Fine Grained Emotion Recognition in Human Dialogs

Meryem M'hamdi, Dr. Valentina Sinstova and Dr. Pearl Pu

School of Computer And Communication Sciences, Swiss Federal Institute of Technology (EPFL)

Email: {meryem.mhamdi, pearl.pu, valentina.sintsova} @epfl.ch

Abstract

Embedding emotional intelligence into the way we build human-machine dialog systems will likely boost user acceptance. Although there is a growing interest in exploring this line of research, fewer work have examined ***fine-grained*** emotion recognition in dialog systems. In this paper, we propose a novel method for recognizing emotions up to 20 categories using the Geneva Emotion Wheel. Our method does not require manual labeling. Instead it uses a small emotion lexicon, and incrementally builds several classifiers using PMI, Word2Vec, and Naive Bayes semi-supervised approach. To further improve the performance and treat imbalances in emotion classes, we then aggregate results from all three classifiers using a set of carefully designed voting rules. We describe offline evaluations as well as an online evaluation with real users. Results show that our method outperforms all other methodologies with an average micro f1-score of 34.0%, which is significant for a 20-category multi-class classifier. This research demonstrates the importance of fine-grained emotion detection in dialog systems and more importantly the feasibility of such an approach by employing the expertise of each classifier in a more elaborate voting model.

1 Introduction

When humans communicate to each other, we pay much attention to adapting our responses to the emotional state of the person with whom we converse. Through our dialogues, we not only communicate ideas but more importantly our emotions. Being emotionally tuned is likely to determine the effectiveness of our interactions and define its success or failure. If understanding and expressing emotions are crucial components to human dialogs, automatic dialog systems need to do the same in order to create more human-like conversations, thus improving user acceptance. Despite the importance of emotion awareness in human dialogues, their artificial counterparts still lack the basic ingredients for appropriately processing emotional signals and for demonstrating a satisfying degree of emotional intelligence. One exception is the recent work by Zhou et al. 2017 where the research aim was to build emotion-aware dialog chatbots. However, using a model of five most frequent emotions (anger, disgust, happiness, like and sadness) is limited although it is a very important first step. More effort is needed towards incorporating more fine-grained emotional states in its modeling. Consider for example how a chatbot may offer re-conciliation in its responses when disappointment is detected in human speakers, how to congratulate them when pride is detected, and how to express compassion when the chatbot needs to show sympathy. Notice all these three emotions are missing in the used model. Ideally a truly emotional conversation agent is capable of detecting and distinguishing not only basic emotions but also complex ones. Such an approach is likely to capture and distinguish the layered subtleties and nuances of human emotions, especially in the way they may communicate to robots in the future. The challenge for machine learning tool design is thus related to developing a high quality multi-class classifier. This task becomes increasingly difficult as more categories of emotions are considered.

Previous work related to domain specific ***fine-grained*** emotion detection in social media text using semi-supervised learning classifiers showed an average F1 score around 20% (Sinstova et al. 2014). One possible limitation of that approach is that the results may depend on the domain on which the classifiers are trained. Furthermore, we believe the current task (ER) is more difficult in dialog text where the corpus is completely unannotated. For these reasons, we adopt a hybrid methodology that aggregates results from three types of classifiers. We start by employing a different feature selection using various natural language processing techniques to refine affective representation of input, and combine it with a lexicon based approach strengthened with syntactic rules. However, this lexicon-based approach can suffer from weak recalls in case emotional keywords don't yet exist in the lexicon. Further, domain-specific lexicon and knowledge rules are required to achieve better accuracy. We therefore extend their performance using distributional semantic methodologies. Semantic similarity scores trained using word embedding are used to select other candidate words for carrying emotional content based on their similarity with lexicon keywords. Secondly, we developed a semi-supervised classifier trained on twitter data and adapted to our dialog data using transfer learning. Thirdly, we defined a voting model that re-assigns emotion labels to reduce the inconsistencies in classification of individual emotion categories.

