Emotion Recognition in Human Dialogs

Meryem M'hamdi, Dr. Pearl Pu, Dr. Valentina Sinstova,
School of Computer And Communication Sciences
Swiss Federal Institute of Technology (EPFL)
Lausane, Switzerland
{meryem.mhamdi, pearl.pu, valentina.sintsova} @epfl.ch

***Abstract*—In this project, we explored fine grained emotion recognition applied to the context of dialogue systems. We based our study on a dataset that consists of two Webly scraped TV Series and Cornell Movie Dialog available online. We aimed for a more fine-grained categorical representation of emotions using Geneva Emotion Wheel (GEW) model of 20 emotions. For that purpose, we followed a hybrid approach that combines rule-based and semantic similarity analysis implemented combining count-based and predictive models with lexicon-based to annotate our dataset with emotion categories. We also run a series of machine learning experiments using annotated dataset from another domain and used domain adaptation to strengthen its ability to generalize on our unlabelled dataset. At the end, we used a voting model between hybrid approach variations and the best semi-supervised approach model that re-assigns labels based on the majority class and increases the recall of certain emotional instances. To evaluate our models, we designed a crowd-sourcing experiment to validate the accuracy of the emotion detection algorithm used. Hybrid lexicon rule-based approach extended with PMI semantic similarity was shown to slightly improve the performance of our initial lexicon-based baseline with an average micro f1-score of 34.0% and area under curve score of 60.2% which accounts for an increase of 0.8% and 1.5% in f1-score and area under curve score respectively and an increase in precision and the recall of the different emotions with better average f1-score of 31.2% for positive categories. It also outperformed Bernoulli Naive Bayes, the best semi-supervised approach even with domain adaptation except for some emotions where the latter increased the recall. By analyzing the relationships between each emotion in the response and emotions that caused it in the context and investigating representative patterns, we reflect on a model of emotional responses that shows the interactions between the emotions of the different actors in dialogue turns and that can be directly embedded, in future work, in a response generation system as a filter that given a context decides on emotionally adequate responses.**

***Keywords*—Emotion Recognition, Hybrid Approach, Natural Language Processing, Semi-Supervised Domain Adaptation, dialogue systems, PMI, word2vec, Compassionate Response Generation**

I. Introduction

Human dialogues are the main form of human-human interaction as they not only involve the exchange of information but also they are the basis of the strength of relationship ties thanks to their rich emotional content. Through dialogue, we not only communicate ideas but most importantly emotions that determine the effectiveness of the interaction and define its success or failure. Despite the importance of emotions in human dialogues, human-computer interaction systems still lack the basic ingredients for processing emotional signals existing in dialogue and demonstrating a satisfying degree of emotional intelligence to the human party. "Contemporary human machine dialog systems always speak with the same unmoved voice and ignore customer's irony and anger or elation". [1]

In the last few years, text-based emotion recognition has gained a considerable amount of attention as it is not only an integral part of human-human communication but it is also a powerful tool in human-computer interaction. Building agents that are able to better understand people emotions and act adequately in emotional situations can help boost an agent capabilities to perform human-related tasks and improve user acceptance. This has been extensively researched but the majority of work focused on either response generation or emotion recognition. Throughout last years, many methodologies have been explored and proposed for sentiment analysis and emotion detection that boost the automatic ability of eliciting and understanding emotions. In the next section, we perform an holistic analysis of those methodologies which are divided into families: keyword-based, learning-based, knowledge-based, rule-based and hybrid-based. Recent work on response generation systems focus more on either generalizing language modelling methodologies to work on open domain tasks employing longer term memory techniques to enhance context awareness. Relatively less effort was done to incorporate emotional awareness in the way the agent adapts its answers to the current emotional state and how the agent behaves depending on the emotion expressed by for example re-conciliating in case of negative emotions, congratulating in case of happiness and calming in case of worrying. In addition to emotion recognition, many efforts have been done to add the ability to express different emotions adapted to pre-defined situations. For example, in e-learning scenarios, "it has been proved to be beneficial for tutoring agents and learning software to show emotional behavior and use strategies based on emotional intelligence". However, most applications are too constrained to the current domain and express emotions according to their goal of development. For example, tutoring agents will more likely motivate students showing hope and interest, while user-frustration systems deal with anger using other emotions and strategies. Fewer efforts have been made to customize a user model that can both infer the current and historical emotional state of the user and define strategic rules to plan communication accordingly.

Clearly, for conversational agents to reach more user acceptance, being able to engage in a compassionate dialogue is what would reduce the gap between a natural conversation settings and an artificial environment. For this purpose, a clearly defined framework that integrates emotion awareness to the way chatbots respond is needed in order to provide a more natural style for communication. There are many applications that can hugely benefit from this enhanced ability to understand, express and deal with emotions in human dialogues. In case of an elderly person who needs physical care, an e-health intelligent agent that is designed to check physical state and recognize illnesses needs to have some emotional intelligence to understand what the person really needs, the gravity of her pain, and whether there is any improvement with medications. In another application, a pyschotherapist agent needs to understand the stress of the other person and be able to sympathize. Even in day-to-day life, emotional intelligence can boost the performance of general purpose conversational agents. As there are many possible ways a conversational agent can respond to a particular context, detecting elicited emotion with more precision can help filter out inadequate answers. For example, responding to an angry person requires more careful conciliation strategies usually by asking more questions to detect the source of the anger as compared to a person feeling sad or despaired which can re-comforted using a different strategy for example by feeling sorry and showing compassion or a worried student who needs encouragement.

In order to do that, a first step consists of building a high quality fine-grained emotion recognition. This is crucial as both the soundness of compassionate communication model and response generation will hugely depend on it. Representative conversational patterns and an understanding of what triggers, reduces or transforms an emotional state into another state are all building blocks of a compassionate response generation system. The goal is to understand how emotional cues can be integrated to orient and enrich open domain conversation.Building systems for recognizing and distinguishing emotions in human dialogues is the center of attention of researchers who explored different approaches. Different techniques were employed depending on the form of data sources relevant to emotion recognition task which are not limited to text but include audio, images, heart and brain signals. Although collecting different sensors can enrich the affective information and help solve the confusion encountered when dealing with input from only one sensor, it can obstructive often leading to contradicting knowledge and requires more effort for pre-processing and combining different inputs to build a common truth. In this work, we address part of those questions contributing to building a compassionate dialogue system by applying and evaluating hybrid approaches to fine-grained emotion recognition in the context of human dialogues and focusing only on text analysis. Based on emotion recognition framework, we provide an analysis of some emotional response patterns and a proposal for integrating emotions in the way the agent responds by training them using neural networks.

In Section II, we review previous work and provide a summary of state-of-the-practice methodologies and concepts in emotion recognition and response generation respectively. Then, we highlight work at the intersection of the two fields. In Section III, we explain our fine-grained emotion recognition approach for annotating the dataset which we also describe the relevance to our emotion response generation task. In Section IV, we present discovered patterns and correlations which represent how a user responds to a specific emotion in the context. Section V and VI is dedicated to architecture of compassionate response generation system and experiments conducted to learn and fine tune embeddings of emotional vectors.

II. Related Work

In this section, we describe previous related work which we divide into two categories: emotion recognition methodologies and emotional dialogue generation. Since this present work aims at studying a task at the intersection of two independent fields and there exists relatively little research that combines emotion recognition applied to response generation, we include a thorough analysis of research in each field separately to explore more possibilities.

A. *Fine-Grained Emotion Recognition*

There are various approaches to text-based emotion recognition at different levels of granularity. While classifying emotions on a fine-grained level conveys more details to distinguish between different emotional states, coarse-grained emotion detection can be achieved more accurately. Although some sentiment analysis techniques can be extended and adapted to emotion recognition, the latter requires more elaborate syntactic and semantic analysis of the sentence as multi-category emotion cues are not as easily extracted as polarity cues. For example, the presence of exclamation marks can point out a subjective sentence while it is not as trivial to distinguish between happiness and enjoyment. Emotion Recognition methodologies adapted from sentiment analysis include corpus-based, knowledge based, lexical affinity methods, lexicon-based, rule-based, statistical approach. In what follows, we cite and discuss research done in each family stressing the importance of a hybrid approach which can build upon weak classifier techniques in order to address the challenges encountered.

*1) **Lexicon Based Detection**:* Keyword based only methodology proved more effective in sentiment analysis than in emotion detection as affective words are more distinguishable from non-emotional words than from each other. This approach relies on a set of emotional keywords or synsets to determine a word emotional orientation. A word is said to be indicative of a particular emotion if it is contained in or if it exhibits a synonymy, antonymy or any other semantic relationship with a set of pre-defined representative words for that emotion. Numerous linguistic resources were developed and made available to achieve this purpose. Based on the assumption that ”it is possible to infer emotion properties from the emotion words” (D’Urso and Trentin, 1998), Strapparava et al. [2] developed WordNet Affect

which extends WordNet by investigating existing correlations between words and concepts and thus maps a set of synsets to their affective concepts. The resource not only determines emotional polarity (positive, negative, neutral) of a word but also determines hierarchically organized emotional labels which distinguish mood, trait, cognitive state, physical state, emotional response, behaviour, attitude, etc. Recently, other lexicons have been proposed to detect emotional categories at a fine grained level namely Geneva Affect Label Coder (GALC) Lexicon which follows the Geneva Emotion Wheel (GEW) model. It comes with database of words commonly used to distinguish up to 36 emotion states constructed by parsing text for these terms and their synonyms.[3] Although this is straightforward and easy to apply this approach, it suffers from (1) the ambiguity of keyword definitions since a word meaning can be interpreted differently depending on context and usage which influences the emotional orientation (for example, complement in ironical context), (2) inability to recognizing emotions of sentences which do not match any keyword in the lexicon, (3) not taking syntactic structure of the sentence into consideration, for example "I am happy" and "I am not happy" will labeled in the same way because they both contain the word happy.

2) ***Human Computation Based Detection****:* This approach focuses on ways to leverage the power of crowdsourcing and human computation to produce emotion annotations with greater level of accuracy. Sintsova et al.[4] worked on the construction of domain-specific multi-category emotion lexicon based on human computation. It has been applied to the data about olympic games gathered from Twitter. However, the work has demonstrated the usefulness of the methodology for different domains. OlympLex lexicon is constructed in many iterations using human annotators from Amazon Mechanical Turk crowdsourcing plateform where annotators were asked to choose the emotion labels corresponding to collected tweets from Geneva Emotion Wheel categories and specify emotion indicators. After further refinements, an emotion recognition model is proposed based on which polarity and multi-label emotion classification can be performed on any new tweets in this domain. The output lexicon outperforms baseline lexicon such as GALC in being context-sensitive.

