood that participants were unlikely to submit such content, given the content could be traced back to them, even if it was stored anonymously in the dataset depending on their privacy settings on Twitter. These screening processes must additionally consider context to preserve non-harmful content which makes reference to such topics in a manner which has low likelihood for harm, as this content is common in sarcastic content and its inclusion ensures a more organic representation of the population. The following examples are taken from the question pool as examples where a harmful subject matter is referenced in a tweet which is deemed very low risk of harm:

*The only thing I got from college was a caffeine addiction.*

*I would kill a man for a forehead kiss rn.*

Following the assessment, the probability and magnitude of psychological harm to participants was determined to be low.

An additional risk associate with collecting data from participants relates to privacy. It is best practice to collect the minimum necessary personal data and ensure storage and reporting comply with legal requirements and good ethical practices. The nature of the work does not benefit from participants being identifiable to the researcher, or anyone else who may view this work thus to preserve privacy, data was collected anonymously. Given there was no data which makes participants directly or indirectly identifiable, GDPR does not apply to this survey, however governance surrounding participant control over their data is still good practice and thus the option to retroactively manage participation was implemented. This was implemented by prompting each participant to input a 6-character code, with inbuilt logic to prevent duplication. The participants were informed that should they wish to withdraw consent, they could reach out to the researcher via the provided email address and provide this code to remove their data from the study. This means that data collected falls into a potentially identifiable classification rather than total anonymity, however identification is controlled solely by the participant. Privacy of underage individuals was additionally managed via in-built logic in the survey. The population was restricted to those over 18, however where a survey is posted on online platforms accessible to those under 18, it was deemed important to add additional measures to ensure data of those under 18 would not be processed. This is addressed in the consent page of the survey which highlights this restriction on participation. A secondary measure involves a question which asks if the participant is over the age of 18. Where the *N*o option is selected for this question, the participant will be routed passed the body of the survey to the final page thanking participants for their responses and the response will not be recorded. These measures establish a framework for data collection which affords participants maximum privacy and minimises associated risks through the collection of information in a manner which does not make individuals identifiable, unless they wish to identify themselves.

This assessment establishes that this survey presents minimal risk of harm to participants. Outlined measure mitigate risk for psychological harm from the presented content, while preserving the purpose of the work. Personal data collected is not excessive and the storage and reporting procedures are deemed sufficient to ensure privacy.

## 11.6.2 Participant Sampling Strategy

The population to be sampled for this survey consists of individuals over 18 years of age who speak English and use emojis. While individuals under 18 present a potentially interesting population for assessment in this problem set due to their high frequency of use of both emojis and social media, there are significant ethical concerns for obtaining data from this population. The Data Protection Act 2018 dictates that data to be collected or processed for those under the age of 16 must be carried out with parental consent. This work expands this limit upwards to the age of 18. Parental consent in the context of a survey distributed via a website link is challenging to verify and therefore cannot be considered sufficient to comply with the law and ethical obligation thus this population must be excluded.

The survey is additionally limited to those who are English speakers, as this is a skill that is required to understand the questions. The final qualification for survey participation is the use of emojis. This is a more challenging qualification to manage, however a question has been added to exclude responses where individuals state that they do not use emoji and responses which have not included a single emoji in their response were additionally excluded as this may indicate that the individual does not use emojis naturally, or they did not correctly follow the instructions for submission. Basic demographic information was recorded to understand potential skew in data and variance in behaviours across the population.

Based on the target population characteristics sampling must be non-probabilistic in nature, as there are criteria which participants must meet for qualify for participation. To obtain participants of varied age and gender, ideally a quota sampling strategy would be implemented to obtain an even distribution of participants with regards to age and gender. This is difficult while avoiding the use of peers, family, and colleagues as participants which is not good practice due to the potential for bias due to their affiliation to the researcher. To mirror quota sampling as closely as possible, surveys were published in several locations with varying demographics of visitors.

## 11.6.3 Sourcing Participants

Participants were sourced using a variety of methods, with a stipulation that no participant should have a close personal relationship with the researcher to avoid bias. Participants were obtained via online forums (e.g. r/SampleSize on Reddit), chat groups containing other course students etc. Selected platforms aimed to ensure a sample which was not skewed towards a particular age or gender. Any platform where the primary visitor is under 18 was excluded.

## 11.6.4 Survey Questions

The survey questions can be broken down into several categories: Consent, qualification, demographic, classification, and emoji-usage questions.