This work is the first of its kind to explore fine grained emotion recognition with 20 emotions and advance methodologies for its application to the context of dialog systems and investigate their performance by comparing a lexicon-based approach to a semi-supervised approach and integrating them into a voting model to show the strengths and weaknesses of each approach both holistically and on a fine-grained level.

2 Related Work

Emotion Recognition

Text-based emotion recognition methodologies can be approximately categorized into the following types: corpus-based, knowledge based, lexical affinity methods, lexicon-based, rule-based and statistical approaches. In this section, we focus on most recent work that use unsupervised and hybrid methodologies to emotion recognition.

Agrawal and An (2012) developed an unsupervised approach where the system first performs data pre-processing and part of speech tagging to extract affective words. It then 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. It works well in general. However, one weakness of this approach is that the accuracy of the semantic relatedness between words largely depends on the size and domain of training dataset. Yang et al. (2012) combined lexicon-based, conditional random field-based and machine learning-based emotion classification using SVM, Naïve Bayesian and Max Entropy. Then, they use a voting model to combine the results from different classifiers. Using a model of 6 categories of emotions, they have achieved an F1-score of 61%, which outperforms the previous methodology (58%).

Another study followed a hierarchical methodology which defined objectives at different levels of granularity at each layer of the hierarchy: at the first level, the objective was separating neutral from emotional sentences, then polarity classification and finally they moved to more fine-grained classification (Ghazi, Inkpen, and Szpakowicz 2010). Different features were used for each classification stage. However, this method did not take into consideration the dependencies between words. Shaheen et al. (2014) 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 (ERR) 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%, the highest of all mentioned approaches.

Sinstova et al. (2014) presented a novel framework for fine-grained emotion classification using distant supervision. It was the first time a 20-class classifier was attempted. There are three main elements in this framework. It shows how to build a domain-specific classifier with little preconditions. Since the seed lexicon exists in multiple languages, including French, German, and Chinese, its applicability is very large and researchers can use it for comparing results across languages. It can distinguish emotions up to 20 categories plus neutral. This approach has been applied to sport tweets and was compared two machine learning classifiers (Naive Bayes and pointwise-mutual information classifier) combined with three initial classifiers. In all cases, the proposed new method outperformed the other ML classifiers for all emotion categories with high F1-scores, with less difference between precision and recall, an overall macro F1-score of 20,5%.

Emotion 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 an overview of some fine-grained emotion recognition techniques developed to generate emotion-aware responses. In particular, we describe data-driven methods challenges and applications.

Holzapfel et al. (2002) proposed 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 whether patterns linking emotions to dialogue state can be learned automatically or is to be hand-coded.

Burkhardt et al. (2009) 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 al. (2013) build and evaluate a dialog system in which the emotion of the response can be controlled to trigger a goal emotion in the addressee. It is accomplished in two steps: predicting the emotion of an addressee based on the context and generating responses that elicit the goal emotion. 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 it, 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. Zhou et al. (2017) aim at accomplishing the same goal of embedding emotion factor in dialog system using an end-to-end framework to incorporate emotion influence in large-scale sequence-to-sequence response generation. Their proposed emotional chatting machine uses recurrent neural network of GRU cells with attention mechanism and given the desired emotion category in the response it comes with an emotional response using three mechanisms: emotion category embedding, internal emotion memory, and external memory. However, this work lacks an emotional model for deciding the most appropriate emotion categories for the response based on the emotion interaction patterns.

3 Dataset

In our approach, we use two sets of data. The first one consists of non-annotated TV series transcripts from the Big Bang Theory and Friends. It contains a total of 111,103 dialogue utterances. In addition to that, we use Cornell Movie-Dialogs Corpus, 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 overall dataset contains 412,826 dialogue turns made up of a vocabulary size of 4,209,560. For our semi-supervised approach, we use tweets posted by spectators of 2012 Olympic games, which is the same dataset that has been treated and annotated in Sinstova et al. 2014. It consists of 2,348,176 tweets annotated using 20 emotions. Its distribution is described in table 1.

