EVE: Emotion Vector Encoding

Towards Modeling Feature Representations for Human Emotional States

Yuya Jeremy Ong
Penn State University
yjo5006@psu.com

Andrew Hankinson
Penn State University
ath5161@psu.edu

ABSTRACT

For centuries, many researchers have been perplexed by the complex and subtle mechanisms underlying human emotional states and has found to be a non-trivial research area. Recent work under Human Computer Interaction (HCI) and Computational Psychology has helped to develop underlying models for recognizing various emotional states from different modalities such as psychological, facial, pose, movement and auditory features. However little attention has been made to model the fundamental representations for human emotional state which allows for both human interpretability and computational interoperability. In this paper, we present a novel feature representation framework to learn a semantically interpretable human emotion feature representations which can be applied to any machine learning task. We utilize a modeling methodology inspired by Kansei Information Processing techniques coupled with the state-of-the-art methods for learning distributed vector representations for emotions and demonstrate their applicability to any machine learning tasks which utilizes human emotional states. Our work helps to improve current methods for representing emotions in a computational setting and allows for both improved human interpretability and computational interoperability for computer systems and further opens a new direction for Human Computer Interaction studies.

1 INTRODUCTION

Humans has had the innate ability to be able to feel, interpret, and understand emotions, often without much thought. Barring any mentally developmental or socialization deficiencies, people tend to be able to take in a lot of context and immediately determine how someone is feeling or are able to interpret the state of how someone may be in. Often, we do this unconsciously and are able to put the state of this person into some qualitative internal representation which we define as "emotions". This very same task is a non-trivial for computer systems and has been a non-trivial effort with ongoing work being performed to emulate this particular human-like phenomena. Computer systems which aims to understand human emotions, or the ability to understand and detect internal and mental states of a person has been an area studied across various domains including Computational Psychology and Human Computer Interaction (HCI) research.

The application and benefits of studying and better improving human emotion recognition and understanding systems would enable computers, robots, and digital assistants to effectively improve understanding on human affective states, enabling them to better empathize and improve the trust between computers to empower decisions and recommendations for an improved experience. With better emotion recognition systems, we can enable better diagnosis systems for patients facing mental illnesses or depression, improve user experiences for digital assistants, and provide additional features for other domains to exploit in analysis of correlation of human behavior to other phenomenas, behaviors, and trends.

The majority of studies in this field primarily focus on the ability to identify emotions. In particular, the area of human emotion modeling has been well explored through the means of classifying and identifying these emotional states given a variety of different input modalities. Facial expressions have been a widely explored area with the works of Essa et al. [9] devising the Facial Action Coding System (FACS) to enable human psychologists to best analyze and study correlations of facial structures to certain affective states. Busso et al. [4], jointly modeled auditory signals along with facial features, demonstrating an improvement over models with facial features independently. In a similar work, Del et al. [5] analyzed which emotions people found dominant if the facial and vocal emotions were mismatched, and Gavrila et al. [10] on the other hand, opts not to examine the fine-grain facial features, but instead the broad strokes of bodily movement. Gunes et al. [11] developed models which jointly learns features from both facial and body expressions. More recently, Kosti et al. [13] utilizes Deep Learning based features for analyzing both body features as well making use of the local semantic visual information to model emotions through a Convolutional Neural Network architecture.

However from a HCI standpoint, these systems are still not capable of empirically showing capacity to "understand" emotions from a natural and intuitive standpoint - a more "human-like" attribute. This raises many concerns and problems towards building systems which require human-to-computer (and vice-versa) interactions, where tangible and sensible representations and understanding are required to better improve the overall interaction between the two entities.

The key challenges we see in modeling emotions lies in the way they are complex, subjective, and dynamic. First emotions are complex, mostly due to their abstract nature, making it a very difficult area to decompose and study from both a qualitative and quantitative nature. This is due to their often due to the semantic subtleness of how the meaning and nuance of these states are highly based on the context of a given state and situation. Furthermore, we find that emotions are very difficult to describe and communicate due to the subjective nature of how individual people associates a particular type of emotion. Finally, we find that emotions are dynamic in the sense that the internal states of a person is never static and are often in a state of flux from a temporal perspective, therefore making it hard to evaluate and understand in a concrete manner.

Despite the recent improvements in the various modeling techniques offered by the power of Deep Learning based systems, they are still subject to difficulty with regards to interpretability due to

the distributed representative nature of the parameters - often said to be treated in a black-box fashion. Furthermore, many state-of-the-art models pertaining to human emotion recognition systems only utilize predominantly two different representations: discrete and continuous states. Although these representations have been widely adopted in various applications as label features for human emotional states, they are considered to be insufficient towards practical applications in building models for human-facing interfaces to interact in a "human-like" manner. However, current research efforts has not worked towards making significant progress in improving addressing these problems. Hence, the fundamental problem that we must address in mitigating this issue is to devise a new feature representation for human emotions and formalize notation, modeling methodologies, and application methods for use in various computational human emotion recognition, understanding, and generative tasks.

In this work, we propose a novel feature representation for human emotions as a distributed vector representation which holds various practical and interesting properties for use in computational processes and human interpretation. Furthermore, we formally define the notation and methodology to construct this representation and devise an encoder-decoder framework which can be applied universally to any models or problem which makes use of human emotion representations. We follow our investigation for the validity of this modeling methodology through constructing this representation using two data sets, the EMOTIC Dataset and the BoLD Dataset. Through constructing our feature representation under two different contexts, we can effectively evaluate the relative empirical ability of the model to better represent emotions in this new fashion. We perform various tasks and develop applications to demonstrate the overall plausibility for use in various computational psychology and HCI tasks.

The rest of this paper is structured as follows. In Section 2 we evaluate the advantages and disadvantages of various types of representations utilize in current state-of-the-art models. We outline some key related work and inspirations which were drawn in Section 3. In section 4 we describe the two datasets we utilized for our study and evaluation of our model. We introduce notation, modeling methodology, and the encoding and decoding framework for the EVE Model in Section 5 and present our empirical findings and results in Section 6 and conclude our paper in Section 7.

