Doctoral Dissertation

Annotation-efficient approaches towards real-time emotion recognition

March 2019

Doctoral Program in
Advanced Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

IHASZ Peter Lajos

Doctoral Dissertation Reviewed by Ritsumeikan University

Annotation-efficient approaches towards real-time emotion recognition

(実時間感情認識向けの効率的な注釈法による機械学習手法)

March 2019
2019 年 3 月

Doctoral Program in Advanced Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

立命館大学大学院情報理工学研究科
情報理工学専攻博士課程後期課程

IHASZ Peter Lajos
イアス　ペーテル　ラヨシュ

Supervisor : Professor KRYSSANOV Victor
研究指導教員 : クリサノフ　ビクター教授

Contents

List of Tables

List of Figures

Abstract

This thesis addresses the question of reducing the training cost of machine learning-based affective classifiers in terms of annotated labels required.
Extracting information from textual and/or audio contents of the users' utterances provides a set of features that would serve as a reliable and inexpensive mean for emotion recognition in commercial software development. By extracting such features from a dialogic context, the interpersonal aspect of verbal utterances would also be analyzed. The latter appears to primarily influence the generation and control of the interlocutors' interdependent affective states.
Supervised machine learning-based classification of the features requires the classifiers to be pre- trained on labeled data before they are deployed for real-time recognition. Owing to the diversity of the vocabulary and audio characteristics, emotion/sentiment recognition in spontaneous dialogues is a very complicated task, typically demanding a large amount of labeled training data to sustain satisfactory recognition accuracy.
In this thesis, a feature set and its corresponding computational application methods are proposed for the improvement of real-time affective state recognition. The proposed methods allow for reasonably accurate classification results while working with small and, for that reason, relatively easy to prepare, or large but unlabeled sets of audio/textual data. As a novel approach, the author argues that emotion-interdependent dialogue acts can improve emotion/sentiment/polarity recognition even on small sets of labeled data, thus making them applicable for the pre-training of commercial games, dialogue systems, and other applications requiring real-time recognition of affective states. Building on appraisal theory definitions of affective states, 'interpersonal relations-controlling' communication functions are defined as 'emotion-sensitive' dialogue acts, and the corresponding model is developed.
The model is tested with:

a) supervised deep learning methods trained on small collections of labeled data, and

b) semi-supervised, deep learning-based multiple instance-learning methods trained on unlabeled data.

Results of the experiments suggest that the proposed dialog act model allows for reliably classifying the affective states of polarity and emotions with algorithms developed for unlabeled data, while also significantly improving sentiment recognition accuracy when applied on labeled data.

Chapter 1

Introduction

Emotions are one of the building pillars of human behavior. Not only they relate to the internal and external stimuli one experiences every minute of every day they also influence how one behaves, or deal with their surroundings. Ultimately, they are the motives behind one‘s goals and actions, even if they are often restricted/ or redressed by rationality.

Naturally, as argued by Robert Plutchik [1], emotions also affect the whole social regulation process of human beings. Conversations are inherently influenced and regulated by direct or subtle emotions [2]. In the field of affective computing, there have been many studies on the nature and characteristics of affective states with the main focus on achieving reliable automatic recognition and accurate imitation of affective states in the form of emotions, sentiments or emotion polarities as both individual and social phenomena [3].

With the advancement of computational technology, automatic recognition became insufficient, and a growing need for real-time recognition emerged. In several commercial software, delayed inference - which would allow for feature engineering and extraction - of the affective states is not available, real-time reaction is needed. In particular, in the field of affective computing, real-time recognition of affective states is expected to be achieved within 100 ms [4]. Dialogue systems and affect-aware games [5] are typical examples of such applications since these systems try to continuously adapt their content according to the perceived affective states of the human interlocutors.

Real-time recognition of affective states has been realized mostly in non-commercial, academic projects, utilizing supervised machine learning methods (as conventional in the case of classification problems). These supervised algorithms are usually trained and tested on physiological features [6] or facial expressions [7], and though showing promising results in a laboratory environment, they rely on carefully positioned, costly sensors. Thus, they cannot yet be applied efficiently in commercial applications. Dialogue systems, for example, are often applied as telephone-customer-service agents and can rely only on audio features. Even in the case of computer games, where facial recognition is often feasible, a shadow on the user‘s face, unleveled position of the camera, or the presence of facial hair could lead to incorrect classification [8]. The use of headphones (which is common in multiplayer gaming sessions) poses even a bigger challenge due to the ‘noisy‘ representation of the player‘s face.

Extracting information from the textual and/or audio content of the users‘utterances would provide a less technology-sensitive, and thus less easily-perturbed set of features that could serve as a reliable and inexpensive mean for emotion/sentiment/polarity recognition, suitable to be applied in commercial software development. Salvaging audio recordings of conversations would provide audio data which is also transcribable into text. Furthermore, in a dialogic environment, the interpersonal aspect of the verbal utterances can be analyzed, which is the main factor in the generation and control of the interlocutors‘interdependent affective states.

As an example of text processing in commercial software, the game, named Facade [9] uses a rule-based approach for real-time emotion recognition. Systems developed lately, however, such as the commercial tool of EmoVoice [10], or the systems proposed by Fayek et al. [11], are achieving significantly better real-time emotion recognition with audio data classifiers that use supervised machine learning methods. Nevertheless, supervised learning necessitates classifiers to be pre-trained on labeled data before they can be deployed for real-time recognition. Owing to the diversity of vocabulary and audio features, recognition of affective states in spontaneous dialogues is a complex task, demanding a large amount of labeled data to ensure satisfactory recognition accuracy. The amount of data needed is significantly larger than it would be for example in the case of physiological (e.g. facial) features [12].

The advancement of machine learning-based emotion/sentiment/polarity recognition is therefore necessary in a way to allow for reasonably accurate classification results while working with small, and therefore relatively easily prepared, or large but sparsely labeled datasets. It is especially important in the case of verbal data, for which there is a lack of large datasets labeled with emotion-related psychological constructs due to the complexity of the cues it provides. Methods applicable on conversational data are preferable, to provide the learning algorithms additional contextual cues. This dissertation concentrates on the problem of achieving accurate recognition of affective states on conversational datasets equipped with a small amount of labeled textual and/or audio datapoints.

One promising approach to advance supervised machine learning-based recognition of affective states in dialogues, is to consider the psychological context (rather than the physiological context) in the form of intentions. The ‘intentional‘ context is conventionally represented as dialogue acts— pragmatic-level linguistic units. The use of dialogue acts for affective recognition was considered in several previous studies [13], [14], [15]; [16]. The acts discussed, however, are mostly related to ‘communication maintenance‘ or ‘domain related‘ intentions, which do not correlate well with emotions. Consequently, the

improvement achieved through the application of dialogue acts in these studies was relatively low [13] even in binary classification scenarios.

In this thesis, the author examines the following questions:

- Can intentional context improve the recognition of affective states in an annotation-efficient way?
- What type of intentional context should be used?
- Through what methods the intentional context can be used?

The author argues that affective state-interdependent intentions can improve emotion/sentiment/polarity classification even on small sets of labeled data, thus being applicable for the pre-training of commercial games, dialogue systems, and other applications requiring real-time recognition of affective states. Building on appraisal theory definitions of affective states, 'interpersonal relations controlling' communicational functions are proposed as 'emotion-sensitive' dialogue acts, and the corresponding model is developed. The model is validated in

a) empirical experiments in comparison to well-known "conventional" dialogue act models

b) computational experiments as a complementary feature set for supervised deep learning methods

c) computational experiments as an indirect target-label set for a multiple instance-learning- based semi-supervised method.

All validation experiments are conducted on conversational data to ensure that the proposed computational algorithms can learn from the sequentiality between the interlocutors'turns, presumably influenced by the interactive affective states. The algorithms are trained with audio and textual features of Japanese and English dialogues.

The rest of the thesis is organized as follows. Chapter II describes previous studies where the classification of affective states was conducted through approaches that could be applied to train classifiers on small amount of labeled data. Chapter III introduces the proposed model of emotion-sensitive dialogue acts and the conceptual basis behind its development. In Chapter IV empirical and computational methods are proposed to validate the applicability of the model. Chapter V elaborates on the experimental setups, implementational details, and results of the empirical and computational validation approaches. Chapter VI discusses the results and compares the validation methods from an applicational perspective while elaborating on their strengths and shortcomings. Chapter VII offers concluding remarks and outlines possible future work.

Chapter 2

Literature survey

This chapter introduces previous work where recognition of the affective states was achieved through methods that could decrease the need for hand-labeling of training data.

2.1 Complementary feature set-based approach

A possible approach is to utilize complementary feature sets to improve the recognition accuracy of affective states, allowing for satisfactory-level classification results even on small sets of training data (not requiring the annotation of large amount of datapoints). As complementary feature sets, the psychological context can be utilized (instead of the physiological one such as audio, visual or textual etc. features). Intention representing dialogue acts are conventionally used as such feature sets.

2.1.1 Content and affective states

Affective states consist of emotions, moods, emotional traits, and sentiments [17]. Affective states can be interpreted according to several different theories, all of which build upon the assumption that any affective state is triggered by a stimulus event. During an interaction, actions of the partner or self typically serve as stimulus events. In the case of conversations, the only directly perceivable stimulus is usually the communication of one‘s thoughts through verbalization and body language. As Ekman [18] pointed out “*Often in civilized life, our emotions occur in response to words, not actions, to events which are complex and indirect, and it is an extended appraisal process which operates with consciousness and deliberation. Then the person is quite aware of what Lazarus calls the ‘meaning analysis‘ which occurs.*”

People, however, react emotionally not to the mere act of word utterance but to the semantic and functional content expressed. Semantic content includes the objects, propositions, and events defined in an utterance. Functional content specifies the communicative function of the utterance, “the way an addressee should use the utterance‘s semantic content to update his information state” [19]. In other words, it specifies the intentions of the speakers behind their utterances.

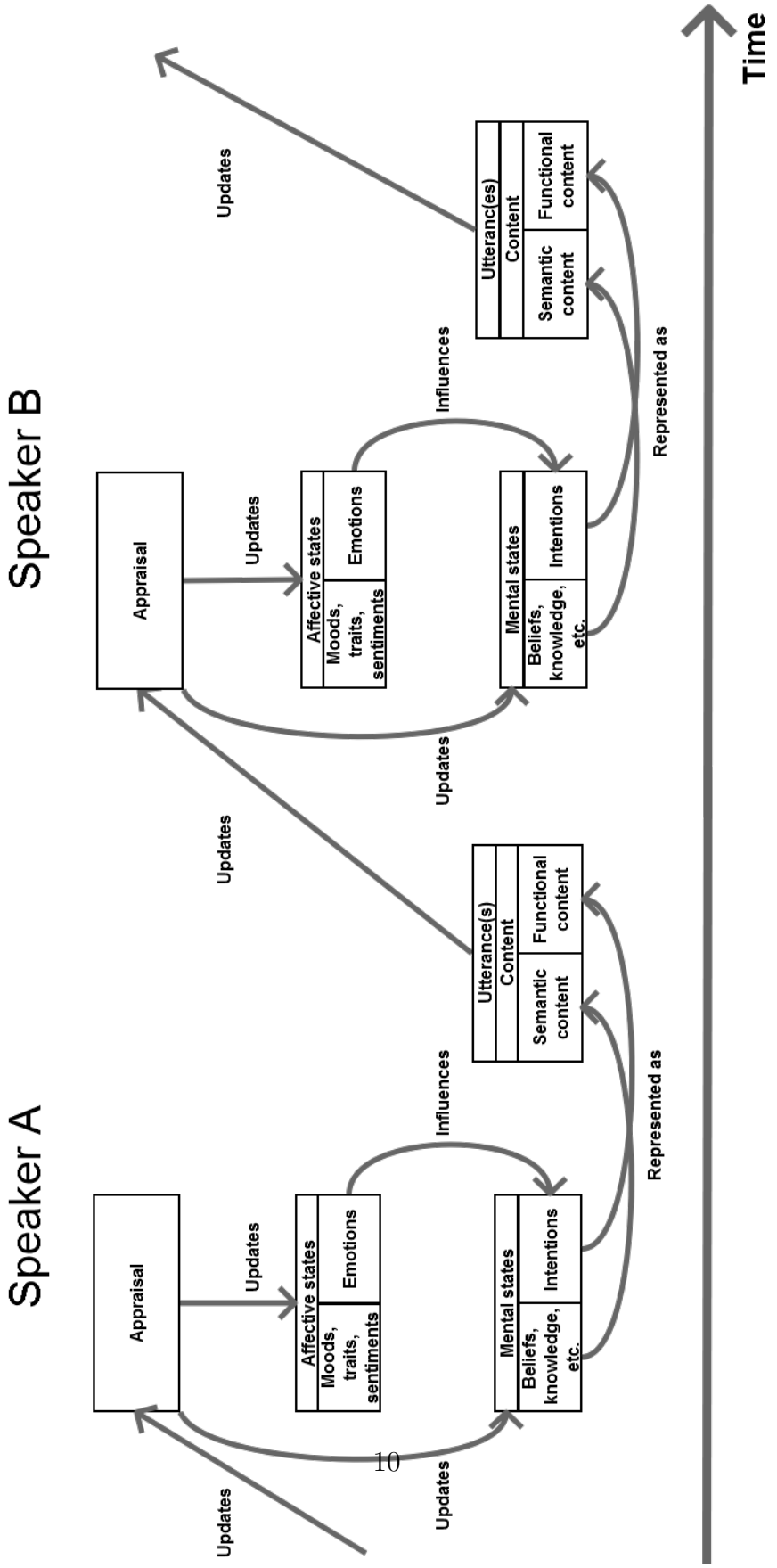

Figure 2.1: Emotion and intention influence loop

The process of emotion (or of any other affective state) elicitation by the utterance content is explained by appraisal theories. The 'meaning analysis'referred to above [20] concerns the examination of content through cognitive processes, based on appraisal criteria. Different combinations of appraisals trigger different cognitive processes, updating mental states and through them eliciting a potentially unlimited range of emotions (or other affective constructs) [21]. On the other hand, it has been found that not only the content of the dialogic unit influence affective states but also that affective states encourage specific action tendencies [22]. Accordingly, the stimulated emotions/sentiments etc. can influence one's intentions, thus - in an indirect manner - the choice of functional and semantic content of the response utterance. The flow of influence between the content of the utterance and affective states is illustrated in Figure 2.1.

2.1.2 Dialogue Acts

According to the findings discussed above, the expressed content of a given utterance can serve both as a cue for the underlying affective state of the speaker and as a cue for the elicited affective state of the addressee. Artificial inference of the affective states from the semantic content, however, is impractical due to the variability in the vocabulary and the multiple, context-dependent meanings of the words. (Even human inference usually necessitates multiple channels of communication for correct interpretation). On the other hand, automatic inference of these states from functional content is more practical, because the possible contents are fewer in number and easier to group into specific 'dimensions'.

Communicative functions (functional contents) are usually represented as dialogue acts—intention-conveying, pragmatic-level dialogic units. Dialogue acts are expressed through utterances and can change the mental states of the interlocutors. Austin [23] defined three levels of dialogue acts (calling them speech acts): locutionary, illocutionary, and perlocutionary. A locutionary act is the speaker's performance of uttering the words and giving meaning to them. An illocutionary act is the speaker's performance of expressing her/his own intentions through the utterance (implicitly or explicitly). A perlocutionary act is the effect of the speaker's utterance on the addressee, reflecting how the listeners reacted to their interpretation (whether correct or incorrect) of the speaker's illocutionary act. Conventionally, most dialogue act categorizations incorporate illocutionary acts, grasping the intentions meant to be conveyed through the utterances. For the above reason, this dissertation refers only to illocutionary dialogue acts, when mentions them.

Dialogue acts can be categorized in many ways, with a particular categorization covering either one communicative function dimension (with mutually-exclusive tags for each annotated segment) or several (with multiple tags for each annotated segment) [24]. Not all categorizations of dialogue acts, however, associates well with affective states. Consider for example, "dialogue management" communicative functions, which includes acts such as "answering", "questioning" and so on. Each of these acts can affect, or be affected by, any possible emotion.

2.1.3 Applications of dialog acts for supervised emotion/sentiment recognition

The use of dialogue acts for affective state recognition purposes was considered in several studies. Dialogue act labels were used in these studies as a complementary feature set - inferred automatically or hand-labeled - input to supervised machine learning algorithms. Although hand-labeling of dialogue acts would further increase the need for annotation, the idea behind their usage is that classifiers can learn from a feature set only to an extent. After a certain amount of training data, the learning ratio on the same feature set will eventually decrease. Utilizing (an easy-to-learn) low-dimensional complementary feature set may provide information from a different perspective. Thus, among two classifiers, one trained on a certain amount of data labeled only with the target labels, and the other on half of that dataset labeled with the target labels and an additional set of labels as a complementary feature set, the latter classifier may perform better. [25]

Supervised machine learning

In the field of machine learning, supervised learning is the task of inferring a function through learning examples. Each example is a pair of a vectorized input object (image, sound, text or other signal) and an output label or value. The supervised learning algorithm infers a function based on the analysis of the training data. The inferred function is updated through each new example. If the learning is successful, the algorithm can - to a certain extent - correctly determine the class labels or corresponding real-valued intervals of unseen instances (also called test instances). To be able to achieve this, the learning algorithm needs to generalize from the training data to unseen test data. [26]

A wide range of supervised learning algorithms is available, with different architectures for function learning and generalization. The most widely utilized algorithms are (it is out of the scope of this thesis to specify them in detail):

- Logistic regression
- Decision trees
- Support Vector Machines
- Multiple instance learning
- Naive Bayes
- Linear regression
- Linear discriminant analysis
- Artificial Neural Networks
- K-nearest neighbor algorithm

Dialogue acts utilizing supervised recognition of affective states

Ang et al. [13] augmented lexical and prosodic features with dialogue acts (repeat, repair, neither) of the current turn to improve emotion recognition. The addition of the dialogue acts resulted in a 4% maximum improvement when classifying the emotional states of annoyance-frustration vs. else (the latter includes all the remaining emotion types), and frustration vs. else through decision trees.

In the study of [14], the emotional salience word score (representing the context-wise appearance likelihood) and dialogue acts (rejection, repeat, rephrase, ask-start over, other) were input together with prosodic and lexical features into a linear discriminant classifier, yielding a 3% improvement in the "binary" classification of negative and non-negative sentiments.

Likewise, Batliner et al. [15] augmented lexical and prosodic data with discourse information of dialogue acts (introduce, request, suggest), obtaining a 1.2% improvement when differentiating between the cognitive states of emotional and neutral through a multilayer perceptron neural network.

Liscombe et al. [16] considered prosodic, lexical, and dialogue act features (65 categories discriminated by call-types of the HMIHY 0300 corpus, e.g. asking for customer representative, requesting information about account balance, etc.), as well as contextual features. Contextual features are the prosodic, lexical, and dialogue act features of the n-1 and n-2 turns of the dialogue. The application of the dialogue acts by themselves led to a 2.6% improvement, while enhanced with the contextual features led to a 4% improvement in the classification of non-negative

vs. negative sentiments. As a classifier, a boosting algorithm was used, where the final classification result is computed from the combination of sub-classification results through several iterations. Sub-classification results are provided by the weak classifiers of one-level decision trees.

These and many other studies with similar goals and results utilized dialogic data only with pre-annotated dialogue act labels to fully evaluate the applicability of the additional feature set.

Recently, there has been a declining interest in the idea of enhancing emotion recognition through the use of dialogue acts. This may be due to, at least in part, the fact that accurate extraction of intentional features is also a task that requires pre-training on labeled data. Annotation of dialogue acts then becomes excessively labor-demanding, especially when contrasted to the rather modest improvements it would yield in emotion recognition. The usage of a dialogue act categorization which is more sensitive to affective states than the categorizations discussed above assumed to yield better improvements which would balance out the annotation cost.

2.2 Big data-based approach

2.2.1 Product review evaluation as sentiment/polarity labels

As an approach different from the previous one, classifiers can be trained on very large sets of which would (even with decreasing learning rate after a certain amount of training) allow for satisfactory level recognition-rate. Several large datasets exist [27], [28] in the form of product reviews labeled with scalable units, indicating user-satisfaction. Satisfaction is conventionally indicated with 1-5 discrete values (usually of stars) or with continuous values of points between [0.0: 5.0]. Since the evaluation of the user is assessing the users‘feeling about the given product, it is often matched with sentiment or polarity. In particular, the discrete or continuous evaluation scores are divided into two or three subsets where each subset are accounting for one of the negative/neutral/positive polarities or sentiment scores. [29]

Product review databases usually contain millions of evaluated reviews. For sentiment/polarity analysis, conventionally neural networks are utilized for they can generalize far better or large sets of data than other supervised methods [30], [31]. For the recognition of affective states in dialogues, however, the monologic product reviews cannot be utilized. Nevertheless, affective classifiers need to be trained on dialogues in order to be applicable in commercial products such as

affect aware games or dialogue systems. Currently, however, large datasets of labeled dialogues are not available. As a possible remedy to this problem multiple instance learning algorithms can be utilized, trainable on large sets of (partly) unlabeled data.

2.2.2 Multiple instance learning

Multiple instance learning is a supervised machine learning method, that requires only sets of datapoints to be labeled instead of all datapoints. Thus, it could significantly reduce the time and effort needed to achieve data-labeling. The next sections introduce the assumptions the conventional variations of multiple instance learning are based on as well as their applications.

Standard assumption

The **standard assumption** behind multiple instance learning (MIL) is that each instance $x \in X$ from the instance space X, has a binary latent label $y \in \{0, 1\}$. Thus, (x, y) is called an "instance-level concept" where an instance is representing an underlying concept $c \in C$ from the concept space C. A 'bag'is a multiset of instance-level concepts, with instances labeled identical to the target class, called positive labels and instances labeled non-identical, called negative labels. A bag is labeled positive if at least one of its instances has a positive label, and negative if all of its instances have negative labels. This assumes, that a bag can be represented by a sole concept. [32] Formally, if a bag is

$$B = \{(x_1, y_1), \ldots, (x_n, y_n)\} \tag{2.1}$$

then a label of B is

$$L = 1 - \prod_{x=1}^{n} (1 - y_x). \tag{2.2}$$

The standard assumption can be applied, for example, to predict molecule activity. A molecule can appear in various shapes, called conformations, If certain receptors bind well to any of the conformations of the molecule, it becomes active. When activated through a certain kind of receptors, the molecule produces a "musky" smell. Thus the MIL concept, in this case, would be the tendency to conformation with certain receptors, which, if large enough, results in the musky smell. Although this is a binary case, the standard assumption can be applied for multi-class cases as well. The standard assumption works well for the certain task described above, however, there are more complex tasks, where the predicted entities cannot be differentiated by one

concept, a bag label should be determined by the simultaneous presence of several concepts. A bag label of a given sickness where the varying symptoms are the concepts is an example of such case. Accordingly, researchers tried to relax the standard assumption, developing other assumptions [32].