Positive	**%**	**Negative**	**%**
Involvement	0.86	Pity	0.43
Amusement	0.71	Sadness	10.63
Pride	6.6	Worry	6.62
Happiness	19.12	Embarrassment	2.2
Pleasure	0.73	Guilt	0.95
Love	29.26	Disappointment	3.37
Awe	0.68	Envy	2.6
Relief	1.2	Disgust	1.11
Surprise	1.12	Contempt	0.05
Nostalgia	1.42	Irritation	9.97

Table 1: Tweets Emotion Distribution.

4 Fine-Grained Emotion Recognition

Figure 1 gives a general overview of the following phases:

Preliminary Data Pre-Processing

We start by performing some preliminary pre-processing on the extracted transcripts including tokenization using Stanford PTBTokenizer, which relies on an efficient, fast and deterministic Finite State Automaton implementation. Given as input the set of regular expressions and corresponding regex operations, it can read the input and run the corresponding operations defined for the regex that matches the input giving the longest match possible if it exists. Since scene metadata draws on the general context without which it is difficult to understand the meaning of the transcript, we incorporate this information as context cues by collating words from scene metadata to transcripts.

Affective Feature Extraction & Data Cleaning

To analyze the syntactic structure of the sentences, we start by performing Part of Speech Tagging (POS) using Stanford POS tagger, which assigns tags from universal tagset. Based on the obtained tags, we keep only NAVA words i.e. words tagged as nouns, verbs, adjectives or adverbs because they are believed to be better candidates of features carrying emotion content compared to other tags (pronouns, articles, etc) which are not relevant in this ER task. After tagging, we construct dependency trees using Stanford Dependency Parser, which reaches the state-of-the-art in terms of accuracy. We observe that the use of Named Entity Recognition using Stanford NER Tagger increases the accuracy by helping filter out objective content such as names of people, monetary values, locations, etc. Then, we convert to lowercase and remove punctuation. In order to normalize the terms and get one common representation of multiple versions of the same word, we perform lemmatization using NLTK WordNet Lemmatizer instead of stemming since it cuts down words. We come up with our customized set of stopwords, which excludes modifiers needed in our syntactic analysis.

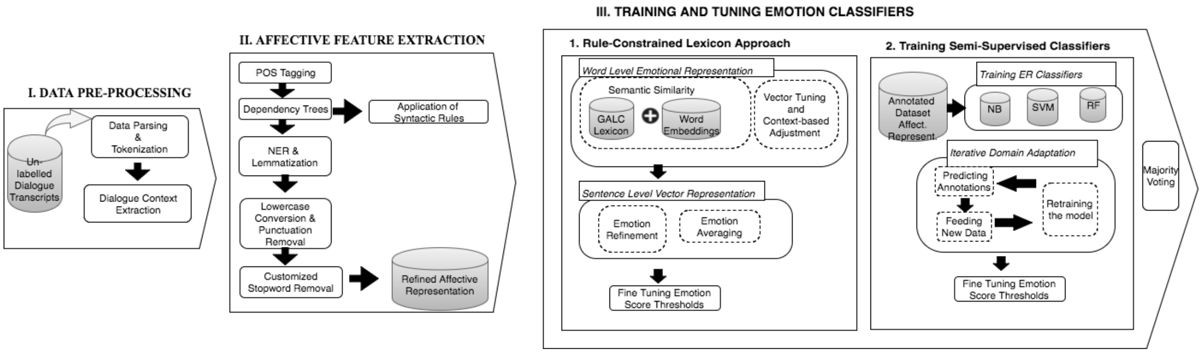
I. DATA PRE-PROCESSING
Un-labelled Dialogue Transcripts
Data Parsing & Tokenization
Dialogue Context Extraction
II. AFFECTIVE FEATURE EXTRACTION
POS Tagging
Dependency Trees
Application of Syntactic Rules
NER & Lemmatization
Lowercase Conversion & Punctuation Removal
Customized Stopword Removal
Refined Affective Representation
III. TRAINING AND TUNING EMOTION CLASSIFIERS
1. Rule-Constrained Lexicon Approach
Word Level Emotional Representation
Semantic Similarity
GALC Lexicon
Word Embeddings
Vector Tuning and Context-based Adjustment
Sentence Level Vector Representation
Emotion Refinement
Emotion Averaging
Fine Tuning Emotion Score Thresholds
2. Training Semi-Supervised Classifiers
Annotated Dataset Affect. Represent.
Training ER Classifiers
NB
SVM
RF
Iterative Domain Adaptation
Predicting Annotations
Feeding New Data
Retraining the model
Fine Tuning Emotion Score Thresholds
Majority Voting