2 BACKGROUND

In this section we introduce some preliminary background in the various types of emotions representations utilized in many of the state-of-the-art models - primarily Discrete and Continuous representations. Each of the models presented in here originate from various models of emotions proposed in For each of the different representations, we evaluate each of the representation based on its advantages as well as its disadvantages.

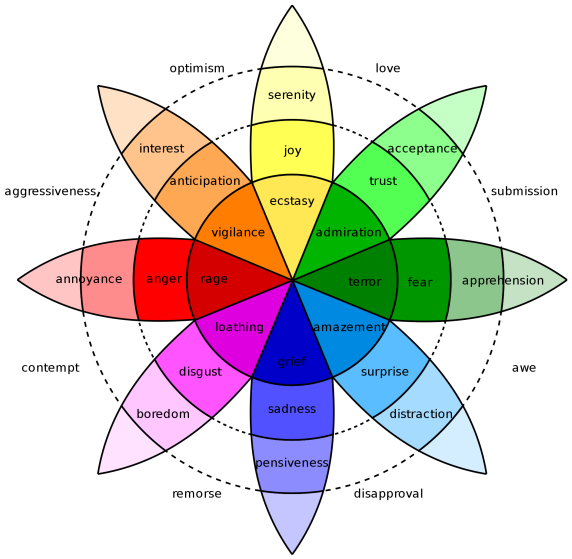
2.1 Discrete Models

Discrete representations are primarily based on a concrete fixed set of adjectives or terms which explicitly describes the human state as some categorical representation. One of the most dominant and common representation used in emotion proposes for a set of six "basic" emotions which are internally hard-wired into the human brain, proposed by Paul Ekman [7]:

ANGER, DISGUST, FEAR, JOY, SADNESS, SURPRISE

Robert Plutchik [18] further extends Ekman's model for emotions, allowing for "complex emotions" to emerge - which entails the emotion of one state to be comprised of a compound of another emotion. This model is often represented in the form of "Plutchik's Wheel of Emotions", where the hierarchy of emotions are built on top of one another as shown in **Figure 1**.

Figure 1: Plutchik's Wheel of Emotion



These emotions, are often thought of to be fundamentally distinct and rudimentary at it's core. Variants of these six different types of emotions are often used in simple classification models, often represented in a one-hot encoded vector representation. While this reduced set of emotions allow for simplicity and ease of interpretability. Bhan et. al [1] notes that some emotions such as "surprise" may be harder to semantically distinguish as one maybe *happily surprised* or *angrily surprised* and adds to the notion that context plays a huge role in some of the terms used in these representations and how linguistic vagueness and relativity is a prominent issue in using such representations for emotions.

However, recent work in affective computing have begun to explore and utilize much more other varieties of emotional words and have attempted to encompass a more words in their categorical labels. Kosti et. al [13] in his work utilized a hand-crafted heuristic to determine new categories of words to utilize and formulated a list of 26 different emotion categories which are also represented as a multi-one-hot encoding schema. Although further building a corpus and adding new emotions to the binary encoded vector schema can work in principle, we find that: 1) hand-crafted features often do not scale well for additional new word terms and are considered fixed once defined, 2) the representational encoding in a

computational setting leads to issues of sparsity making it difficult from a computational and mathematical perspective, and 3) suffers from semantic interpretations and interoperability as it is difficult to compare one emotion from another quantitatively.

2.2 Continuous Models

Another prominent feature representation method for emotthere are ion representation is the continuous or dimensional model. In this model, instead of representing each emotion as a categorically distinct word or state, we instead attribute each element in the vector as a continuous value representing the presence of a given feature for a particular emotional state.

The most widely adopted representation for continuous values adopted from Russell [19] include: **Pleasure** (other times referred to as valance), **Arousal**, and **Dominance**. Pleasure refers to the measurement of how positive or pleasant an emotion is, typically ranging from some negative to positive value. Arousal measures the agitation level of the person, ranging from non-active to ready to act. Finally, Dominance measures the control level of the situation by the person, ranging from submissive (non-control of the situation) to dominant (in-control of the situation).

Typically each of the dimensions are normalized between some defined range, such as 1 to 10 or between 0 to 1. In this representation, we find the vector representation to be much more dense and interoperable for computers to perform calculations on. However, in the development of such dimensional representations we come across two issues: 1) the representation used in this manner is not entirely easy and intuitive to understand and 2) each dimension is hand-crafted based on some feature which is often hard to scale and consistently acquire with regards to annotating datasets using this notation. For instance, when describing how someone feels, we would not say, "Jim is feeling a valence of 2, a arousal of 3, and dominance of 4". Instead, we would most likely use phrases such as "Jim is feeling sad and depressed because his grandmother recently passed away" - which is a much more naturalistic and easier way for humans to interpret emotions.

2.3 Representation Synthesis

Having evaluated and identify each of the model representation's strengths and weaknesses, we look towards combining the best properties from the representations while also mitigating weaknesses of the particular representation.

A desirable model representation would possess both these key properties where we can manipulate the inputs and outputs easily in a computational and a human-interpretable form. In short, we can distill the above these properties of a desirable emotion model as mostly described by Bhan et. al [2]:

- Emotion representations are to have bipolar relations accounting for much of the variance in the model [19].
- Emotions represents a form of "qualia" which can be dimensionally represented in a vector space [2].
- A emotional state can be comprised of a linear combination of other emotional state representations [19][18].
- Emotional states have a semantically dependent structure to them in which they are not defined within an independent construct, but in a relative manner in conjunction with other emotions.

For this, we look to using a vector space based approach proposed by Bann et. al [1][2], who has previously explored the use of *semantically distinct emotions* which is based on the theory that emotions are communicated by language and that they are mostly informed by the linguistic context in which they use the words to describe their language.