Extended assumptions

A **presence-based assumption** is the generalized extension of the standard assumption, where a bag is labeled positive only if it contains at least one instance of n several different concepts. Formally, a concept is a function $v_{PB} : 2_{\mathrm{x}} \Rightarrow \omega$, where for a set of required concepts $C \subset \mathbb{C}$,

$$v_{PB}(X) \Leftrightarrow \forall c \in C : \triangle(X, c) \geq 1 \tag{2.3}$$

In the **threshold-based assumption**, a bag is labeled positive only if a certain number of instances of each concept are present simultaneously.Thus, to each required instance-level concept a threshold is associated:

$$v_{TB}(X) \Leftrightarrow \forall c_i \in C : \triangle(X, c) \geq t_i \tag{2.4}$$

where $t_i \in N$ is the lower threshold for concept i.

The **count-based assumption** defines a minimum and maximum number of required instances for each concept. Each concept thus has a lower threshold $t_i \in N$ and an upper threshold $z_i \in N$:

$$v_{CB}(X) \Leftrightarrow \forall c_i \in C : t_i \leq \triangle(X, c) \leq z_i \tag{2.5}$$

The **generalized multiple instance learning assumption** defines a set of required instances $Q \subseteq X$. The number of instances sufficiently close to the required instances Q needs to reach a certain limit n in order to label a bag postiive. Scott et. al. [33] further generalized this assumption defining attraction points $Q \subseteq X$ and repulsion points $\overline{Q} \subseteq X$. A bag, then, is labeled positive if and only if it contains instances which are sufficiently close to at least n of the attraction points and does not contain instances that are sufficiently close to repulsion point more than m .

Conventional application

The standard assumption has been mainly applied through iterated discrimination algorithms, which usually contains two phases. An axis parallel rectangle (APR) is populated in the first phase. The population is achieved in an

iterative manner until it contains at least one instance from each and every positive bag and excludes all instances from any of the negative bags. Then, a relevance metric is rendered to each instance x_i indicating the number of the excluded negative instances if removed from the APR. Then candidate representative instances are selected in decreasing order of their relevance. The process is repeated until no instance of the negative bag remains in the APR.

As the result of the first phase, the APR supposed to contain only instances from the positive bags. A looser APR is drawn in the second phase. The looser APR is based on Gaussian distributions centered at each attribute. From the second APR positive instances with fixed probability will fall outside. [34]

As mentioned above, this method assumes that a bag‘s label is determined based on the presence of a single concept. There are some complex problems, however, where a bag label is determined by the simultaneous presence of several concepts.

The Two-Level Classification (TLC) algorithm, proposed by Weidmann [35], learns multiple concepts under the count-based assumption. It tries to learn instance-level concepts in the first step. In particular, a decision tree is built from each instance of every bag of the training set. The bags then are mapped to a feature vector based on the output of the decision tree. In the second step, the underlying concept of the instances is learnt through running a single-instance algorithm on the feature vectors.

Scott et. al [33] proposed an algorithm, called GMIL-1, to learn concepts under the GMIL assumption. GMIL-1 enumerates all axis-parallel rectangles $\{R_i\}_{i\in I}$ in the original space of instances, and defines a new feature space of boolean vectors. A bag B is mapped to a vector $\mathbf{b} = (b_i)_{i\in I}$ in this boolean vector-based feature space, where $b_i = 1$ if APR R_i covers B, and $b_i = 0$ otherwise. A single-instance algorithm is applied to learn the concept in this new feature space.

Most MIL methods, however, are applied for image/molecule activity/ document recognition, where the recognized entities are the bags. Thus those methods - including the ones described above - are concentrating on the prediction of unseen bags, instead of the prediction of unlabelled instances they contain. In the case of emotion/sentiment recognition, however, the MIL would be used to train instance-level predictors based on the bag labels.

Application for sentiment analysis

In the approach proposed by Kotzias et al. [36] instance labels were inferred through propagating information from the bag labels to the instances. In particular, the unknown label aggregation function on the training data was inverted. The approach used K similarity measure to compute a

group-structure-compatible label assignment algorithm, which can be used to assign the same label to similar train instances. The developed algorithm was also used to classify instances not found in the training set. The predicted labels then were aggregated and were used to classify unseen bags.

The approach is based on an objective function that allows for smooth inference of the instance-level labels. The function considers instance-level similarity, with respect to group-level label constraints at the same time:

$$J(\theta) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \triangle_1 (\hat{y_i}, \hat{y_j}) + \frac{\lambda}{K} \sum_{k=1}^{K} \triangle_2(\hat{l_k}, l_k) \tag{2.6}$$

where:

- $K(x_i, x_j) \in [0, 1]$ is the similarity measure between instances $\mathrm{x_i}$, $\mathrm{x_j}$
- $\triangle_1(\hat{y}_\theta(x_i), \hat{y}_\theta(x_j))$ is a non-negative penalty on the prediction differences for instances i and j;
- $\triangle_1(_i, _j)$ is a non-negative penalty on the prediction errors for group k.
- $l_k = A(G_k, \theta) \in [0, 1]$ is a real-valued scalar representing the output of the aggregation function for all instance-level label predictions in a group G_k.
- $\lambda > 0$ can be selected via cross-validation on a validation set. It balances the contributions between the two sums.

Trained and tested on the Amazon, IMDB and Yelp datasets the proposed method classified three sentiments with an accuracy of 86%-88%.

As another method that learns to predict the polarity of text segments from bag-level labels, Angelidis and Lapata [37] reduce each segment's class probability distribution $\mathrm{p_i}$ to a single real-valued polarity score. To achieve this, they first define a real-valued classweight vector $w = \left\langle w^1, ..., w^C | w^c \in [-1, 1] \right\rangle$ that assigns uniformly-spaced weights to the ordered labelset, such that $w^{(c+1)} - w^{(c)} = \frac{2}{C-1}$. For example, in a 5-class scenario, the class weight vector would be $w = \left\langle -1, -0.5, 0, 0.5, 1 \right\rangle$ Then, the polarity score of a segment is computed as the dot-product of the bag probability distribution p_i with vector w:

$$polarity(s_i) = \sum_c p_i^{(c)} w^{(c)} \in [-1, 1] \tag{2.7}$$

with a gated extension:

$$gatedpolarity(s_i) = a_i \cdot polarity(s_i) \tag{2.8}$$

where a_i is an attention weight assigned to the i-th segment.

Trained and tested on the IMDB and Yelp datasets it yielded classification results between 91% and 94% for three polarities. A significant deficiency of this method, however, is that class probability distributions p_i are gained through a supervised feature map classifier pre-trained on sentence-level labels.

Although learning methods described above are trained only on a set (bag) of instances (instead of each and every instance), selection and hand labeling of the bags is usually still necessary (in [37], sentence-level labels were also needed for pre-training). In the case of the above review-based datasets, the bags were already given in the form of reviews, where for each review there was a scale-based evaluation of the reviewer attached that could serve as a bag label. Reviews with similar level evaluation could then be aggregated into bigger bags.

In the case of conversations, however, selection of bags becomes problematic, necessitating the partition of dialogues by time, topic, or interlocutor. In the specific case of emotion recognition, where a bag supposed to represent one particular emotion, partitioning/labeling becomes even more complex. For this reason, to the author‘s best knowledge, there exist no dialogic datasets with labeled subsets of affective constructs.

For affective state recognition through MIL, sub-sets of dialogues expressing a particular emotion are assumed to be efficient bags. Section-labeling, with the additional task of reason-based sub-sectioning, is difficult to automate. Section labeling and sub-sectioning by hand, on the other hand, is a labor and time demanding task, which can easily nullify the benefits MIL would provide in contrast to segment-level labeling. Accordingly, a method capable of finding sub-sections applicable as emotion-related bags needs to be developed. This thesis describes an approach of mining dialogic videos through emotion sensitive intentions-based search, that not only makes the videos directly applicable for MIL-based emotion recognition but also completely eliminates the need for their hand labeling.

2.3 Related work summary

There are two main approaches towards the improvement in recognition of affective states for audio and/or textual features:

a) Utilizing the intentional context in the form of dialogue act labels and use them as complementary features to predict the output labels: suffers from the deficiency, that the dialogue acts used are not yielding significant improvement in recognition accuracy

b) Using large sets of data labeled with sentiments/polarity tags to train

- supervised neural-networks: suffers from the problem that in the case of dialogues all utterances of the dialogues would need to be labeled
- multiple instance learning algorithms: the target labels are used as bag labels, which significantly (in the case of reviews, completely) reduces the need for instance labeling. For dialogic data, however, definition and labeling of bags would be necessary which is a difficult task, demanding manual labor.

The author argues that the usage of emotion sensitive intentions representing dialogue acts, associable with certain affective states would

- boost the recognition accuracy of emotion/sentiment recognition if used as a complementary feature set: it would help the supervised machine learning methods to learn from the low dimensional, easy-to-learn features of additional labels, firmly associated with the output labels
- serve as a basis to define emotion indicating search phrases for the selection of videos concentrating on emotional dialogues. Such videos could serve as bags containing instances of emotion representing utterances, while the emotions the search phrases correspond to could serve as the bag labels. The labels thus can be utilized for the training of multiple instance learning algorithms in a semi-supervised way.

Chapter 3

An emotion sensitive dialog act model

This chapter describes an emotion sensitive dialogue act model and the conceptual basis behind its development.

3.1 An emotion sensitive dimension of dialogue acts

Certain dimensions of communicative functions - represented through dialogue acts - assumed to show a stronger correlation with affective states than others. The author assumes that the main reason behind the moderate improvements the usage of dialogue acts yielded in previous work (see 2.1.3) is that the taxonomies in question did not utilize dialogue acts that show a strong correlation with affective states. Thriving to discover such a dimension, the author turned to the emotion stimuli explaining appraisal theories.

Several categorizations of appraisal criteria have been proposed. One of the most well- known is due to Ellsworth and Scherrer [21], who classified appraisals along:

A. novelty and pleasantness – how novel and safe/pleasant is the situation for the addressee based on the content;

B. conduciveness to needs, goals, and values – how desirable is the outcome for the addressee as indicated by the content;

C. power and coping – how feasible it is for the addressee to control the situation based on the content; and

D. social identity, norms, values, and justice – how much the content is in accordance with the addressee‘s image of social etiquette, expectations, and so on.

Among the appraisal criteria listed above, it appears reasonable to assume that those of (B) and (D) are stimulated mainly by the functional content. For example, the communicative function behind the utterance “You are a gentleman!” could be

‘praising‘. Once the meaning of the utterance is understood, the function ‘praising‘ itself would potentially satisfy the speaker‘s needs (item B), updating his social identity while satisfying the likely expectations of social norms, values, and justice (item D). In the same situation, any other ‘praising; function would have the same effect on the two dimensions, with only the magnitude varying in accordance with the semantic content. In the case of (A), however, understanding the underlying function ‘praising‘ behind the speaker‘s utterance would, most probably, be not sufficient to affect an appraisal about the novelty of the situation.

Unlike the cases of (B) and (D), appraisals using (A) and (C) can only be correctly comprehended when one takes into account the semantic content (magnitude and modality) of the utterance. To appraise the situation‘s novelty to a satisfactory level, for instance, in addition to the four criteria ‘(have I ever been/am I usually) praised by (this person/people)‘, the addressee must also assess the semantic content with four more criteria ‘(have I ever been/am I usually) called a gentleman by (this person/people)‘. A key assumption is, therefore, formulated as follows:

Assumption 1. The appraisal of conduciveness to needs, goals, and values, and the appraisal of social identity, norms, values, and justice are affected by the functional rather than the semantic content.

At the same time, it is understood that the dimension of functional content needs to be defined, which can serve as a triggering stimulus to the appraisal criteria of (B) and (D). For example, the functional content of the utterance “You are a gentleman!” could be understood from the viewpoint of communication management as ‘making a statement‘. In that case, the likelihood of the functional content helping to assess the desirability of the outcome (B) or the accordance with social expectations (D) is low. Thus, the author adopts the idea that criteria (B) and (D) are mainly influenced by the ‘social status‘ and ‘identity-controlling‘ dimensions of the content‘s communicative function that is formulated as follows:

Assumption 2. Communicative functions that control interpersonal relations can stimulate affective states via the appraisal criteria of (B) (conduciveness to needs, goals, and values) and (D) (expectations of social identity, norms, values, and justice) without relying on semantic content.

Following from Assumption 2, that interpersonal relations controlling communicative functions can directly stimulate affective states through the appraisal criteria of Assumption 1, presumes a direct link between the two concepts, leading to the third and final assumption of this study:

Assumption 3. Communicative functions that directly stimulate the appraisals influencing emotions/sentiments can also serve as indicators of the

affective states that originally influenced these communicative functions. Communicative functions that control interpersonal relations are presumed to be such indicators.

Thus, it is assumed by the author that ‘interpersonal relations-directing intentions‘ such as ‘criticizing‘ or ‘empathizing‘ would provide an ‘emotion sensitive‘ communicative function dimension, which is likely to affect or be affected by only a limited range of emotions/sentiments.

An act of ‘criticizing‘, for example, often affects the addressee‘s self-esteem negatively and would probably elicit an emotional reaction of negative valence, such as ‘fear‘ or ‘anger‘. Furthermore, ‘criticizing‘ is likely to be expressed under the influence of a negative valence emotion, such as ‘anger‘ or ‘disgust‘. It then appears natural to expect that the emotion ‘anger‘ should be associated with a dialogue act representing ‘interpersonal relation control‘, such as ‘criticizing‘ more consistently than with a dialogue act representing ‘dialogue/turn management‘ such as ‘repeat‘ [13] or ‘task-oriented actions‘ such as ‘requesting information about account balance‘ [16].

3.2 The proposed model

3.2.1 Interpersonal relations-managing acts

Accordingly, a dialogue act model representing ‘interpersonal relations managing‘ communicative functions has been developed, called the IA (interpersonal acts) model (see Table 3.1).

Several social factors can be considered when defining communicative functions that control interpersonal relations. The presented study employs Brown and Levinson‘s (B&L) politeness framework [38], which accounts for the interlocutors‘ interdependent social status, and which can be summarized as follows. Face, the “public self-image that every member wants for himself” [39], is divided into ‘positive‘ and ‘negative‘ faces. The positive face can be thought of as the ‘social‘ face, desiring approval by others; the negative face is “the basic claim to territories, personal preserves, and rights to non- distraction” [38]. According to B&L, every person has a negative face, which desires that her or his actions are unimpeded by others, as well as a positive face, which desires that their wants and thoughts are desirable to others. To build and maintain smooth interpersonal relationships, each interlocutor must constantly preserve each other‘s face wants. Nevertheless, there are situations where harming each other‘s face wants, through a face-threatening act (FTA), is inevitable. The model accounts for such acts through the main and sub-categories of positive and

negative face-threatening acts potentially damaging the positive/negative face wants of the addressee.

When FTAs are performed, strategies may be applied to 'save' the face of the speaker or/and the addressee. Such redressive strategies include positive politeness (easing the FTA by satisfying the addressee's positive face wants with jokes and other verbal and non-verbal signals, indicating companionship and common ground), and negative politeness (easing the FTA by satisfying the addressee's negative face wants by paying deference, or through other verbal or non-verbal signals that indicate respect). The more the speaker tries to save the addressee's face, the more she or he damages her/his own face. To ensure an optimal number of categories for annotation (i.e. not too many to deal with for the annotators but still sufficient for the study's purposes), face-saving strategies were not included in the presented IA categorization. Nevertheless, the aforementioned strategies can be incorporated into the model by extending each type of FTA with subcategories corresponding to the possible redressive strategies (Section 2.4.) the given FTA allow for.

Matsumoto [40] argued that B&L's model ignores communicative functions that are not associated with FTAs or subsequent redressal. This inadequacy was later addressed through the introduction of "face enchantment" by Hernandez [41]. Hernandez showed that there are cases where certain politeness strategies work not as a redressive force but as a communicative function that maintains the social relationship when the speaker enchants the addressee with 'gifts' to the addressee's positive face. Hernandez also introduced the notion of self-face work, that is "focusing on one's own face without directly affecting the addressee's face" to ease the harm it suffered (inflicted by either interlocutor) during the interaction or to make a favorable impression. Finding excuses or changing the topic are examples of such strategies. Self-face work is incorporated in the IA model as a subcategory of Non-Face Threatening acts.

The IA categorization is intended to be used as a one-dimensional, independent model when needed, as well as an extension of other multi-dimensional models. When used as a one-dimensional tagset, it can be used as a target label for the recognition of 'interpersonal relations managing' intentions or as an additional feature set for the recognition of affective states. When incorporated into a multi-dimensional tagset, the IA model serves as a dimension incorporating 'interpersonal relations managing' communicative functions.

Table 3.1: Taxonomy of interpersonal acts

Categories, subcategories			
Interpersonal acts	Non-face threatening acts		Paying attention
			Empathizing
			Agreeing
			Self-image improving
	Face threatening acts	Positive FTA	Criticizing
			Indiscrete commenting
		Negative FTA	Indebting partner
			Commanding/requesting

The IA model conforms to the ISO Standard for Dialogue Act Annotation 24617-2 [42] in the following:

- The categorization differentiates between semantic and functional content.
- The dialogue acts defined in the model represent communicative functions.
- The represented communicative functions can be associated with a specific dimension. Therefore, the proposed tagset can be used as a function-specific dimension and can be combined with all general and function-specific acts defined by the standard or by its representative tagset, the DIT++. For example, the utterance “うん” (“Mhmm/Yep”) can be regarded as a ‘paying attention‘ interpersonal act, and as either the DIT++ <‘general-purpose, answer‘> act or the function-specific <‘auto feedback, auto-positive feedback‘> act (for further details on the DIT ++ dialogue acts, see 3.2.2).

- The defined acts are intended to correspond to functional-segments (minimal stretches of behavior having one or more communicative functions).

The IA model does not conform to the following aspects of the ISO 24617-2 standard:

- Functional dependency relations, feedback dependency relations, and rhetorical relations are not accounted for by the proposed model. The responsive interpersonal acts of 'paying attention' and 'empathizing' are typically used to represent functional- and feedback-dependency relations. However, their purpose in the developed taxonomy is solely to account for the dimension of interpersonal relations-managing actions, which includes these actions of responsive nature. All other acts can have functional- and feedback-dependency relations with each other, and all 12 acts can perform rhetorical functions, based on the dialogic situation.
- Although qualifiers (e.g., certainty, conditionality, partiality, or sentiment) can be attached to the proposed tags, the IA tagset used in the study does not assume the use of qualifiers, owing to the following:
  - Allowing for qualifiers would lead to a large number of possible tags that would make the annotation process unnecessarily confusing and prohibitively time-consuming.
  - Due to the large number of possible tags, each particular tag would get associated rarely, if at all, with a given emotion, even when a larger corpus was used.
  - Interpersonal acts have already been defined, with the intention to serve as indirect sentiment qualifiers themselves. Interpersonal acts could be used to represent the previous turns' stimuli for emotions/sentiments, as well as results of the cognitive process influenced by the affective states of the current turn. Accordingly, the model is assumed to be applicable to improve the recognition of affective states of emotions/sentiments.

3.2.2 Extending the model for validation purposes

The IA model is developed to advance real-time emotion/sentiment recognition in commercial products such as dialogue systems or affect aware games. Accordingly, in two of the proposed validation methods, its emotion-sensitivity and applicability as a complementary feature set are to be tested out in a dialogic, cooperative gaming environment. The model can be

Table 3.2: Extended taxonomy of interpersonal acts

Categories, subcategories				Examples
Interpersonal acts	Non-face threatening acts	Partner unrelated commenting	P-u. positive commenting	“よーし見つけた”(“Finally! I found it!”)
			P-u. negative commenting	“やばい”(“That looks bad!”)
			P-u. neutral commenting	“どこだ？”(“Where is it?”)
		Trying to gorund/maintain good relationship	Paying attention	“うん”(“Mhmm/Yep”)
			Empathizing	“マジで？”(“Seriously?”, in reaction to statement)
			Accepting as superior (showing deference)	“わかりました”(“Understood!”)
			Agreeing	“そうそうそうそう {笑}”(“Yes, yes, yes, yes! [laughter]”)
			Self-image improving	“何もしなくても倒せるから大丈夫” (“Even if you don‘t do anything I can defeat it, it‘s all right!”)
	Face threatening acts	Positive FTA	Criticizing	“え、あれでいいの? {笑}”(“Are you sure you will be alright like that? [laughter]”)
			Indiscrete commenting	“また死んだ？”(“You died again?”)
		Negative FTA	Indebting partner	“取っていってやる”(“I will take it for you.”)
			Commanding/requesting	“じゃたまり場来て”(“Come to the gathering spot!”)

extended with ‘Partner-unrelated commenting‘ dialogue acts which allow for the partner-unrelated in-game utterances (reacting to the game contents) to be distinguished from partner-related (reacting to the partner‘s deeds) ones.
‘Interpersonal relations managing‘ intentions are complex constructs, accounting

for the hierarchy and intimacy between the interlocutors. There are cultures where the social practice is more rigorous than in others, and the aforementioned two aspects manifest in the form of additional functional content. Japanese, for example, is a language, where additional acts may needed to properly cover the interpersonal relations managing intentions. In particular, in the case of Japanese the model can be further contemplated with the culture-specific part of "accepting as superior" , accounting for a culture where deference is often displayed not only through conjugational forms but through the use of specific phrases as well. Two of the proposed validation methods utilizes the extended model on Japanese in-game data.

Table 3.2 summarizes an extended model, tailored to fit Japanese conversations in a cooperative gaming environment and illustrated through Japanese sample utterances (gathered from Japanese in-game dialogues).

Chapter 4

Proposed validation methods

This chapter describes empirical and application-oriented methods for the validation of the proposed IA model.

4.1 Empirical validation

Empirical validation methods are proposed to measure the association strength between the affective states and the tags of the proposed dialogue act tagset. These methods are to prove the general applicability of emotion sensitive intentional context for the recognition of affective states. In particular, emotions are chosen to be the target of analysis, for they serve as a fine-grained-enough metric to meaningfully assess the proposed tagset.

