**Depth discussion of word vector selection**

*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.

*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.

**Word Vector Optimisation Process**

*Model Architecture*

*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 challenges 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.

*Model Tuning*

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 X* 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.

*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.

\*\*\*Add in 3D surface plots showing the process and discuss\*\*\*

**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.

*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.

*Models Assessed:*

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.

*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.

*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.

*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.

*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.

*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.

*Performance 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.

*Neural Network Evaluation*

*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.

*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 X* 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 |

**Selection of pre-trained models for dimensional emotion theory parameter regression tasks**

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.

*Evaluation of common architectures in literature for sentiment regression task*

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 X* 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.

**Survey Components**

*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.

*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.

*Sourcing of 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.

*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 X* 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 X* 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 X* 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 X* 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.

*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 X* 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 X* 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.

**Selection of Statistical Tests**

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 X* 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 X* 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 |  |

**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.

**Features Evaluated during Statistical Evaluation of Pragmatic Features**

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.

**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.

*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.

*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.

*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.

*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 X* 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.

**Survey raw results**