3) ***Learning Based Detection****:* Learning-based methodology trains machine learning classifiers on annotated dataset to learn hidden patterns in the data. Earlier work heavily exploits machine learning but lacks syntatic or semantic analysis of the sentence. Strapparava et al. [5] developed several variations of knowledge-based and corpus-based methodologies. In their knowledge-based approach, they used Latent Semantic Analysis to represent affective words homogeneously in a vector space and experimented with different weighting schemas to sum up vectors of terms contained in the document. They compared emotion detection accuracy against corpus-based approach which uses Naive Bayes classifier trained on blog data. However, this work didn't make use of natural language pre-processing and syntactic analysis. Burget et al. [6] assessed more machine learning algorithms namely SVM with different kernels, k-neast neighbours, decision trees, bayes networks, and linear discriminant analysis. They performed more pre-processing by converting to lowercase, removing stopwords, punctuation and tokenization and used tf-idf as weighting schema. Although they managed to achieve an average accuracy of 80% on Czech news headlines, it is not clear whether their approach can achieve similar performance if applied to open domain data in another language. Yan et al. [7] examined different machine learning algorithms to determine which techniques proved to perform well in coarse-grained emotion classification can be adapted to a fine-grained open domain level. After annotating a representative sample of twitter data and performing some natural language pre-processing (tokenization, normalization, stemming, filtering), machine learning algorithms including support vector machine, bayesian networks, decision trees and k-nearest neighbor on tweets and their results are trained and evaluated using cross validation. Pre-processing steps limited to tokenization, case normalization, stemming and filtering using minimum word frequency threshold aim at reducing sentences into a unigram bag of word representation. The task was divided into three subtasks: emotion presence (Emotion, None), emotion valence (Positive, Negative, Neutral, Multiple Valence, None) using multi-class-single classification, and emotion categorization into GEW 28 categories using both one versus one and one versus all strategies. A performance of 58% was achieved using Sequential Minimal Optimization and Bayesian Networks with SVM being the top performing classifier in multi-class-single and multi-class-binary respectively. All those papers made the assumption that the emotion of group of words can be infered by looking at the emotions of each word separately without taking into consideration the impact of the relationships and dependencies between words as it is the case in negation, metaphor. Also, little effort has been made to explore new ways of representing features in a vector space that captures best their semantic similarity.

Learning based methodology requires substantial amount of annotated data which is expensive not only in terms of time and money but also training. Recognizing precise emotions can be challenging to humans as they can be interpreted differently depending on the understanding of the context and knowing the person, and often reflects how the annotators themselves express the emotion. So, emotional classification performed in a supervised manner can be highly biased by the quality of human annotation.

4) ***Hybrid Detection****:* Hybrid detection tries to come up with a compromised solution to emotion detection by combining emotional keywords with patterns learned from the data using machine learning and natural language processing syntactic and semantic rules in addition to different theories from psychology and cognitive science. This approach aims to overcome the shortcomings of previously described methodologies, however relatively few research has been conducted in this regard as following this direction often requires more rigorous text analysis and complex algorithmic framework.

Agrawal et al. [8] defined an unsupervised approach to

emotion recognition which performs pre-processing and part of speech tagging to extract affective words and computes emotion vectors for each word based on pointwise mutual information (PMI) similarity measure against a reduced lexicon. It also uses a set of syntactic rules to rebalance and refine the total emotion score of the group of affect words. One of the weaknesses of this approach is that the accuracy of semantic relatedness between words largely depends on the size and domain of the dataset from which they are trained. On six emotion-dataset, this methodology achieved an accuracy of 58% and applied to a four emotion dataset, it reached 62%.

Yang et al. [9] combines lexicon-based, CRF-based (conditional random field) and machine learning-based emotion classification using SVM, Nave Bayesian and Max Entropy. Then, they use a voting model to combine the results from different classifiers. On a model of 6 emotions, they achieved an F-score of 61% with precision 58% and recall 64%.

Ghazi et al. [10] trained a multi-class emotion recognition using an hierarchical approach. They defined objectives at different levels of granularity at each layer of the hierarchy: at a first level, the objective was separating neutral from emotional sentences, then polarity classification and finally they moved to more fine-grained classification. Different features are used for each classification stage, however, this method doesn't take into consideration the dependencies between words.

Shaheen et al. [11] exploited more syntactic dependencies by applying a set of separation and deletion rules. They defined a novel framework for emotion classification which transforms a given input into an intermediate emotional representation (Emotion Recognition Rule) using complex syntactic and semantic analysis and generalizes this representation using various ontologies such as WordNet and ConceptNet. Those generated emotion recognition rules are compared to a set of reference ERRs extracted from a training set using k-nearest neighbors and point mutual information. Applied to blog and twitter data annotated with six Ekman emotion model, this approach reached an average F-score of 84%.

In [12], Sintsova et al. presented Dystemo, a novel framework for semi-supervised emotion classification. Starting from initial classifiers based on general purpose emotion lexicons, a classifier that overcomes classification imbalances can be constructed using Balanced Weighted Voting Algorithm. This approach aims to minimize the amount of annotated data or knowledge required to build a robust classifier. In order to detect the dominant emotion, lexicon-based approach determines a set of pseudo labels. Different lexicons come with different pseudo labels which lead to imbalances of emotion distributions as each lexicon is biased towards dominant emotions. This approach has been applied to sport tweets and was compared with two machine learning classifiers (Naive Bayes and pointwise-mutual information classifier) combined with three initial classifiers. In all cases, BWV outperformed the other ML classifiers for all emotion categories with higher F1-scores and with less difference between precision and recall both in coarse and finer grained classification tasks. The best performance achieved for a model of 20 emotions is 20,5% macro F1-score. We consider this method as the state-of-the-art since we are working on a fine grained emotion classification task which uses the same emotional model and the same direction.

B. Emotional Response Generation

Related work treating emotion recognition in the context of human dialogues vary with respect to the form of the dataset used (text, audio, video) and granularity level of emotion models. In this part, we give a general overview of some fine-grained emotion recognition techniques developed to generate compassionate responses, in particular, we describe some psychological and data-driven framework, challenges and applications.

Holzapfel et al. proposed in [13] a framework for integrating emotional cues to an existing dialogue system. Their strategy for building a compassionate response consists of a layered dialogue manager: on a lower level, it analyzes the sentence structure to understand the goal of the dialogue, the middle layer represents the dialogue in an abstract state and defines interaction patterns/strategies to transition between dialogue states, then on a higher level, it integrates emotional cues to define high level decisions. This framework can make use of any emotional model and can be used in a wide range of applications , however, it is not clear how patterns linking emotions to dialogue state can be learned automatically or need to be handcoded.

Burkhardt et al. [14] focused on detecting customer dissatisfaction and anger from speech dialogue systems and how human conciliation strategies can be transferred into dialog system. It uses two different approaches to detect emotion in human machine communication: crowdsourcing and acoustic features. Then, they aggregated those annotations and trained algorithms such as classification by regression, decision-tree approach and logistic regression. Although this work proposed an approach to detect emotions in the context of a computer human interaction system, a framework that maps detected emotional state to adequate conciliation strategies has not been tested.

Unlike the two previous works, Hasegawa et el. in [15] build and evaluate a dialog system in which the emotion of the response can be controlled to trigger a goal emotional state in the addresse. It is accomplished in two steps: predicting the emotion of an addressee based on the context and generating responses that elicits the goal emotion. A dialogue corpus is extracted from microblog posts and tagged using emotional expressions clues. For predicting the emotion elicited in the addressee, an online passive-aggressive algorithm is used to train the eight binary classifiers corresponding to Plutchik emotion model. For eliciting the best response to a given utterance, a statistical approach based on SMT Moses decoder is used. On top of this framework, an adaptation model was developed to elicit the response that elicits the response that corresponds to the goal emotion. Eight translation models and language models are learned from the emotion-tagged dialogue corpus, each of which is specialized in generating the response that elicits one of the eight emotions.

Compared to emotion dialogue research based on acoustic or visual data, relatively less work have focused on text-based techniques. Based on some studies, it is clear that the

inclusion of emotional features collected from different sensors and dialogue features can greatly enhance the quality of compassionate dialog systems. With the advent of social media, microblogs, the same frameworks developed for other forms of data can be extended with text-based emotion recognition techniques to boost the accuracy of the system.

III. Fine-Grained Emotion Recognition

A. *Dataset*

Our approach makes use of non-annotated movie transcripts extracted from two webly scraped publicly available series: the Big Bang Theory and Friends which consists of a total of 111,103 dialogue utterances. In addition to that, we use Cornell Movie Dialogs Corpus. [1] which consists of a large metadata-rich collection of fictional conversations extracted from raw movie scripts. This corpus has 220,579 conversational exchanges between 10,292 pairs of movie characters in 617 movies. The full dataset contains 412,826 dialogue turns made up of a vocabulary size of 4,209,560.

We designed a crowdsourcing experiment to annotate 250 transcripts randomly chosen from the dataset for validation purposes. To label the transcripts, we use Geneva Emotion Wheel (GEW) model which consists of 20 emotion categories as shown in figure 1.

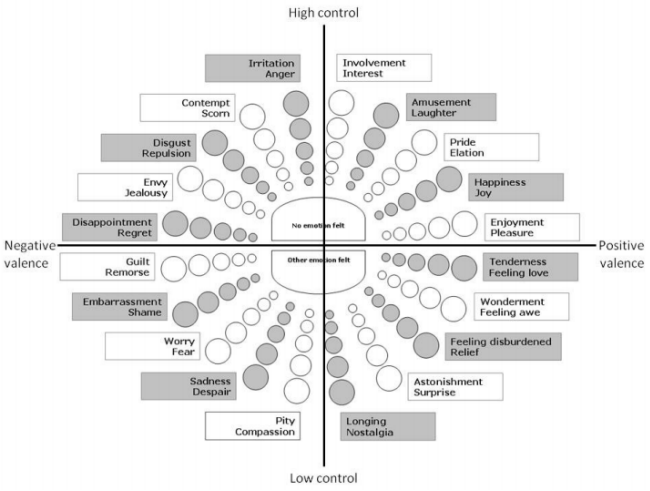


Fig. 1. Fine Grained Geneva Emotion Wheel

Figure 2 shows the distribution of emotions in the validation dataset.

In our semi-supervised approach, we make use of tweets posted by spectators of 2012 Olympic games since they more likely cover 20 emotions which is the same dataset treated and has been previously annotated in [4]. In addition to that we add a It consists of 2,348,176 tweets where the distribution of emotions is described in table I.