*Consent:* The purposes of these questions are to establish the participant consenting and eligible to participate in the survey. Valid consent involves potential participants fully understanding the purpose of the survey, the data which will be collected and how it will be used and stored, and any risks associated with their participation. It is also important that there is a clear statement to ensure participants are aligned with the target population. To achieve this anywhere the link for the survey was shared, the following text was added before the link:

*Hello, I am searching for participants to complete a survey for me! The purpose of this survey is to gather data about how emojis are used in sarcastic content online, and how this differs from non-sarcastic content. The data will be used in part to develop a model for sarcasm detection. The results will form part of my dissertation to be submitted in partial fulfilment of the requirements of the degree 'MSc in Data Analytics'. Your responses will not be identifiable to you. I am looking for participants 18 or older who speak English and use emojis. If you have any more questions, feel free to reach out to me for help via my email sba22224@student.cct.ie. Thank you in advance!*

The landing page of the survey contains all relevant information for participation in greater detail. The page contains a consent clause at the bottom stating that the participant has read and understood the content and is happy to participate in the survey. The survey landing page reads as follows:

*Thank you for taking the time to complete this survey! The purpose of this survey is to gather data about how emojis are used in sarcastic content online, and how this differs from non-sarcastic content. The data will be used in part to develop a model for sarcasm detection. The results will form part of my dissertation to be submitted in partial fulfilment of the requirements of the degree 'MSc in Data Analytics'.*

*Are there any requirements to participate?*

*Yes. You must be 18 years or older to participate. You must be an English speaker who uses emojis. If this does not describe you, please do not submit a response.*

*What will I be asked?*

*You will be asked basic questions about your demographics, and to provide information about how you use emojis in online communication. As you may need to use emojis in some of your answers, this survey is easiest to complete on a mobile device, however for computers there is an emoji keyboard available within the survey!*

*How is my data stored and managed?*

*All responses are anonymous. However, if for any reason after submitting your response you would like to withdraw your data from consideration in this work this is possible. On the next page of this survey, you will be prompted to generate a 6-character code which can be used to manage your participation in this survey by reaching out to myself (email is below). You are also welcome to contact me with questions without quoting this code if you have any questions. Data will be accessible by myself and any academic faculty who may require access (for supervision and/or grading of the work).*

*How can I reach out?*

*If you have any questions about this work, or how your data will be used, feel free to reach out to me through my student email: sba22224@student.cct.ie.*

*I have read and understand the above content and I am happy to proceed:*

*Yes/No*

*Qualification questions:* Once consent has been obtained, the participant is routed to questions to verify they fall within the target population for the survey. Each of these questions are mandatory to ensure that prevent events where someone is not identified as ineligible by not providing a response to the disqualifying question. These questions are as follows:

Table 24 Survey questions to verify participant eligibility.

|  |  |  |
| --- | --- | --- |
| **Prompt** | **Response options and format** | **Embedded logic** |
| Qualifying Questions  These questions make sure you are within the target population for my survey. | Static text field  No response required |  |
| Create a 6-digit ID code.  This can be letters or numbers (e.g. AB1234). You can use this to reach out regarding the use of your data. | Free text box  Mandatory field | No duplicated responses allowed.  Response must be exactly six digits.  Alpha-numeric characters only |
| Are you over 18 years of age? | Yes/No  Single select radio buttons  Mandatory field | Where ‘No’ is selected, the participant is routed out of the survey to the end thank you message. No data is retained. |
| Do you use emojis? | Yes/No  Single select radio buttons  Mandatory field | Where ‘No’ is selected, the participant is routed out of the survey to the end thank you message. |

The goal of the strict formatting and logic is to minimise data cleaning necessary to remove invalid responses and avoid the collection of data from underage individuals.

*Demographic questions:* The next section establishes basic demographic information about the participants. These questions aim to ensure the data is not disproportionately representative of certain subsets of the population. It is important to establish this information as communication style differs based on background, and conclusions drawn from any subsequent work can only be associated with populations which have been assessed in the research. Future research may expand upon these features as a manner to improve upon sarcasm detection, accounting for your demographic information, background, or interests to make more individualised assessments however this is outside the scope of this work and significant privacy and ethical considerations may arise as the quantity of personal information collected increases. The questions in this section are as follows:

Table 25 Survey questions to establish demographics of participants.

|  |  |  |
| --- | --- | --- |
| **Prompt** | **Response Options and Format** | **Embedded Logic** |
| Demographics  This section will ask you basic demographic questions about yourself. The purpose of these questions is to determine the role (if any) that your background may play in how you use emojis, and the way you express sarcasm. | Static text field  No response required |  |
| What is your gender? | Male/ Female/  I don’t want to say/ Other  Single select radio buttons  Mandatory field | Other option enables an optional free text field to input custom response |
| What age are you? | 18-24/ 25-34 / 35-44/  45-64/ 65+/ I don’t want to say  Single select radio buttons  Mandatory field |  |

These questions are constructed in a manner which ensures every eligible participant can provide an accurate response which is most applicable to them or provide no information for a given question. The goal in this case is to prevent discomfort of the candidates and gain responses which are as accurate as possible.