Following Assumption 3 (see 3.1), interpersonal relations affecting communicative functions should indicate, by their very nature, the affective states that influence them. For instance, ‘criticizing‘ should, perhaps, co-occur mostly with the emotion ‘anger‘ or the with ‘negative‘sentiment.

The more consistently the affective context is indicated by the interpersonal acts (and vice-versa), the better the proposed tagset can be considered to represent affective states and affective states-related interpersonal interactions. The proposed IA model‘s performance was evaluated according to the consistency, with which interpersonal acts were paired with, and distinguished by, certain subsets of emotions manifested in the same utterance. The model‘s adequacy was determined in relation to the tagsets of two widely-used dialogue act models, SWBD-DAMSL [43] and DIT ++ [44]. For a detailed description of the models see 5.1.1.

4.1.1 Analytic framework

The greater the dependency between the acts and emotions the more suitable should be the given model (out of the three in comparison) for representing affective states and the interdependent social relations. However, the presence of mere dependency (i.e. correlation or correspondence) between these variables would, in itself, not necessarily indicate that the given model accounts for a wide range of emotions (and social situations).

Consider the extreme case where, for example, an emotion category with large

frequency counts corresponds to more than one-third of the total number of acts from a given dialogue act categorization. The correlation would be strong between the two variables, even though the dialogue act model in question would generally serve as a poor indicator for specific emotions.

The assumption that interpersonal relations influence affective states, while also being influenced by them, suggests that different aspects of interpersonal relation oriented communicative functions are likely manifested in different affective contexts. A model that does not account for communicative functions which are also (to a certain extent) distinguishable by their affective context does not allow for uncovering interpersonal relations to a satisfactory degree. Hence, the variance in emotion-dialogue act correspondence should also be examined, with special attention paid to the strength and exclusivity of the associations between each emotion and the dialogue acts of the models compared. The main steps of the analysis are, therefore, as follows:

A. Compute and summarize the occurrence counts of emotions and dialogue acts in a two-way contingency table for each model. As only the co-occurrences are used throughout the analysis, the sequential nature of the data is unimportant, and the frequency counts obtained for the five conversations are aggregated.

B. Determine whether there is a significant overall dependency between the variables of dialogue acts (of a given model) and emotions, using Fischer's exact test [45].

C. Compute normalized pointwise mutual information (npmi) for the co-occurring dialogue acts (of each model) and emotion categories to assess the element-wise magnitude of the association between the two dimensions. For every speech act and its paired emotion, this measure is defined as follows [46]:

$$npmi = \frac{ln\frac{p(x,y)}{p(x)p(y)}}{-ln(p(x,y))} \tag{4.1}$$

where p(x) is the marginal probability to observe dialogue act x, p(y) is the marginal probability to observe emotion y, p(x,y) is the joint probability to observe x and y at the same time, -ln p(x,y) is the normalizing coefficient, so that [-1.00, 1.00]. The npmi measure does not, therefore, depend on the total number of occurrences of a given emotion or dialogue act but reflects the consistency of co-occurrence, meaning that an act with few occurrences can still have a strong association with an emotion when the two co-occur relatively frequently.

In the context of this thesis, npmi is used to estimate the extent, to which the occurrence of a given dialogue act would indicate the simultaneous observation of a given emotion, and vice-versa. An npmi value of -1.00 indicates that the two events have not been observed together, a value close to 0.00 – that the events are decoupled (their simultaneous occurrence is a random coincidence), and a value close to 1.00 – that the events have always been observed together. For each model, positive npmi values are summed up and, then, normalized by the number of the model‘s degrees of freedom. The resulting statistics serve as comparable metrics between models for assessing the overall strength of association between dialogue acts and emotions.

Positive npmi values are grouped, based on ‘association strength‘ intervals subjectively defined as follows: [0.0, 0.20) for ‘weak‘ associations, [0.20, 0.40) for ‘medium,‘ [0.40, 0.60) for ‘strong,‘ [0.60, 0.80) for ‘very strong,‘ and [0.80, 1.00] for ‘extremely strong‘ associations. The models are compared for how consistently and comprehensively their dialogue acts are related to particular emotions.

4.2 Computational validation - supervised learning

The applicability of the developed tagset is proposed to be tested in supervised sentiment classification experiments. The experiments are to prove that the utilization of the complementary feature set of emotion sensitive dialogue act labels can significantly improve the recognition accuracy even on moderate size training sets. For applications that utilize real-time recognition of affective states like affect-aware games or customer-service dialogue systems, the recognition of sentiments, covering several discrete emotions is preferred to the more fine-grained, but less reliable emotion recognition. In the works of [13], [14], [15], and [16], the proposed dialogue acts were also tested in binary-classification scenarios.

In the validation experiments to be conducted, the interpersonal act tags will be used to enhance a text and audio based sentiment classifier (detailed in 3.2.2) in comparison with

- two identical classifiers which utilize acts of other well-known dialogue act models,
- and with a baseline classifier which does not utilize dialogue acts as an additional feature set for sentiment classification

The four classifiers (one baseline classifier and three augmented classifiers, processing dialogue acts) is to be trained and tested on the same dataset of

audio streams of dialogues and their transcriptions. The dataset is to be compiled from cooperative in-game conversations, in order to test the proposed taxonomy‘s applicability on recordings similar to the real-life application it is developed to improve.

Each utterance in the transcriptions is to be annotated with four labels in total: one sentiment label, and three dialogue act labels (one of the proposed tagset and two of the other dialogue act models used for comparison). Interpersonal acts are proposed as acts to be automatically inferred or to be hand-labeled beforehand their application for emotion/sentiment recognition. Following the line of previous studies [13], [14], [15], and [16], in the experiments hand-labeled dialogue act tags are to be used, concentrating only on measuring the proposed model‘s adequacy for augmenting sentiment recognition. Thus, the augmenting-performance of the dialogue act models in comparison could be fully assessed, unhindered by their varying recognizability by automatic means.

Figure 4.1 shows the overall design of the four classification scenarios.

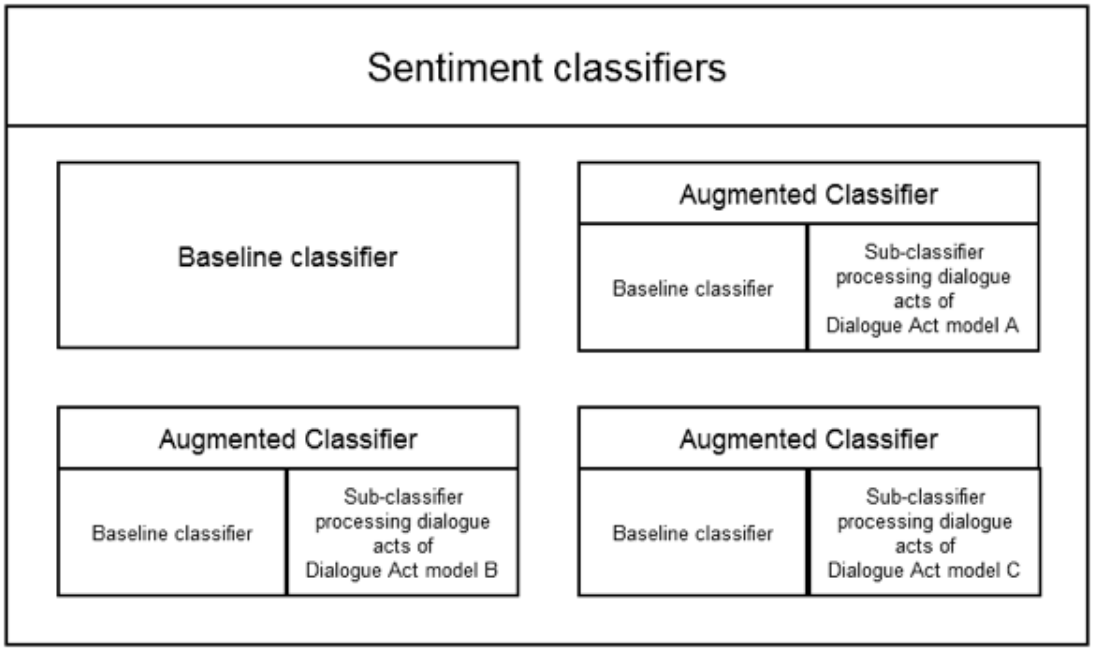

Figure 4.1: Sentiment classification scenarios

4.2.1 Classification procedure

Four classification scenarios are to be conducted to verify the applicability of the interpersonal acts for the improvement of sentiment analysis. In one

scenario, the sentiments are classified by a baseline classifier, while in the other three scenarios, the classification will be achieved by augmented classifiers, each processing a different set of dialogue act labels as an additional feature set (see Figure 4.1).

In the scenario, when the baseline classifier is utilized, it processes only the audio streamings of the dialogues and their textual transcriptions. Figure 4.2 depicts the architecture of the classifier consisting of two sub-classifiers.

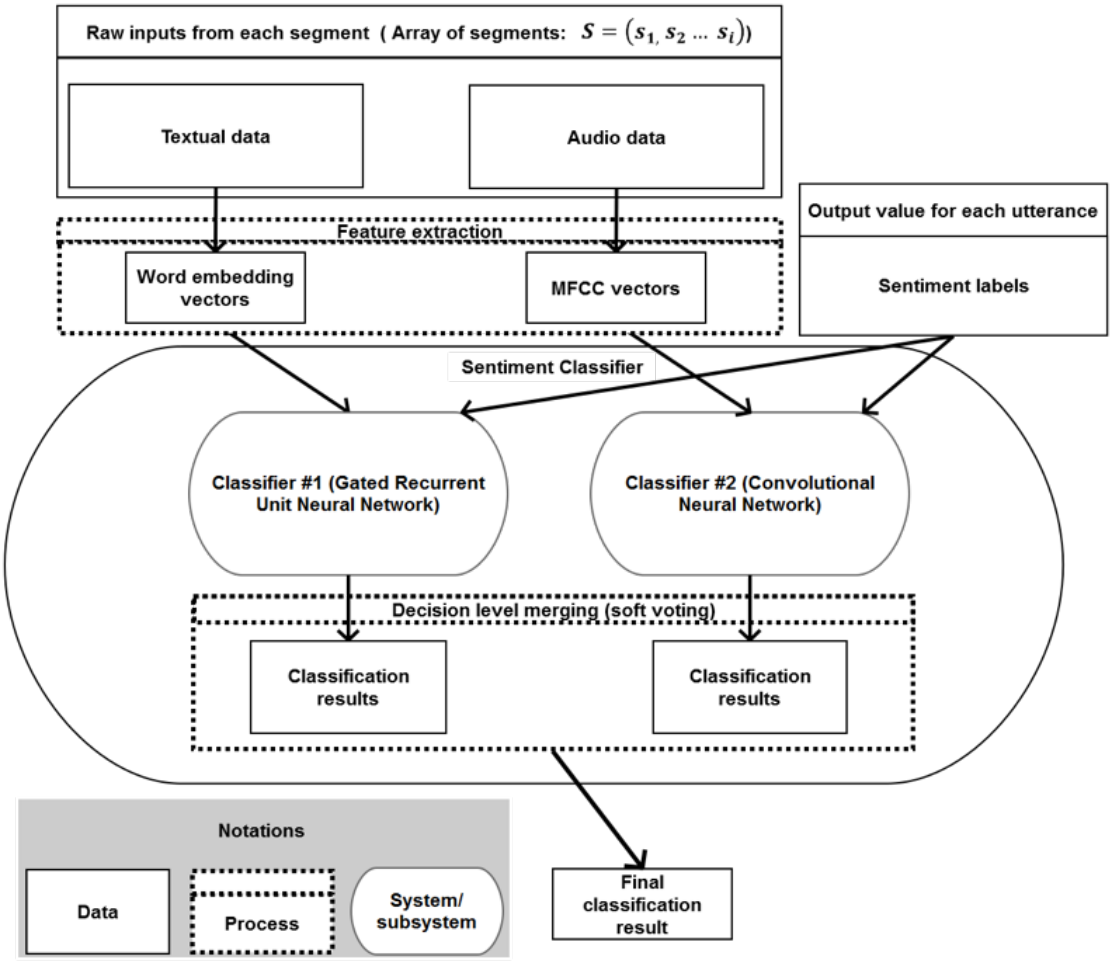

Figure 4.2: Baseline sentiment classifier

Sub-classifier #1 is a Gated Recurrent Unit (GRU) Neural Network which processes the textual transcriptions of the audio streams. A GRU can recall its previous internal states to process sequences of inputs and find possible dependencies within long sequences of embedded utterances (functional-segments) [47].

Sub-classifier #2 processes the audio data. The audio feature vectors are processed by a one-dimensional Convolutional Neural Network (CNN) which can effectively extract the important vectors among a large number of others through its several convolutional and pooling layers [48]. Both sub-classifiers are trained and tested on the same sentiment labels. The audio and textual features extracted from the functional-segments are processed in the order they occurred in the conversation, to help the GRU find meaningful dependencies between them. Using the ensemble learning method of soft-voting [49], the results of the two independently-trained sub-classifiers are to be merged at the decision-level.

In the other three scenarios, the baseline classifier is augmented with a third sub-classifier, another GRU, processing dialogue act labels as textual data (to find the possible dependencies in their sequence). In each scenario, the sub-classifier is processing dialogue act labels from one of the dialogue act models of IA, DIT++ or SWBD DAMSL.

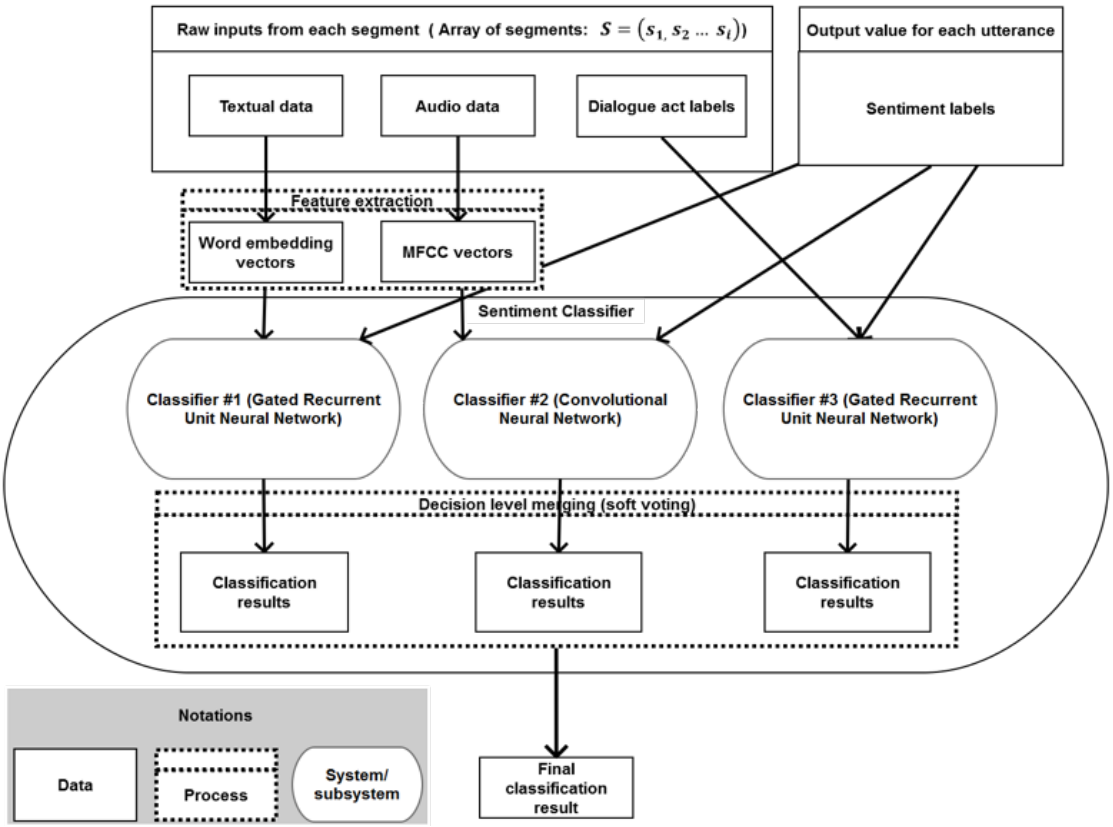

Figure 4.3: Augmented sentiment classifier

Similarly to the baseline method, the classification results of each sub-classifiers are weighted and merged through soft-voting. Through weighting the results, the system could learn which sub-classifier is contributing the most to the correct classification result. The dialogue act classifying sub-classifier #3 for example, assumed to contribute less than sub-classifier #1 or #2. Majority voting would not allow for such learning, it would simply output the result, chosen by at least two sub-classifiers, or choose among the results arbitrarily if all outputs of the sub-classifiers are different. Decision-level merging was chosen instead of feature-level merging (where the audio feature vectors, word embedding‘s of the transcriptions and digitalized dialogue act labels would be merged into one tensor before fed into the network) because in accordance with the study of Planet and Iriondo [50], preliminary experiments showed that decision-level merging yield better classification results.

The target labels for the training and testing of sub-classifier #3 are also sentiment labels. Figure 4.3 depicts the architecture of the augmented baseline method used in the other three scenarios.

4.3 Computational validation - semi-supervised learning

The applicability of the developed tagset is also proposed to be tested in semi-supervised polarity classification experiments. The experiments are to prove that the utilization of the developed IA model as a basis for search phrases in a semi-supervised multiple instance learning based method can yield satisfactory level polarity recognition accuracy, while not requiring any hand-made labeling.

4.3.1 Movie scene features as labeled bags

In this era of big data, there is an abundance of unsalvaged data on the world-wide-web. With the proper techniques, however, structured data can be mined, gathered in a way that renders hand-made annotation unnecessary. When one performs a search on the web, the search engine in use selects and returns data (let it be websites, videos or other content) according to a given search phrase. Thus the outputted data is structured in the sense that each datatpoint has a certain level of relevance to the search phrase.

In the case of affective state recognition, one would need data relevant to sentiments or emotions. Thus thinking in the level of websites would be too broad, a conversation-level search is needed. As a vast resource of dialogues, movie scenes can be harvested. YouTube, for example, contains millions of videos with movie

scene contents. Not only YouTube contains large amount of dialogic scenes, but these scenes are also categorized by the titles they are uploaded with. Several of the videos also contain tags, describing their content. Thus a search query on videos, related to a certain emotion (e.g. ‘angry scenes‘) would supposedly return videos containing dialogic scenes where at some point the emotion of e.g. ‘anger‘ is expressed by at least one of the interlocutors. Not only direct search on emotion categories but categories of emotion-indicating communicative functions (e.g. for the emotion of ‘anger‘the intention of ‘criticize‘, ‘argue‘or ‘despise‘; see 4.1) could yield similar results.

Relevant videos to ‘anger‘ will not only contain verbal and/or non-verbal expressions of anger but will have them as their base concept.

The YouTube 8M dataset (Y8M) [51] contains frame- and video-level audio and image features of 6.1 millions of YouTube videos in total, with thousands of movie scene contents (at the present). The dataset is free to download under the Creative Commons Attribution 4.0 International license [52]. Since the original audio-visual versions of the feature-sets in the Y8M are also available and trackable online on YouTube, an indirect YouTube search can be conducted in the Y8M for audio features of emotion - or emotion-related intention-expressing videos. As the number of relevant videos for each search phrase is restricted in the Y8M dataset, experiments are to be conducted utilizing multiple dimensions of search-phrases (emotions and communicative functions as well) alternatively as well as simultaneously to reach better recognition accuracy.

Each feature set of every video selected from the Y8M would contain features of a dialogic scene, focusing on the direct and/or indirect (through emotion sensitive communicative functions) expression of a certain emotion (but probably containing several other emotions as well). Thus each feature set would serve as “weakly” labeled bags for emotions. As the author thrives to find a method easily applicable for commercial use, the technology-sensitive visual features (requiring the application of visual sensors) contained in the Y8M are not processed. From the frame-level and video-level audio features, only frame-level features are to be applied since video-level features would not be applicable for instance-level training.

The MIL task is to group the frame-level audio features of each bag into processable instances and automatically label them with emotion labels based on the emotion / emotion-indicating intention containing search phrases the videos correspond to. An instance is to be labeled with the emotion label of the bag (labeled positive in standard MIL) if it represents the concept (an emotion) identical to the bag; otherwise, it is not labeled (labeled negative in standard MIL). Labeling would be achieved through latent variable-based unsupervised clustering,

discussed in the next section. Feature instances of several videos can be grouped together along the emotions they are labeled to represent.

4.3.2 Instance-level polarity detection

To advance real-time emotion recognition through the proposed weakly supervised method, polarity classification assumed to be more applicable than emotion recognition. The bags of emotions, mined from "weakly labeled" YouTube video features would be merged into positive and negative polarity-representing bags. Choosing polarity detection over emotion recognition is in accordance with the fact that commercial applications requiring real-time affective recognition would benefit more from less-fine-grained but more reliable classification results. The reason for choosing polarity detection instead of the neutral aspect-including sentiment classification is that the authors could not find a reliable method to find videos where 'neutral'emotions are expressed in a dialogic environment.

Accordingly, an aggregated set of audio feature instances, affiliated with the same basic emotion would constitute a bag of a certain emotion, and an aggregated set of emotion bags with the same polarity would constitute a final bag of positive or negative polarity.

Formally, a polarity bag is defined as follows:

$$B = \{(x_1, y_1), \ldots, (x_n, y_n)\} \tag{4.2}$$

where y_i is a polarity label and instance $x_i \not\subset X \not\subset A$ of X bag of emotion in A bag of aggregated positive or negative polarity emotions.