[1]Cristian Danescu-Niculescu-Mizil. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011ff

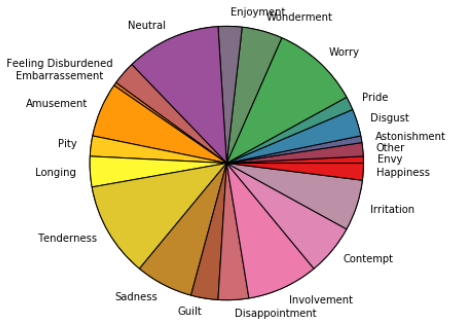


Fig. 2. Emotion Distribution of Human Validation Sample

TABLE I. Tweets Emotion Distribution

Emotion Category	of tweets	%
Involvement-Interest	20204	0.86
Amusement-Laughter	16674	0.71
Pride-Elation	155636	6.6
Happiness-Joy	450662	19.12
Enjoyment-Pleasure	17094	0.73
Tenderness-Feeling Love	689714	29.26
Wonderment-Feeling Awe	15996	0.68
Feeling Disburdened- Relief	28229	1.2
Astonishment- Surprise	26398	1.12
Longing- Nostalgia	33501	1.42
Pity-Compassion	10189	0.43
Sadness-Despair	250690	10.63
Worry-Fear	155989	6.62
Embarrassement-Shame	51846	2.2
Guilt-Remorse	22416	0.95
Disappointment- Regret	79383	3.37
Envy-Jealousy	61327	2.6
Disgust-Repulsion	26059	1.11
Contempt-Scorn	1072	0.05
Irritation-Anger	235097	9.97
Neutral	9172	0.39

B. *Description of the Methodology*

*1) **Motivation and Overview of the Framework**:* As movie dialogue data is not annotated with emotions, we propose a hybrid methodology that combines a lexicon based approach strengthened with syntactic and semantic analysis techniques along with a semi-supervised classifier trained on twitter data and adapted to movie transcripts using techniques from transfer learning. We decide not to follow a fully supervised approach where the cost, time and expertise required for annotation make it quite unfeasible to accomplish within the time frame. Although we designed an approach to annotate a small subset of the dataset, this annotation is just for validation purposes and we cannot use it to train a supervised or even a semi-supervised learning-based classifier as there are not enough instances for each category and because of this skew the classifier will be highly biased towards the dominant categories. We also are aware of the limitations of a knowledge-based or lexicon based only which can suffer from weaker recall of emotional instances in case emotional keywords don't exist in lexicon and especially since domain-specific lexicon and knowledge rules are required to achieve better accuracy. For all those reasons, we got encouraged to build upon promising coarse-grained ER research treating hybrid emotion recognition by 1) testing their adaptivity to our fine-grained emotion recognition task 2)

trying different variations of natural languages analysis techniques 3) exploring semi-supervised and domain adaptation in order to build a clear framework and motivation for substituting supervised methods which are the state-of-the-art in terms of performance.

In our methodology, we use the following principles to orient us throughout the classification pipeline. Before feeding data into any classifier, we focus on building an intermediate representation of a given instance using rigorous natural language techniques. For that purpose, we perform term normalization in order to improve the quality and expressiveness of extracted features. We also pay special attention to noisy words by customizing the list of stopwords to be removed through the use of Named Entity Recognition to detect entities like proper nouns that are less likely to carry emotions. In order to better understand the sentence and keep only content that carry emotions or structural relationships between words that are crucial to understanding the emotion, we also need to analyze the syntactic structure of the sentence. It is only by studying the syntactic dependencies that we can assign different weights since there are specific modifiers or context that impact the emotionality of a word. For example, we need to analyze the sentence "I was never happy" to determine that "never" is a negation modifier of "happy" and thus happy will not have the same emotionality and the sentence should be labelled as sadness.

The diagram in figure 3 gives a general overview of the followed procedure which consists of 4 phases: I) preliminary data pre-processing through parsing, stop word removal and lemmatization, II) affective feature extraction through syntactic analysis, the application of syntactic rules and named entity recognition, III) training both lexicon-based and rule-based classifiers extended with semantic analysis and semi-supervised approach using domain adaptation and IV) fine tuning parameters and evaluation methodology.

2) ***Data Parsing and Natural Language Processing****:* As dialogue data was scraped from the web, we start by performing some parsing and utf-8 encoding and storing it into a dataframe that holds transcripts separated by scene and actor. It was later observed that scene metadata draws on the general context without which it is difficult to elicit the emotion of the transcript. Therefore, we decided to incorporate this information as context cues by collating words from scene metadata. Then, we tokenize the transcripts using RegexpTokenizer which splits a string into substrings using regular expression. It forms tokens out of alphabetic sequences, money expressions and any other non-whitespaces. The downside of this tokenizer is that it doesn't exclude punctuation and numbers but this is not a problem since we remove them. We also detect non-english words and discard them. We will not perform lowercase conversion, punctuation removal, lemmatization and stopword removal at this point since this will lead to inaccurate part of speech tags.

3) ***Affective Feature Extraction****:* To analyze the syntactic structure of the sentences, we start by determining POS tags for each word in each transcript. For that purpose, we use the nltk POS tagger which is assigns tags from universal tagset. After tagging, we construct dependency trees using Stanford Dependency Parser which reaches the state-of-the-art in terms of accuracy. The goal is to learn the relationships between the words and the various constituencies of the sentence. In order to recognize particular entities, we perform Named Entity Recognition using StanfordNERTagger. This will allow us to filter out inaffective contents based on the type of entity (entities like names of people, organizations, monetary values, locations, expressions of time, quantities are not emotional). Since a transcript can consist of more than one sentence, we analyze sentence by sentence based on punctuation marks, we extract features for each sentence separately then we combine them. In order to determine the intermediate affective representation of a particular sentence, we apply a set of rules based on the analyzed syntax:

- Deletion of named entities like names of people, organizations, monetary values, locations, expressions of time, quantities. For example, in sentence like "When we went to Roma last August, we had a lot of fun", entities "Roma" and "August" are discarded.
- Ignoring the part of the sentence before concession indicated by "but", "nevertheless", "however" as the information before is contradictory and can lead to false prediction. For example, in "I used to be happy but now I am miserable", we keep only "I am miserable".
- Detecting negation modifiers such as "no", "not", "never" and combining them with the word or part of sentence depending on it.
- Removing usual words that are used too often like "do", "be", "have", "go", "get", etc and removal of non-affective verbs which can be determined by looking in WordNet Affect.
- Keeping only words tagged as nouns, verbs, adjectives, adverbs, and some modifiers. This implies the removal of pronouns, particles, conjunctions, etc

4) ***Further Data Cleaning****:* We then convert to lowercase and remove punctuation. In order to normalize the terms and get one common representation of multiple versions of the same word which will increase the accuracy of semantic similarity scores, we decided to perform lemmatization using WordNetLemmatizer which is performant and highly accurate. It can reduce any word into its radical form (for example plural (mothers) to singular nouns (mother), conjugated verbs (procrastinating) to infinitive (procrastinate)) as long as it knows its part of speech tag to apply lemmatization rules relevant to this specific type. We decided not to use stemming since this will cut down words and generate words that don't exist. Instead of removing standard nltk stopwords, we come up with our own customized version which excludes modifiers important in our syntactic analysis.

5) ***Emotion Recognition Classifiers****:* We use rule-constrained extended lexicon approach based on Geneva Affect Label (GALC). GALC is a lexicon which enumerates a set of representative words for each one of its 20 emotion categories. However, this version includes stemmed terms which would require us to stem all words in the dataset

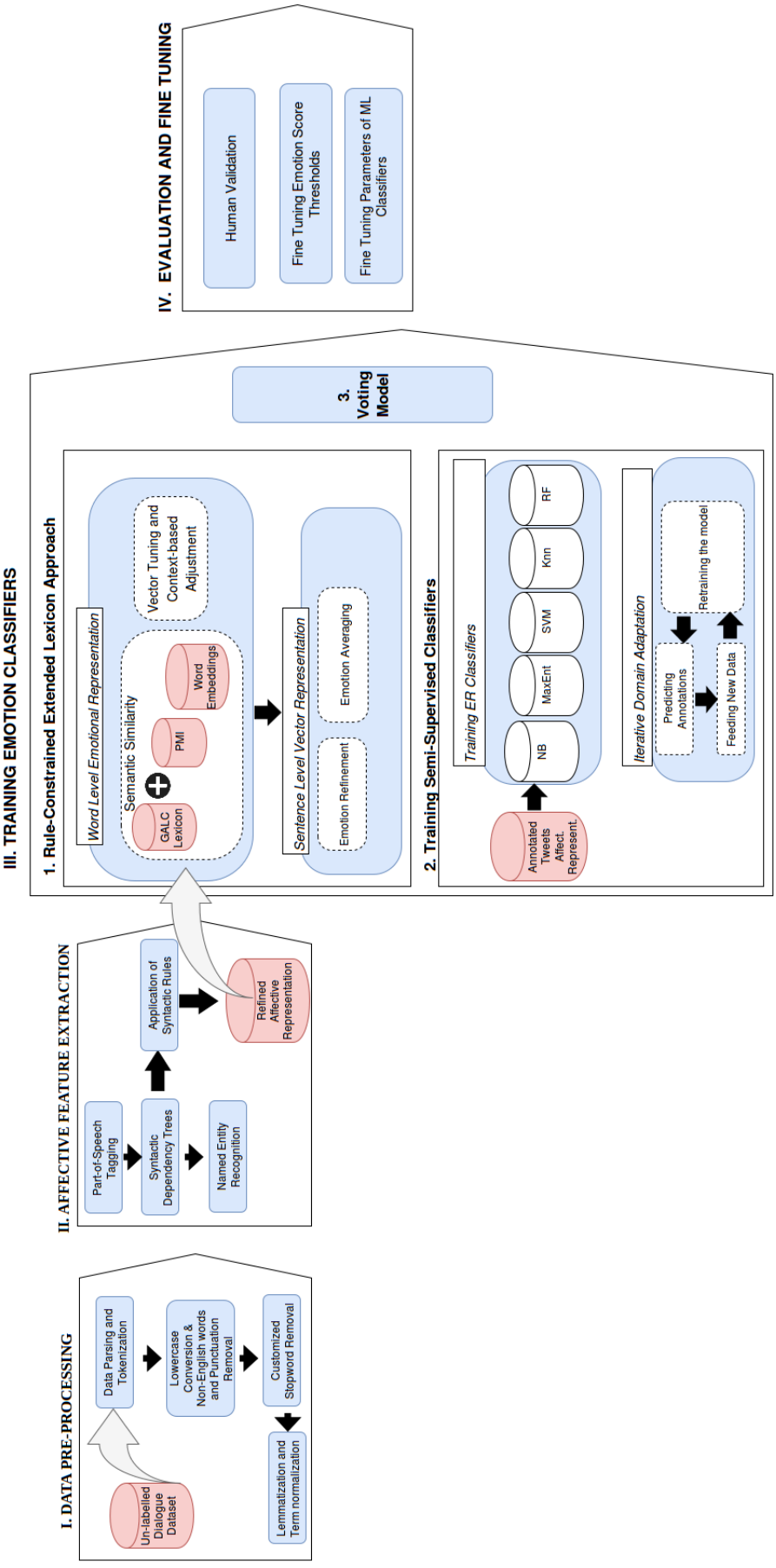


Fig. 3. An Overview to Emotion Recognition Framework

in the same way. Since the current state of stemmer is not guaranteed to lead us to perfect match with the stemmed terms, we prefer to use the extended version of GALC which replaced stemmed terms with word instantiated using tweet data. It consists of 1026 terms with 52.9 on average per emotion category. We also try semi-supervised approach where we train different machine learning classifiers on annotated dataset. At the end, we define a majority voting algorithm as an attempt to increase the recall of certain emotions.