*Classification and emoji use questions:* The body of the survey consists of questions aimed to understand how individuals use emojis in sarcastic and non-sarcastic content. There are two primary question types within this section; classification questions which prompt the participant to classify a string of text as sarcastic or non-sarcastic and emoji usage questions which aim to understand how the participant would use emojis when they believe the content is sarcastic or non-sarcastic. These questions appear in pairs where they are related to the same string of text. These questions are repeated 10 times for each participant. The questions in this section are as follows:

Table 26 Survey questions about emoji usage.

|  |  |  |
| --- | --- | --- |
| **Prompt** | **Response Options and Format** | **Embedded Logic** |
| Emoji Usage  In the next section you will be presented with a number of tweets. For each tweet you will be asked to decide if you think it is sarcastic or not and then fill in the emojis you would use in the tweet, if any. There will be 10 tweets in total. If you feel any of the tweets do not apply to you, you can skip them. | Static text field  No response required |  |
| What is sarcasm? | Free text field  Mandatory field |  |
| Text string appears at the top of the prompt.  What is true about this tweet? | It is sarcastic/  It is not sarcastic/  I don’t know  Single select radio buttons  Optional field |  |
| What emojis (if any) would you add to this tweet?  Edit the below text to add the emojis you would use. | Free text field  Optional field | Default value is identical to the string in the previous classification question |
| Do you have any comments you would like to make? | Free text field  Optional field |  |

The first question additionally serves to monitor the quality of responses. Where the response indicates a lack of understanding of what sarcasm is, the quality of the response can be assessed with increased scrutiny. The prompt selection for the paired usage questions varied between participants to obtain a more comprehensive understanding of how sarcasm is conveyed in different contexts. The prompts for each pair of questions were selected from subsets obtained from the iSarcasm dataset of sarcastic and non-sarcastic tweets which contained no emojis. The sampling strategy can be described as follows:

Figure 17 Outline of sampling for emoji usage survey questions.

Sampling of questions is broken down into several stages for this survey, with a goal of ensuring that bias due to question ordering, sampling or decision fatigue are avoided. Quality control questions were randomly selected from both sarcastic and non-sarcastic subsets. These questions were changed intermittently while the survey was live to ensure that a question which results in unrepresentative behaviour was selected for the duration of the work. For each survey, the remaining questions were randomly sampled from the two subsets to yield a 1:1 ratio of sarcastic to non-sarcastic tweets, accounting for the label for the quality control question. This sampling method aims to ensure that each participant is likely to encounter tweets that they will classify as sarcastic *and* non-sarcastic while completing the exercise. The annotations from the original dataset are considered the ‘correct’ response as context is vital to the classification and the context two individuals project onto the same text may lead them to contradictory, yet both valid conclusions. They instead serve as speculative assignments which hope to ensure that there is a likelihood that both classifications will be utilised in each survey. The pattern of appearance of the speculative sarcastic and non-sarcastic prompts distributed at random throughout the survey. In all cases, sampling is performed prior to the initialisation of the survey and is independent of any information about the participant to avoid bias in the questions presented to each participant.

*Closing window:* Finally, the survey is routed to a thank you screen which thanks the individual for their participation and displays the researchers email address again for any questions the respondent may have.

## 11.6.5 Survey Design

The survey design was constructed in a manner which aimed to promote optimal outcomes with regards to response rate, homogeneity of data between participants and response quality. The Jotform platform was selected as the survey builder tool based upon its broad range of conditional formatting options and end-user friendly interface which is compatible across computer and mobile devices. Where a survey appears visually appealing and the interface is easy to use, participants are more likely to complete their responses thus this was an important consideration.

The selection of question formatting was considered based on the requirements of the question, with an overall goal of making the response as simple as possible, with preference to options which increase homogeneity between responses. Single select radio buttons were determined to be the optimal approach to collect data which falls into discrete classes given their simplicity of use for participants and high uniformity between responses. Free text fields were appropriate in some cases, however where possible conditional formatting was used to ensure the user input abided by the instructions.

A screenshot of a survey

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer code

Description automatically generated

Figure 18 Single select radio buttons to collect discrete data. In built logic ensures that responses yield valid output using prompts to ensure only one option can be selected. Text fields may also contain conditional formatting to ensure instructions are followed.

The question type which requires the most effort to complete for participants is that which prompts them to insert emojis into the prompt tweets. As response quality is likely to decrease with increasing work requirements, this has been formatted to reduce the work requirement as much as possible:

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Figure 19 Free text field for emoji use questions is populated with a default value of the text string the participant has previously classified as sarcastic or non-sarcastic to improve response quality. This enables the participant to insert emojis with ease into the text prompt. The figure displays the desktop version of the survey which features a drop-down menu containing emojis to make responses easier where an emoji keyboard is not available.

## 11.7 Statistical Tests Used

## 11.7.1 Selection Methodology

Statistical tests serve to provide insights as to whether the characteristic under evaluation influences the population or if two groups are different to one another. The correct selection of statistical test is essential to obtain results which are accurate and reliable. This selection is performed based on characteristics of the underlying data. A primary criterion for selection is the determination of whether data is parametric or non-parametric, which can be determined through the assessment of data with regards to the assumptions made about its distribution:

Figure 20 Flow chart to determine is data under evaluation is parametric or non-parametric.

An additional assumption make for parametric tests are that the observations in each group are independent of observations in every other group. Given survey questions were randomly sampled for inclusion, this assumption is met in all cases.