4.3.3 Unsupervised-clustering based classification

This study proposes a novel approach towards multiple instance learning in the form of unsupervised clustering of bag instances, mapped into the feature space of their latent variables. The approach is depicted in Figure 4.4

Training procedure:

1) Frame-level audio feature sets of YouTube videos, edited (by the uploaders of the video) to focus on the expression of a certain basic emotion and/or emotion-sensitive communicative function is to be selected from the Y8M (several videos for each emotion). In particular, the titles attached to each frame-level audio feature set are to be extracted and matched to the titles

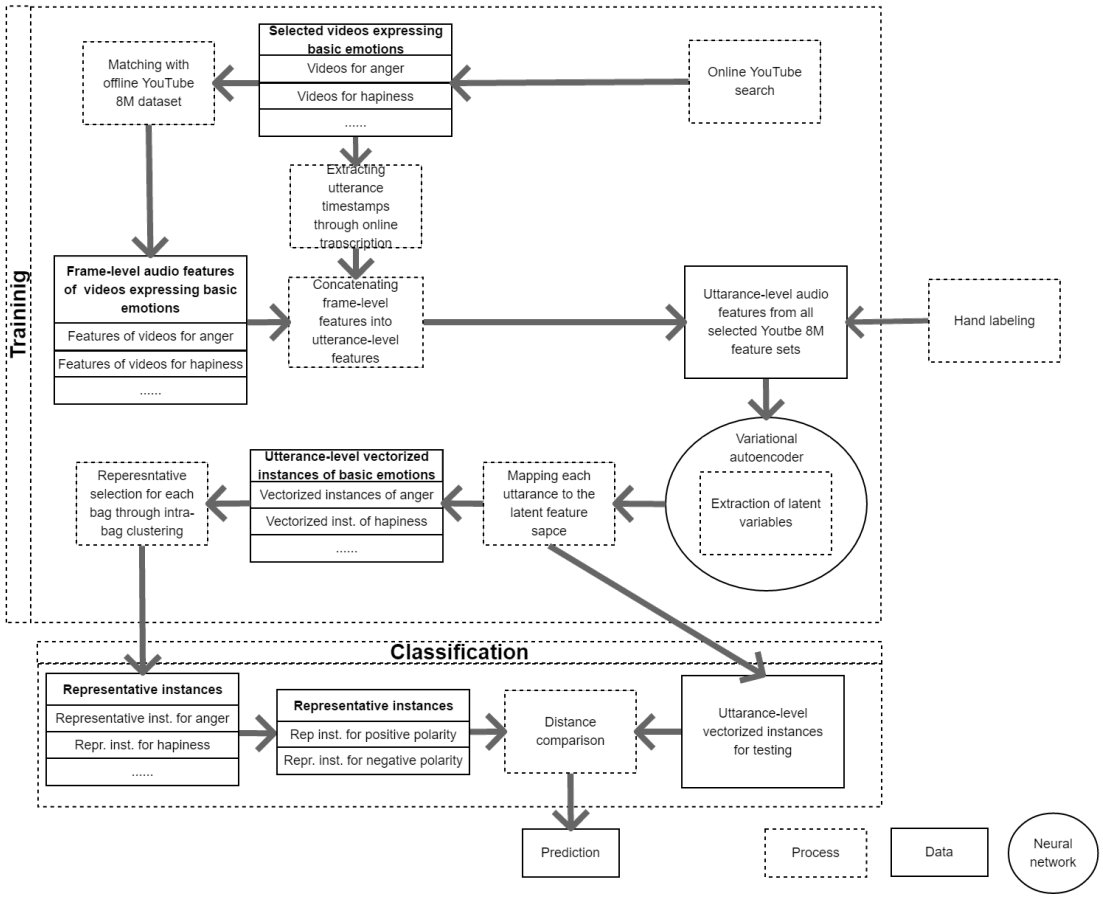

Figure 4.4: Latent variable-based unsupervised clustering

of online YouTube search. The search is thus conducted on all of the YouTube videos including those contained in the Y8M dataset in their feature-extracted form. Although only videos of the YouTube 8M dataset are to be utilized, through this method, the synonym- and relevance- measures of the YouTube search would be salvaged. The feature sets in the Y8M dataset were also categorized by the entities/topics of the videos they are extracted from, but since these entities were not afflicted with affective constructs, this categorization was not utilized.

2) The frame-level audio feature sets from Y8M of each selected video will be grouped into utterance-level features. The grouping is conducted based on timestamps provided by an online text converter applied on the online version of the videos (textual transcriptions are not to be used in accordance with YouTube‘s terms of service [53]).

3) Latent variables of all utterance-level features are to be extracted through a Variational Autoencoder (VAE) [54] and population of the MIL bags will be performed based on the instances‘positions in the latent feature space. Building the proposed MIL on latent variables is to counteract the large amount of noise expected in the bags of dialogic utterances.

   VAEs extract latent variables through encoder layers, transforming the input data into abstract variables. Then, through decoder layers, the variables are transformed back into a predicted input form. During training, the abstract variables are constantly updated according to the loss between the original and predicted input. In VAEs, constraints are added to force the generation of latent vectors to roughly follow a unit Gaussian distribution, usually set to be a centered isotropic multivariate Gaussian. The isotropic Gaussian priors allow each latent dimension in the representation to push itself as far as possible from the other factors [54]. Thus, VAEs are known to give representations with disentangled factors which attribute is crucial in a noisy dataset gathered form features of “weakly” labeled YouTube videos. Through training the VAE on all of the features (not separately for each video), the extracted latent feature space would include all concepts possible for the selected videos.

4) The utterance-level features can then be mapped to the latent feature space via transforming each audio feature vector into a vector indicating its affiliation towards each extracted latent variable (i.e. the datapoint‘s position in the latent feature space).

5) All transformed vectors are grouped by the emotions the video (they originate from) was selected by. In the case of vectors of communicative functions representing videos, the vectors are to be grouped along the emotions they correspond to. Thus emotion-bags are created, consisting of utterance-level instances of several videos, supposedly focusing on the expression of the same emotion and/or emotion-related communicative function.

6) The instances are clustered within each bag with an unsupervised clustering method. Once stable clusters are found, instances in the largest cluster are selected as representatives for the given emotion bag.

7) After removing non-representative instances from each emotion bag, the polarity bags can be populated with representative instances of the corresponding emotions. These bags are to serve as the base of a similarity-measure based classification. Emotion bags can also be used before aggregation to train basic emotion classifiers.

Classification

1) To compile a test set, similarly to the training procedure, first basic emotion-oriented movie scenes need to be selected and matched with their frame-level audio feature sets in the Y8M.

2) After grouping the frame-level features into utterance-level instances, each instance needs to be hand-labeled with basic emotion labels and polarity labels (in accordance with the corresponding emotion label).

3) The VAE pre-trained on the audio features of the training set extracts the latent variables from the audio features of the labeled test instances.

4) The instances can be transformed into vectors representing their position in the latent feature space.

5) The similarity between the transformed test instances and the instances populating both polarity bags is to be measured to predict the label of each test instance. Comparing the predicted labels with the original hand-made labels yields the classification result.

Chapter 5

Experiments

This chapter describes the empirical and computational experiments conducted to verify the adequacy of the proposed IA model. It details the experiments in terms of the labeled datasets, the dialogue act models used for comparison, the specific setups for the computational experiments, the computational methods and environment the experiments were conducted in, and the experimental results.

5.1 Empirical validation

5.1.1 Dialogue act taxonomies used for comparison

Empirical validation experiments to measure the emotion-sensitivity of the proposed IA model was conducted on cooperative gaming data, to measure the proposed model‘s adequacy in an environment it meant to be applied. The gathered dataset consists of Japanese conversations, thus the extended IA model was used (see 3.2.2).

Two other dialogue act taxonomies were used for the purpose of mutual comparison regarding emotion-sensitivity. The SWBD-DAMSL and DIT++ dialogue act models were selected because they are widely known and used, and represent communicative functions using one-dimensional and multi-dimensional approaches, respectively.

The SWBD-DAMSL tagset defines dialogue acts for 42 one-dimensional (mutually-exclusive) intentions. Although it contains a few dialogue acts for social obligations, it does not include acts accounting for social status or intentions that would influence self-esteem, and a wide range of subsequent emotions.

The multi-dimensional DIT++ consists of one set of ‘general-purpose communicative functions‘(intentions) and nine tagsets constituting ‘dimension-specific communicative functions‘such as ‘auto-feedback‘, ‘allo-feedback‘, and so on (for the dimension of ‘task/activity‘no tagset of communicative functions are defined). Well-formed tags on functional-segments are pairings of <D,F> where D is one of the ten dimensions and F is a communicative function of the corresponding dimension (e.g. <‘autoFeedback, request‘>). DIT++ assumes that every functional-segment of the dialogue is initially annotated with one tag from the dimension of ‘general purpose communicative functions‘. In addition,

every functional-segment can be optionally tagged with up to nine tags, one for each ‘dimension specific communicative function‘ dimension. As each dimension contains mutually-exclusive tags, one segment can be annotated with between one and ten tags in total. DIT++ has a function-specific dimension of ‘social obligations‘, containing communicative functions such as ‘greeting‘, but its acts do not account for non-obligatory interpersonal relation management. This study considered 22 ‘general purpose‘ acts, including all specifications described by Bunt [2009], and ten acts from the dimension of ‘social obligations‘. However, since the acts from the dimension of ‘social obligations‘ (e.g. <‘social obligations, greeting‘>) defined in DIT++ can only co-occur with one ‘general purpose‘ act, a linearized tagset would contain 32 different acts in total, considering all possible combinations among the two dimensions.

5.1.2 Annotated corpus

Five natural language dialogues from the Online Gaming Voice Chat Corpus with Emotional Labels (OGVC) [55] were selected to validate the IA model. The conversations were performed in Japanese during massively multiplayer online role-playing game (MMORPG) sessions; the specific games involved were Ragnarok Online, Monster Hunter Frontier, and Red Stone. The in-game context of these dialogues demands co-operation, is rich in stimuli, and potentially provides for a wide range of emotions. The dialogs were performed in Japanese, for which no large datasets labeled with interpersonal relation-indicating (or other emotion-related) tags exist.

The five conversations consist of a total of 6,902 spontaneous utterances. Three dialogues were performed by three pairs of male players (4,397 utterances), and two dialogues by two pairs of female players (2,505 utterances). The corpus contains both transcriptions and audio recordings of each conversation. The conversations are initially segmented into individual utterances, with each interlocutor identified.

To provide a proper context for the annotators, the textual and audio data was reassembled into a dialogic form based on the timestamps in the monaural sound files. (These files do not include network delay data, and for that reason, the exact timing of the conversations could not be fully reproduced). The conversations were re-segmented into functional-segments (not necessarily corresponding to one utterance) by the author and a native Japanese speaker, yielding 6,934 functional-segments in total.

For the purpose of evaluation, each segment had to be annotated with tags from DIT++, SWBD-DAMSL and IA. It has also been tagged with the emotions experienced by the interlocutors. Annotators were employed to assign tags from the three models, and to complete the emotion tagging. (The corpus creators

annotated only 80% of the original utterances with basic emotion tags.)

The size of the functional-segments differs in each model. For example, the functional-segment for the act <'general-purpose, answer'> from DIT++ was often expressed in a single utterance segment, while 'empathizing' from the IA model tended to be expressed through two or three utterances. The smallest possible segments were therefore chosen during functional-segmentation, considering all the three models. In the case when a functional-segment of a given model (typically IA) covered several smaller segments, all those segments were annotated with the same tag. On average, a dialogue contained 1,386 segments corresponding to approximately 62 minutes of audio.

5.1.3 Emotion annotation

Eight of the ten emotion tags employed by the compilers of the corpus are identical to the basic emotions defined by Plutchik [2]: joy, sadness, anger, fear, acceptance, disgust, surprise, and anticipation. The remaining two tags, 'neutral' and 'other', are complementary; their purpose is to account for emotions that cannot be classified into any of the original eight categories. In the case of segments that had already been tagged with emotions by the corpus compilers, only tags assigned by at least two of the compilers were retained. When all three compilers assigned different tags, one tag was selected (based on the judgment of a fourth native speaker hired for this task) and retained.

No original tags were provided for 1,355 of the functional-segments. Tags were added to them by three native Japanese speakers, employed for the experiments using the ten emotion labels. The annotators were three male university students between the age of 20-23 with over 100 hours of online-gaming experience each. Transcripts and audio recordings of the dialogues were provided to the annotators, who were asked, to determine the underlying emotion type of the interlocutors for each segment. All segments, therefore, received one emotion tag. Before the actual tagging procedure, each annotator participated in a brief training session, where 150 consecutive example segments (from the same corpus, not used in the study) were shown with suggested emotion labels. The segments were for the experiments to show how the labels should be attached in common and uncommon cases (e.g. a certain emotion type was expressed through not one but several segments). The inter-annotator agreement for emotion tags assessed with Fleiss' Kappa [56] was 68.3%.

5.1.4 Dialogue Act Annotation

The corpus was then annotated with the dialogue act tags of SWBD-DAMSL and DIT++, and with the interpersonal act tags of the IA model. Three

native Japanese speakers (different from those who annotated the emotions) were employed. Each added tags from one of the three tag sets to the transcriptions of the five dialogues while listening to the corresponding audio recordings. Since three models were used, the annotation was conducted in three iterations, each iteration for a different dialogue act model.

The annotators were two male and one female university students between age 21-25, with more than 80 hours of online gaming experience each. They were instructed to determine the interlocutor‘s intention for each segment and to label it with the most appropriate dialogue act tag from each dialogue act model. The annotators participated in a training session similar to the one conducted for the emotion labeling. This training session involved the same 150 consecutive functional-segments, repeated three times. Each time the 150 segments were labeled with the tags of one of the dialogue act models, showing how the dialogue acts of the given model could be expressed through one or several segments. The annotators were cautioned that certain acts of certain models (typically the acts of the IA model) tend to be expressed through several functional-segments. The inter-annotator agreement (estimated again with Fleiss‘ Kappa) was 69.1% for the DAMSL SWBD tagset, 71.7% for DIT++, and 66.2% for the IA model. The IA model had the lowest ratio because the functional content dimension it covers permits more subjectivity than do the other models.

Similarly to the case of emotion tags, a single dialogue act tag from each taxonomy was assigned to each segment (hence three tags per segment). Any tag selected by at least two annotators was retained for the analysis; otherwise, one of the three different tags assigned by the annotators was retained (based on the judgment of a fourth native speaker).

A preliminary analysis revealed that the DIT++ tagset is over-specified for the given experimental data. To compensate for the data set‘s limited size, several optional class specifications were omitted. This was considered reasonable as it improved the performance of DIT++ in the experiments. The number of acts in each taxonomy considered during analysis was further decreased by disregarding those not assigned to any functional-segment. For the analysis, 28 acts of SWBD-DAMSL:
3rd party, Acknowledge, Affirmative non-yes, Agree, Apology, Appreciation, Backchannel question, Conventional close, Declarative question, Directive, Hedging, Maybe, Negative answer, No-answer, Non-verbal, Non-understandable signal, Offer, Open-question, Or-clause, Other answers, Response acknowledgement, Statement, Statement-opinion, Summarize, reformulate, Tag-question, Thanking, Wh-question, Yes-no-question
and 17 acts of DIT++:
Address request, Address suggestion, Agreement, Answer, Apology, Check-question,

Confirm, Disagreement, Disconfirm, Inform, Instruct, Offer, Propositional question, Request, Set question, Suggestion, Thanking were retained.
All 12 acts of the extended IA model occurred in at least one of the five dialogues: *Partner-unrelated positive commenting, Partner-unrelated neutral commenting, Partner-unrelated negative commenting, Paying attention, Empathizing, Accepting as superior, Agreeing, Self-image improving, Criticizing, Indiscrete commenting, Indebting partner, Commanding/requesting*

5.1.5 Experimental results

Occurrence of emotions and dialogue acts

The aggregated occurrence and co-occurrence ratio of emotions and dialogue acts in the five dialogues are summarized in Figure 5.1 for the SWBD DAMSL tagset, in Figure 5.2 for the DIT ++ tagset, and in Figure 5.3 for the IA model. The figures show the empirical distributions of the ten emotions differentiated by random color assignment (the horizontal axes) across the model‘s dialogue acts (the vertical axes). The area of each unit is proportional to the corresponding emotion-speech act co-occurrence frequency (the width is directly proportional to the emotion frequency, and the height – to the dialogue act frequency).

The most-frequently observed emotions are ‘acceptance‘ (18.2% of the total number of segments) and ‘neutral‘ (24.8%). On the other hand, emotions with negative valence, such as ‘sadness‘ (5.3%) and ‘anger‘ (2.4%) were observed relatively rarely.

In the case of dialogue acts, the data reveals larger diversity in their distribution than in the case of emotions. From the SWBD-DAMSL tagset, 13 acts (of the 28 retained for analysis) occurred in fewer than 1% of the total number of functional-segments. Observing that other 14 acts from the original tagset (see 5.1.4) never occurred in the conversations, the tags of the SWBD-DAMSL may be considered over-specified for the given data.

In the case of the DIT ++ tagset, the occurrence ratio of speech acts is better balanced. Of the 17 acts, eight occurred in less than 1% of the whole data (other 16 were earlier excluded from the analysis due to their complete absence in the annotated data – see 5.1.4). This suggests that the DIT ++ tagset is also over-specified for the data. The most frequently observed acts are ‘Inform‘ (28.5%), ‘Answer‘ (25.2%), and ‘Confirm‘ (17.5%) that mainly cover gameplay-related utterances.