Figure 1: An Overview to Emotion Recognition Framework

Word Level Emotional Representation

Let s be a sentence and E be in the set of m emotion labels excluding neutral $E = \{e_i | i \in [1, m]\}$ which is GEW 20 categories. Let $R = \{r_j | j \in [1, k]\}$ be the set of representative words for an emotion e_i. 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.

Semantic Similarity

In order to come up with semantic scores, we follow a hierarchical methodology where we start by keyword spotting using lexicon, if not found, we search for closest match by computing the semantic similarity and trying both a count-based approach (such as PMI) and predictive approach (word2vec) between words in the lexicon and words in the sentence:

- Keyword Spotting: We follow Geneva Affect Label (GALC), a lexicon which enumerates a set of representative words for each one of its 20 emotion categories. We use the extended version of GALC, which replaces stemmed terms with words instantiated using Twitter data (Valentina, Musata, and Pu 2014). It consists of 1026 terms with 52.9 on average per emotion.
- Pointwise Mutual Information: Pointwise Mutual Information (PMI) is a statistical measure based on co-occurrence. We followed the assumption that two words co-occurring together with a frequency exceeding a certain threshold are more likely to exhibit a higher semantic similarity. PMI similarity between two words $w_1$and w_2 can be calculated as follows:

$$PMI(w_1, w_2) = co(w_1, w_2)/occ(w_1) * occ(w_2) \quad (1)$$

where occ (w_1) is the number of times $w_1$occurs in the dataset, co(w_1, w_2) is the number of times w_1 and w_2 co-occur within a window of size 10.

- Word2Vec Model: Word2Vec is a predictive model trained using 2 layer neural networks which learns word embeddings from raw text. We obtain already trained Continuous Bag of Words Model (CBOW) 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. We prefer CBOW over Skip-Gram since it smoothes over distributional statistics by turning context data into observation without needing data as large.

Both PMI and word2vec were trained on the 1 Billion words Benchmark corpus (Chelba et al. 2013). We use those obtained similarities to compute the emotional vector: $\sigma(w_i) = [SIM(w_i, e_1) \dots SIM(w_i, e_m)]$ (2) where each element of the vector is an emotional score equal to the geometric mean of semantic similarities between the word and each representative term in the lexicon.

Emotion Refinement

We determine syntactic dependencies between words relevant to the context of emotion detection to refine the vectors of dependent words as defined by these 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, instead, we averaged the two words out, unfortunate having a sadness emotion and luck having happiness emotion would lead to a neutral emotion, whereas taking into account 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 the adjective. For example, the emotion of "I feel depressed" is sadness (emotion of depressed).
- Adverbial Modifier: when a verb is followed by an adverb complement, the emotion of the verb depends on the adjective as in sentence "I struggled happily".
- Nominal, adjectival or Verbal Negation Modifier: when a noun or an adjective (or a verb) is preceded (or followed) by a negation modifier, for example, "This is not funny" the emotion of the word is neutralized.