Where the nature of the data distribution has been determined, the secondary consideration is the relationship between the data to be evaluated. The quantity of groups to be compared, and the dependence or independence of the data must be understood to identify the most appropriate test. Independent data are data sets which are not in any way influenced by each other, and the converse is true for dependent data.

Table 27 Selection of Statistical Tests.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measurement Scale of the Dependent Variable** | **One Independent Variable** | | | | **Two Independent Variables** | |
| **Two Levels** | | **More Than Two Levels** | | **Factorial Designs** | |
| **Two Independent Groups** | **Two Dependent Groups** | **Multiple Independent Groups** | **Multiple Dependent Groups** | **Independent Groups** | **Dependent Groups** |
| **Interval or Ratio** | Independent t-test | Dependent t-test | One-Way ANOVA | Repeated Measures ANOVA | Two-Factor ANOVA | Two-Factor ANOVA Repeated Measures |
| **Ordinal** | Mann-Whitney U | Wilcoxon Signed Rank | Kruskal-Wallis | Friedman |  |  |
| **Nominal** | Chi-Squared |  | Chi-Squared |  | Chi-Squared |  |

## 11.7.2 Statistical Tests Used

*Shapiro-Wilk Test:* The purpose of the Shapiro-Wilk test is to determine if the data under evaluation is normally distributed. The test computes the test statistic, *W*, which is based on the correlation between the observed data distribution and the assumed normal distribution where a greater value implies greater correlation. The test defines the null hypothesis as the assumption that the data is normally distributed. The test requires data to be independent, continuous, have at least three observations, no significant skew, and no outliers. This test provides a simple initial evaluation for normality, however as a one-tailed test, it may result in false positive outcomes in some cases. Where a positive result is obtained, to ensure accuracy of the result a Q-Q plot will be used to confirm or reject the result.

*Q-Q plot:* The Q-Q plot is a qualitative method to evaluate normality in data distribution which plots the quantiles of the data against that a normal distribution. Where the distribution is normal the result is data which aligns with the identity line *x=y*, implying a linear relationship and thus normal distribution of the data.

*Levene’s Test:* Levene’s test assess the equality of variance of a parameter between groups. This is achieved by comparing the absolute difference between each value and the respective mean and performs an ANOVA test on the resulting differences. The null hypothesis is defined by an assumption of homoscedasticity. The test requires that data is independent and numeric.

*Outliers:* Interquartile range is a measure of the spread of the central 50% of values in a dataset. In this case outliers were defined as values which fell outside of the range of:

Which represents values greater than 2.025 standard deviations from the mean, where approximately 4% of data resides given Gaussian distribution.

## 11.8 Pragmatic Features Evaluated

This section outlines parameters evaluated to identify potential differences between sarcastic and non-sarcastic text.

*Frequency and position of emojis in text:* These parameters were assessed as potential structural features that may be indicative of sarcasm. Frequency of emoji use was defined by the average number of emojis per text string, which was not controlled for text length in this case as the text length of the prompts between each subset displayed the same distribution. Emoji position was defined as a value within range 0 to 1, indicating the relative location of the emoji to the other characters within the string.

*Sentiment Score Features:* Sentiment score was evaluated across all available metrics for both dimensional and emotional theory models. A comprehensive approach was used to ensure any possible factors which may have relevance to the problem set could be identified. Reported figures for each feature represent the average across each value for the given subset; such an approach has limitations given this is averaging values which may be generated from prompts of opposing sentiment. Where the distribution of positive and negative content is even between the subsets any potential insights may cancel out observed effects. Secondary reporting displays the same metrics, distinguished by their overall polar classification to reduce this impact.

*Sentiment Skew:* Compares the relative positivity and negativity of emojis found in the text, with the purpose of understanding if either sentiment is disproportionately represented in either subset.

*Sentiment Congruence:* Aims to evaluate the alignment of sentiment between the emojis and text within a given text prompt. The metrics reported compare the sentiment of the text at the sentence level to the emoji sentiment and incongruence is defined as the difference between these values. Reported values are averages across the subsets for these values.

*Text based markers:* Instances of defined text features were defined by the count of the given feature in the prompt. Average quantities per text prompt were the reportable values. Markers evaluated were defined for this task as follows:

* Hashtags: Words or phrases preceded by the # symbol.
* Laughter indicators: Words starting with ‘haha’ or ‘LOL’. The quantity of o’s in LOL can vary to identify as many iterations as possible with the same intended meaning. Search was case insensitive to ensure as many instances as possible could be identified. Given these markers add insight to the intended tone, evaluation of their relative representation across subsets may provide value.
* Capitalised words: Words which begin with an upper-case letter. Given postulations that good punctuation practices are not as important in sarcastic content, this may provide a structural indicator for potentially sarcastic content.
* User mentions: Words or phrases starting with the @ symbol.
* Pragmatically relevant punctuation: Aims to identify and punctuation which indicates tone or intent of a given text for example which may be of value in identifying sarcastic text. The punctuation marks defined as relevant were ‘!’, ‘?’ and ‘…’.
* Affirmatives: Words which imply agreement to a given message.
* Negations: Words which contradict or deny the associated topic or idea.
* Intensifiers: Words which add emphasis to the associated topic or idea.
* Interjections: Words used to interrupt the flow of the speech or text.
* Mid-word capitalisations: Words where capital letters are found in any position after the first. This includes screaming, where entire words are capitalised and interval capitalisation within words, both structures are employed to add insight to the intended tone of the message.