C. *Word Level Emotional Representation*

Let s be a sentence and E be in the set of m emotion labels excluding neutral E = $\{e_i$ — i $\in [1, m]\}$. This is the set of emotion categories we want our sentences to be classified to. Let R_i = $\{r_j$ — j $\in [1, k]\}$ be the set of representative words for an emotion e_i. In our case, these are 20 categories of emotions in GEW model. Then our classification task is reduced to the problem of assigning a dominant emotion label $l_s \in E$ to sentence s. Let F = $\{f_i \in [1, n]\}$ be the set of words or features in a sentence s. We represent each feature f_i with an m dimensional emotional vector where each element corresponds to the semantic relatedness score of the word with the vector space of the corresponding emotion label.

D. *Semantic Similarity*

In order to come up with a semantic score, we tried two different measures in order to decrease the chance of mismatches and as an attempt to reduce the inaccuracies resulting from training some measures on smaller datasets. Relying for example only on PMI measure can hugely bias our scores resulting in co-occurence scores specific to the context in hand but not necessarily reflective of the common patterns that reflect emotional similarity between pairs of words. To increase the accuracy of the scores obtained, we following a hierarchical methodology where we start by keyword spotting using lexicon, if not found, we search for closest match by computing the semantic similarity using PMI or word2vec between words in the lexicon and words in the sentence. Recent work such as the work of Baroni et al. [16] shows that a predictive approach(such as word2vec) gives better distributional semantic representation than count-based approach(such as PMI). But, we decided to try the two different approaches to compare them and our strategy is more collaborative than competitive since we want to define a framework where relatively weak classifiers can help each other out to increase the recall and precision of emotion classification on a fine-grained level.

Since emotions are not always explicitly expressed in the form of emotional cues but can be indirectly inferred depending on the context in which they appear. Also, emotional cues don't always refer to the same emotion. Depending on the context, those cues can evoke different emotions in certain contexts.

*1) **Pointwise Mutual Information**:* We use in addition to direct keyword spotting, Pointwise Mutual Information (PMI) which is a statistical measure based on co-occurence. We follow the assumption that two words co-occuring together with a frequency exceeding a certain threshold are more likely to exhibit a higher semantic similarity. In mathematical terms, PMI between two words w_1 and w_2 can be calculated using the following formula:

$$\text{PMI } (w_1, w_2) = \frac{co-occurrence(w_1,w_2)}{occurence(w_1)*occurence(w_2)}$$

where occurence (w_1) is the number of times w_1 occurs in the dataset, co-ocurrence(w_1, w_2) is the number of times w_1 and w_2 co-occur within a window of a specified size.

*2) **Word2Vec Model**:* Word2Vec is a predictive model trained using neural networks which learns efficiently word embeddings from raw text. We follow a Continuous Bag of Words Model which learns to predict the word given a context within a symmetric window based on the sum of the vector representations of the words in the window. The motivation behind using this new generation of distributional semantics models is that vectors can be trained in a supervised manner unlike in count-based approaches (like pmi) where vector counts need to be reweighted and smoothed using dimensionality reduction techniques in an unsupervised manner in order to achieve higher performance. In word2Vec, "the vector weights are directly set to optimally predict the contexts in which the corresponding words tend to appear" [17]. In other words, a count-based methodology first represent the words after computing statistics of their co-occurences within a context then normalizes those statistics by mapping them to small, dense vector for each word by applying re-weighting and smoothing techniques. On the other hand, a predictive model like word2vec can directly learn those weights by training them through backpropagation. There are two main variations of this model: Continuous Bag of Words Model (CBOW) and Skip-Gram Model. Figure 4 shows the differences between the two models in which the first predicts the word given a context and the second one predicts the context given the word. We experimented with both models and visualized the results using t-SNE dimensionality reduction technique. After evaluating word clusters, it seems that Skip-gram is more relevant to our ER task since it has the ability to learn more fine-grained vector representations but only when more data is available which is not our case. Therefore, we use CBOW which smooths over a lot of distributional statistics without needing data as large by turning context data into observation.

We surveyed each one of the above techniques and we combined them in the following manner to calculate the semantic similarity relatedness score between a word and an emotion category:

- **1st Stage: Lexicon-based Keyword Spotting:** If there is a match between the word and at least one representative word for a specific emotion, then the similarity score of the word for that emotion is set to 1.
- **2nd Stage: WordNet Affect:** If the word is not found in the set of the representative words for the lexicon, it

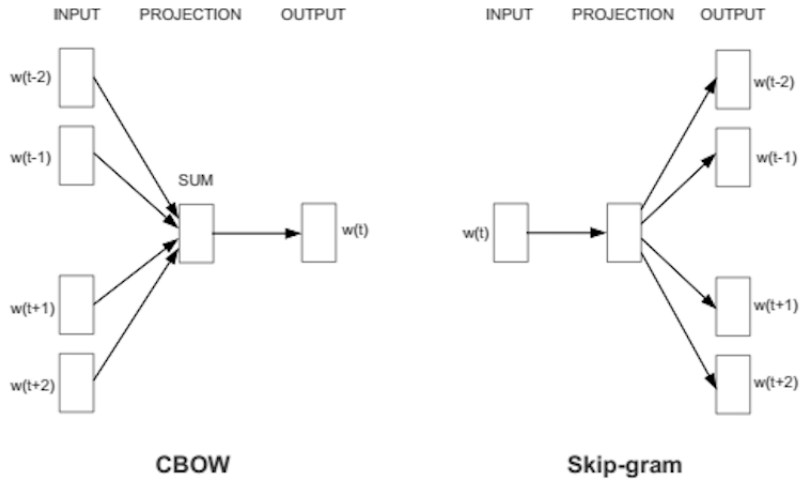


Fig. 4. Continuous Bag of Words Vs Skip-gram Model

is searched in WordNet Affect which affective concept is closest to it, then a mapping is done to an emotion category and the similarity score is increased to 1

- **3rd Stage: Average of PMI and Word2Vec Similarity Scores:** We compute PMI scores for each pairs of words that co-occur together within a context window of 10 (it gives us the best results) and train our word2vec model to get word embedding based similarity scores then we add up the mean scores between the word in question and each representative word then we normalize using arithmetic mean as shown in the formula below since we have unbalanced number of representative words per each category in our lexicon:

SIM $(w_i, e_i) = \sqrt[k]{\frac{PMI(w_i,r_j)+word2VecSim(w_i,r_j)}{2}}$

Based on this, each word feature has an emotional vector representation $\sigma \mathrm{w}_i = \begin{bmatrix} SIM(\mathrm{w}_i, e_1) \\ SIM(\mathrm{w}_i, e_2) \\ SIM(\mathrm{w}_i, e_3) \\ ... \\ SIM(\mathrm{w}_i, e_m) \end{bmatrix}$

E. Emotion Refinement

We take syntactic dependencies between words relevant in the context of emotion detection to redefine dependent words as defined by these handcoded rules:

- **Adjectival Modifier:** when a noun is preceded by an adjective, the emotion of the adjectival phrase depends on the emotion of the adjective as it cancels out the emotion of the noun. For example, in the sentence "What an unfortunate luck!!" noun "luck" depends on adjective "unfortunate" which defines emotion of this sentence. If we were to average the two words out, unfortunate having a sadness emotion and luck having happiness emotion would lead to a neutral emotion, whereas taking into account an interesting case of inter-relationships between words will assign Sadness to this phrase.
- **Adjectival Complement:** when a verb is followed by an adjectival complement, the emotion of the verb depends on that of the adjective. For example, in the sentence "I feel depressed", the emotion of the sentence is that of the adjective which is sadness.
- **Adverbial Modifier:** when a verb is followed by an adverb complement, the emotion of the verb depends on that of the adjective. For example, in the sentence "I struggled happily"
- **Nominal or adjectival Negation Modifier:** when a noun or an adjective is preceded by a negation modifier, for example, "This is not funny" the emotion of the noun or the adjective is neutralized.
- **Verbal Negation Modifier:** when a verb is followed by a negation modifier, as in "I don't suffer", the emotion of the noun or the adjective is neutralized.

F. Sentence Level Vector Representation

Now that we have transformed word features into 20 dimensional vector representation and applied syntactic rules, we can calculate the dominant emotion of the sentence by averaging the emotional vectors of the words that make up the sentence as shown in the formula below:

$$\sigma(s) = \begin{bmatrix} \frac{\sum_{i=1}^{n} \sigma_1(w_i)}{n} \\ \frac{\sum_{i=1}^{n} \sigma_2(w_i)}{n} \\ ... \\ \frac{\sum_{i=1}^{n} \sigma_m(w_i)}{n} \end{bmatrix}$$

where m is the number of emotion categories, n is the number of words in the sentence. As a result, we will obtain the emotional vector for each sentence. To get the dominant emotion, we take the index of the element with maximum score provided that it exceeds a certain threshold t, otherwise it is categorized as neutral.