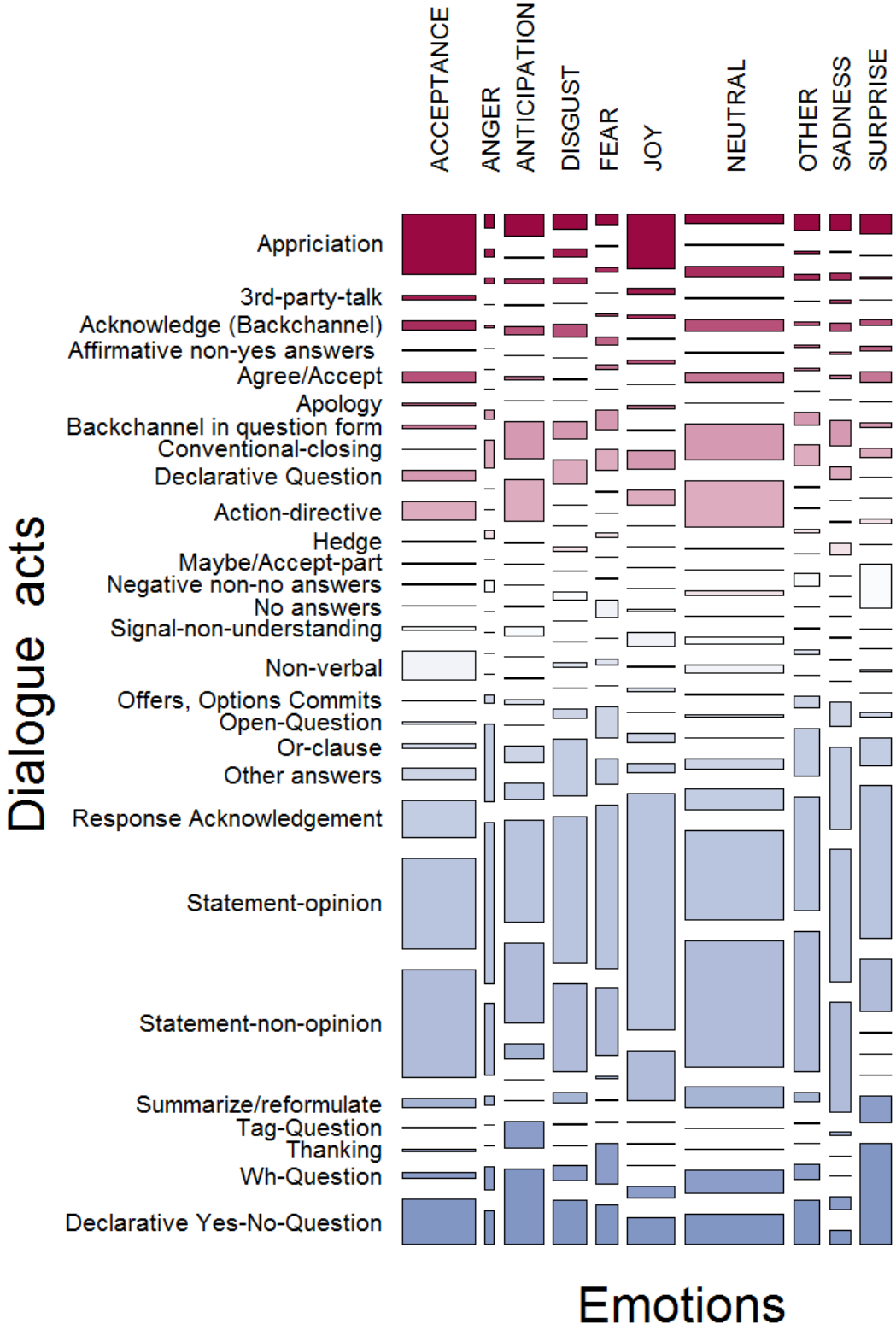

Figure 5.1: Co-occurrence ratio of basic emotions and the SWBD-DAMSL acts

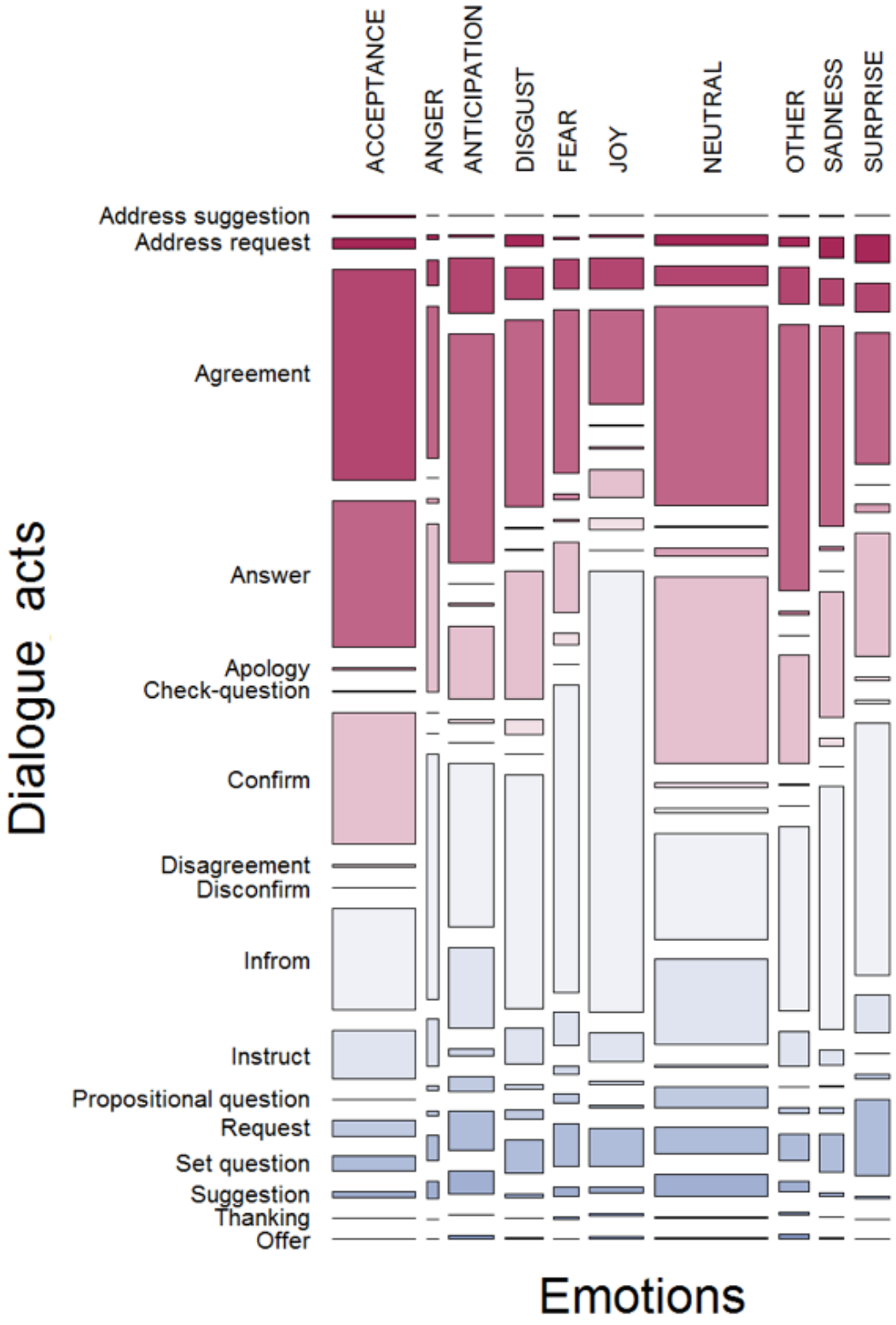

Figure 5.2: Co-occurrence ratio of basic emotions and the DIT++ acts

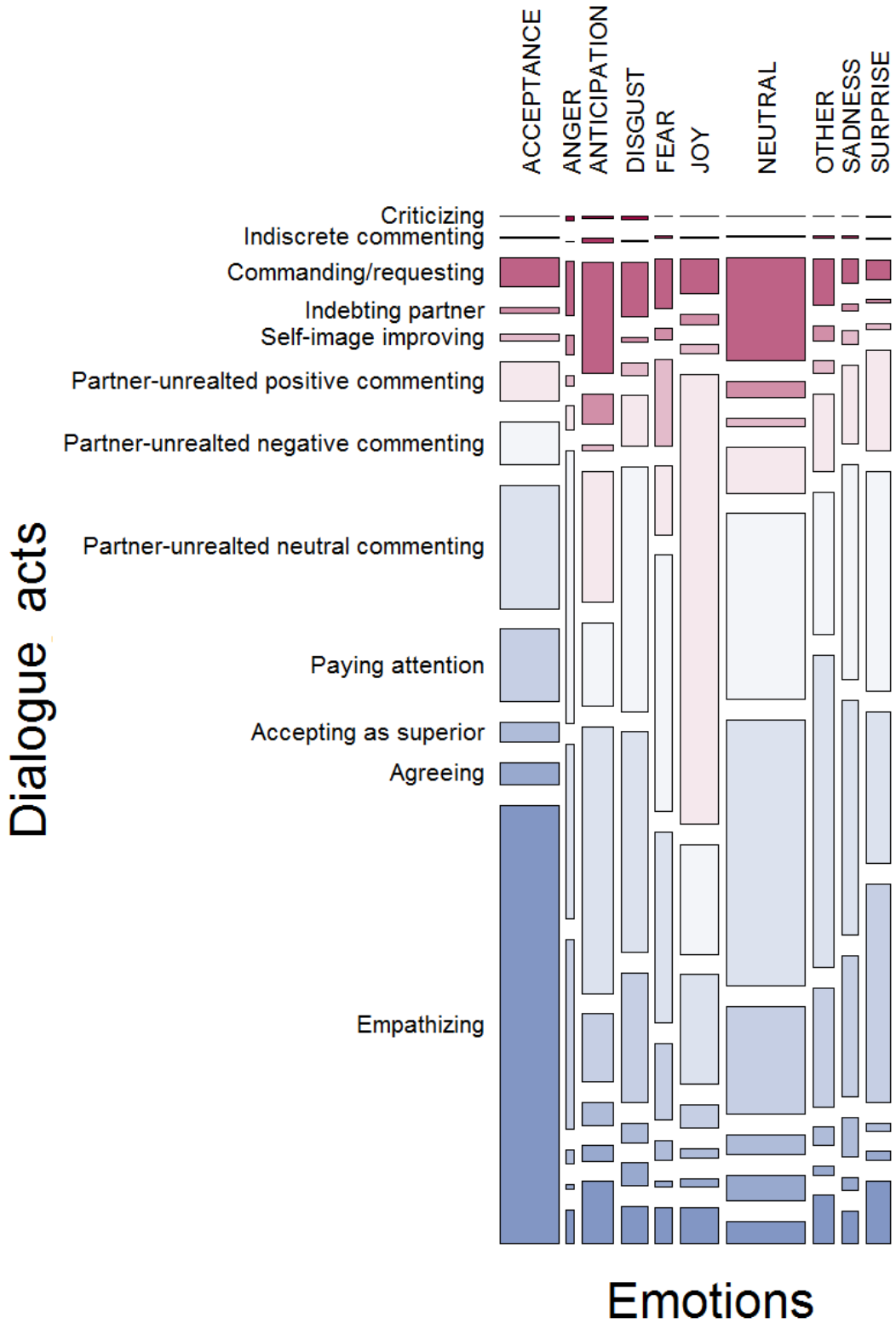

Figure 5.3: Co-occurrence ratio of basic emotions and the IA model acts

In the case of the IA model, all the originally defined 12 acts were retained after labeling, implying that the number of defined categories fits the data well (the number of social acts was intentionally kept low, but their specification was completed without pre-analysis of the data.) Of the 12 acts, two occurred in less than 1% of the utterances: 'Indiscrete commenting' (0.2%) and 'Criticizing' (0.1%). It is understood that further experiments would be required to validate the universal applicability of the IA definitions to dialogues from other domains, especially in the cases of 'Criticizing' and 'Indiscrete commenting'. The most frequently occurring dialogue acts are 'Partner-unrelated negative commenting' (19.1%), 'Partner-unrelated positive commenting' (13.7%), and 'Partner-unrelated neutral commenting' (25.4%).

Overall dependency between emotions and dialogue acts

A Fischer's exact test [45] showed that the dialogue acts of all the three models have significant ($p < 0.001$) dependence with emotions. Analysis of the sums of positive npmi values normalized by the degrees of freedom (see Table 5.1) revealed that the IA taxonomy has the strongest overall correlation with emotions, compared to the other two models.

Table 5.1: Dependency between emotions and the dialogue acts

	Degrees of freedom (df)	Sum of positive npmi values	Sum of positive npmi values / df
SWBD-DAMSL	28	9.08	0.324
DIT++	17	5.87	0.345
IA	12	4.74	0.395

Association pairs

Based on the npmi values each dialogue act has with each emotion, association values were obtained, indicating their co-occurrence frequency, with regards to their marginal probability. Considering the ranges of the obtained values, there are no very strong [0.60, 0.80) or extremely strong [0.80, 1.00] associations between any of the three models' dialogue acts and emotions. Although most dialogue acts are weakly associated with more than one emotion, several of them have a single medium-strength association with an emotion in [0.20, 0.40). There also was one act of the IA model that was positively (strongly, calculated npmi[0.40, 0.60)) associated with each and every occurrence of a given emotion and that, therefore, would serve as an exclusive indicator for that emotion.

Figures 5.4, 5.5, and 5.6 depict the associations detected with positive npmi values. In the figures, dialogue acts having medium-strength association with a given emotion are indicated with blue borders along with the corresponding npmi values. Strongly associated dialogue acts and their npmi values are indicated with red borders, and the exclusive indicator act of the IA model is with green borders. Non-positive npmi values are not displayed.

SWBD-DAMSL has pairs the following medium-strength associations:

- ‘Anticipation‘-‘Acceptance‘(npmi: 0.21)
- ‘Non-verbal‘- ‘Acceptance‘(npmi: 0.24)
- ‘Or-clause‘- ‘Acceptance‘(npmi: 0.25)
- ‘Apology‘- ‘Fear‘(npmi: 0.21)
- ‘Neutral‘- ‘Joy‘(npmi: 0.22)
- ‘Signal-non understanding‘- ‘Surprise‘(npmi: 0.32)
- ‘Declarative yes-no question‘- ‘Surprise‘(npmi: 0.20)

DIT++ has pairs the following medium-strength associations:

- ‘Disconfirm‘-‘Neutral‘(npmi: 0.31)
- ‘Inform‘- ‘Joy‘(npmi: 0.20)

and a pair of strong association between ‘Agreement‘and ‘Acceptance‘(npmi: 0.41).

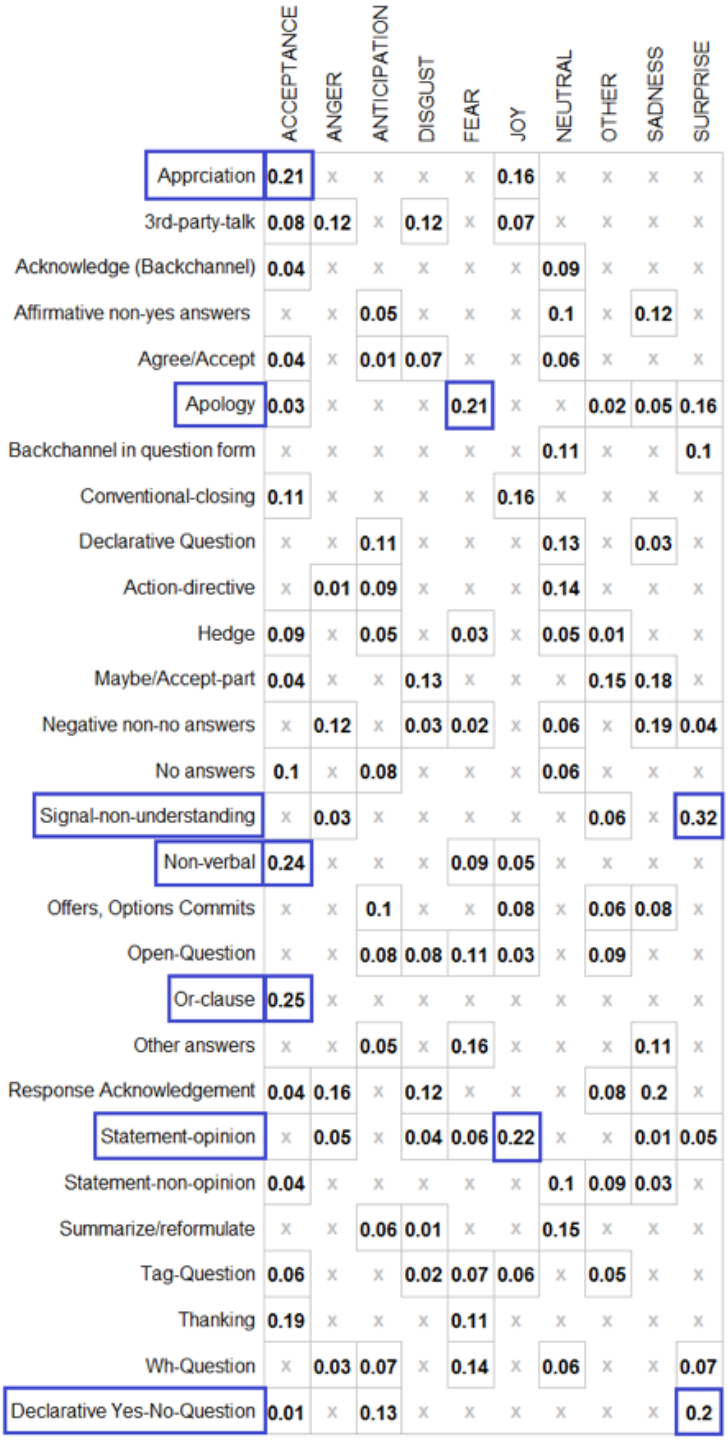

Figure 5.4: Associations between basic emotion tags and the SWBD-DAMSL acts

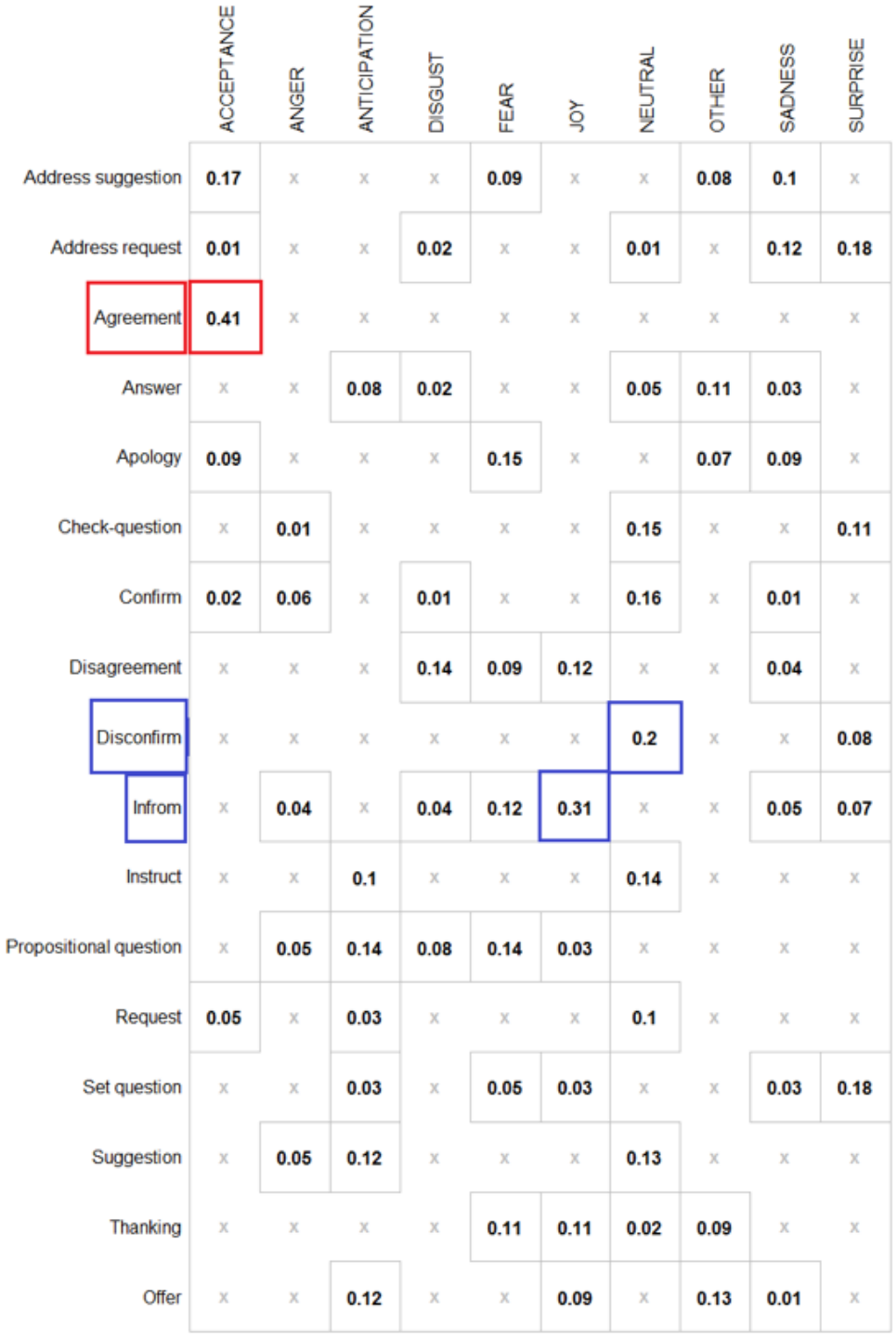

Figure 5.5: Associations between basic emotion tags and the DIT++ acts

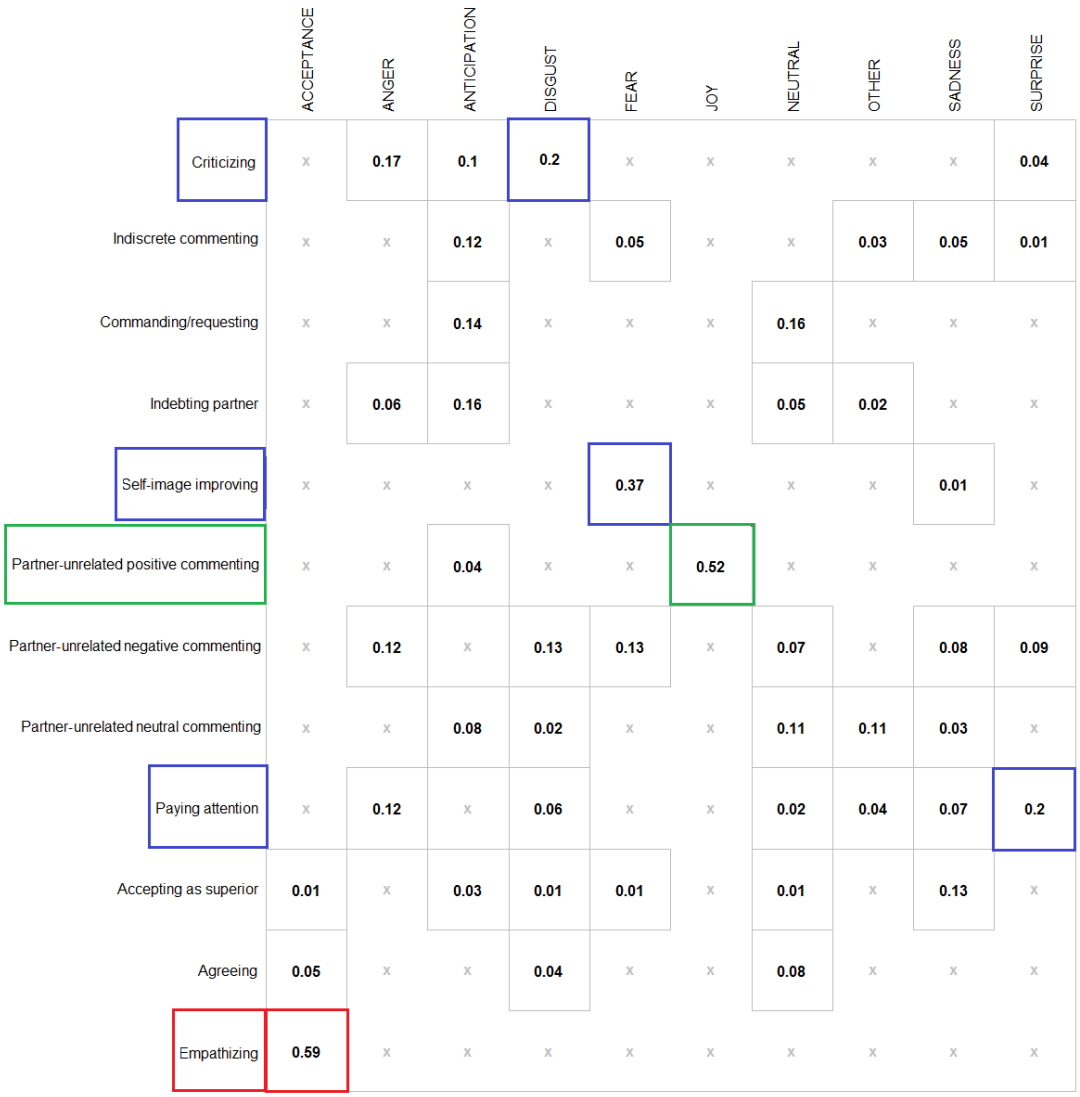

Figure 5.6: Associations between basic emotion tags and the IA model acts

The IA tagset has the following medium-strength associations:

- 'Disgust'- 'Criticizing'(npmi: 0.20)
- 'Self-image improving'- 'Fear (npmi: 0.37)
- 'Paying attention'- 'Surprise'(npmi: 0.20)

It also has one pair of strongly associated pairing of 'Empathizing'- 'Acceptance'(npmi: 0.59), and a pairing of strong association where the emotion

is associated with the interpersonal act exclusively: ‘Partner-unrelated positive commenting‘- ‘Joy‘.

Indicative power

The overall emotion-indicative power of the models was assessed by the model‘s element-wise ratio of the good, strong, and exclusive indicators, and the results are summarized in Table 5.2 The ratio is computed by dividing the frequency of each indicator type by the number of tags (retained for annotation) in the given taxonomy. (An indicator that is exclusive can also be strong, and an indicator that is both exclusive and strong can also be a good indicator. These intersections are accounted for in each group.)

Table 5.2: Element-wise ratio of the emotion-indicative power of dialogue acts

	Ratio of medium-strength associations	Ratio of strong associations	Ratio of exclusive indicators
SWBD-DAMSL	7/28=0.25	0	0
DIT++	3/17=0.15	1/17=0.05	0
IA	5/12=0.42	2/12=0.16	1/12=0.08

5.2 Computational validation - supervised learning

5.2.1 Data

The corpus used for the empirical experiments was already partly labeled with emotion labels, transferable to sentiment labels, and contains utterance-level textual and dialogue-level (chunkable to utterance-level) audio features. For the above reasons, it fits well the purpose of training and testing of supervised machine learning algorithms. Accordingly, computational validation of the proposed IA model through supervised sentiment recognition was tested on the same conversational in-game corpus.

Dialogue act tags of the extended IA (see 3.2.2), SWB-DMSL and DIT++ tagsets were kept for the same functional-segments. The same basic emotion labels used in the empirical experiments were collapsed into negative (consisting of anger, fear, sadness, and disgust), positive (consisting of surprise, joy, acceptance, and

anticipation), and neutral (consisting of neutral and other) sentiment labels. These labels correspond to the valance-categories of negative (consisting of angry, afraid, sad and annoyed), positive (consisting of astonished, happy, pleased/satisfied and excited) and neutral (consisting of neutral) from Russel's circumplex of emotions [57], with the addition of the emotion other to the neutral category.

For the training of the proposed method 80% of the all functional-segments (5547 segments) were used while the remaining 20% (1386 segments) constituted the test set.