As opposed to the orthodox top-down methodology utilized in psychology, where the objective is to model emotions down to a set of core components of minimalistic emotions, our modeling technique takes a bottom-up approach by embracing the quantitative variety of emotions - theoretically giving these models a dynamic capacity to describe almost every possible emotions at such spectrum of high fidelity.

We build on these fundamental models towards further evaluating and developing a robust framework for devising a semantically better representation with a higher fidelity of emotional states as a discrete spectrum. In the next section, we further build upon other alternative and relevant theories of emotions as well as a modeling framework to formally define a working framework to best model these representations.

3 RELATED WORK

In this section, we evaluate additional areas of inspiration and contrast against similar methodologies in modeling based on "qualia"-like features. We draw fundamental inspiration from linguistic and color theory supported further by Kansei Information Modeling methodology, a modeling technique developed for mapping qualitative states to numerical parameters often utilized in user product development processes.

3.1 Color Theory

Emotion and color are very linked in the sense that both share properties of being continuous and discrete. On one hand, color can be seen based on a spectrum of continuous values, where there are close to infinite color states on some broad subspace. Color also holds various dimensional properties such as contrast, intensity, and different spectrum. However when communicating these color states, we often point to one shade and often treat them in some discrete state.

For instance, given only one of hundreds of shades of red, we often attribute the color as being discretely red. However, given ten different shades of red they can be placed in some ordinal degree. Emotions are similar, in the sense that one can be happy at any one of many different levels, but we often attribute them as either happy or perhaps joy or ecstatic dependent on our subjective capacity to describe these emotions.

Various studies [16] [8] [21] have demonstrated this strong overlap of the fundamental properties shared between color and emotion. Primarily, we see that they pose both similar attributes in their representation, dimensional degree, as well as implications of contextual differences both situationally and culturally [17].

3.2 Linguistic Theory

In essence, language has been used as one method to tangibly best describe and approximate the experience or mental state of the given person, given the linguistic capacity of the said language

the person has available to them. Kazemzadeh [12] evaluates the relationships between the fuzzy-sets, one in English, the other in Spanish of the different emotional states. As both fuzzy-sets were presented in context, there was a great deal of similarity found between specific Spanish emotions and more general English emotional categories.

Recently, however, many works have begun to migrate away from linguistic representations of emotion, and instead evaluate other tangible modalities such as pictorial representations - notably emojis. With the prominent use of these symbolic representations used to describe emotional states, works such as that of Wijeratne et. al [22] [23].

More relevantly, the capability of such distributed vector representational methods for describing emotions using these pictorial symbols to best encapsulate semantic represent these emotional states. Eisner et al [6] proposes emoji2vec, a model which learns sentiments of emojis by vectorizing the words in the descriptions which come baked into the encoding scheme. This is based on the Word2Vec model, which creates vectors of probabilities that, given a word and a distance from it, the likelihood that a certain word will fill that role. Each word therefore also has a vectorized representation of its likelihood to be the target word given a distance from the target and the word in that spot. As a result, these models have shown high potential for use in modeling emotional states and have shown great success in encapsulating the subtle semantics of emotional states within a pictorial context.

Our model attempts devise a generalized framework which formalizes a method to construct learned feature representations to produce a distributed vector model to model these various emotional properties from a linguistic standpoint. In doing this, our goal is to circumvent the language barrier - both quantitatively and qualitatively, improve subtle semantic interpretations of emotions, as well as enabling the model to adaptively and learn and improve its own representation in a continuous manner.

3.3 Kansei Information Processing

One of the challenges in building models which translate these qualitative attributes or often referred to as "qualia" or the *relative placement of emotional states based on an individual threshold*, to quantitative parameters lies in the robust methodology that is required to build these types of models. For this, we look into a methodology known as **Kansei Engineering** or more generally **Kansei Information Processing**, which is a statistical process developed by Nagamachi [15] for Nissan Motors to help translate producers translate consumer's implicit qualitative needs to physical parameters pertaining to the target products that they are designing for the consumer. The term "Kansei" is based on a Japanese term used to describe the overall *impression* of a particular entity or state - often this is translated as "sensitivity", "sensibility", feeling, and emotion in various domains of psychology.

The Kansei Engineering (KE) framework provides a guiding reference for formulating a formal methodology to construct these models. We utilize a similar methodology to model the human emotional state as they have shown to be very successful in successfully modeling products which meets the requirements and demands of a consumer qualitatively Schütte [20]. Notably the Kansei Methodology helps to produce an end model such that one can translate between the user's qualitative domain to the numerical parameters it attempts to estimate.

Towards developing such Kansei Information Processing model, the model development process is phased as follows:

First, we start with deciding the *Kansei Domain*, or the content of what we are attempting to measure must be selected. In this case, we are evaluating "human emotions" as our primary domain.

Prior to understanding what we are attempting to measure, we must then devise a set of fixed corpus known as *Kansei Words*. In this phase we determine the semantic spanning of the topics and formulate a set of corpus that we will use, usually in the form of adjectives or a noun. For the EVE Model, we define this set of corpus as any emotion words that can humans use to describe their own feelings. In particular, we do not impose an upper limit to how many words we use to model this in our framework.

Third, in the Kansei Information Processing model, we then look to evaluation of how to measure the various Kansei Words and how they associate to the targeted measured parameters - a process often referred to as *Feature Space Spanning*. The key idea behind this process is to provide as much sample independent instances to show various relationships between the inter-relationships of each of the Kansei Words and how they are related to each other. In our modeling process, this would be highly dependent on the model we are building. In particular we propose two different method, one which requires a collection of discrete to continuous valued mappings, and another which only requires a simple discrete valued representation.