G. Semi-Supervised Approach

In this approach, we follow the same pre-processing and syntactic procedure opting for word2vec representation of words. Using word2vec, we represent each word using 300 dimensional vectors. Then, we take the average over the vector representations to get the sentence vector representation. After that, we ran a series of experiments and variations of machine learning algorithms to train on another annotated dataset in order to determine what works well for our task. In order to increase the capability of machine learning model trained on one dataset to generalize well on dataset from another domain, we opted for domain adaptation approach where random instances from unlabelled dataset are gradually and iteratively added to the training process treating them as golden truth.

1) ***Machine Learning Pipeline****:* In what follows, we explain details of the followed machine learning pipeline and reasons for certain decisions. The diagram in figure 5 summarizes the procedure followed.

2) ***Data Treatment and Vector Representation****:* We apply the same procedure explained above for coming up with a refined affective feature representation of each tweet in addition to some minor data cleaning such as finding and removing usernames preceded by @, replacing emoticons with their corresponding word meaning, removing hashtag signs, etc. We train a gensim CBOW word2vec model on tweets in tokenized and lemmatized format before applying it on the tweets in affective feature format to get a vector per each word. Using context of 10, minimal word occurence of 4 and dimensionality of 300, we get the best results after evaluating it on the validation dataset. To obtain the vector representation of each tweet, we average over the vector representation of each word in the tweet. Then, we randomly split the dataset into training and testing dataset and train the model only on the training dataset and a part of the target unlabelled dialogue dataset.

3) ***Dealing with Emotion Class Imbalances****:* Before applying any machine learning algorithm, we start by addressing the problem of unbalanced distribution of emotion categories. If not done properly, this can highly bias classifiers towards dominant classes, thus we decided to take the weights of the distribution of different classes into consideration and pass them as a parameter to training our scikit learning algorithm whenever it is possible.

4) ***Feature Covariance Matrix and Dimension Reduction****:* Because of the curse of high dimensionality, fitting machine learning models with more dimensions tend to result in less predictive power. We tried applying dimensionality reduction technique to keep only a few principal components that capture most of the variability of the data. For that purpose, we used PCA which is a linear orthogonal transformation that transforms the data to a new coordinate system. Since we have vector representation with a dimensionality of 300, we tried to reduce the dimensionality by keeping fewer principal components (10,15,300), for some algorithms such as random forest, we get the best results with 10 dimensions.

5) ***Training Classifiers****:* We tried the following algorithms with different implementation variations and parameters:

- SVM with different kernels: linear, radial, polynomial
- Naive Bayes: bernoulli, gaussian, multinomial
- K-Nearest Neighbors
- Random Forest

Some of them didn't give us fine grained results, predicting only a subset of target emotions which didn't encourage us to pursue them any further. In what follows we include the details of training bernoulli Naive Bayes which gives us the best results in terms of training and testing errors ratio, less overfitting, better recall of emotions needed for our fine grained task. We also include a summary comparison of the different algorithms.

6) ***Domain Adaptation****:* Since we are planning to train machine learning model on annotated dataset(source) and apply it to predict labels for dataset belonging to another domain(target), domain adaptation can be used in order to improve the accuracy of the predictions. Depending on the target task, there are many different approaches that can be applied: reweighting, iterative, searching for common representation space and so. Since surveying more than one technique is beyond the scope of this project, we go for a simple iterative auto-labelling approach which is also better suited as it doesn't need any target instances already labelled. The algorithm can be described as follows:

- An initial model M is learned based on only source labelled dataset.
- Model M is used to label a small subset of the target dataset.
- The newly labelled target instances are used to retrain another model M'.

The effectiveness of this approach largely depends on the size of the datasets and the number of iterations as the more it is trained the more impact it has on bringing closer the two space representations. In our experiment, we use 10 iterations to gradually add more instances from the target dataset into the training processing.

7) ***Parameter Tuning using Cross Validation****:* In order to fine tune the parameters of our model, we use grid search in order to optimize number of trees, and size of subset of tried features. We then use cross validation with 10 iterations to train and test on different splits of the dataset which has two advantages: first it reduces the likelihood of overfitting as the model is trained on more data splits, second it returns the mean scores which gives a more accurate idea of the performance of the model.

8) ***Learning Curves and Variance and Bias Reduction****:* To visualize how the classifier performs as we increase the size of data and to determine whether we are suffering from a model with high variance or high bias, we plot the learning curves

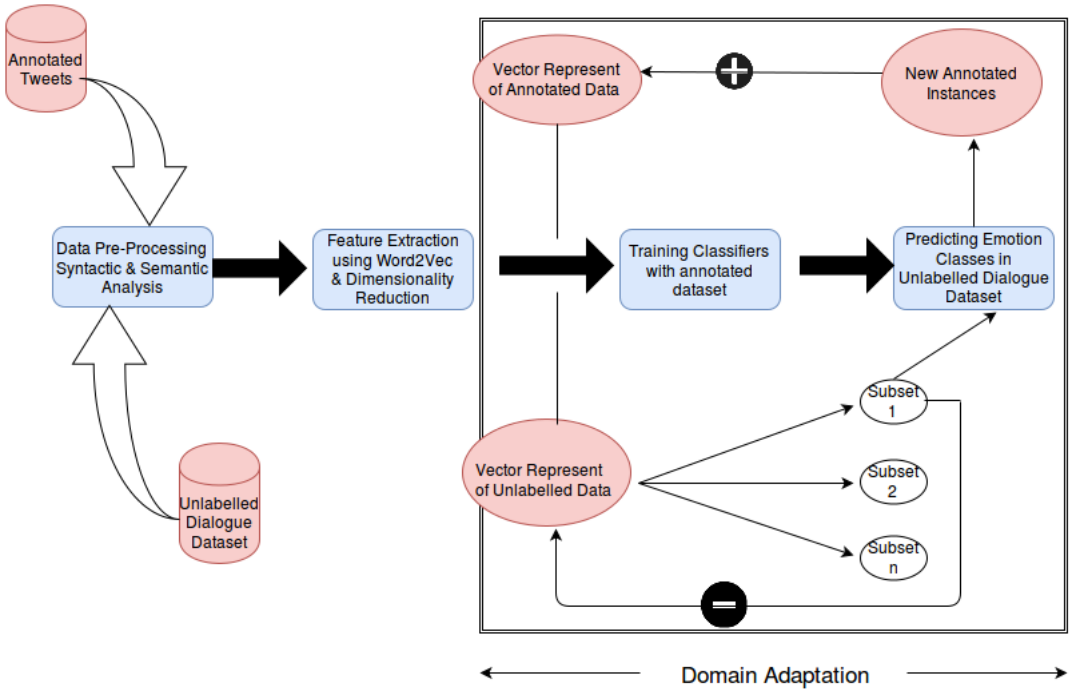


Fig. 5. Semi-Supervised Pipeline

for different splits of the training dataset. As shown in figure 6, naive bayes classifier benefits from adding more data as the testing performance increases towards the end from 10% to 20% and the variance between the training and testing shrinks (low variance and low bias). So, as more data is added, we observe reduction of overfitting and better predictions.

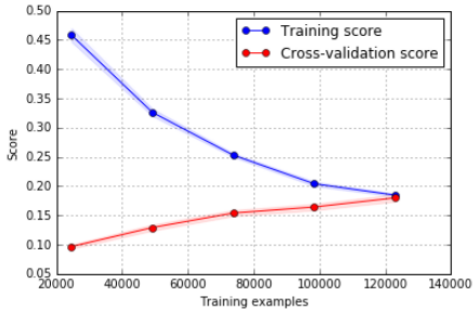
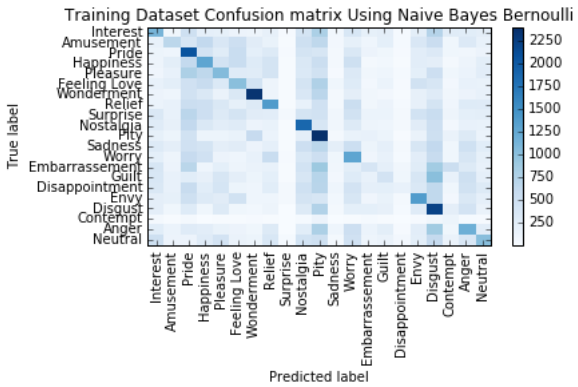


Fig. 6. Learning Curve for Different Splits of Training Dataset

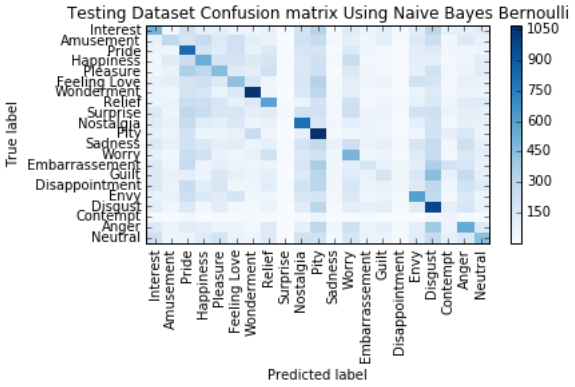
*9) **Confusion Matrices**:* Using the combination we got after fine tuning, we plot the confusion matrices of training, testing and validation datasets to observe the disparities of the performance of the model with respect to the different emotions and how well it machine learning model. Figures 7(a), 7(b) and 7(c) show confusion matrices for training, testing and validation datasets respectively. We also include in this section, for comparison purpose, the confusion matrix using GALC-Extended with PMI which gives us the best performance.

By investigating the different confusion matrices, we observe that the classifier does a great job finding true positives especially for emotions: pride, wonderment, pity and disgust where it returns approximately 100% accuracy. However, this ability to recognize those emotions well somehow biases its predictions which tend to more centered around those emotions. By looking at the confusion matrix of GALC extended with PMI, we can see that it has less false negatives for most emotions but not for all emotions there are strong true positives. We can conclude that although GALC extended with PMI outperforms Naive Bayes, the latter still does better job recognizing some emotions especially for relief and disappointment.