## 11.9 Topic Modelling Process

Topic modelling was carried out to supplement quantitative feature-based analysis approaches to identify topics or themes in the dataset corpus. As the quantitative statistical analysis carried out is limited by the scope to represent the features numerically, nuances are likely to be overlooked. Topic modelling may provide a deeper comprehension of the underlying pragmatics which characterise sarcastic text. Such a task has several challenges in the context of the problem set in question; the short form nature of Tweets limits the available context, additionally sarcasm detection applies across a broad range of subject matters and the dataset for modelling reflects this. This makes the task of generating coherent topics which can generate insight to the problem set more challenging to obtain.

## 11.9.1 Data Preparation

The capacity of topic modelling methods to generate coherent insights into the corpus is limited by the quality of the data it is trained on. Where noise is removed and words are grouped based on their root, greater coherence can be achieved in the training process. To achieve the best outcomes the following steps were taken:

*User Mentions and URLs:* User mentions and URLs were removed from the text as these features do not generally represent useful information in the context of their text only. While insights may be gleaned from alternative information such as the relationship between the author and the account referenced, or the content of link within the URL, this falls outside the scope of topic modelling and thus is regarded as noise.

*Spelling:* Where words are misspelled, the information the intended word provides to the model is lost, resulting in increased noise and loss of potential important information. Work conducted prior to the topic modelling process indicates that good grammar and punctuation practices are less prominent in sarcastic text compared to non-sarcastic counterparts which is also the case for informal content such as social media posts more broadly, thus it is reasonable to assume that there may be spelling errors in the dataset. A spelling check was applied to the text to correct misspelled words to reduce noise from the dataset and enable the consideration of misspelled words in the topic modelling process.

*Contractions and slang:* Linguistic devices such as contractions and slang provide valuable information pertaining to pragmatics, however given their high variability between individuals, their high representation in the corpus may be of detriment for topic modelling. For this reason, dictionaries which map these words to their more formalised meanings. It should be noted that for a final sarcasm detection model, the consideration of the slang terms in the form which was found in the original text may be of greater value to understanding relevant patterns, provided their vector forms can be found in the embedding, however for this use case, the conversion was implemented to reduce granularity of the corpus to provide more coherent topics.

*Punctuation:* Punctuation was removed as it does not provide useful information for the purposes of topic modelling. Where the feature does not provide data which can be used to improve pattern understanding, it instead increases noise. For this reason, punctuation was removed from the data.

*Stopwords:* Stopwords can be defined as words which do not provide any information to determine the underlying message within language. By definition, these words cannot provide any insights into the topic modelling task and thus were removed from the text.

*N-grams:* N-grams are continuous sequences of words within a document. They aim to capture insights from phrases through the retention of context, which is obtained from specific word combinations, rather than individual words only which can be of benefit to topic modelling. Due to the relatively small average word count for each text string, the evaluation of n-grams was limited to bigrams (n=2) and trigrams (n=3). The model was trained using a variety of combinations of n-gram configurations to determine the optimal approach. Omission of n-grams was additionally considered as they possess notable limitations with regards to sparsity within the dataset. During manual evaluation of text prompts to filter potential harmful content, an additional observation was the variety of topics discussed within the dataset. This observation leads to several assumptions about the output of topics; the topics are likely to be more generalised than specific and the corpus will have few frequently occurring n-grams. This conclusion implies that this pre-processing step may not be optimal, however quantitative evaluation provides greater reliability thus an assessment of their outcomes will be conducted.

*Stemming and Lemmatisation:* Stemming and lemmatisation refer to text normalisation processes which identify canonical representations for a set of related words where the former is a more simplistic process which reduces words to pseudo-stems while the latter considers context to return linguistically valid outputs known as lemma. These distinctions yield different outcomes in certain use cases:

Lemmatisation of the word *caring* returns *care* given it accounts for word meaning, however the more simplistic approach of stemming does not and outputs *car*, which is erroneous.

Lemmatisation with an input of *stripes* will return *strip* for a verb and *stripe* for a noun, whereas stemming cannot discrimination based on part of speech tag thus always returns *strip*.

In some cases, outputs are the same for example both models return *run* for an input of *running*, however stemming will return output more efficiently due to its more simplistic approach.

As the corpus is not especially extensive for the process, the computational cost is not an important factor in the selection, but rather performance. The goal of the process is to reduce granularity between words which provide the same meaning to improve topic comprehension, however it is difficult to assess the extent to which this is valuable before the smoothing effect is too great to generate topics which provide the most optimal topic outcomes. Given a case where lemmatisation returns *good* given an input of *better*, there is some contextual information lost. However, stemming returns the word unchanged preserving this context. Given the advantages and limitations of each option, the model output was assessed using, one, both or none of these options.