Finally, once we define the Domain, Kansei Words, and the Feature Space instances, we then build our model in a process known as *Synthesis*. The goal of the synthesis process is to build a relational model which correlates between the Kansei Words and the parameters we are attempting to measure. Some of the most common modeling methodologies often use include: Fuzzy Logic, Neural Networks, Genetic Algorithms, Rough-Set Analysis, Regression Algorithms, and Partial Least-Square Analysis. In our modeling process we introduce two methodologies, given the different type of Feature Space instances we are provided. In this model, we aim to develop a vector space model which allows to map discrete space on a multidimensional subspace which can allow us to quantitatively measure the semantics between two discrete emotion words geometrically.

4 DATASET

In this section, we introduce the two datasets we have utilized in our experiments to validate the EVE modeling representation and methodology: the EMOTIC Dataset and the BoLD (Body Language Dataset). Both datasets are curated within the context of understanding the emotions of human body language. For the EMOTIC, this is represented as simply static images, while for the BoLD, we annotate the data based on short video clips.

In both sets, we take only properly labeled, non corrupted data. Both sets contain just over approximately 25,000 annotations samples, however each within a similar but slightly different context such that one annotates human emotions from a static image, while

another evaluates them on a short video clip. In our following experiments, we develop two EVE models for each of these different datasets, such that we collect instances of the emotion words and their co-occurring relationships in various contexts based on a fixed collection of emotion corpus we have defined.

As for the type of labels we use to annotate the dataset both use the same 26 discrete emotion categories, as well as the same continuous metrics of Valence, Arousal, and Dominance (VAD).

Specifically, the 26 discrete emotions are described to those labeling as follows:

- **Peace**: well being and relaxed; no worry; having positive thoughts or sensations; satisfied.
- **Affection**: fond feelings; love; tenderness
- **Esteem**: feelings of favorable opinion or judgment; respect; admiration; gratefulness
- **Anticipation**: state of looking forward; hoping on or getting prepared for possible future events
- **Engagement**: paying attention to something; absorbed into something; curious; interested
- **Confidence**: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
- **Happiness**: feeling delighted; feeling enjoyment or amusement
- **Pleasure**: feeling of delight in the senses
- **Excitement**: feeling enthusiasm; stimulated; energetic
- **Surprise**: sudden discovery of something unexpected
- **Sympathy**: state of sharing othersâĂŹ emotions, goals or troubles; supportive; compassionate
- **Doubt/Confusion**: difficulty to understand or decide; thinking about different options
- **Disconnection**: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
- **Fatigue**: weariness; tiredness; sleepy
- **Embarrassment**: feeling ashamed or guilty
- **Yearning**: strong desire to have something; jealous; envious; lust
- **Disapproval**: feeling that something is wrong or reprehensible; contempt; hostile
- **Aversion**: feeling disgust, dislike, repulsion; feeling hate
- **Annoyance**: bothered by something or someone; irritated; impatient; frustrated
- **Anger**: intense displeasure or rage; furious; resentful
- **Sensitivity**: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
- **Sadness**: feeling unhappy, sorrow, disappointed, or discouraged
- **Disquietment**: nervous; worried; upset; anxious; tense; pressured; alarmed
- **Fear**: feeling suspicious or afraid of danger, threat, evil or pain; horror
- **Pain**: physical suffering
- **Suffering**: psychological or emotional pain; distressed; anguished

4.1 EMOTIC Dataset

The first dataset is the EMOTIC dataset, provided publicly through the University of Oberta [13]. This dataset contains static images of various human subjects, with each a bounding box defined around them (provided to us by x, y, width, height). All three categories are tagged on the continuous VAD scale, as well as 26 discrete emotional categories. The data was annotated through a study they have curated through a paid study conducted on the Amazon Mechanical Turk (AMT) platform.

The emotions listed are varied widely across the three subsections of this set, as the training set is singly labeled. On average, no more than 3 emotions are put on any given image. When we look at the test set, however, each of these has been labeled by multiple respondents, thus resulting in an increased average of unique emotions per images.

This dataset was developed towards a trained a Deep Learning based CNN model on which the researchers report a low error rate on predicting more than 25% of emotional categories. The majority of the remaining categorical emotions, they posit, are being predicted well, but not as well as the top quarter emotional states.

4.2 BoLD Dataset

The BoLD or **Bo**dy **L**anguage **D**ataset is another dataset we have utilized for this project. This dataset was curated within the same evaluational context as the EMOTIC dataset, where we utilized the same annotation schema - except we now evaluate the data based on short video clips instead of static images.

This dataset contains short videos, and a bounding box on each person of interest who is visible throughout the video. There are also tags available which indicate that the person is obstructed or if the bounding box partway through the video switches focus from one person to another.

5 EVE MODEL

In this section we formally introduce the Emotion Vector Encoding (EVE) Model along with the key notation which we will be referencing throughout the paper, then presenting the type of models and methodologies we utilized in the EVE model for formulating this new representation.

The EVE model is comprised of the model, which is based on a mapping between a symbolic representation of an emotional state a to vector space representation in some subspace. Based on this model, we have an encoder and decoder function which would be defined to convert between the symbolic representation to a vector representation and vice-versa. Specifically, we describe two encoding methods: the *Mean Vectorization Model* and the *Word Embedding Model* variant. As for the decoder model, we introduce the KD-Tree method which is used as a method to convert back the vector based representation to a discrete symbolic form - even if the corresponding vector form is not exactly the same value as defined by the original encoder value.

5.1 Notation

Let $\mathcal{M}$ be defined as an emotion vector encoding model, which is constructed based on some defined model or algorithm based on set of emotion corpus set C, which contains a set of words w, hence forming the set $C = \{w_1, ..., w_N\}$, where N is the number of words contained in the emotion corpus.

Given model $\mathcal{M}$, we define $\mathcal{M}_e(X)$ as our encoder function such that X is a symbolic input - usually a word term, w, where $w \in C$. The resulting value of $\mathcal{M}_e$ is a vector based representation of the symbolic representation of the emotion.