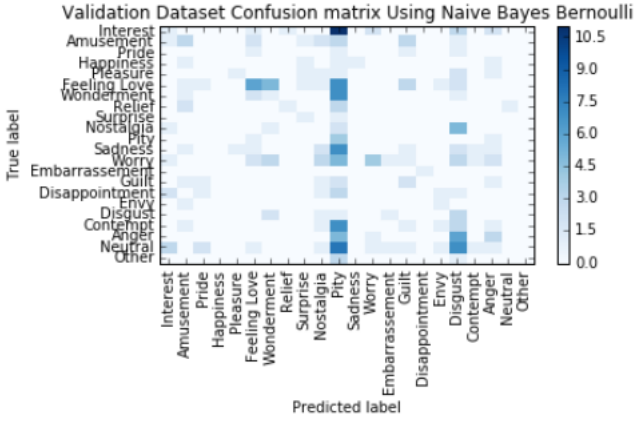
*10) **Summary of Machine Learning Algorithms**:* Table II shows a comparison between the different algorithms tried in terms of their performance on the training, testing and validation. After trying many different algorithms with different algorithms and using cross-validation over different subsets of training and testing datasets, we come at the conclusion that although some algorithms did better in training and testing dataset, they generalized poorly even with the use of domain adaptation. For instance, with random forest we get a high f1-score of 70% on both training and testing datasets, however on the validation dataset, it perform very badly with a score of only 4.9% which tells us that this algorithm overfitted on validation dataset although it performs evenly good on testing dataset chosen randomly. On the other hand, naive bayes, the most basic algorithm we tried performed evenly on training, testing and validation. Although Naive Bayes macro f1-score for training and testing doesn't exceed 20% it still generalizes evenly on the validation dataset making it the best algorithm that we compare later on with our other approaches.



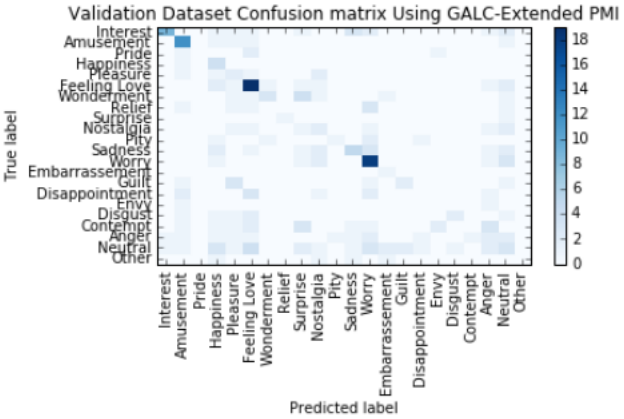
(a) Training Confusion Matrix with Naive Bayes



(b) Testing Confusion Matrix with Naive Bayes



(c) Validation Confusion Matrix with Naive Bayes



(d) Validation Confusion Matrix with GALC-Extended with PMI

TABLE II. SUMMARY OF MACHINE LEARNING ALGORITHMS

	Training (Labelled Only)				**Testing**				**Validation**			
	Macro-P	Macro-R	Macro-F1	AUC	Macro-P	Macro-R	Macro-F1	AUC	Macro-P	Macro-R	Macro-F1	AUC
RF	**70.3**	**72.4**	**70.9**	**78.8**	**69.5**	**72.0**	**70.3**	**78.0**	4.9	6.2	4.9	53.3
SVM (Linear)	21.5	17.0	18.9	50.2	20.6	16.4	16.5	47.6	11.4	6.8	5.9	48.2
NB (Bernoulli)	18.3	17.5	16.4	58.7	18.5	17.4	16.4	58.2	**13.3**	**14.4**	**10.4**	**54.5**

H. Voted Model Between Hybrid Based and Semi-Supervised

At the final stage and after evaluating our different classifiers, a voting model was applied to aggregate the predictions from relatively weak ER classifiers towards building stronger one. We largely take into consideration the ranking of the classifiers to impact weights and importance in the voting process. The idea is to take the majority when it is possible, otherwise favor the classifier with the best performance. The purpose of this model is to further improve our rule-based pmi classifier which gives us the best results on average and in most emotions. So, our rational is to re-assign an emotion label only if the other two classifiers: word2vec rule-based and semi-supervised agree on a different emotion. To improve the recall, in case pmi returns neutral class, our voting approach checks if word2vec returns an emotion and uses it otherwise, it takes the one of semi-supervised.

I. Validation Approach

*1) **Rational**:* In order to evaluate the performance of our ER methodologies applied in an unsupervised and semi-supervised manner, we designed a crowdsourcing experiment for a sample of the dataset. Due to lack of time, money and people, it is not possible to annotate the whole dataset. Therefore, it is crucial to use a sample that is representative enough to be able to generalize with higher confidence on the accuracy of the ER framework. For that purpose, we need to chose subtitles randomnly and to make sure the distribution of emotion classes in the sample dataset follows the same distribution in the whole dataset. Rather than going for a completely random approach to select sample dataset, we started by separating subtitles classified as non-emotional (70%) from emotional ones and filtered out texts with less than 30 characters to have more chance of having subtitles containing emotional contents (11%) of the whole data and 40% of emotional

data). Since classification into 20 different categories using our unsupervised learning methods resulted in skewed distributions of emotions, we choose to take that into consideration by following a stratified sampling strategy. For example, if we would like to have a sample size of 1000 sentences and we have 30% of sentences classified as emotion A and 20% sentences classified as emotion B and 50% of sentences classified as emotion C, then we need to take 300 random sentences from emotion A, 200 random sentences from emotion B and 500 random sentences from emotion C.

To reduce labelling by chance and mistakes, we have thought about two-fold validation strategy where the sample dataset is divided into subsamples which are given to two people. So for each subtitle we will have the labels from two people and we can calculate the agreement by simply taking an unweighted or weighted average if there is a way to assess the expertise of the annotator before the start of the experiment. For instance, a way to assess it is by designing a training homework (quiz) on a small dataset for which we already have the golden truth. Depending on the correctness, precision of annotators in the training homework, their labels in the unannotated dataset will be weighted accordingly. However, this cross validation strategy is not feasible, given the insufficient number of volunteers and lack of financial and time resources. Thus, we used an alternative strategy where annotators are asked to justify their annotations which will enable them to rethink an annotation in case they find it hard to justify. For this purpose, we have included justification questions where the annotator is asked to indicate textual indicators in the subtitle which justify the chosen emotion and how else they could have expressed it. This approach not only gained us better annotation accuracy but also we crowdsourced domain-specific textual features that could be used to extend the lexicon and an opportunity to build reflexion on how we express and understand emotions.

2) ***Experiment and Reflexion****:* We aimed for annotating in sum 250 subtitles by 12 annotators. For the purpose of collecting responses from people, we have created 20 Google forms with two versions: shorter version with 10 subtitles, and the second one with 15. Depending on time, some annotators were asked to annotate 25 subtitles and the separation of the two versions enabled them to do it in two steps to ease the workload. We asked for volunteers among friends and people who work in the laboratory and gave them a description of the emotion model used and the general scope of the Emotion Recognition task. To avoid giving more/less advantage to humans annotators vs our ER system, we provided the subtitles in the same textual form as it was provided, i.e: without any contextual clues linking events from other scenes to the current one and without video or audio from the movie from where they were extracted. They were asked to focus more on general witnessed dominant emotion than the one expressed by a specific character or author since in our ER system we did not take into consideration the emotion from a particular scope or point of view. For example, there is no difference between ”I am happy”, ”they are happy” or ”there is a lot of happiness” as all of them can be annotated as ”Happiness-Joy”. Also, if there are two emotions one from the past and the other from the present or future we asked to select the most recent and dominant one.

TABLE III. HUMAN VALIDATION EMOTION DISTRIBUTION

Emotion Category	# of subtitles	%
Involvement-Interest	21	8.4
Amusement-Laughter	16	6.4
Pride-Elation	4	1.6
Happiness-Joy	5	2
Enjoyment-Pleasure	7	2.8
Tenderness-Feeling Love	28	11.2
Wonderment-Feeling Awe	12	4.8
Feeling Disburdened- Relief	7	2.8
Astonishment- Surprise	2	0.8
Longing- Nostalgia	9	3.6
Pity-Compassion	6	2.4
Sadness-Despair	17	6.8
Worry-Fear	26	10.4
Embarrassement-Shame	1	4
Guilt-Remorse	8	3.2
Disappointment- Regret	9	3.6
Envy-Jealousy	2	0.8
Disgust-Repulsion	8	3.2
Contempt-Scorn	15	6
Irritation-Anger	15	6
Neutral	28	0.112
Other	4	1.6

TABLE IV. STATISTICS OF SUBTITLES WITH AMBIGUOUS OR MULTIPLE EMOTIONS

	# of subtitles	%
Ambiguous	27	10.8
Multiple Emotions	16	6.4

Table III shows the distribution of emotions in the validation sample. The main challenges encountered during this experiment can be summarized in two points:

- **Emotion Ambiguity:** This is due to lack of context which makes it hard to understand the emotionality of the sentence. It can also be caused by different ways of interpreting the emotion which leads to other kinds of discovered emotions.
- **Multiple Dominant Emotions:** Sometimes when expressing complex feelings, multiple emotions are used which makes it hard to narrow down to one predefined category. This is a peculiar problem with human dialogs unlike tweets whose length is limited and are much more likely to cover one idea and thus carry one dominant emotion. In human dialogs, people can talk more freely and not constrained to one emotion by utterance as they can express many emotions simultaneously.
  This can be solved by defining new emotion models based on clusters of emotion categories where every cluster contains closely related emotions or a weighted emotional model in which multiple emotions can contribute to one complex emotion. Table IV shows the distribution of instances with ambiguous and mutiple emotions in the validation sample.