5.2.2 Implementation

5.2.2.1 Pre-processing

Sub-classifier #1 process the word embeddings of textual transcriptions of the audio streams. The functional-segment-level text chunks were further chunked to words by MeCab [58] a morphological analysis engine designated for the Japanese language. After stripping the nouns, verbs, and adjectives from their conjugations and particles, word embeddings were created with the GloVe embedding algorithm [59] (for further explanation about GloVe see 5.2.2.3) which was trained on the Wikipedia dump data [60]. Through the algorithm, each word was transformed into a vector of 200 dimensions based on the co-occurrence probability with other adjacent words. The word embeddings then were re-ordered into their original position they had in the functional-segments constituting embedded sequences.

Algorithm 1 Preparing textual input

```
1: listofembeddedsequences(textualinputdata) ← empty
2: for all dialogue ∈dialogues do
3:     for all segment ∈dialogue do
4:         embedded sequence ←empty
5:         words ←chunking with Mecab(segment)
6:         for all word ∈words do
7:             word embedding ←Glove(word)
8:             append to embedded sequence(word embedding)
9:             padding with zeros(embedded sequence)
10:        end for
11:    end for
12:    append to embedded sequences (embedded sequence)
13: end for
       return list of embedded sequences
```

Each embedded segment was padded with zeros into a uniform length of 150 words, forming a three-dimensional input-array of the shape [number of datapoints, uniform sequence length, number of embedding dimensions].

Sub-classifier #2, processes the audio data. The audio files, containing each of the five dialogues, were partitioned into functional-segments. Then, every functional-segment was saved as a three-second length monaural wav file, lengthening the original segments with silence or shortening them in the case they were longer than 3 seconds. The segments were transformed by the OpenEar software [61] into 20 dimensions of low-level audio features incorporating: *signal energy, FFT-spectrum coefficients, mel-spectrum coefficients, mel-frequency cepstral coefficients, pitch, voice quality, LPC coefficients, PLP coefficients, formants, time-signals, and spectral bands.* The vectorized audio-input arrays have the shape of [number of datapoints, uniform sequence length(in milliseconds), number of embedding dimensions]

Algorithm 2 Preparing audio input

```
list of audio features (audio input data) ←empty
for all dialogue ∈dialogues do
    for all segment ∈dialogue do
        uniformized segment ←trimming/padding to 3 seconds(segment)
        set of low level audio features ←OpenEar(uniformized segment)
        append to list of audio features (set of low level audio features)
    end for
end for
    return list of audio features
```

5.2.2.2 Architecture

The word embeddings were processed by two consecutive layers or Gated Recurrent Unit (GRU) Neural Network, ending in a fully-connected, feedforward neural network layer with ‘softmax‘activation function (for classification of multiple classes), constituting sub-classifier #1. GRUs were used for they can recall their previous internal states to process sequences of inputs, and find possible dependencies within long sequences of embedded utterances [47]. (For further explanation about GRUs see 5.2.2.3).

The audio feature vectors were processed by sub-classifier #2, a One-dimensional Convolutional Neural Network (CNN), consisting of three convolutional and two pooling layers, and a final, fully-connected softmax layer

Algorithm 3 Sub-classifier #1

output of first layer ←GRU(*list of embedded sequences*)
output of second layer ←GRU(*output of first layer*)
classification result ←fully-connected softmax layer(*output of second layer*)
return classification result of sub-classifier #1

[62]. CNNs can effectively extract the important ones among the large number of vectors through its several convolutional and pooling layers [48]. (For further explanation about CNNs see the 5.2.2.3).

Algorithm 4 Sub-classifier #2

output of first layer ←1D CNN(*list of audio features*)
output of second layer ←pooling layer(*output of first layer*)
output of third layer ←1D CNN(*output of second layer*)
output of fourth layer ←pooling layer(*output of third layer*)
output of fifth layer ←1D CNN(*output of fourth layer*)
classification result ←fully-connected softmax layer(*output of fifth layer*)
return classification result of sub-classifier #2

The supplementary feature set of dialogue acts (of the given dialogue act model) was processed as textual data by a GRU and a feedforward softmax layer (to find the possible dependencies in their sequence) constituting sub-classifier #3.

Algorithm 5 Sub-classifier #3

output of first layer ←GRU(*list of dialogue act labels*)
classification result ←fully-connected softmax layer(*output of first layer*)
return classification result of sub-classifier #3

Using the ensemble learning method of soft-voting [49], the results of the two independently-trained classifiers were merged at the decision-level. Specifically, three fully-connected feed-forward network layers were trained on the classification results to acquire weights for them. The average of the sums of the weighted results was then computed.

All three sub-classifiers were trained and tested on the same sentiment output labels. The audio and textual features extracted from the functional-segments were processed in the order they occurred in the conversation, to help sub-classifier #1 and sub-classifier #3‘s GRUs to find meaningful dependencies within their sequences.

Algorithm 6 Soft-voting

1: output of first layer ←fully-connected softmax layer(*list of classification result of sub-classifier #1, #2, and #3*)
2: output of second layer ←fully-connected softmax layer(*output of first layer*)
3: classification result←fully-connected softmax layer(*output of second layer*)
return final classification result

5.2.2.3 Technical details

Word embedding

Language modeling and feature learning techniques vectorizing words or phrases from the vocabulary of a given natural language are collectively called word embedding techniques. The base concept is to computationally embed a high-dimensional space -with a dimension for each word - to a continuous vector space of a much lower dimension. Dimensionality reduction on the word co-occurrence matrix, neural networks, or probabilistic models are considered conventional word embedding methods. [63]

GloVe [59] is an unsupervised word embedding algorithm. It encodes an abstarctized form of semantic meaning based on word-to-word co-occurrence probabilities. Table 5.3 is an example considering the co-occurrence probabilities for the target words ‘steam‘and ‘ice‘with various words from the vocabulary of a 6 billion word English corpus (based on the table displayed in [59]).

It can be seen that ‘ice‘co-occurs more frequently with ‘solid‘than it does with ‘gas‘, whereas ‘steam‘co-occurs more frequently with ‘gas‘than it does with ‘solid‘. Both words co-occur with their shared property ‘water‘frequently, and both co-occur with the unrelated word ‘fashion‘infrequently. The noise from non-discriminative words like ‘water‘and ‘fashion‘cancel out in the ratio of probabilities. Accordingly, large values (much greater than one) correlate well with properties specific to ‘ice‘, and small values (much less than one) correlate well with properties specific to ‘steam‘.

Table 5.3: Probability and ratio of word co-occurrences

Probaility and ratio	k = solid	k = gas	k = water	k = fashion
P(k|ice)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(k|steam)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P(k|ice) / P(k|steam)	8.9	8.5×10^{-2}	1.36	0.96

GloVe utilizes a log-bilinear model with a weighted least-squares objective. Its objective is to learn word vectors such that their dot product equals the logarithm of the words'co-occurrence probability. Since the logarithm of a ratio equals the difference of logarithms, this objective connects (the logarithm of) ratios of co-occurrence probabilities to the vector differences in the word vector space. Because these ratios can encode some form of meaning, this information gets encoded as vector differences as well.

Recurrent neural networks

Recurrent neural networks (RNNs) are deep artificial neural networks designed to recognize patterns in sequences of textual, speech, handwriting, time series etc. data. RNNs utilize a temporal dimension to learn through chronologic sequences [64].

In particular, the prediction a recurrent net makes at time step $t-1$ affects its prediction about time step t. RNNs learn from two input sources: the present and the recent past, both of which are taken into account in the prediction of unseen data. RNNs, thus, are distinguished from feedforward networks by the continuous utilization of their own outputs as secondary input.

Sequential information is preserved in a hidden state, which can stretch over several time steps, affecting the prediction of each new datapoint. Accordingly, RNNs can find correlations between events separated by many moments as well. These correlations are called "long-term dependencies", described mathematically as the following:

$$h_t = \theta(W_{x_t} + U_{h_{t-1}}) \tag{5.1}$$

The hidden state at time step t is h_t. It is a function accounting for the input at time step x_t, modified by a weight matrix W of the hidden state of the previous time step h_{t-1}, and multiplied by its own hidden-state-to-hidden-state matrix U. The importance of the present input in relation to the past hidden state and vice-versa is represented through the weight matrices. Weight matrices are adjusted throughout the training of the network via backpropagation. The sum of the weight input and hidden state is squashed by the function θ– either a logistic sigmoid function or a tanh.

In every time step a feedback loop occurs, thus each hidden state contains traces of all preceding hidden states h_{t-1}, for as long as memory can persist. For the above reason, RNNs utilize an extension of backpropagation called 'backpropagation through time'(BPTT). In RNNs, time is expressed through an ordered series of calculations linking one time step to the next, to enable backpropagation. However,

since RNNs are seeking to establish connections between a final output and events several time steps before, during the backpropagation process -where the layers and time steps of deep neural networks relate to each other through multiplication- derivatives tend to vanish or explode.

Long-Short Term Memory RNNs (LSTMs) are designed to preserve the error so it can be efficiently backpropagated through time. [65] By maintaining a more constant error, LSTMs can link causes and effects remotely allowing them to effectively learn over several time steps.

LSTMs contain information outside the normal flow of the recurrent network in a gated cell. Information can be 'stored in', 'written to', or 'read from'a cell. The cell makes decisions about what to store, and when to allow 'reads', 'writes'and 'erasures'via input and output gates. Similarly to the neural network's nodes, the gates act on the signals they receive and block or pass the information based on their strength and importance, according to the gates own sets of weights.

The weights are adjusted via the LSTMs learning process: the cells learn when to allow data to enter, leave or be deleted through the iterative process of randomizing weights, backpropagating error, and adjusting weights via gradient descent.

A Gated Recurrent Unit Network (GRU) is a simplified LSTM without an output gate. At each time step, the contents from the GRU's memory cell is fully written into the larger net. GRUs are more suitable to be used on smaller datasets on which LSTMs may over-filter information [47].

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep artificial neural networks that are used primarily to classify images [48]. In mathematical terms, a convolution is the integral measuring how much two functions overlap as one passes over the other.

Each convolutional layer of a CNN consist of a set of adjustable filters. Every filter is small spatially but extends through the full depth of the input volume. For example, a filter applied on an image, with size 5x5x3 is five pixels long in width and height, and its depth is three, for e.g the three color channels. During the convolution, the images are sliced along their depth, and each filter is slid across the width and height of the given slice's input until it covers the whole slice spatially. The dot products between the entries of the filter (which are the weights) and the input at any position are computed, producing a two-dimensional activation map that reflects the responses of that filter on every spatial position. The spatial position checked by the filter at a given time is called local region. Thus, convolutional layers are called locally connected layers, since they process only a portion of the whole input data at a given time. Finally, the activation

maps are stacked along the depth dimension and producing the output.

When another convolutional layer is added to the network, the output of the previous layer becomes the input of the next one. The next layer then processes an input of activations, describing lower level features. Through the filters applied in the next layer, activations representing higher-level features will be outputted. The addition of several layers may result in getting activation maps that represent very complex features, and filters that activate when there is e.g. handwriting in the image. In the case when CNNs are used for classification tasks, once high-level features are obtained they are input into a fully connected layer at the end of the CNN, which performs the classification task.

With the usage of pooling layers, every depth slice‘s input volume can be downsampled along both width and height, discarding a significant proportion of the activations, to avoid overfitting.

A One-dimensional Convolutional Neural Network (1D CNN) works similarly to conventional CNNs, with the modification that the filter is sliding along a one-dimensional row of numbers [66]. In this case, the filter has only two dimensions. E.g. in the case of a 3x3 filter, the input data would be sliced into three channels where on each channel the filter would slide along the row of numbers checking three numbers at a time and mapping their dot product with the filter weights onto an activation map. Then, the three activation maps would be stacked to produce the output volume. This kind of layers is used in the case when two-dimensional input data is not available, typically in the case vectorized audio streams.

Soft-voting

Soft-voting is a type of ensemble method [49]. Ensemble methods combine the classification predictions of similar or conceptually different machine learning classifiers via majority or soft-voting. In the case of majority voting, the predicted final class label $\hat{y}$ is the label that has been predicted most frequently by each classification model C_j:

$$\widehat{y} = mode\{C_1(x), C_2(x), ..., C_m(x)\} \tag{5.2}$$

Thus, in the example case of three classifiers with the binary outputs of

- $C_1(x) : \widehat{y} = 0$
- $C_2(x) : \widehat{y} = 0$
- $C_3(x) : \widehat{y} = 1$

$\widehat{y}$ would be 0. Naturally, this method cannot be applied in the case of less than three classifiers.

Weighted majority vote associates a weight w_j with classifier C_j:

$$\widehat{y} = \arg\max_i \sum_{j=1}^{m} w_{j\chi_A}(C_j(x) = i) \tag{5.3}$$

where χ_A is the characteristic function $[C_j(x) \in A]$, and A is a set of unique class labels. Thus, with weighted majority voting the same outputs as above

- $C_1(x) : \widehat{y} = 0$
- $C_2(x) : \widehat{y} = 0$
- $C_3(x) : \widehat{y} = 1$

would yield $\widehat{y} = 1$.

In soft-voting, final prediction is the average the of the predicted class-probabilities from each classifier.

$$\widehat{y} = \arg\max_i \sum_{j=1}^{m} w_j(p_{ij}) \tag{5.4}$$

where w_j is the weight that can be assigned to the jth classifier. In the example case of non-weighted outputs of

- C_1(x): [0.9,0.1] (the predicted probabilities for class 0 | 1)
- C_2(x): [0.8,0.2]
- C_3(x): [0.4,0.6]

the averaged outcome would be [0.7,0.3], resulting in $\widehat{y} = 0$. However, assigning the weights of

- C_1(x) -> 0.1
- C_2(x) -> 0.1
- C_3(x) -> 0.8

the averaged outcome would be [0.25,0.35], resulting in $\widehat{y} = 1$.

5.2.2.4 Computing environment

The proposed approach of latent variable extraction and clustering were coded in *Python*, and the code is compatible with Python version 3.6. The code was run on Spyder IDE 3.2.6 on Ubuntu 16.04.2 LTS. The workstation used is equipped with an Intel Xeon E5-1650 v4 3.60GHz CPU and 128GB DDR4 RAM. In the current environment, the trained computational method performs the prediction of one datapoint between 0.18ms to 0.23ms (depending on the experimental setup).

The most important Python modules utilized in the implementation are listed below:

- **theano**: a library and optimizing compiler for manipulating and evaluating mathematical expressions, especially matrix-valued ones. Used for the implementation of CNN and RNN networks.
- **keras**: high-level neural networks API, written running on top of TensorFlow, CNTK, or Theano. In the experiments conducted, it was used with Theano backend.
- **pandas**: Pandas is a library for data manipulation and analysis. Used to load the textual transcriptions.
- **numpy**: NumPy is a library for scientific computing, especially for matrix transformations. Used to create, reshape save and load matrices of textual and audio inputs.
- **sklearn**: Scikit-learn is a machine learning library, used for the implementation of soft-voting.
- **matplotlib**: Matplotlib is a plotting library, used to plot co-occurrence analytics.

5.2.3 Experimental setups

Three experimental setups for sentiment recognition were developed based on the architecture proposed in 4.2.1 and used separately for all three augmented classifiers, processing a given dialogue act model. In each setup the input labels of sub-classifier #3 were different, to account for the variability in the number of functional-segments the dialogue acts of each given model tend to be expressed. The author tried to find the most optimal setups in the case of each dialogue act model, which resulted in the best sentiment classification accuracy of each sentiment classifiers‘ sub-classifier #3. Similarly to previous work, the dialogue acts were not parsed by a sub-system but used ‘as-is‘, to reveal the maximal extent of their applicability for sentiment recognition (see 2.1.3).

Setup 1: The dialogue act of the preceding functional-segment (s_{i-1}), performed by Speaker-A can represent the intention-level stimuli for the sentiment of the i^{th} segment (s_i) performed by Speaker-B (especially an interpersonal act). Subsequently, the dialogue act in Speaker-B's s_i assumed to represent the outcome of a cognitive process influenced by their affective states (see 2.1.1). To account for the causational connection between the consecutive utterances, in Setup 1, each batch processed by sub-classifier #3 consists of the dialogue acts of $s_{i-1}$and s_i labeled with the sentiment of s_i.

Since functional-segment lengths are not consistent among the three different dialogue act taxonomies, intentions are sometimes expressed through several consecutive functional-segments. In such cases s_{i-1} and s_i are performed by the same speaker. The interpersonal act of s_{i-1} and s_i then represents an ongoing mental state (intentional context). This can still serve as a cue for the affective state of the same speaker, expressed in the current segment. To help sub-classifier #3 differentiate between these scenarios, each dialogue act label in Setup 1 indicates the performer as Speaker-A or Speaker-B. The number of possible dialogue acts is therefore doubled for each taxonomy. (In the case of the IA tagset, for example, 'Criticizing' is subdivided into 'Criticizing_A' and 'Criticizing_B'.) Figure 5.7 shows the input labels for sub-classifier #3 used in Setup1. A shift between topics may negate causative or continuation relationships between consecutive utterances. However, because accounting for topic-shifts would further increase the number of training labels, the author elected not to consider them in the context of such a small dataset.

Raw input to Classifier #3 for each segment (Array of segments: $S = (s_1, s_2 \dots s_i)$)

Dialogue act labels of	
s_{i-1}	s_i
$Speaker_A / Speaker_B$	$Speaker_A / Speaker_B$

Figure 5.7: Setup 1: Feeding speaker-specified dialogue act labels of the current and previous functional-segment into sub-classifier #3

Setup 2: Dialogue acts of the SWDB-DAMSL and DIT++ models are presumed to have weaker sensitivity to affective states, and may not serve as a stimulus for them. An experiment was therefore conducted with batches containing only the dialogue act labels and sentiment labels of s_i, where the models above may perform better. Speaker-A and Speaker-B were, however, still differentiated. Figure 5.8 depicts the input labels for Classifier #3 used in Setup 2.

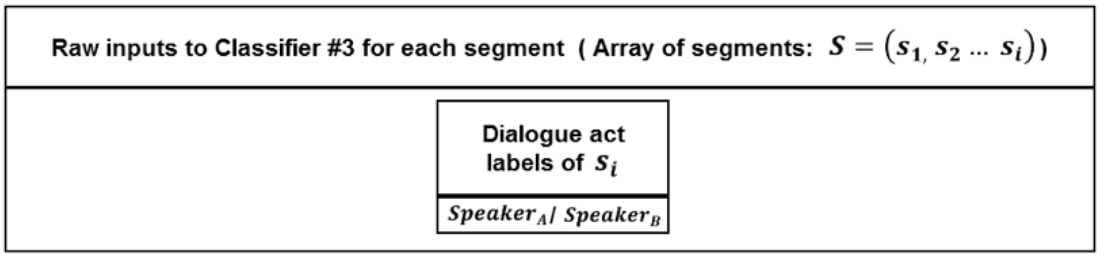

Figure 5.8: Setup 2: Feeding speaker-specified dialogue act labels of the current functional-segments into sub-classifier #3

Setup 3: It is possible, that by increasing the number of dialogue acts through differentiating between a dialogue acts performed by Speaker-A and Speaker-B, too much noise would be added to the data. Setup 3 is indented as a setup with minimized noise, where the dialogue act tags were not differentiated by the speaker, and each batch contained only the dialogue act and sentiment label of s_i Figure 5.9 displays the input labels for sub-classifier #3 used in Setup 3.

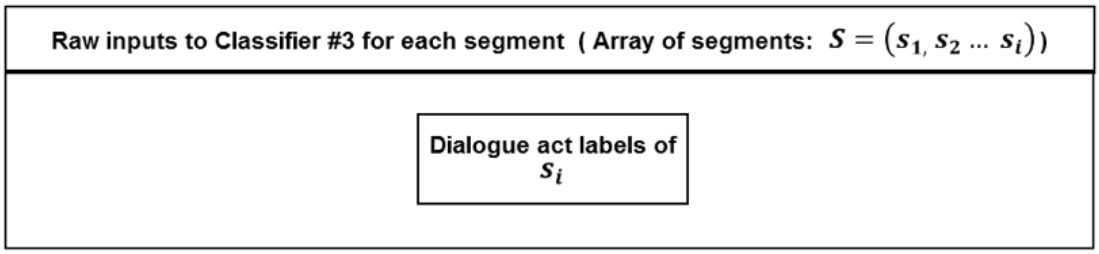

Figure 5.9: Setup 3: Feeding speaker-unspecified dialogue act labels of the current functional-segment into sub-classifier #3

When comparing the final classification results of the four sentiment classifiers, (the baseline classifier and the three augmented classifiers utilizing a given dialogue act model) each augmented classifier uses their own sub-classifier#3 with the setup most optimal to exploit the attributes of the dialogue act categorization it utilizes. Each sentiment classifier used in the experiment was trained and tested on the same sets of functional-segments, through 10-fold cross-validation. Specifically, the data were randomly partitioned into 10 equal sized subsamples, from which a single subsample is retained for testing the model, while the remaining 9 subsamples are used as training data (6240 segments) [67]. (The order of the functional-segments within each subsample and the order of the subsamples themselves were retained to ensure the neural networks can learn from the structure of the conversations.) To reduce variability, the testing was achieved in 10 iterations, each time using

a subsample as testset that have not been used previously, and producing 10 recognition accuracy results. The average of the 10 recognition results indicates a less biased overall recognition accuracy.

5.2.4 Experimental results

Table 5.4 lists the experimental results obtained for machine learning - based sentiment recognition through the three sub-classifiers, separately. In the case of sub-classifier #3, all cases of processing the three dialogue act models are scrutinized separately along the proposed three experimental setups.