With an encoder function to convert between the symbolic notation to a vector based representation, we also define $\mathcal{M}_d(\hat{x})$ as a decoder function which performs the inverse operation of converting a vector representation, $\hat{x}$, back to the symbolic representation.

However, it must be noted that model $\mathcal{M}$ is not a one-to-one correspondence and therefore the decoder function would be a function which attempts to approximate or retrieves the symbolic representation with some degree of error, which can be defined as a threshold parameter of the decoder function or can be defined as a top k-nearest approximation to the given vector input $\hat{X}$. From this, we then would formally define this fact as $\mathcal{M}_d(\mathcal{M}_e(x)) \approx x$.

5.2 Encoder

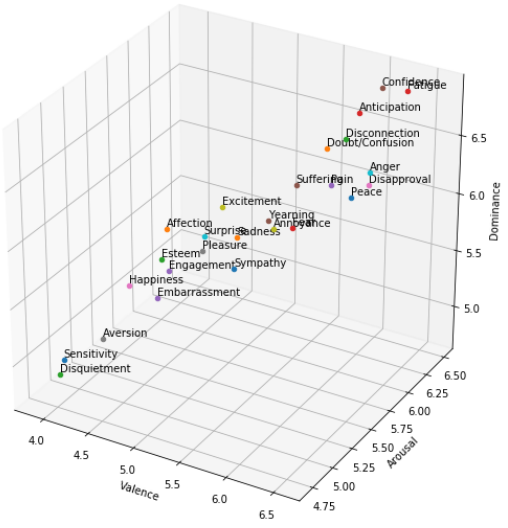
In this subsection we introduce two encoder variants: the **Mean Vectorization Model**, a simple model which synthesizes both discrete and continuous value paired labels and the **Word Embedding Model** which learns a dense vector representation from a collection discrete value co-occurrences.

5.2.1 ***Mean Vectorization Model (MVM).*** Our initial approach to evaluate the plausibility of synthesizing the discrete and continuous representation, we first utilize a simple and naive approach where we obtain the cumulative *independent* average of the valence, arousal, and dominance averages for each discrete emotion state. Specifically we make the naive assumption that the VAD for each emotion is independent to simplify the computation behind defining the vector space location.

However, in both EMOTIC and BoLD, the discrete labels are mutli-hot encoded vectors, or vectors which contains more than single binary values activated within the same bit-vector representation. Therefore, for each multi-hot encoded and VAD pair, we decompose the vector into multiple one-hot encoded VAD pairs for every activated bit in the multi-hot encoded value.

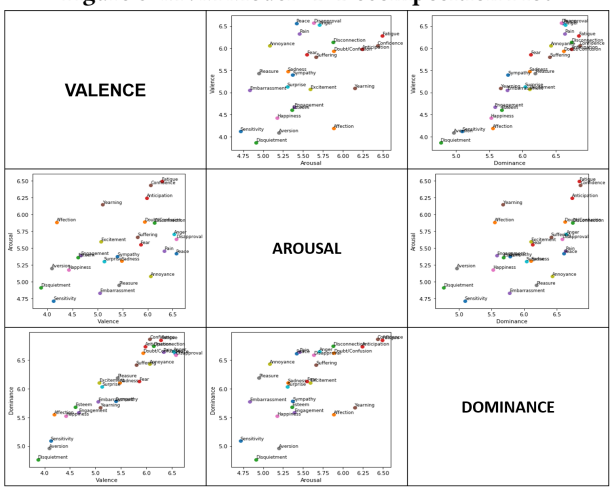
Hence for each of the discrete N emotion value, we compute the mean values for each of the corresponding dimension in the continuous labels. As a result, this would give us back a projection mapping of the discrete emotion on a vector space as demonstrated in **Figure 1** and **Figure 2**.

Figure 2: MVM Model 3D Plot



5.2.2 ***Word Embedding Model.*** Another word vector model representation that we can leverage is based on the distributed vector representation model proposed by Mikolov et. al [14] or often referred to as word embeddings. Recent advancements in Natural Language Processing has shown significant progress with the use of word embeddings as feature extraction processes for

Figure 3: MVM Model 2D Decomposition Plot



modeling word semantic features. We find that this representation works very effectively well and supports many of the desirable features which enable a much better representation for emotions.

Word Embeddings are based on a Deep Learning based model representation for learning term co-occurrence probabilities found in a given document set for a given corpus. By modeling the likelihood that a given word would appear within the same context, we can develop a high-dimensional and distributed embedded vector representation for each term - such that words that tend to appear often together would geometrically cluster within the same regions while words that often do not appear within the same context would separate from each other.

In this work, we follow a very similar method to construct these word embeddings as proposed by Mikolov et. al. However, subtle attention to training must be made as the raw data used to train on this data are not sentences, and instead based on a very small and fixed-size multi-hot encoded vector where the order of the entities are not important in this model. Therefore, we outline some key practices we followed when we trained our word embedding model for the EVE model.

To model a word embedding based model in the context of emotions, we model the occurrence for how likely certain emotions would co-occur with each other. We model a conditional probability based on the Softmax probability function as defined below:

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{\exp {v'_{w_O}}^{\top} v_{w_I}}{\sum_{w=1}^{W} \exp {v'_{w_O}}^{\top} v_{w_I}}$$

This function describes how likely a emotion is to occur given another particular emotion. This is formulated by all the ways that the specific conditional probability of the given emotion can co-occur, divided by the sum of all likelihoods of that emotion occurring in all instances.

Given this probability function, we perform stochastic gradient descent optimization based on our cost function that we try to minimize:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

To train these embedding models, there are several hyperparameters that we must tune, with the given constraints and assumptions we can make about the dataset we are training our model on. First, one of the most important parameter would be the dimension of the latent variable that we are training our models on. Heuristics for training word embeddings in many applications usually choose a dimension of 150 to 200. However, since the dimensionality of our corpus, N, and sample size would be significantly smaller than a typical corpus set used for normal Natural Language Processing tasks, we would have to proportionally scale our latent vector dimensions to control the amount of sparsity present within the distributed vector representation.