IV. RESULTS AND ANALYSIS

Let TP is the number of true positives, FP the number of false positives and FN the number of false negatives. Then

TABLE V. SUMMARY OF PERFORMANCE OF THE DIFFERENT METHODOLOGIES

Methodology	**Micro**			**A**	**Macro**			**AUC**	**rank**
	P	R	F1		P	R	F1		
GALC-Plus	33.2	33.2	33.2	33.2	24.1	25.6	24.8	58.7	2
GALC-Extended PMI	**34.0**	**34.0**	**34.0**	**34.0**	**24.7**	**26.3**	**25.42**	**60.2**	**1**
GALC-Extended Word2Vec	32.0	32.0	32.0	32.0	21.8	25.6	20.6	57.4	3
Semi-Supervised with Naive Bayes	12.4	12.4	12.4	12.4	13.3	14.4	10.4	54.5	5
Voting Model	33.2	33.2	33.2	33.2	22.8	26.4	21.2	56.5	4

TABLE VI. EVALUATION RESULTS PER POSITIVE EMOTION CATEGORY

Positive Emotions	**GALC -Plus (Extended)**			**GALC-Extended PMI**			**GALC-Extended Word2Vec**			**Semi NB**			**Voting Model**		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Involvement	64.3	42.8	51.4	81.8	42.8	56.2	75.0	42.8	54.5	25.0	18.7	21.4	**50.0**	**75.0**	**59.9**
Amusement	38.7	75.0	51.1	50.0	75.0	59.9	**52.1**	**75.0**	**61.5**	0.0	0.0	0.0	0.0	0.0	-
Pride	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	**20.0**	**80.0**	**32.0**
Happiness	13.3	80	22.8	20.0	80.0	32.0	**22.0**	**80.0**	**34.8**	50.0	14.3	22.2	12.5	28.6	17.4
Pleasure	16.6	14.3	15.4	13.3	28.6	18.2	12.5	28.6	17.4	35.3	21.4	26.6	**46.3**	**67.8**	**55.1**
Love	43.5	71.4	54.1	**47.5**	**67.8**	**55.8**	45.0	64.3	52.9	8.3	8.3	8.3	59.9	25.0	35.3
Awe	**42.9**	**50.0**	**46.2**	59.9	25.0	35.3	50.0	16.7	25.0	50.0	14.3	22.2	0.0	0.0	-
Relief	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	**20.0**	**50.0**	**28.6**	0.0	0.0	-
Surprise	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	**15.4**	2 **2.2**	**18.2**
Nostalgia	18.2	22.2	19.9	15.4	22.2	18.2	14.3	22.2	17.4	4.9	6.7	9.0	**50.0**	**16.7**	**25.0**
Average over positive emotions	23.7	35.6	28.5	**28.8**	**34.14**	**31.2**	27.1	32.9	29.7	19.3	13.4	15.8	25.4	31.5	28.1

TABLE VII. EVALUATION RESULTS PER NEGATIVE EMOTION CATEGORY

Negative Emotions	**GALC -Plus (Extended)**			**GALC-Extended PMI**			**GALC-Extended Word2Vec**			**Semi NB**			**Voting Model**		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Pity	**100**	**16.7**	**28.6**	50.0	16.7	25.0	50.0	16.7	25.0	0.0	0.0	-	41.7	29.4	34.5
Sadness	40.0	23.5	29.6	41.6	29.4	34.5	33.3	23.5	27.6	50.0	15.4	23.5	**47.4**	**69.2**	**56.2**
Worry	**51.7**	**57.7**	**54.5**	43.9	69.2	53.7	0.0	0.0	-	0.0	0.0	-	20.0	100	33.3
Embarrassement	20.0	100	33.3	20.0	100	33.3	25.0	100	40.0	15.4	25.0	19.0	**59.9**	**37.5**	**46.2**
Guilt	50.0	25.0	33.3	50.0	25.0	33.3	**59.9**	**37.5**	**46.2**	0.0	0.0	-	0.0	0.0	-
Disappointment	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	**20.0**	**50.0**	**28.6**	0.0	0.0	-
Envy	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	-	7.3	37.5	12.2	**28.6**	**25.0**	**26.7**
Disgust	50.0	12.5	20.0	**66.7**	**25.0**	**36.4**	28.6	25.0	26.7	0.0	0.0	-	0.0	0.0	-
Contempt	0.0	0.0	-	0.0	0.0	-	0.0	0.0	-	**21.4**	**20.0**	**20.7**	14.3	13.3	13.8
Irritation	**18.2**	**13.3**	**15.4**	14.3	13.3	13.8	12.5	13.3	12.9	0.0	0.0	-	0.0	0.0	-
Average over negative emotions	32.9	24.9	28.3	**28.65**	**27.86**	**28.4**	25.3	28.5	26.8	11.4	14.8	12.9	21.2	27.4	23.9

precision, recall and F-score are defined as:

Precision P: P = $\frac{TP}{TP+FP}$

Recall R: R = $\frac{TP}{TP+FN}$

F1-score = $2 \times \frac{PR}{P+R}$

In evaluating the performance of multi-emotion classification, we use macro version of F1-score as it measures more effectively performance in the presence of unbalanced classes as we don't want large classes to dominate smaller ones. The formulas for calculating precision, recall and F1-score are given below:

where π_{macro} is macro-averaged precision defined as:

$\pi_{macro} = \frac{1}{m} \times \sum_{i=1}^{m} \frac{TP_i}{TP_i+FP_i}$

and ρ_{macro} is macro-averaged recall defined as:

$\rho_{macro} = \frac{1}{m} \times \sum_{i=1}^{m} \frac{TP_i}{TP_i+FN_i}$

and macro F1-score = $2 \times \frac{\pi \times \rho}{\pi+\rho}$

We also consider micro-averaged precision which is defined as :

$\pi_{micro} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+FP_i}$

micro-averaged recall defined as:

$\rho_{micro} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i+FN_i}$

and micro F1-score defined as:

micro-f1 = $2 \times \frac{\pi_{micro} \times \rho_{micro}}{\pi_{micro}+\rho_{micro}}$

In addition to those two measures, we also use area under curve(AUC) which takes into consideration the unbalances in the number of instances for each class. The formula for AUC is as follows: Let c be a fixed classifier. Let x_1, x_2,..., x_m be the output of c on the positive examples and y_1, y_2,..., y_n its

output on the negative examples. Then, the AUC associated to c is given by:

$$\text{AUC} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{x_i > x_j}}{mn}$$

To show how close are classifiers predictions to crowdsourced truth labels, we use accuracy (A) which is defined as:

$$\text{A} = \frac{TP+TN}{TP+TN+FP+FN}$$

Based on both the micro-averaged and macro-averaged F1-score shown in table V, the two approaches that yields the best performance is rule-based GALC-Extended with PMI with an micro-f1 of 34.0%, a macro-f1 of 25.4% and AUC score of 60.2% which accounts for an increase over the baseline GALC-Plus of 0.8%, 0.6% and 1.5% in micro-f1, macro-f1 and AUC respectively. On classifying positive and negative emotions GALC-Extended with PMI also outperforms the baseline with average f1-scores of 31.4% and 28.4% which is a larger increase of 1.8% for positive emotions and slight improvement of 0.1%. Higher recall of GALC-Extended PMI can be explained by the fact that it doesn't rely only on specific keywords in the lexicon but looks for other semantically similar implicit keywords. The fact that the baseline model used in this comparison performs fairly good is due to the use of rule-based syntactic approach and rigorous natural language processing techniques for cleaning and analyzing the affective weight of certain parts of the sentence which enriches the methodology and goes beyond a simple lexicon-based approach. We wanted to show the role of applying pre-processing to get rid of noise in the data and extract only relevant features which can hugely impact the accuracy of any classifier.

In the third rank comes GALC-Extended Word2Vec with a micro-f1 of 32.0%, macro-f1 of 20.6% and AUC of 57.4% which is a slight decrease compared to the two methods. The reason why a predictive semantic similarity model didn't perform as well as expected is due to the fact that its implementation uses neural networks which would require longer training time and more rigorous parameter tuning to come up with better vector estimation unlike PMI which could be trained easily on larger datasets and still can perform better without requires a lot of time. The best selected semi-supervised approach with no matter how many iterations used in domain adaptation achieves on average the lowest performance with a micro f1-score of 12.4% and macro f1-score of 10.4 % which is less than half of the performance performed by the other rule-based hybrid approaches and baseline. However, for some emotion classes, it still performs better f1-score namely 28.6% for relief, 28.6% for disappointment, 20.7% for which all other classifiers gave a f1-score of 0.0% as shown in tables VI and VII.

In the fourth place comes the weighted voting model which still does much better than the semi-supervised approach, but against our expectations, on average it doesn't improve the recall of the best methodology. Nevertheless, there are some emotions for which it exhibits the best predictive power: interest, elation, pleasure, relief and surprise. This is quite useful especially for emotions like pride and surprise all other classifiers gave us 0.0 f1-score and the voting model interestingly corrects that with higher f1-scores of 32.0% and 18.2% respectively as shown in tables VI and VII. This is an interesting finding as it tells us that some classifiers are better in classifying certain emotions and this fact can inspire future work to come up with an approach that takes into consideration the expertise of each classifier to assign weights in a more elaborate voting model.

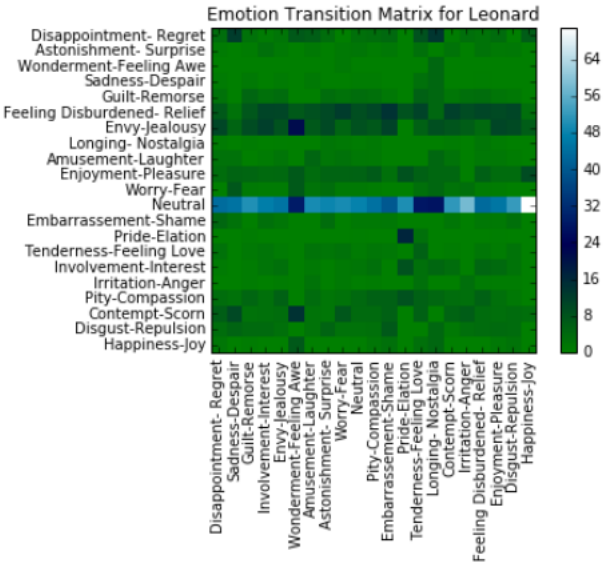
V. Discussion and Conclusion

This paper surveys hybrid fine-grained emotion recognition techniques and adapts them to the domain of human dialogues. We combined both lexicon-based, syntactic rule based, semantic analysis using both count-based (PMI) and predictive word embeddings (word2vec) and semi-supervised machine learning approach to leverage the capability of our model to detect emotions with better recall and precision. By refining the affective representation of subtitles through the use of various techniques such Named entity recognition, lemmatization, and the application of some handcoded syntactic rules, we managed to improve the accuracy of our predictions. In order to evaluate our framework, we designed a human validation experiment where we annotated 250 out of the whole dataset. We got the best results using GALC-Extended PMI on average and high recall for certain emotions not found by most other classifiers using either semi-supervised approach or voting model. The fact that combination of rule-based and lexicon-based with PMI performs relatively well tells that people not only use explicit keywords to express an emotion but use implicit expressed that can only understood by referring back to the context. All in all, we conclude that all methodologies can work together to increase the recall of emotions as each one has a strength to detect certain emotions that others sometimes cannot detect in a satisfying way.