*Transformation order:* The order in which the outlined preprocessing steps are conducted has an impact on the outcomes of the work. The order selected was based on the consideration to the impact they make have on other steps. For example, where a spellcheck is performed prior to mapping slang and contractions to their equivalents, the mapping is likely to identify a greater number of instances to be mapped, making this a preferable configuration. Where the optimal order is more ambiguous, an iterative approach was implemented to determine the best sequence.

## 11.9.2 Models Assessed

Given the goal of the model to identify topics which can broadly describe prominent themes for a given set of text inputs, the use of labels is counterproductive to the purpose of the task thus unsupervised learning methods are appropriate for this task. The text is known to have relatively short word length and have a vast range of topics when viewed manually. The dataset is populated with a significant proportion of sarcastic content at this point which is known to have many ‘layers’ to the topics it contains.

*LDA model:* A probabilistic generative model which explains observations based on their co-occurrence patterns. The model is popular due to its simplicity of implementation and interpretability. However generally underperforms on shorter text and smaller corpa. Additionally, as it assumes document exchangeability, evolution of topics over time cannot be accounted for which would be problematic in the context of any application of a detection pipeline which uses the model over extended periods as online content evolves rapidly. Several instances of this model have been identified for broad subject matters using Twitter content, therefore the model was assessed for this task.

NMF model: A model which learns topics by decomposing highly dimensional vectors into non-negative matrices of fewer dimensions representing the topics and their respective weights. The process of non-negative matrix factorisation breaks down a matrix into two:

Where the matrices W and H are composed of k and m rows respectively:

Based upon the factorised matrices, , the weighted sum of a set of components can be determined, given the rows in H are components and rows in W are their weights:

The restriction of the components for consideration to non-negative weights facilitates a unique manner to identify topics; decomposition of the document-term matrix (where columns represent documents and rows represent weights) yields topics, which can be streamlined using the weighted sum function. Such a method would not be possible where negative weights are permissible, as a negative topic is uninterpretable. The architecture of the NMF model addresses sparsity and noise directly through the decomposition of the document term matrix into a document-topic and topic-term matrices, representing the distribution of topics and terms in each document respectively making it particularly effective for short text.

## 11.9.3 Model Selection and Evaluation

To ensure the best outcome was identified, both models were evaluated in their optimised forms. Given the challenges associated with obtaining insightful results in this context, a comprehensive approach to model selection was necessary.

To optimise each model, hyperparameter tuning in combination with optimisation of pre-processing steps was carried out iteratively to obtain the optimal result. Hyperparameter tuning utilised a grid search approach as it provided the most comprehensive assessment of hyperparameter within the grid. Hyperparameters assessed in each case were as follows:

*Number of Topics:* Topic numbers were evaluated across all four tests carried out, with the goal of obtaining a value which yielded the most optimal outcome across each of the tests. The number of topics was kept constant across each test to aid in judgement-based comparison work. The range for assessed topics was low as this is preferable in short text due to the difficulties in generating more granular topics where data is sparse and noisy.

*Passes:* The number of passes control the number of times the model processes the corpus. This value must be balanced such that it is sufficiently large to learn and refine the topics in the corpus, however the increase in performance is not indefinite; too many passes lead to overfitting. The values were varied from small to relatively large, as the data has a broad range of topics based on assessments during survey data screening for harmful content. This indicates that the results of the modelling may be broad and conceptual, therefore the data is particularly susceptible to overfitting. However, the benefits of granularity are also of value to consider, given the process is not excessively computationally expensive.

*Chunk Size:* Chunk size determines the number of documents that are processed together during analysis. With increasing chunk size, processing time decreases however this also may reduce generalisation abilities. Smaller chunk sizes were therefore represented in greater proportions in the hyperparameter tuning landscape based upon this effect, however larger chunk sizes were also assessed to ensure a comprehensive approach to tuning.

*Alpha:* This hyperparameter was varied during the NMF model tuning process to adjust the impact of the regularisation parameter. Specifically, this hyperparameter controls the sparsity of the basis matrix. Where the selected alpha value is too low, poor generalisation is observed and the converse is true where an excessively large value is selected.

*Beta Loss:* Beta loss is a regularisation parameter which controls the sparsity of the coefficient matrix. This operates alongside the alpha hyperparameter in the objective function. The objective function consists of reconstruction error and regularisation terms. The regularisation term is proportional to the Frobenius norm of the factor matrices raised to the power of beta for the coefficient matrix.

*Solver:* This parameter controls the minimisation of the objective function during the factorisation process within NMF modelling. The assessed options were a multiplicative update and coordinate descent solvers. The former is generally suitable for data with more topics and it more robust to noise. The latter is generally suitable for data with fewer topics, however, is more sensitive to noise. Given the characteristics of the data overlaps with optimal options for each solver to some extent, both options were evaluated during tuning.