For each epoch of training, we take all of the emotions given the given sample set, and perform a random shuffling of the order. Since the raw representation of the model encountered these emotions independently of order, we considered the whole process agnostic of the order that they were presented in. We also find that this process empirically helps prevent our model from over-fitting.

Hence, another key hyperparameter we would have to focus on was the window size for the model. This was also another parameter which careful attention must be made, as a key underlying assumption was that the order of the word terms do not matter. The parameter tuned for this model would need to be carefully adjusted based on the average proportion of activated bits in the overall model in order to consider the likelihood that the model would equally likely encounter the co-occurring terms.

Therefore, our window size is set to be 2 as our corpus size dimension was fairly small, and we require that each training notation have at least 2 unique emotions present. In this way, our training is faster from not needing to look from $Word_{t-w}$ all the way through to $Word_{t+w}$ and adjust the vector locations for double the width of the window words, where t is the target emotion being predicted and w is the context emotions used as a feature to predict the target.

5.3 Decoder

In this subsection we introduce a decoder function which we can use to reverse the vector based dense representation of the emotions back to a discrete and tangible symbolic representation, which we have defined as $\mathcal{M}_d$. We made note earlier in the paper of a property of the EVE model that our modeling function does not hold a one-to-one correspondence between the emotion representation, and therefore the returned value would need to account for a certain degree of error tolerance when reconstructing back the discrete emotional state - described by the notation: $\mathcal{M}_d(\mathcal{M}_e(x)) \approx x$.

One algorithm we can make use of to efficiently index and to perform a reversal function from a vector representation to a discrete point is by using a KD-Tree [3]. A KD-Tree, as the name implies, is similar to a typical tree based data structure, except is much more generalized to a higher dimensional subspace. KD-Trees are often used as a spatial partitioning algorithm often used in state-of-the-art search algorithms to perform searches in higher dimensional spaces. Using this algorithm would prove to be faster with regards to runtime and improve the overall computational runtime of the decoder model we have built. Given N elements within a k dimensional subspace, the worst case theoretical runtime of the algorithm is said to be of the order of $O(k \cdot N^{1-1/k})$ making it an acceptably fast search time for large subspaces.

Concretely, this decoder function can be utilized as a method to also find the k-Nearest emotions given a vector embedding generated by a regression-based predictive model. Hence, this decoder module can be appended as a function which can be used as an interface to help assist humans communicate and interpret results for building interpretable Deep Learning models by providing them with a semantic suggestion of emotional states the particular entity might be feeling in the provided input for the model. This as a result helps to improve the overall interpretable recall value for the potential candidate of responses of a model, hence providing sensible results that humans can make better sense of. We later demonstrate this capability in under **Section 6** of our paper.

5.4 Model Similarity Metric

To evaluate the plausibility of the universality of the EVE model, some similarity metric must be utilized to compare between two sets of emotion model pairs to evaluate how similar the co-occurring probabilities are between the inter-relational pairs of N emotions. However, in the study of psychology this is often very difficult to achieve and standardize, but significant to ensure that experiments in this domain must be consistent across similar samples. For comparing similarity for a pair of vector vector representations of emotion words w_1 and w_2, we can find it's similarity by the cosine similarity function:

$$cos(w_1, w_2) = \frac{w_1 \cdot w_2}{||w_1||\ ||w_2||}$$

If emotion co-occurrence is universal, the models should be everywhere similar in their representations of emotion words. However to perform a cosine similarity metric across an entire corpus, we would need to define a two-fold similarity measurement process which can compare the similarity between two EVE models. Hence, we define a new metric which will help us to compute the similarity between two EVE models, represented as $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$. We let superscript k denote our model index value, either 1 or 2 for the two models we are computing the similarities against.

Given a corpus set of N unique emotion words and a single EVE model, $\mathcal{M}$, we first construct a inter-relational similarity matrix for the EVE model $\mathcal{M}$. This process is then repeated for all i and j in the $N \times N$ matrix S and for both models.

$$S_{ij}{}^{k} = \frac{\mathcal{M}_e{}^{k}(w_i) \cdot \mathcal{M}_e{}^{k}(w_j)}{||\mathcal{M}_e{}^{k}(w_i)||\ ||\mathcal{M}_e{}^{k}(w_j)||}$$

The result of this process for both models will be $S^{(1)}$ and $S^{(2)}$, symmetric matrices describing the cosine similarities between every encoded emotion word in our corpus.

We then normalize the similarity matrix by the mean and standard deviation to appropriately scale the scores as the sparsity of the matrix may differ from one model to the other. Then, we compute the row-wise cosine similarity between both models for each corresponding word N in the matrix and compute the average similarity for each similarity for each of the inter-relational similarities

between the two EVE models. We therefore define that to be the similarity metric of two EVE models as defined:

$$sim(\mathcal{M}_1, \mathcal{M}_2) = \frac{1}{N}\sum_{i=1}^{N} cos(S_i^{(1)}, S_i^{(2)})$$

Therefore, the resulting range of this function should return a value between 0 and 1. The closer this similarity is to 1, the more likely that our models are similar and that they are describing emotions in a very similar manner - hence implying the plausibility for our model to describe emotions in a consistent manner (given a similar contextual collection of emotion states). Conversely, the closer it is to 0, the less likely that our models are similar to each other and that they do not exhibit a similar type of behavior in regards to modeling emotions, hence not displaying universality between two contextual sample spaces.

6 RESULTS AND EVALUATION

In this section we present our results and evaluations for each of the models we have presented. To evaluate the EVE models we have defined, we present various experiments to evaluate our models. We perform two key evaluation processes entailing both a qualitative as well as a quantitative assessment for the performance of our models.