Table 5.4: Recognition accuracy of the separate sub-classifiers obtained in the experiments

Method	Precision			Recall			F1-score			Overall
	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	acc.
Sub-classifier #1	0.38	0.45	0.69	0.27	0.63	0.60	0.32	0.42	0.64	56.18%
Sub-classifier #2	0.41	0.34	0.64	0.29	0.52	0.56	0.34	0.41	0.60	51.91%
Sub-classifier #3 processing labels of DIT ++, Setup1	0.14	0.17	0.23	0.07	0.20	0.24	0.09	0.18	0.23	19.60%
Sub-classifier #3 processing labels of DIT ++, Setup2	0.18	0.21	0.25	0.11	0.25	0.29	0.14	0.23	0.27	23.72%
Sub-classifier #3 processing labels of DIT ++, Setup3	0.13	0.16	0.21	0.06	0.20	0.23	0.08	0.18	0.22	18.84%
Sub-classifier #3 processing labels of SWBD-DAMSL, Setup1	0.15	0.21	0.23	0.07	0.24	0.25	0.10	0.22	0.24	20.42%
Sub-classifier #3 processing labels of SWBD-DAMSL, Setup2	0.15	0.18	0.24	0.09	0.24	0.26	0.11	0.21	0.25	21.06%
Sub-classifier #3 processing labels of SWBD-DAMSL, Setup3	0.14	0.18	0.23	0.07	0.25	0.27	0.09	0.21	0.25	20.84%
Sub-classifier #3 processing labels of IA model, Setup1	0.26	0.25	0.37	0.14	0.31	0.34	0.18	0.28	0.36	32.16%
Sub-classifier #3 processing labels of IA model, Setup2	0.24	0.24	0.32	0.12	0.29	0.33	0.16	0.26	0.32	28.89%
Sub-classifier #3 processing labels of IA model, Setup3	0.20	0.21	0.29	0.11	0.26	0.32	0.14	0.25	0.30	26.06%

Table 5.5 shows the overall experimental results obtained for sentiment recognition of the baseline classifier, and the three augmented classifiers utilizing different dialogue act models (in their best-performing setups). It

Table 5.5: Recognition accuracy of the separate baseline and augmented-classifiers obtained in the experiments

Method	Precision			Recall			F1-score			Overall acc.
	NEG	NEU	POS	NEG	NEU	POS	NEG	NEU	POS	
Baseline method (Sub-classifier #1 and #2)	0.42	0.50	0.77	0.30	0.70	0.67	0.35	0.58	0.72	62.33%
Baseline method + Sub-classifier #3 processing labels from DIT ++ (with best performing setup)	0.45	0.58	0.76	0.23	0.70	0.81	0.30	0.63	0.78	66.10%
Baseline method + Sub-classifier #3 processing labels from SWBD-DAMSL (with best performing setup)	0.46	0.57	0.75	0.22	0.69	0.81	0.30	0.62	0.78	65.20%
Baseline method + Sub-classifier #3 processing labels from IA model (with best performing setup)	0.60	0.62	0.82	0.31	0.77	0.90	0.41	0.69	0.83	74.42%

5.3 Computational validation - Semi-supervised learning

5.3.1 Data

Data mining

As detailed above, frame-level audio feature sets from the YouTube 8M dataset were selected through indirect YouTube search for semi-supervised polarity recognition. The search was conducted through the YouTube API [68]. Search-phrase inputs to the API were

a) subjectively selected English synonyms (based on the Collins Thesaurus [69]) of the eight basic emotions defined by Plutchik [2]

   - ‘happy‘for ‘joy'
   - ‘sad‘for ‘sadness'

- ‘angry‘for ‘anger'
- ‘scared‘for ‘fear'
- ‘surprised‘for ‘surprise'
- ‘expectant‘for ‘anticipation‘
- ‘disgusted‘for ‘disgust'
- ‘impressed‘for ‘acceptance'

b) seven English synonyms (based on the Collins Thesaurus) of communicative functions (intentions) corresponding to six basic emotions mentioned above. The communicative functions used were the ones defined in the core IA model (see 3.2.1) tailored for dialogic data in general. Correspondence was decided based on the results of the empirical experiments detailed in 5.1.5.

- ‘criticizing‘for ‘criticizing‘corresponding to ‘disgust'
- ‘being indiscrete‘for ‘indiscrete commenting‘corresponding to ‘anger'
- ‘being there‘for ‘empathizing‘corresponding to ‘acceptance'
- ‘commanding‘for ‘commanding/requesting‘corresponding to ‘anticipation'
- ‘requesting‘for ‘commanding/requesting‘corresponding ‘anticipation'
- ‘talking himself out‘/ ‘talking herself out‘for ‘self-image improving‘corresponding to ‘fear'
- ‘noticing‘for ‘paying attention‘corresponding to ‘surprise'

The communicative functions of ‘indebting partner‘and ‘agreeing‘did not have enough emotion-indicative power to be utilized in this method.

Since the word phrases were in English, the targeted videos were also of English language. Non-English videos with English titles were filtered out from the search results. The reason for choosing the English language is twofold:

- The number of English videos on YouTube are much larger than videos of any other language, providing large sets of data to select from
- Since the larger part of the movies created in general have English speaking versions (original or dubbed version) it is more likely to find the necessary amount of scenes to populate each aggregated emotion bag.

The selected videos (having an extracted audio feature set in the Y8M) were filtered through the API to be between one and five minutes in length, in order to get videos of dialogues focused on the expression of one particular emotion/ communicative function. Based on the basic emotion search phrases 225 minutes of related dialogic videos were found, providing an average of 45 videos per emotions, and an average of 956 utterance-level feature sets retrieved from the Y8M. For each emotion-indicating communicative function, another 206 minutes of dialogic videos were found, resulting in 37 videos per communicative functions and an average of 872 utterance-level audio feature sets. The utterance-level features were compiled from frame-level audio features. Each frame contains a 128-dimensional feature vector, extracted from a deep convolutional neural network, trained on log-mel spectrogram patches as described in [70]. For basic emotions 361 feature sets, for emotion-indicating communicative functions, 337 feature sets were selected in total, leading to the extraction of a total of 7650 and 7429 utterances separately from the Y8M.

Aggregating frame-level instances into utterance-level instances was conducted along time-stamps. Time stamps were provided by the online text converter of Cloud Converter [71], applied on online YouTube video streams of the corresponding videos of the audio feature sets of Y8M .

Data structuring and annotation

Although the training of the proposed method does not require annotation, testing the method for the purpose of the study necessitated the compilation of a test set. 80% of the utterance – level features were used for training the proposed system, while 20% for testing it. In particular, for training 6120 and 5943 instances, for testing 1530 and 1485 instances were extracted from the feature sets selected by basic emotion search-phrases and by communicative function search-phrases, separately. All test instances have been annotated with basic emotion tags by three native English male speakers of age between 27 and 32.

The annotators were asked to determine the underlying emotion of the interlocutors for each utterance while watching the selected YouTube videos online. All utterance-level test instances received one tag. The inter-annotator agreement for emotion tags assessed with Fleiss‘ Kappa [56] was 68.2%. The emotion labels were then transformed into polarity labels based on their valences defined in Russels‘s circumplex of emotions [57]. The reason for not having the utterances annotated with polarity tags from the beginning is that recognition of specific emotions is also analyzed in the study.

5.3.2 Implementation

5.3.2.1 Architecture

Latent variable extraction

The audio data arrays were fed into the encoder layers of the VAE. Since the encoder consists of three consecutive GRU neural network layers ending in a fully-connected layer, the decoder also consisted of three GRU layers, where the first layer is input with the output of the decoder. The latent variables are the output of the encoder‘s fully connected layer. Their number was decided as follows: for each emotion representing bag, eight latent variables were extracted under the assumption that in an ideal case each latent variable represents a different emotion (or emotion-related concept that can occur in any bag). In the case when the latent concepts are not covering the emotions but other abstract constructs, the number of eight latent variables may prove to be too large. However, the latter would mean that only a few of the variables would account for concepts significant enough to differentiate between emotions in the form of audio features. Having fewer latent variables than hidden concepts, on the other hand, can drastically affect the results of the clustering-based classification. Eight variables were hence assumed to be enough to account for all the possible significant variations.

Algorithm 7 Latent variable extraction

while learning **do**
    textbfEncoding
    output of first layer ←GRU(*list of utterance-level audio features*)
    output of second layer ←GRU(*output of first layer*)
    output of third layer ←GRU(*output of second layer*)
    compressed form of input = latent variables #1 ←fully-connected relu layer(*output of third layer*)
    textbfDecoding
    output of first layer ←fully-connected relu layer(*output of encoding*)
    output of second layer ←GRU(*output of first layer*)
    output of third layer ←GRU(*output of second layer*)
    list of regenerated utterance-level audio features ←GRU(*output of third layer*)
    updating the encoding/decoding algorithm(*loss function*(*list of regenerated utterance-level audio features*))
end while
    return Latent variables

Unsupervised Clustering

Once the eight latent variables got extracted, each emotion bag instance receives the corresponding vector representing its position in the latent feature space. The vectors of each instance were clustered by Expectation Maximization (EM) [72] through Gaussian Mixture Models (GMMs) [73]. GMMs assume that the data points are Gaussian distributed; this is a less restrictive assumption than assuming that they are circular by using the mean (like other clustering methods, such as k-means). As eight-dimensional datapoints were clustered, flexibility in the terms of cluster shape was of utmost importance. Each instance was regarded as being generated by a mixture of Gaussians. To find the parameters of the Gaussians that best explain the data, a conventional EM was used, computing a matrix where the rows are the data point and the columns are the Gaussians [72]:

$$W_j^{(i)} = \frac{\phi_j \mathcal{N}(x^{(i)}; \mu_j, \Sigma_j)}{\sum_{q=1}^{k} \phi_q \mathcal{N}(x^{(i)}; \mu_q, \Sigma_q)} \tag{5.5}$$

where ϕ_j is the weight for each Guassian, μ_j is the mean of the Guassians, and Σ_j is the co-variance of each Guassian. In matrix W an element at row i, column j is the probability that $x^{(i)}$ was generated by Gaussian j. The probability of a given Gaussian is computed in the numerator and is normalized along k Guassians in the denominator.

The number of stable clusters was decided based on the Dunn Index cluster validation metric [74] applied through several iterations with varying cluster sizes. The Dunn Index is the ratio of the smallest inter-cluster distance and the largest intra-cluster distance. It is computed as

$$D(\zeta) = \frac{C_k, C_1 \in \zeta, C_k \neq C_1(\min_{i \in C_k, j \in C_1} dist(i,j))}{\max_{C_m \in \zeta} diam(C_m)} \tag{5.6}$$

where $diam(Cm)$ is the maximum distance between observations in cluster (Cm). The Dunn Index has a value between zero and ∞, and should be maximized.

Once stable clusters are found, the cluster containing the largest number of vectors is selected as a representative cluster for the given emotion. Then, polarity bags are populated with the vectors of the representative clusters of each corresponding (negative or positive valence) emotion.

Vector comparison with unseen data

In the case of unseen data, latent variables were extracted by the VAE pre-trained on the training set. Similarly to the training data set, the unseen data must consist of utterance-level audio features. After the test instances were mapped to the latent feature space, they were compared with the emotion representative

Algorithm 8 Unsupervised clustering

```
for all bag ∈emotion bags do
    for all instance ∈bag do
        mapping to latent feature space(utterance-level audio features)
    end for
    for  i ≠number of latent variables do
        Number of stable clusters ←0
        Current strongest indicator ←0
        i++
        Cluster stability indicator ←Dunn Index(EM(GMM(instance, i)))
        if Cluster stability indicator > Current strongest indicator then
            Number of stable clusters ←i
            Current strongest indicator ←Cluster stability indicator
        end if
    end for
    for all instance ∈bag do
        largest cluster ←EM(GMM(instance), Number of stable clusters)
    end for
    final emotion bags ←leave only instances of largest cluster(bag, largest cluster)
end for
polarity bags ←aggregated along valence(final emotion bags)
    return polarity bags
```

instances populating the negative and positive polarity bags. As a similarity measure, cosine-similarity [75] was used, for it is judging the orientation of the vectors. Cosine similarity is computed as

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.7}$$

where A_i and B_i are components of vector A and B respectively. The resulting similarity ranges from [-1:1], where -1 indicates exact opposition, 1 indicates exact matching, and 0 indicates orthogonality or decorrelation; in-between values indicate intermediate similarity or dissimilarity. The test instances are labeled, based on the similarity their vectorized latent variable representations have with the vectors populating the positive and negative polarity bags.

Algorithm 9 Classification

```
list of mapped test instances ←empty
for all utterance-level instance ∈list of test instances do
    mapped test instance ←mapping to latent feature space(utterance-level instance(of audio features))
    list of mapped test instances ←append(mapped test instance)
end for
for all test instance ∈list of mapped test instances do
    decision on polarity ←0
    best cosine similarity ←0
    for all polarity bag ∈polarity bags do
        for all instance ∈polarity bag do
            current cosine similarity ←cosine similarity(instances, test instances)
            if current cosine similarity > best cosine similarity then
                decision on polarity ←polarity bag
                best cosine similarity ←current cosine similarity
            end if
        end for
    end for
     return decision on polarity
end for
```

5.3.2.2 Technical details

Auto-encoders

Every auto encoder consists of an encoder $\phi : \mathcal{X} \rightarrow \mathcal{F}$ and a decoder $\psi : \mathcal{F} \rightarrow \mathcal{X}$ part. The encoder encodes the high dimensional feature set it is input with into lower dimensionality, compressed features called latent variables: $\mathbf{z} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$. $\mathbf{z}$ is a latent variable, σ is an element-wise activation function such as a sigmoid function or a rectified linear unit, $\mathbf{W}$, is a weight matrix and $\mathbf{b}$ is a bias vector. Then, from the hidden layer, the decoder reconstructs the input into the same high level dimensionality: $\mathbf{x}' = \sigma'(\mathbf{W}'\mathbf{z} + \mathbf{b}')$. Decoding and encoding are done by neural networks, learning from the loss of the reconstructed vs. original datapoints through back-propagation and trying to minimize reconstruction errors. [76]

$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2 = \|\mathbf{x} - \sigma'(\mathbf{W}'(\sigma(\mathbf{W}\mathbf{x} + \mathbf{b})) + \mathbf{b}')\|^2 \tag{5.8}$$

The possible neural networks of encoders and decoders vary in type (feedforward, RNN, CNN etc.) and depth according to the nature and complexity of the input data. Latent variables $\mathbf{z}$ are the outputs of the hidden layer in the middle, initialized and updated through the encoding-decoding process of the autoencoder. Latent variables represent a few basic concepts the features of the input data can be grouped by. The hidden layer is always the last fully connected layer of the encoder ϕ. Accordingly, the (first) layer of the decoder ψ processes a two-dimensional input generated by the fully connected layer. The number of latent variables is often equivalent to the number of nodes in the fully hidden layer. [76]

Variational autoencoders

Variational autoencoders (VAE) make strong assumptions concerning the distribution of latent variables $\mathbf{z} \in \mathbf{Z}$. In VAEs, constraints are added that forces the generation of latent vectors to roughly follow a unit Gaussian distribution. As the Gaussian distribution, conventionally a centered isotropic multivariate Gaussian $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is selected. The isotropic Gaussian priors allow each latent dimension in the representation to push themselves as farther as possible from the other factors. Thus, VAEs are known to give representations with disentangled factors. [54] The above constrain requires an additional loss component which measures how closely the latent variables match a unit gaussian and a specific training algorithm called Stochastic Gradient Variational Bayes (SGVB) [77].

VAEs build on the assumption that the data is generated by a directed graphical model $p(\mathbf{x}|\mathbf{z})$ and that the encoder is learning an approximation $q_\phi(\mathbf{z}|\mathbf{x})$

to the posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ Thus, the loss function of a VAE has the following form:

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big(\log p_\theta(\mathbf{x}|\mathbf{z})\Big) \tag{5.9}$$

Here, D_{KL} stands for the Kullback – Leibler divergence [78].

5.3.2.3 Computing environment

The proposed approach of latent variable extraction and clustering were coded in *Python*, and the code is compatible with Python version 3.6. The code was run on Spyder IDE 3.2.6 on Ubuntu 16.04.2 LTS. The workstation used is equipped with an Intel Xeon E5-1650 v4 3.60GHz CPU and 128GB DDR4 RAM. In the current environment, the trained computational method performs the prediction of one datapoint between 0.18ms to 0.23ms (depending on the experimental setup).

The most important Python modules utilized in the implementation are listed below:

- **theano**: a library and optimizing compiler for manipulating and evaluating mathematical expressions, especially matrix-valued ones. Used for the implementation of variational autoencoder layers.
- **keras**: high-level neural networks API, written running on top of TensorFlow, CNTK, or Theano. In the experiments conducted, it was used with Thenao backend.
- **pandas**: Pandas is a library for data manipulation and analysis. Used to load the textual transcriptions.
- **numpy**: NumPy is a library for scientific computing, especially for matrix transformations. Used to create, reshape save and load matrices of audio inputs.
- **sklearn**: Scikit-learn is a machine learning library, used for the implementation of Gaussian Mixture Models with expectation maximization.

5.3.3 Experimental setups

To test the efficiency of the latent variable extraction-based polarity classification, three experimental setups were designed with identical architectures but different test and training sets. The 8:2 train:test ratio was set in all cases. In Setup 1, the train and test sets used were the audio feature sets selected by basic emotion search-phrases (6120:1530 train:test instances).

In Setup 2, the sets used were the audio feature sets selected by emotion-indicating communicative function search-phrases (5943:1485 train:test instances) Setup 2 accounted for only six basic emotions and accordingly only six emotion bags were populated with representative instances. Thus, each polarity bag was populated bags with the instances of only three emotion bags.

In setup 3, the test and train sets contained audio feature sets gathered from both search phrase dimensions (12063: 3015 train:test instances) to acquire a larger set of training data: each emotion bag was populated with the instances corresponding to the same emotion of basic emotion - and communicative function-based search phrases. (In the case of sadness and joy the bags were populated only with the instances gathered through basic emotion search phrases.) In all setups, the test labels were polarity labels transformed from basic emotion labels.

5.3.4 Experimental results

Table 5.6 summarizes the polarity classification results for each experimental setups separately.

Table 5.6: Polarity classification results

Setup	Precision		Recall		F1-score		Overall
	NEG	POS	NEG	POS	NEG	POS	acc.
Setup 1: Basic emotion search phrases	0.52	0.85	0.63	0.76	0.58	0.81	71.22%
Setup 2: Comm. function search phrases	0.42	0.67	0.50	0.55	0.48	0.60	58.17%
Setup 3: Mixed search phrases	0.68	0.88	0.74	0.82	0.73	0.87	79.08%

Table 5.7 further elaborates on the performance of the proposed method through emotion-level classification.

Table 5.7: Basic emotion classification results

Emotions	F1-score		
	Setup 1	Setup 2	Setup 3
Anger	0.38	0.24	0.40
Fear	0.30	0.17	0.25
Joy	0.44	–	0.42
Sadness	0.28	–	0.34
Anticipation	0.12	0.11	0.28
Surprise	0.08	0.28	0.12
Acceptance	0.22	0.22	0.20
Disgust	0.36	0.42	0.37
Overall acc.	29.40%	19.81%	31.63%

Chapter 6

Discussion

6.1 Discussion on the results of the empirical analysis

6.1.1 Occurrence of emotions and dialogue acts

The most frequently observed emotions are ‘acceptance‘and ‘anticipation.‘On the other hand, ‘sadness‘and ‘anger,‘, and other emotions of negative valence were observed relatively rarely. This implies a generally excited and friendly dialogic atmosphere, anticipated and required in a co-operative gaming context.

Considering that ‘partner-unrelated‘acts are the most frequent within the three categorizations compared, it became clear that, by the very nature of in-game conversations, most of the utterances were related to in-game events rather than to the conversational partners. The co-operative nature of the dialogue further manifests in the fact that the IA model‘s acts of ‘indiscrete commenting‘, ‘criticizing‘, and ‘self-image improving‘were observed very rarely. This result also assumes that the interlocutors‘ focus on the game rather than on each other. It also relates to the cooperation demanding ”friendly atmosphere” that would easily be ruined with the excessive use of those three acts. The cooperation demand can also be the reason for frequent occurrence of the IA acts of ‘empathizing‘, ‘paying attention‘, and ‘commanding/requesting‘throughout the conversations.

6.1.2 Association pairs based indicative power

As it can be seen from the table 5.1 and figures 5.4 to 5.6, the IA model not only has stronger overall correlation with basic emotions, it also has a better ratio of good, strong, and exclusive indicators than do the other models.

In the case of SWBD-DAMSL, three of the medium strength association pairs are connected to the same emotion of ‘acceptance‘(see Figure 5.4). Obviously, however, a given dialogue act having medium-strength, or better associations with multiple emotions cannot be considered as a good indicator for an emotion and vice-versa. Thus, the three dialogue acts mentioned above are not as good indicators for ‘acceptance‘as for example the dialogue act of ‘apology‘for the emotion ‘fear‘, where ‘apology‘has medium-strength association only with ‘fear‘and ‘fear‘has with ‘apology‘(and no stronger association is present). In the case of the

DIT++ and IA models multiple associations with the same emotion or dialogue act were not detected (Figure 5.5 and Figure 5.6).

The above results imply that the IA model would provide for a more consistent classification in an affective context, compared to the other models. By invoking Assumption 3 (see 3.1), these results would also mean that the IA model tags better correspond to the social status and self-esteem - managing interpersonal actions than the other models. In other words, the proposed classification appears to be appropriate to reflect the interpersonal relations managing communicative functions. The dialogue act-indicative power of emotions was not assessed but is left for future study.

6.2 Discussion on the results of supervised computational approach

6.2.1 Recognition accuracy of the separate sub-classifiers

The overall low recall in the negative sentiment indicates bias in the distribution of emotion types within the dataset. As mentioned in 6.1.1, this may be due to the fact that online gameplay requires cooperation that would easily be ruined if negative emotions were to be expressed excessively. Audio vectors seem to be slightly better indicators of negative sentiment while word embeddings are of neutral and positive sentiments.

As shown in Table 5.4, sub-classifier #1 processing the textual data produced 56.18% recognition accuracy, while sub-classifier #2, processing the audio data achieved 51.91%, separately. This implicates that the word vectors trained on GloVe (see 5.2.2.3), served as a more consistent cue for sentiment recognition than the low-level audio feature vectors. Presumably, the convolutional neural network of sub-classifier #2 was not able to generalize well enough on such a small dataset. As expected, sub-classifier #3, processing only the one-dimensional textual data of dialogue acts, performed significantly weaker.

The setups containing the best overall recognition accuracy for each dialogue act model (processed by sub-classifier #3) are highlighted in bold type. With the best performing setups for processing the given dialogue act model sub-classifier #3 achieved 21.06%, 23.72%, and 32.16% recognition accuracy if trained on the SWBD-DAMSL, DIT++ and IA tags separately. The use of the IA tagset (in the best-performing setup) yielded 11.10% and 8.44% better recognition accuracy compared to the best performances of the SWBD-DAMSL and DIT++ tagsets, respectively.