*L1 ratio:* This parameter controls the ratio of L1 and L2 regularisation in the factorisation process. Regularisation in the NMF model consists of two terms; L1 which encourages sparsity in the factorisation process by promoting some of the coefficients to be zero and L2 which acts to oppose this effect by encouraging small values for all coefficients. The two regularisation terms must be balanced to obtain the optimal point which minimises over and underfitting.

Optimal models in each case were assessed using only one quantitative metric, c\_v coherence which is calculated using a sliding window, a one-set segmentation of top words and an indirect confirmation measure which uses normalised pointwise mutual information and cosine similarity:

Where the coherence score represents the arithmetic mean of these similarities. The score effectively identifies topics that are coherent and interpretable however it does not in all cases correlate well to human judgement. Given the nature of topic modelling, quantitative metrics cannot provide an output which considers the entire picture of the performance as they lack capacity to assess coherence with the degree of nuance of human judgement. For this reason, a series of quantitatively high performing models were identified, and a judgement-based assessment was carried out to characterise the interpretability and relevance of the topics generated. The combination of these two methods aims to offset bias associated with human judgement while still benefiting from the greater ability of human linguistic interpretation to assess such topics compared to quantitative metrics. Finally models with the greatest performance were subjected to a final quality assessment where a random sample of text where the topic was determined to be the dominant feature were generated and judgement was applied to assess their relevance to the respective topics.

## 11.9.4 Results

Based on quantitative coherence metrics, the NMF model was found to outperform the LDA model for both the overall dataset and the sarcastic only subset. Given the improved capacity of NMF models to perform with shorter text strings, this result is in line with intuition and can be attributed directly to the greater suitability of the NMF model architecture to handle short text. Assessment of the topics based on judgement between the two models yielded a contrasting outcome. While both models generated topics which were generally conceptual, the LDA model yielded outcomes which were easier to interpret. Notably the differences between the topics generated for the entire dataset and the sarcastic subset were explainable in the case of the LDA model, however this was not the case for the NMF model where topics were harder to interpret and had fewer distinctions between the two subsets. Where an analysis of a sample of text prompts compared to their dominant topic was carried out, there was significantly greater alignment in the case of the LDA model based on human judgement. Assessment of a random sample of documents most aligned with each topic corroborated observations that topics generated by the LDA models generated greater insight than those obtained from the NMF model making it the optimal model for the task.

Table 28 Summary of results from topic modelling for best models.

|  |  |
| --- | --- |
| **Top Words for Topic** | **Interpretation** |
| **LDA Model- All Tweets (Coherence Score=0.604)** | |
| Like, Love, Year, Get, Tri, Make, Old, Let, Also, Think | Expression of Preference |
| Name, Want, Would, Life, Answer, Miss, Come, You, Jazz, Anyway | Reflection or Contemplation of Past Events |
| Get, Time, Look, Can’t, See, Account, Wait, Show, Lot, Happy | Personal Experiences or Expectations |
| People, Day, Every, One, Give, Four, Could, Always, Live, Help | Routine Life |
| **LDA Model- Sarcastic Subset (Coherence Score=0.560)** | |
| Kill, Love, Yeah, Like, Chill, Busy, Soda, Room, Spider, Spill | Leisure Activities |
| First, Would, Let, One, Film, Age, People, Forget, Name Close | Opinions of Others (In the Media) |
| Casual, Aware, Appropriate, Attire, Croc, Work, Show, Wow, Get, Nice | Appearance and Clothing |
| Don’t, Think, There, Tell, Start, Enough, Suppose, Game, Ye, Get | Opinions and Thoughts |
| **NMF Model- All Tweets (Coherence score=0.707)** | |
| Kill, Love, Yeah, Spider, Soda, Room, Spill, Don’t, Keyboard, There’s | Undetermined |
| Like, Time, Really, Game, Day, Miss, Hurt, Thing, Love Effort | Personal Experiences |
| Super, City, Polite, Film, Confine, Villian, Awfully, Invasion, Particular, Hero | Fiction |
| Good, Guess, Collect, Change, End, Crime, Highly, Series, Episode, Usually | Evaluation and Opinions |
| **NMF Model- Sarcastic Subset (Coherence score=0.689)** | |
| Kill, Love, Yeah, Room, Soda, Spider, Don’t Keyboard, Chilling | Undetermined |
| Super, City, Polite, Film, Villian, Awfully, Hero, Confine, Invasion, Particular | Fiction |
| Like Love, Stop, Smell, Sudden, Everybody, Train, Sandwich, Board, Philly | Preferences for Food |
| Collect, Good, Guess, Wait, Partner, Stanford, Unnecessarily, Ready, Order, Slot | Decision Making |

Based on these assessments, the LDA method was determined to generate topics of greater value to the evaluation. This work additionally serves to demonstrate the limitations of such a method where the topic is too broad; while some insights can be gained from the results, the topics are too broad to provide significant weight to associated findings, where they are not corroborated by more robust evidence.