We first evaluate the plots generated by the Mean Vectorization Model to evaluate visual clusters and emerging patterns and observations we can draw from development of such vectorization processes. For qualitatively evaluating both our models, we use a k-Nearest Neighbors approach to provide us with the top 5 nearest emotions using the KD-Tree decoder model we have proposed in the previous section. Then we perform a model similarity evaluation using the metric that we have also defined in the previous section across both the Mean Vectorization and Word Embedding Model for Both the EMOTIC and BoLD Datasets.

6.1 Visual Observations

In the previous section, we introduced two plots generated by plotting our Mean Vectorized Model - one in a three-dimensional euclidean plot and another one based on a two-dimensional projection plot of each feature separately as shown in **Figure 2** and **Figure 3**.

As shown in **Figure 2**, we can see two general clusters based on the valence of the emotion as suggested in the properties of our emotion representation. We can see that the inner clusters of points are mostly positive emotions, while the outer cluster of emotions mostly comprise of negative emotions. Hence, we can see that even such simple averaging models help us to formulate sensible semantics of emotions in relation to other emotions present.

Looking at **Figure 3**, we can decompose the plot and observe several other key trends and patterns which emerge from the dataset. First, one of the two salient properties we can see in the plots are the amount of variance each pair of dimensional pairs generate. In particular we see that the variance between *Valence* and *Arousal* are the greatest and the variance between *Valence* and *Dominance* is the smallest. These trends correlate to some of the previous studies and theoretical formulations of the previous models that we have evaluated in the previous sections of the paper.

6.2 K-Nearest Neighbor Evaluation

For qualitatively evaluating the emotion qualia, we use the K-Nearest Neighbor method to show us the top emotions the model renders given a emotion word. In a practical application, such as classification for identification or emotions, posing a single emotion may not be sufficient in certain context and therefore a proposal for several different emotional states should be provided. Furthermore, from an HCI standpoint, posing various relevant candidate emotions than a single fixed emotion can further bridge a naturalistic interaction interface between the user and the computer system.

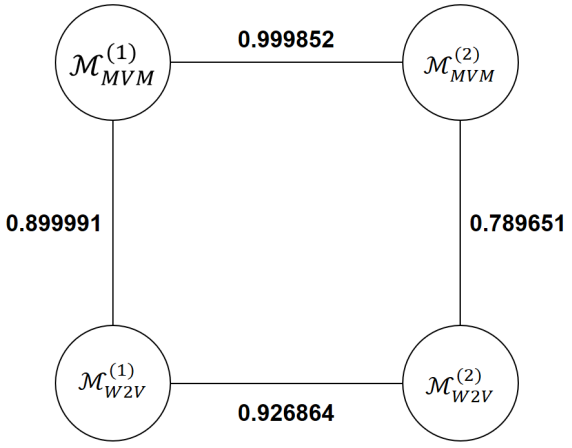
We demonstrate our empirical results in both **Table 1** and **Table 2**. We find that much of the similarities in Table 1 have relatively high similarity rates due to the small dimensionality of the emotion representation. However, we still find that their relative results to be relatively decent, though with an exception of emotions like Pleasure having Sadness which does not really make sense from a semantic standpoint. We find that in both sets of the results between the EMOTIC and BoLD dataset, we do not find much overlap in the emotions that show in the k-Nearest Neighbor results.

On the other hand, the word embedding model shows a much better and sensible result with a better qualitative sense in which results appear at the top. In particular, for the word embedding model, even with the vector summation of the two emotions ("Fear" + "Sadness") as an input to the encoder provides relevant results which makes sense - implying our model being able to capture even compositional and subtle properties of emotion qualia as proposed earlier in the ideal model representation for human emotions.

6.3 Model Similarity Evaluation

In the previous section we defined a similarity metric which can be used to quantitatively compare the similarity between two given models. In this evaluation, we will use this metric to compare the similarity between the models generated between the two different datasets and the modeling encoder variations.

Figure 4: EVE Model Similarity Graph



We represent the similarity relationship as a graphical model demonstrated in **Figure 4**, where each vertex represents a model $\mathcal{M}^{(k)}$ trained on a particular dataset, k, where $k = 1$ is the EMOTIC

Dataset and $k = 2$ is the BoLD Dataset and the edges representing the EVE model similarity metric we defined earlier. The horizontal edges along the graph represents the relationship between the models which are trained on different datasets, but utilizing the same encoder algorithms. On the other hand, the vertical edges on the graph represent the relationship between two models which have been encoded on different encoders but on the same dataset.

In this graph, there are several interesting properties we can interpret from the results we have obtained. First, the vertical relationships between the two datasets implies that the EMOTIC dataset has fewer variance as opposed to the BoLD which has a much higher variance in the dataset since the EVE similarity is weaker. However, the high cosine similarity for each of the two horizontal edges implies that even with a different encoding algorithm, the relative semantic structure for these emotion vectors are placed in a relatively similar way.

Looking at the horizontal edges on the graph, we see how the similarities for both edges are relatively high. For the Mean Vectorization Methods - we find that the similarity of this to be very high due to the dense nature of the representation. However, the relatively high cosine similarity still validates the fundamental psychological theory of Valence, Arousal, and Dominance being a key feature in emotion modeling.

However, what we find more interesting is the relationship between $\mathcal{M}^{(1)}_{W2V}$ and $\mathcal{M}^{(2)}_{W2V}$, which are the two models encoded based on the word embedding model, but on two different dataset contexts. Given that word embeddings are primarily a unsupervised approach for learning semantic features from a collection of samples, what this implies is that even from two different sample spaces, we have found that our model is able to effectively model human emotions achieving a relatively high degree of semantic similarity. This implies that even in a different semantic feature space, our emotion vectorization encoder is able to effectively learn and encode subtle nuances of emotions as a vector space representation.