Furthermore, in the case of the IA model, the best performance was

achieved through Setup 1 (considering dialogue acts of preceding utterances and differentiating between speakers). In the case of the SWBD and DIT++, however, best performances were achieved with Setup 2 (differentiating between speakers but not considering preceding dialogue acts), which implies that dialogue acts that are unrelated to affective states (and cannot serve as a stimuli for them) are less adequate for harvesting contextual information during sentiment/emotion classification tasks.

In general, sub-classifier #3 shows similar performance when trained on the SWBD-DAMSL and DIT++ tags in terms of precision, recall and accuracy of the given sentiments. However, training the sub-classifier on the IA tagset resulted in a noticeably higher precision in negative sentiment. This fact indicates that the IA tagset, accounting for 'Face-threatening'interpersonal verbal-actions, has more acts consistently co-occurring with negative sentiments, thus, is more adequate to serve as a cue for them than the other two dialogue act models. The above results are strong evidence in favor of the definition and use of emotion-sensitive dialogue acts specifically for augmenting emotion/sentiment recognition systems.

6.2.2 Recognition accuracy of the baseline and augmented-classifiers

As shown it has been shown in Table 5.5, the merged output through the soft-voting (see 5.2.2.3) process could predict the correct sentiments with a 62.33% accuracy. This moderate accuracy reflects well the complexity of the task to recognize affective states when working with a small dataset.

Although, as sole features, dialogue act tags appear to be poor indicators for sentiment classification, as a complementary feature set (through decision-level merging) they improved the baseline model's recognition accuracy. The usage of SWBD-DAMSL improved overall recognition accuracy by a maximum of 2.87%, of DIT++ by 3.77%, and of IA by 12.09%. Only the setup in which sub-classifier #3 performs the best (related to the given dialogue act model) was selected for augmenting the baseline method.

In previous studies, the addition of dialogue act labels resulted in an improvement of 4% at most [13] using only two affective-types and larger data sets. Thus, in the context of such a small dataset, these result is considered to be meaningful, indicating the usefulness of cognitive context (in the form of dialogue acts) for sentiment/emotion recognition.

A single factor Anova test, computed from the validation scores of each classifier's 10-fold cross-validation process, shows that the improvement yielded by the usage of the IA model is indeed significant. Table 6.1 shows the results of the Anova tests computed through the results of each classifier's best-performing

recognition setup, in comparison to the results of the best performing setup of the classifier that applies the IA tagset.

Table 6.1: Significance of improvement yielded by the application of IA acts

Classifier	p-value	F-score	F-critical
Baseline-method vs. Augmented method processing IA - best-performing setup (bfs)	<0.005	151.80	4.41
Augmented method processing IA -bfs. vs. Augmented method processing DIT++ - bfs	<0.005	21.77	4.41
Augmented method processing IA - bfs. vs. Augmented method processing SWBD-DAMSL - bfs.	<0.005	43.21	4.41

To further scrutinize the applicability of the proposed tagset, an IA act processing sub-classifier #3 (with the best performing setup1) was used to augment the text processing sub-classifier #1 and audio-processing sub-classifier #2 separately, improving their recognition accuracy by 5.14% and 7.01 % (see Table 6.2). Since the merging of the sub-classifiers was done by soft-voting, weighting and averaging mid-classification results these results are not surprising. Both sub-classifier #1 and #2 performs better than #3, getting stronger weights in the soft-voting process. sub-classifier #1 however, processing the (at least in this dataset) more reliable word-embedding vectors, performs better than #2, thus getting even stronger weights during the soft-voting process, not letting the interpersonal acts to heavily influence the final classification result. Sub-classifier #2, on the other hand, is a slightly weaker classifier, letting sub-classifier #3‘s results dominate more. Thus the usage of interpersonal acts improvers sub-classifier #2 even more than it does sub-classifier #1, but results in a weaker overall classification accuracy. In the case of merging the outputs of all three sub-classifiers, the weighting is more balanced, advancing the final classification result further.

Table 6.2: Improving the separate sub-classifiers by the application of IA acts

Method	Improvement	Overall acc.
Sub-classifier #1 augmented with Sub-classifier #3 processing IA - best performing setup (bfs)	5.14%	61.32%
Sub-classifier #2 augmented with Sub-classifier #3 processing IA - best performing setup (bfs)	8.01%	59.92%

In the case of larger datasets with more labeled utterances, it can be expected that the difference in performance of the various dialogue act models (for augmenting sentiment/emotion classification) would diminish. However, annotating large datasets with labels of emotion-related constructs is a highly labor-intensive task. Also, in the gaming domain, no such large datasets for the Japanese language currently exist. Automatic classification of the proposed IA acts and the amount of training data required for their satisfactory-level recognition is to be tested and analyzed in future studies.

6.3 Discussion on the results of semi-supervised computational approach

6.3.1 Recognition accuracy of polarity classification

In the light of the results of related studies [36] and [37], the classification results (at best, 79.08% for two categories) indicated in Table 5.6 may appear moderate. Nevertheless, the proposed classifier was trained on only 12063 datapoints at maximum, in contrast to the above studies, which were trained on hundreds of thousands of instances. The Y8M dataset provided only a maximum of 15078 relevant utterances produced according to the developed method) from which 3015 (20% of the total utterances) was used as labeled test instances. According to the proposed method, however, the training set could be enlarged manifold, limited only by the number of the datasource in use (and the number of instances that can be labeled for testing).

The fact that classification results of Setup 1 (overall accuracy: 71.22%) is noticeably higher than of Setup 2 (overall accuracy: 58.17%) indicates that basic emotion search phrases allow for the selection of more relevant videos (and their audio feature sets) for the proposed multiple instance learning method than search phrases of emotion-indicating communicative functions. The results of Setup 3 (overall accuracy:79.08%), however, show that simultaneous use of

both search phrase-based audio sets results in improved recognition accuracy. Having the classification result improved by the additional usage of audio feature sets that matches communicative function-based search results, presumes that the proposed communicative functions introduce more information about basic emotions than noise. In particular, these search phrases cover additional videos focused around the expression of a certain basic emotion (through the corresponding communicative function) which could not be found through basic emotion search phrases owing to the limited size of the Y8M dataset. In other words, communicative functions serve as a quasi-alternative/complementary feature set for the proposed semi-supervised method.

6.3.2 Recognition accuracy of basic emotion classification

Table 5.7 further scrutinizes the classification results in the more fine-grained case of basic emotion recognition. From the F1 score values, it can be seen that negative polarity has been more accurately predicted in all three setups. The lower recognition accuracy on the positive polarity stipulates that the audio features of the Y8M contain cues that show stronger association with negative emotions.

Setup 1, utilizing basic emotions as search phrases, allows for stable recognition of 'joy'(0.44) and 'anger'(0.38), while Setup 2, utilizing communicative functions as search phrases, results in the relatively reliable recognition of 'disgust'(0.42) and 'anger'(0.24). While the overall recognition accuracy of Setup 2 (19.81%) is lower than in Setup 1 (29.40%), the mixed usage of communicative functions - and basic emotions - based feature sets improve recognition results as indicated by the results of Setup 3 (31.63%). Enlarging the training set through the usage of both search phrase dimensions - expanding the variance of the latent space, and the hidden parameters of the unsupervised clustering - allows for better generalization, and yields better recognition results.

As videos - directly/indirectly selected to concentrate on the expression of basic emotion - tend to contain several utterances representing various emotions, the classifier has to deal with a large amount of noise. In the proposed method, the audio feature inputs were mapped to the latent feature space then clustered with an unsupervised Gaussian Mixture Model through Expectation Maximalization to select the corresponding instances for each emotion bag. Experiments involving other unsupervised clustering methods and/or different hyperparameters for the latent variable extracting Variational Autoencoder may deal better with noise and yield better classification results, especially in the case of the more fine-grained emotion recognition task.

6.4 Comparison of the proposed computational methods

The proposed IA tagset has been applied through two computational methods, a supervised learning based sentiment classifier and a semi-supervised polarity classifier. Both classifiers are utilizing the proposed dialogue act model as intentional context representing pragmatic-level linguistic units. In the supervised approach, the proposed interpersonal acts are used as a complementary feature set beside the audio and textual features. In the semi-supervised approach, the acts serve as alternative or complementary search phases beside basic emotion tags, used in the selection of audio features for training. Thus, even in the latter approach, the interpersonal acts and the basic emotion tags can also be thought of as indirect complementary feature sets.

The supervised method uses the emotion sensitive dialogue acts to improve sentiment recognition in a way to allow satisfactory-level recognition accuracy even on smaller sets of labeled data. As the experimental results suggest, this can be successfully achieved through the proposed tagset of interpersonal acts. The approach, however, is supervised, thus a certain amount of labeled data will always be required. Furthermore, to ensure satisfactory level improvement, the training data need to be hand-labeled or automatically classified (and thus pre-trained on an additional dataset) with dialogue acts. In the experiments conducted, only hand-labeled acts were used (following the line of similar dialogue act utilizing work [13], [14], [15]; [16]), to discover the maximum potential of the proposed tagset. This is a deficiency of the proposed method, allowing it to be used only on small-sized data, where the labour required for the hand labeling/classification of the training data with dialogue acts may prove to be a good tradeoff for the improvement it yields in recognition accuracy.

The semi-supervised approach uses basic emotion and/or interpersonal act tags as YouTube search phrases to populate emotion bags of a multiple instance learning approach. Thus the results of the search are used as indirect, ”weak labels” making hand-made notation completely unnecessary. This method yielded promising results on larger data, especially when trained based on both basic emotion and dialogue act search phrases. It also has the ability to enlarge the training set manyfold, restricted only by the size of the dataset in use (and not the by the cost the labeling would require). The approach, however, does not generalize well on a smaller dataset and assumed to be outperformed by the former supervised approach trained on similar size of data. Table 6.3 summarizes the recognition accuracy results of the separate approaches detailed along the size of training data used.

Table 6.3: Recognition accuracy of the developed supervised and semi-supervised approaches in relation to training data size

Approach	Training segments	Overall acc.
Supervised sentiment classification	6240	74.42%
Semi-supervised polarity classification emotion search phrases	6120	71.22%
Semi-supervised polarity classification comm. func. search phrases	5943	58.17%
Semi-supervised polarity classification mixed search phrases	12063	79.08%

Thus for smaller datasets, the supervised approach is more fitting, while in the case of larger datasets the semi-supervised approach seems to be a highly applicable solution.

Chapter 7

Conclusions

In this concluding chapter contributions of the thesis are summarized with their possible impact and important directions of future work are described.

7.1 Contributions

The central problem addressed in this thesis is the training cost of affective classifiers in terms of the requirement of hand-made annotation for the preparation of training sets. Recently there is a growing demand for real-time affect awareness - the ability to infer and react to the affective states of the user - in commercial softwares, such as in dialogue systems or games. As real-time recognition is often achieved through the pre-training of supervised classifiers, the annotation of large sets of training data with the target labels is necessary. In the case of the easily extracted and often used audio and textual training sets, even thousands of labeled datapoints would not yield satisfactory-level recognition results. For improved recognition results the following two approaches are used in general for audio and/or textual feature-based methods:

- Utilizing the intentional context in the form of dialogue act labels and use them as complementary features to predict the output labels
- Using large sets of review data, labeled with sentiments/polarity tags and use them as output labels of supervised neural-networks or multiple instance learning algorithms

These methods, however, suffer from the deficiency, that the improvement they would yield does not worth the labeling-cost their training requires. As a possible remedy to this problem, the author proposes the usage of emotion sensitive intentions representing dialogue acts, associable with certain emotions/sentiments as

- as a complementary feature set to help the supervised machine learning methods to learn from label features firmly associated with the output labels
- as a basis to define search phrases for the selection of videos concentrating on emotional dialogues. Such videos could serve as bags containing instances

of utterances, while the search phrases they correspond to as the bag labels. The labels then can be utilized for the training of multiple instance learning algorithms in a semi-supervised way.

As emotion-sensitive dialogue acts, interpersonal relations directing communicative functions were proposed, constituting the one-dimensional dialogue model of interpersonal acts. The acts were defined based on appraisal theories [21] and the Politeness theory of Brown and Levinson [38]. The applicability of the model was verified in

- empirical experiments where the model proved to have significantly higher sensitivity to emotions in comparison to two well-known dialogue act models of SWBD-DMSL and DIT++
- computational experiments through supervised sentiment classification where the utilization of the IA tagset as a complementary feature set improved the classification results significantly higher than the utilization of the SWBD-DMSL or DIT++ tagsets
- computational experiments through multiple instance-based semi-supervised polarity classification where the utilization of the IA tagset as a complementary set for the labeling of emotion bags improved classification results

Among the computational methods developed, the supervised approach fits smaller datasets (easy to annotate with emotion and IA tags) while the semi-supervised approach shows promising results to be applicable on large and unlabeled datasets.

The developed IA model thus has the potential to significantly improve the recognition of affective states. The developed empirical and computational methods verify its applicability on small and large datasets as well.

7.2 Future work

7.2.1 Extending the dialogue act model

The core model (see 3.2.1) developed can be extended/modified in accordance with the language and/or context it is intended to be used on. In this thesis, it has been extended to fit the conversational data used in the supervised experiments (see 3.2.2).

Further extensions are also possible, for more detailed modeling of the dialogic environment: differentiating between illocutionary and perlocutionary - acts as

response illocutionary acts - (see 2.1.2), for example, would further represent the causational relation between interpersonal relations directing intentions and affective states. Over-specification of the model, however, would result in dialogue acts having low association rates with each basic emotion. Reformulation of the proposed acts, while keeping their number may also bring better results, which need to be tested through further empirical experiments.

As the association between emotions and interpersonal relations directing communicative functions are two-sided, the interpersonal act indicating power of emotions can also be analyzed. interpersonal acts could be utilized by dialogue systems or commercial games for the natural language understanding of user input and/or for natural language response generation.

7.2.2 Extending the supervised method

As mentioned in 5.2.3, differentiating between topic-shifts could further improve the efficiency of the proposed supervised classifier, which is able to learn from the sequentiality of the data through its GRU layers. Different topics/sub-topics within the dialogues could be fed into variational-size batches to help the system learn.

Experimenting with different architectures in terms of the numbers and hyperparameters of GRU and CNN layers of the sub-classifiers may further improve recognition accuracy. Applicability of the model may also be tested on other languages and gaming/conversational corpora, with a complementary feature set of a modified IA model, tailored to fit the target context.

The experiments conducted, utilized dialogic data with pre-annotated dialogue act labels to fully evaluate the applicability of the additional feature sets for sentiment recognition. To measure the applicability of the IA model more thoroughly, however, the amount of training data needed for its satisfactory-level automatic classification needs to be assessed. In particular, satisfactory-level recognition in this scenario would point to a minimum level of classification accuracy that ensures that the automatically annotated interpersonal act labels would be able to improve sentiment recognition as a complementary feature set. Assessing the tradeoff between annotation-cost and the yielded improvement is a complex task. Computational experiments are required to measure the learning rate while utilizing a complementary feature set of interpersonal act labels, opposed to using only output (sentiment) labels on larger datasets.

7.2.3 Extending the semi-supervised method

Further experiments are needed to validate the proposed approach on larger datasets, containing more relevant videos applicable for the definition and

population of emotion/polarity bags. The proposed method believed to yield better results if trained on larger sets of data (big data).

As videos - directly/indirectly selected to concentrate on the expression of basic emotion - tend to contain several utterances representing various emotions, the classifier has to deal with a large amount of noise. Experiments involving other unsupervised clustering methods and/or hyperparameters for the latent variable extracting Variational Autoencoder may yield better classification results.

The proposed method can be further extended with the utilization of other emotion-indicating dimensions instead/beside interpersonal act- and basic emotion tag-based search phrases.

Bibliography

[1] R. Plutchik, "Evolutionary bases of empathy," *Empathy and its development*, vol. 1, pp. 38–46, 1987.

[2] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[4] R. B. Miller, "Response time in man-computer conversational transactions," in *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, pp. 267–277, ACM, 1968.

[5] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges 6*, pp. 227–236, Springer, 2015.

[6] H. Yoon, S.-W. Park, Y.-K. Lee, and J.-H. Jang, "Emotion recognition of serious game players using a simple brain computer interface," in *ICT Convergence (ICTC), 2013 International Conference on*, pp. 783–786, IEEE, 2013.

[7] M. Obaid, C. Han, and M. Billinghurst, "Feed the fish: an affect-aware game," in *Proceedings of the 5th Australasian Conference on Interactive Entertainment*, p. 6, ACM, 2008.

[8] D. Duncan, G. Shine, and C. English, "Facial emotion recognition in real time," 2016.

[9] M. Mateas and A. Stern, "Structuring content in the façade interactive drama architecture.," in *AIIDE*, pp. 93–98, 2005.

[10] T. Vogt, E. André, and N. Bee, "Emovoice—a framework for online recognition of emotions from voice," in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, pp. 188–199, Springer, 2008.

[11] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on*, pp. 1–5, IEEE, 2015.

[12] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 698–704, IEEE, 2015.

[13] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Seventh International Conference on Spoken Language Processing*, 2002.

[14] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[15] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech communication*, vol. 40, no. 1-2, pp. 117–143, 2003.

[16] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[17] N. H. Frijda, "Moods, emotion episodes, and emotions.," 1993.

[18] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, University of Nebraska Press, 1971.

[19] H. Bunt, "Multifunctionality in dialogue," *Computer Speech & Language*, vol. 25, no. 2, pp. 222–245, 2011.

[20] R. S. Lazarus and R. S. Lazarus, *Emotion and adaptation*. Oxford University Press on Demand, 1991.

[21] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[22] N. H. Frijda, "Emotion, cognitive structure, and action tendency," *Cognition and emotion*, vol. 1, no. 2, pp. 115–143, 1987.

[23] J. L. Austin, *How to do things with words*, vol. 88. Oxford university press, 1975.

[24] A. Popescu-Belis, "Dimensionality of dialogue act tagsets," *Language Resources and Evaluation*, vol. 42, no. 1, pp. 99–107, 2008.

[25] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011.

[26] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[27] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, pp. 1188–1196, 2014.

[28] J. Huang, S. Rogers, and E. Joo, "Improving restaurants by extracting subtopics from yelp reviews," *iConference 2014 (Social Media Expo)*, 2014.

[29] M. Fan and M. Khademi, "Predicting a business star in yelp from its reviews text alone," *arXiv preprint arXiv:1401.0864*, 2014.

[30] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78, 2014.

[31] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," *Expert Systems with applications*, vol. 40, no. 16, pp. 6266–6282, 2013.

[32] B. Babenko, "Multiple instance learning: algorithms and applications," *View Article PubMed/NCBI Google Scholar*, pp. 1–19, 2008.

[33] S. Scott, J. Zhang, and J. Brown, "On generalized multiple-instance learning," *International Journal of Computational Intelligence and Applications*, vol. 5, no. 01, pp. 21–35, 2005.

[34] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[35] N. Weidmann, *Two-level classification for generalized multi-instance data*. PhD thesis, Master' s thesis, Albert-Ludwigs-Universität Freiburg, Germany, 2003.

[36] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597–606, ACM, 2015.

[37] S. Angelidis and M. Lapata, "Multiple instance learning networks for fine-grained sentiment analysis," *Transactions of the Association of Computational Linguistics*, vol. 6, pp. 17–31, 2018.

[38] P. Levinson, P. Brown, S. C. Levinson, and S. C. Levinson, *Politeness: Some universals in language usage*, vol. 4. Cambridge university press, 1987.

[39] E. Goffman, "On face-work: An analysis of ritual elements in social interaction," *Psychiatry*, vol. 18, no. 3, pp. 213–231, 1955.

[40] Y. Matsumoto, "Reexamination of the universality of face: Politeness phenomena in japanese," *Journal of pragmatics*, vol. 12, no. 4, pp. 403–426, 1988.

[41] N. Hernández-Flores, "Politeness as face enhancement an analysis of spanish conversations," *Current trends in the pragmatics of Spanish*, vol. 123, p. 265, 2004.

[42] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. R. Traum, "Iso 24617-2: A semantically-based standard for dialogue annotation.," in *LREC*, pp. 430–437, Citeseer, 2012.

[43] D. Jurafsky, "Switchboard swbd-damsl shallow-discourse-function annotation coders manual," *Institute of Cognitive Science Technical Report*, 1997.

[44] H. Bunt, "The dit++ taxonomy for functional dialogue markup," in *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 13–24, 2009.

[45] G. J. Upton, "Fisher's exact test," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 395–402, 1992.

[46] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.

[47] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[48] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring convolutional neural network structures and optimization techniques for speech recognition.," in *Interspeech*, pp. 1173–5, 2013.

[49] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.

[50] S. Planet and I. Iriondo, "Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition," in *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on*, pp. 1–6, IEEE, 2012.

[51] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[52] C. Commons, "Attribution 4.0 international (cc by 4.0)," *Retrieved from*, 2018.

[53] "Youtube terms of sevice." `https://www.youtube.com/static?template=terms`. Accessed: 2018-10-31.

[54] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[55] Y. Arimoto and H. Mori, "Emotion category mapping to emotional space by cross-corpus emotion labeling," *Proc. Interspeech 2017*, pp. 3276–3280, 2017.

[56] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.

[57] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[58] T. Kudo, "Mecab: Yet another part-of-speech and morphological analyzer," *http://mecab. sourceforge. jp*, 2006.

[59] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[60] "Wikimedia database dump of the japanese wikipedia on july 20, 2016." `https://archive.org/details/jawiki-20160720`. Accessed: 2017-09-04.

[61] F. Eyben, M. Wöllmer, and B. Schuller, "Openear—introducing the munich open-source emotion and affect recognition toolkit," in *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*, pp. 1–6, IEEE, 2009.

[62] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[63] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016.

[64] F. J. Pineda, "Generalization of back-propagation to recurrent neural networks," *Physical review letters*, vol. 59, no. 19, p. 2229, 1987.

[65] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[66] Y. LeCun, Y. Bengio, *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[67] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[68] "Youtube api." `https://developers.google.com/youtube/v3/`. Accessed: 2018-10-31.

[69] C. Thesaurus, "Collins thesaurus of the english language complete and unabridged," 2002.

[70] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 131–135, IEEE, 2017.

[71] "Cloudconverter." `https://www.360converter.com`. Accessed: 2018-10-31.

[72] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[73] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

[74] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.

[75] M. Steinbach, G. Karypis, V. Kumar, *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, pp. 525–526, Boston, 2000.

[76] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49, 2012.

[77] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[78] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.