## 11.10 Sarcasm Detection Model

## 11.10.1 Data Preparation – Twitter Specific Text Features

*Hashtags:* To mitigate the impact of hashtag-based annotation strategies, hashtags relating to the words sarcasm and irony were eliminated from text strings. Remaining hashtags may contain useful information for the model thus a strategy to retain the words contained within the hashtags was implemented. The # symbol was separated from the associated text using a space. Compound word hashtags were separated using a camel-case split method where possible and a secondary strategy which extracted words from the string through comparison to a corpus of words. Observation of the data indicates that hashtags are a frequent source of sentiment incongruence between string components thus it is important to retain such information for later analysis.

Before hashtag splitting:

#LivingMyBestLife

Result:

# Living My Best Life

*URLs:* URLs do not convey useful information in the same manner as hashtags and thus they were removed from the data.

*User mentions:* User mentions do not generally contain semantic or sentiment information, however their usage frequency was shown to have a significant difference in earlier work. For this reason, each user mention was converted into *@user* to homogenise their use and enable their consideration in model training.

## 11.10.2 Data Preparation – Emoji Features

*Emoji separation:* Emojis are regularly added to text consecutively, which presents a challenge in this context, where models may recognise the string of emojis as a single word. As embeddings are present for individual emojis, rather than combinations, emojis were separated using a space.

*Emoticon to emoji conversion:* Emoticons are series of characters which convey emotion, in a similar way to emojis. A dictionary which maps emoticons to their equivalent emoji was used to convert any instances identified for emoticons within the dataset so that the sentiment they wish to convey could be considered by the model.

## 11.10.3 Data Preparation – Grammatical Features

*Special Characters:* Some special characters, or combinations of special characters were deemed to be relevant for the purposes of sarcasm detection during previous analysis. These characters were retained and separated from adjacent text based for consideration during model training. Characters determined to be irrelevant were removed from the text. Consideration to grammar implemented in this process; characters such as hyphens and apostrophes within words were retained to preserve good grammar and align words with their presentation in the embedding.

*Word case:* All words were converted to lower case to ensure that instances of the same word were identified to be the same as word processing is case-sensitive.

*Stopwords:* Stopwords were removed as they do not offer insight in terms of sentiment or semantics to the models therefore are likely to primarily contribute noise to the model for this problem set.

*Lemmatisation:* Lemmatisation was carried out to reduce the granularity of similar words within the corpus. For example, where *joy* and *joyful* are both present in the data vocabulary, lemmatisation enables pattern recognition between these iterations of the same base word thus improving the models’ ability to identify patterns in the text.

Contractions: Based on a similar principal to that described during lemmatisation, contractions serve to increase granularity in the dataset vocabulary, resulting in a more complex classification task. A dictionary to map contracted words to their long form equivalent was implemented to reduce this effect.

*Slang:* As slang is often not contained in word-vector corpus’ and is highly prevalent in the dataset due to the nature of both sarcasm and social media user-generated content more broadly, it is important to implement methods for the model to consider the information within these words. A dictionary which maps slang to a more formal representation of the same information was used to convert the slang in the text to a format which could be considered within the boundaries of the word embeddings available.

*Repeating characters:* Repeating characters are similarly prevalent in sarcastic content on Twitter where they are used to convey information outside of the literal meaning of the words such as tone. Given these instances cannot be considered in their present form, these words were condensed into a more grammatically correct form.

*Validity of Tokens:* Tokens were defined as valid where they had at least one character, any token which did not meet this definition was removed from the data.

## 11.10.4 Model Selection for Sarcasm Detection

The purpose of this section is to evaluate the identified neural network architectures for sarcasm detection based on the characteristics known about the sarcastic data from work carried out in previous chapters. Work to date indicates that patterns characterising sarcasm are highly complex and incongruence across between word sentiment is important. The position of emojis in the string is a relevant consideration however this work has not evaluated if this also applies to words. As sarcasm can be represented in many ways, the model must be able to generalise well.

*CNN:* Convolutional neural networks are more commonly used in image processing, however, have been reported to yield good results in literature for NLP tasks, particularly classification tasks including sarcasm detection. In contrast to architectures such as GRU and LSTM, the model learns patterns across space rather than time. This results in a model which identifies commonly co-occurring words regardless of their position in the string as there is no mechanism for capturing long term dependencies. Given the relevance of emoji position in text, this is a limitation of this model. CNNs are however capable of learning complex features and have good generalisation abilities due to their pooling layers which reduces the dimensionality of the output, which may ensure only the most relevant information is retained.

*LSTM:* LSTM models process words sequentially, enabling them to consider word order for text prompts. Their ability to capture contextual information, facilitated by their memory cells discussed in section X. The forget state within the architecture may facilitate good generalisation to unseen data, which is important in the context of the task in question however the efficacy of this component to improve generalisation is situation dependent based on observations in literature thus it is important to test the model to determine the viability LSTM models for the task in question.

GRU: GRU models have similar mechanisms to facilitate sequential data processing and capture of long-term dependencies as LSTM models, which are detailed in section X. Observations in literature indicate that the differences in the architecture of its memory mechanism make it more suitable for shorter text sequences, indicating that they may be a viable option despite their lesser representation in sarcasm detection literature at present.