7 CONCLUSION

Our contributions in this paper includes:

- A novel framework for encoding and decoding human emotions as a distributed embedded vector representation.
- A modeling methodology for emotion representation using distributed vector representations.
- Proposed algorithms which can be utilized to both encode and decode between these different representations.
- Devised a feature representation which embodies various properties from both continuous and discrete emotion representations as well from other areas in color and linguistic theory.
- Visually empirical, qualitative, and quantitative experiments which demonstrates the feasibility of this representation to be used in various affective computing models.

In this paper we have introduced and formalized a framework for developing an encoder and decoder model for human emotional states - opening a new research direction towards affective computing to improve human emotion state representations.

REFERENCES

[1] Eugene Yuta Bann. 2012. Discovering Basic Emotion Sets via Semantic Clustering on a Twitter Corpus. *CoRR* abs/1212.6527 (2012). arXiv:1212.6527 http://arxiv.org/abs/1212.6527

[2] Eugene Yuta Bann and Joanna J. Bryson. 2013. Measuring Cultural Relativity of Emotional Valence and Arousal using Semantic Clustering and Twitter. *CoRR* abs/1304.7507 (2013). arXiv:1304.7507 http://arxiv.org/abs/1304.7507

[3] Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM* 18, 9 (1975), 509–517.

[4] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 205–211.

[5] Liyanage C De Silva, Tsutomu Miyasato, and Ryohei Nakatsu. 1997. Facial emotion recognition using multi-modal information. In *Information, Communications and Signal Processing, 1997. ICICS., Proceedings of 1997 International Conference on*, Vol. 1. IEEE, 397–401.

[6] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359* (2016).

[7] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[8] Andrew J. Elliot and Markus A. Maier. 2014. Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. *Annual Review of Psychology* 65, 1 (2014), 95–120. https://doi.org/10.1146/annurev-psych-010213-115035 arXiv:https://doi.org/10.1146/annurev-psych-010213-115035 PMID: 23808916.

[9] Irfan A. Essa and Alex Paul Pentland. 1997. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 19, 7 (1997), 757–763.

[10] Dariu M Gavrila. 1999. The visual analysis of human movement: A survey. *Computer vision and image understanding* 73, 1 (1999), 82–98.

[11] Hatice Gunes and Massimo Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30, 4 (2007), 1334 – 1345. https://doi.org/10.1016/j.jnca.2006.09.007 Special issue on Information technology.

[12] Abe Kazemzadeh, Sungbok Lee, and Shrikanth Narayanan. 2013. Fuzzy logic models for the meaning of emotion words. *IEEE Computational intelligence magazine* 8, 2 (2013), 34–49.

[13] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] Mitsuo Nagamichi. 1992. Kansei Engeering and Fuzzy Theory. *Journal of Human Engineering* 28, Supplement (1992), 32–33.

[16] Niels A Nijdam. [n. d.]. Mapping emotion to color. ([n. d.]).

[17] Li-Chen Ou, M Ronnier Luo, Andrée Woodcock, and Angela Wright. 2004. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application* 29, 3 (2004), 232–240.

[18] ROBERT PLUTCHIK. 1990. Chapter 1 - {EMOTIONS} {AND} PSYCHOTHERAPY: A {PSYCHOEVOLUTIONARY} {PERSPECTIVE}. In *Emotion, Psychopathology, and Psychotherapy*, Robert Plutchik and Henry Kellerman (Eds.). Academic Press, 3 – 41. https://doi.org/10.1016/B978-0-12-558705-1.50007-5

[19] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[20] Simon Schütte. 2005. *Engineering emotional values in product design: Kansei engineering in development*. Ph.D. Dissertation. Institutionen för konstruktions- och produktionsteknik.

[21] Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of experimental psychology: General* 123, 4 (1994), 394.

[22] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. Emojinet: An open service and api for emoji sense discovery. *arXiv preprint arXiv:1707.04652* (2017).

[23] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. A semantics-based measure of emoji similarity. *arXiv preprint arXiv:1707.04653* (2017).

Table 1: Top 5-Nearest Emotions from MVM [EMOTIC (Top) & BoLD (Bottom)]

Anger		Pleasure		Excitement	
Disapproval	0.999964	Embarrassment	0.999774	Sensitivity	0.999898
Pain	0.999842	Sadness	0.999774	Esteem	0.999898
Peace	0.999730	Disconnection	0.999391	Happiness	0.999734
Fear	0.999477	Annoyance	0.999359	Engagement	0.999719
Annoyance	0.999467	Pain	0.999358	Confidence	0.999688

Anger		Pleasure		Excitement	
Aversion	0.999189	Affection	0.999872	Anticipation	0.998858
Disapproval	0.997003	Happiness	0.999858	Engagement	0.998493
Annoyance	0.996613	Esteem	0.999110	Esteem	0.997461
Suffering	0.994039	Peace	0.998402	Sympathy	0.997130
Disquietment	0.991150	Excitement	0.994050	Affection	0.995656

Table 2: Top 5-Nearest Emotions from W2V [EMOTIC (Top) & BoLD (Bottom)]

Anger		Pleasure		Fear + Sadness	
Aversion	0.88505	Esteem	0.827299	Fatigue	0.895345
Embarrassment	0.85801	Sympathy	0.563033	Pain	0.894816
Disapproval	0.83252	Anticipation	0.542841	Embarrassment	0.888998
Doubt/Confusion	0.77493	Confidence	0.506502	Sensitivity	0.840702
Disconnection	0.71646	Yearning	0.500476	Disapproval	0.720077

Anger		Pleasure		Fear + Sadness	
Disconnection	0.41355	Esteem	0.523179	Pain	0.520315
Doubt/Confusion	0.38639	Peace	0.487809	Embarrassment	0.518160
Disquietment	0.38140	Happiness	0.477360	Yearning	0.498758
Fatigue	0.37158	Anticipation	0.439886	Fatigue	0.493225
Fear	0.35472	Affection	0.430348	Suffering	0.481661