**Chapter 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.

**Dataset selection**

*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.

*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.

**Sentiment classification:**

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.

*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. Analysis of patterns discussed in section \*\*\* indicates that this is a highly complex task requiring computationally-expensive models to achieve due to the subtleties in sentiment differences. Such a task is challenging due to limitations associated with feature extraction of emoji. 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.

\*\* Add in smoothing optimisation discussion here\*\*

*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 more sparse. 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 X*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.

**Feature extraction of emoji**

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.

*Comparison of Embeddings of Emotion Vocabulary to Emoji:*

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.

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 X* Cosine Similarity.

*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.

*Binary Word-Association Lexicon*

This method draws upon the EmoLex lexicon which assigns binary labels to convey association to basic emotions for each word. For vocabulary tagged for association with a basic emotion (), association to the basic emotion is ranked 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 highly associated vocabulary.

*Word-Emotion Intensity Lexicon*

This method proposes alternative approach to address skewed weighting from word associations, using the NRC-EIL lexicon containing emotion intensity information. The top-k words are determined as previously described however the similarity score is redefined as the average emotion intensity () from the lexicon.

This approach more closely aligns with the definition attached to the human-annotated data where the emotion score is the average relative affinity of the emoji to the basic emotion, however equal weighting is assigned to each top-*k* word. As *k* increases, the range for sim(ve, vbe) increases, diluting the influence of more relevant vocabulary. Weightings based upon cosine similarity to the basic emotions may be considered to mitigate this effect. Possible relevant vocabulary not included in the lexicon is also not considered using this method.

*Results*

*Table X* Pearson Correlation scores for all considered prediction methods. Bolded Scores represent the highest correlation observed for the emotion in the respective column, excluding human agreement scores.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variant | Anger | Anticip | Disgust | Fear | Joy | Sadness | Surprise | Trust | Average |
| *Word Vectors of Basic Emotion Words* | | | | | | | | | |
|  | **0.51** | 0.24 | 0.40 | **0.54** | **0.55** | 0.30 | 0.30 | **0.30** |  |
| *Binary Word-Association Lexicon* | | | | | | | | | |
| *k*=5 | 0.50 | **0.26** | 0.38 | **0.54** | 0.48 | 0.42 | 0.24 | 0.28 |  |
| *k*=10 | 0.47 | **0.26** | 0.38 | 0.53 | 0.51 | 0.41 | **0.40** | 0.29 |  |
| *k*=50 | 0.45 | 0.25 | **0.44** | 0.49 | 0.51 | 0.42 | 0.32 | 0.28 |  |
| *k*=100 | 0.45 | 0.24 | **0.44** | 0.46 | 0.51 | **0.44** | 0.28 | 0.27 |  |
| *Word-Emotion Intensity Lexicon* | | | | | | | | | |
| *k*=5 | 0.00 | -0.32 | 0.06 | 0.12 | -0.07 | 0.12 | -0.15 | 0.01 |  |
| *k*=10 | 0.08 | -0.32 | 0.11 | 0.12 | -0.08 | 0.14 | -0.11 | -0.11 |  |
| *k*=50 | -0.01 | -0.29 | 0.00 | 0.08 | -0.06 | 0.07 | -0.11 | -0.16 |  |
| *k*=100 | 0.00 | -0.23 | -0.07 | 0.06 | -0.04 | 0.05 | -0.04 | -0.16 |  |
| *Human Annotation* | | | | | | | | | |
|  | 0.67 | 0.74 | 0.65 | 0.59 | 0.78 | 0.69 | 0.64 | 0.72 | 0.69 |