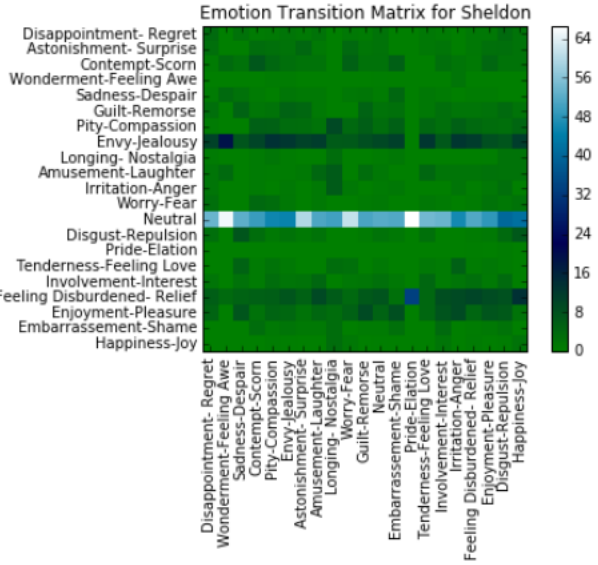
VI. Analysis of Emotional Responses

In order to find representative patterns in emotional responses and to show its impact in a dialogue system, we look at the distributions of emotions in contexts and responses for different actors and scenes. For this, we started by computing the transition matrices that capture causality relationships between emotional contexts and their respective responses in both directions (probability distributions for emotions in responses by fixing the emotion in the context and probability distributions for emotions in contexts by fixing the emotion in the response). Figures 7(e), 7(f) and 7(g) show transition matrices for three different actors in BBT: Leonard, Sheldon and Penny respectively.

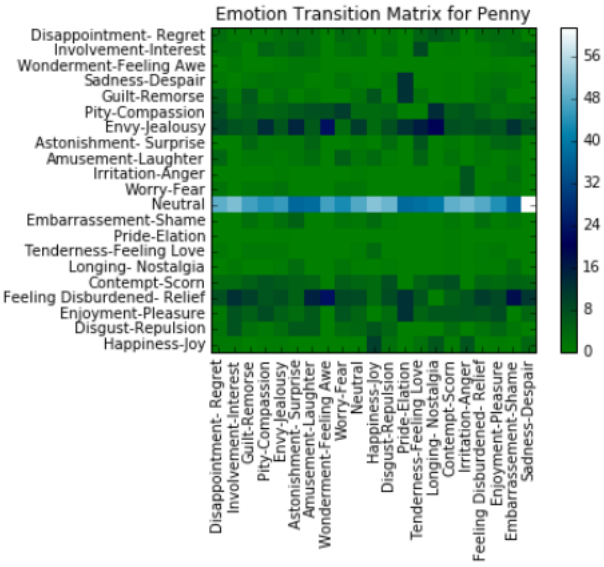
By analyzing those transitions, we can say that different actors respond differently to different emotions where Leonard is more curious and is attentive to details, Sheldon more likely to with contempt-scorn to all emotions, and Penny is more compassionate which somehow goes inline with the real characters in the TV series. This can be exploited in a



(e) Transition Matrix of Emotional Responses of Leonard



(f) Transition Matrix of Emotional Responses of Sheldon



(g) Transition Matrix of Emotional Responses of Penny

compassionate dialogue system by learning by incorporating the idea of personas since it is hard to come up with a general compassionate system that works for all situations. As each person can respond differently to an emotional situation depending on their goals, if the dialogue system can take the position and personality of that person and does exactly like how this person reacts to different people, then dialogue system can engage in more natural conversations. For example, if the person is your friend then she will most likely enjoy time with you, else if this person is your doctor then she should show more interest in helping you and more compassion towards you, etc.

In particular, we have tried to focus on three kinds of patterns: imitation, polarity, indifference between the emotions in contexts and responses. As figures 7(h) and 7(i) show, strong imitation patterns can be observed especially in positive emotions like pity-compassion, feeling-disburdened-relief, amusement-laughter, happiness-joy, tenderness-feeling love and wonderment-feeling awe and sometimes for negative emotions like irritation-anger. And this sounds pretty reasonable as a positive feeling is more likely to propagate and generate positiveness. If the goal of a compassionate system is to increase the happiness and acceptance of its user, mirroring or mimicking emotions can be a successful strategy to boost

certain emotions or countering negative emotions to alleviate their effect.

VII. Challenges Encountered and Future Work

Through crowdsourcing, we gained better understanding of the challenges and venues to be looked at when dealing with such fine-grained emotion recognition task. In real world, more specific methodologies need to be developed to deal with ambiguity and mutiple emotions to satisfy a specific goal. Now if the goal of emotion recognition is to help conversational agent be aware of the emotionality of the speaker, learn from it and show a certain level of emotional intelligence (not plain responses that are not tailored to the emotional state of the person spoken to), then this agent can benefit from a system that can disambiguate the important emotional state to be dealt with. If we think of it as a psychotherapist who listens to the patient, people with psychological problems or stress usually confuse a lot of things and are full of emotions that they don't know how to express well. Yet, although it is a promising venue, it still cannot be achieved using the current state of natural language processing tools.

In future work, more work can be done to analyze syntactic dependencies and more towards supervised approach. We have mainly focused on adjectival modifier, adjectival complement, adverbial complement and negation modifiers but there exists many other types of modifiers such as intensifiers, dimishers, modality, interrogations and conditionality. We could also study and customize hand-coded rules depending on the emotion category

If we can detect fine-grained emotions with a high accuracy and analyze emotional responses patterns, the next step would be to think about a way of how to integrate those emotional cues into dialogue using current state of the art techniques for response generation. In Appendix A, we include a proposal of the architecture of a neural Seq2Seq system and explain how learned emotions further constrain the responses. This proposal is for interested readers and is open to further reflexion as although it has been thought of carefully, it has not been tested.

VIII. Acknowledgment

First of all, I would like to express my gratitude to my supervisor Professor Dr. Pearl Pu for her patient guidance, enthusiastic encouragement and immeasurable support. I also would like to thank to Dr. Valentina Sintsova for suggesting and explaining this project to me and for her very useful suggestions and kind support. My special thanks are extended to all people and friends who spent time and effort annotating the validation dataset.

References

[1] Felix Burkhardt, Markus van Ballegooy, Klaus-Peter Engelbrecht, Tim Polzehl, Joachim Stegmann. Emotion Detection in Dialog Systems: Applications, Strategies and Challenges. Deutsche Telekom Laboratories.

[2] Carlo Strapparava, and Alessandro Valitutti. "WordNet Affect: an Affective Extension of WordNet." LREC. Vol. 4. 2004.

[3] Scherer, K.R. (2005). What are emotions? And how should they be measured? Social Science Information, 44(4), 695-729

[4] Valentina Sintsova, Claudiu Musata and Pearl Pu. 2013. Fine-Grained Emotion Recognition in Olympic Tweets Based on Human Computation. In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 12-20. Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W13-1603.pdf

[5] Carlos Strapparava, and Rada Mihalcea. "Learning to identify emotions in text." Proceedings of the 2008 ACM symposium on Applied computing. ACM, 2008.

[6] Radim Burget, Jan Karasek, and Zdenek Smekal. "Recognition of emotions in Czech newspaper headlines." Radioengineering 20.1 (2011): 39-47.

[7] Jasy Liew Suet Yan and Howard R. Turtle. 2016. Exploring Fine-Grained Emotion Detection in Tweets. In Proceedings of NAACL-HLT, pages 73-80. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N/N16/N16-2011.pdf

[8] Ameeta Agrawal and Aijun An. 2012. Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations. Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, WI-IAT 2012, vol. 01, pp. 346353

[9] Hui Yang, et al. "A hybrid model for automatic emotion recognition in suicide notes." Biomedical informatics insights 5. Suppl 1 (2012): 17

[10] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. "Hierarchical versus flat classification of emotions in text." Proceedings of NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics, 2010.

[11] Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. Proceedings of IEEE International Conference onData Mining Workshop (ICDMW), pp. 383392.

[12] Valentina Sintsova, Claudiu Musat and Pearl Pu. 2014. Semi-Supervised Method for Multi-Category Emotion Recognition in Tweets. In IEEE International Conference on Data Mining Workshop. Retrieved from http://sentic.net/sentire2014sintsova.pdf

[13] H. Holzapfel, C. Fuegen, M. Denecke and A. Waibel, "Integrating emotional cues into a framework for dialogue management," Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, 2002, pp. 141-146. doi: 10.1109/ICMI.2002.1166983

[14] Burkhardt, F., van Ballegooy, M., Engelbrecht, K.-P., Polzehl, T., Stegmann, J., 2009. Emotion detection in dialog systems: applications, strategies and challenges. In: Proc. ACII, Amsterdam, Netherlands, pp. 16.

[15] Hasegawa, Takayuki, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. Predicting and Eliciting Addressees Emotion in Online Dialogue. In ACL (1), 96472, 2013. http://www.anthology.aclweb.org/P/P13/P13-1095.pdf

[16] Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Dont count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR

	Emotion Category	Probability
20	Neutral	24.583762
0	Pity-Compassion	9.715158
19	Feeling Disburdened- Relief	9.369369
10	Irritation-Anger	9.141104
13	Amusement-Laughter	9.025033
15	Happiness-Joy	8.835341
17	Tenderness-Feeling Love	7.203390
16	Wonderment-Feeling Awe	5.210918
9	Disgust-Repulsion	5.124451
11	Involvement-Interest	5.111821
7	Disappointment- Regret	5.035971
14	Enjoyment-Pleasure	4.818092

(h)

12	Pride-Elation	4.761905
8	Contempt-Scorn	4.084720
4	Guilt-Remorse	4.013378
5	Embarrassement-Shame	3.909465
1	Longing- Nostalgia	3.470716
2	Worry-Fear	3.270224
6	Envy-Jealousy	2.409639
3	Sadness-Despair	1.060071
18	Astonishment- Surprise	0.000000

(i)

Fig. 7. Emotion Immitation Patterns Percentages per each emotion

Workshop, 2013.

APPENDIX A
PROPOSAL FOR EMBEDDING EMOTIONS IN DIALOGUE SYSTEMS USING NEURAL SEQ2SEQ

The following proposal is for integrating emotions into the architecture of Neural Seq2Seq chatbot:

- We can associate the space of responses for each emotion (keywords, representative terms) with a vector representation. This vector can be initialized randomly but will be learned throughout the training process by back-propagating word prediction errors to each neural component. Based on that, we build a matrix M(i,j) in which i is the emotion in the context and j is the emotion in the response.
- We incorporate the Matrix in the LSTM model by simply adding it as a component in the model. More specifically, we add it as a component to calculate the output of the hidden layer at each time stamp which is so far only impacted by input token and previous output of hidden layer.
- This system can give different candidates for each message depending on the emotion detected. It is now the role of the emotional response model to act as a filter that will define which pattern maximizes the possibility of making the whole system behave as a compassionate system that feels the emotion expressed and respond showing an adequate emotion. The pattern could be imitating emotions towards building more positive emotions if we want a happy dialogue system or any other emotional response model depending on the goal of the dialogue system.