Sentence Level Vector Representation

After transforming word features into vectors of 20 elements and applying syntactic rules, we calculate the dominant emotion of the sentence by averaging over the emotional vectors of the words that make up the sentence as shown in the formula below:

$$\sigma(s) = [\frac{\sum_{i=1}^{n} \sigma_1(w_i)}{n} \dots \frac{\sum_{i=1}^{n} \sigma_m(w_i)}{n}] \qquad (3)$$

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

Semi-Supervised Approach

In this approach, we apply the same pre-processing and syntactic procedure on the annotated dataset in addition to some minor data cleaning such as removing usernames and replacing emoticons. Using CBOW word2vec model, we get the best results according to validation experiment by representing each word using 300 dimensional vectors using context of 10, minimal word occurrence of 4. Then, we average over the vector representations to get the sentence vector representation. Then, we randomly split the annotated dataset into training and testing dataset and train and compare different machine learning algorithms including SVM with different kernels, K-Nearest Neighbors and Random Forest using Scikit-learn implementation.

We split annotated into 70% for training and 30% for testing. Since we train them on annotated dataset (source) and apply them to predict labels for dataset belonging to another domain (target), we use domain adaptation method using a simple iterative auto-labeling approach which is better suited to our case as it doesn't need any target instances already labeled to increase the capability of machine learning model trained on one dataset to generalize well on dataset from another domain. In a first step, an initial model M is learned based on only source labeled dataset. Then, the model M is used to label a small subset of the target dataset. The newly labeled target instances are used to retrain another model M'. The effectiveness of this approach 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 process. In the results section, we compare the performance of the different algorithms and shows that Naive Bayes performs the best with increased ability to generalize on test data.

Voting Model between Hybrid Based and Semi-Supervised

After evaluating our different methodologies, we realize that no one stands out as a strong classifier as they don't perform consistently across all emotion categories. Some emotion categories can be better learned by some classifiers, which leads us to think that each classifier has its strengths and weaknesses. In this fine-grained emotion recognition problem, it is normal not to have classifiers which can learn equally on all emotions. Some of them lack information to distinguish between one emotion and another, others are more biased towards dominant emotions. Inspired by the improved performance of the different voting strategies employed in Yang et al. (2012), we tested a majority voting approach customized towards a better collaboration between the different methodologies in order to increase both the overall recall of emotional instances and the fine-grained recall per each emotion.

Given the emotion labels obtained by the three classifiers (PMI, word2vec and semi-supervised), we take the majority when it is possible, otherwise we favor the classifier with the best performance overall. In case of total disagreement, we take into consideration the ranking of the classifiers to impact the importance in the voting process:

- If the emotion category returned by PMI is not neutral and word2vec and semi-supervised agree on an emotion different from the one returned by PMI, return the emotion agreed by word2vec and semi-supervised
- If all three classifiers return an emotion other than neutral but they don't agree, return the emotion of PMI
- If PMI returns neutral then check word2vec if it returns also neutral, just return semi-supervised.

Validation Approach

In order to evaluate the performance of our ER methodologies, we design a crowdsourcing experiment on a random sample of the dataset. We start by separating transcripts classified as non-emotional from emotional ones and filter out texts with less than 30 characters to increase the chance of selecting subtitles containing emotional contents. We take into consideration the classification imbalances and skews 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 70% sentences classified as emotion B, then we need to take 300 random sentences from emotion A, 700 random sentences from emotion B. To reduce labeling by chance and mistakes, we ask annotators to justify their annotations by supplying textual indicators in the subtitle and how else they could have expressed it. This approach not only gained us better annotation accuracy but we also crowd-sourced domain-specific textual features that could be used to extend the lexicon.

We aim for annotating a total of 250 subtitles by 12 annotators. Table 3 shows the distribution of emotions in the validation sample. For the purpose of collecting responses from people, we have created 20 Google forms with two versions: shorter version with 10 subtitles, and a longer one with 15 to ease the workload for some annotators who can annotate in two steps depending on time. We have asked for volunteers among friends and people who work in the laboratory and have given them a description of the emotion model used and the general scope of the Emotion Recognition task. To avoid giving more or less advantage to humans annotators versus our ER system, we provided the subtitles in the same textual form as they were provided, i.e: 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. For e.g., 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 ask to select the most recent and dominant one.

Positive	#	**Negative**	#
Involvement	21	Pity	6
Amusement	16	Sadness	17
Pride	4	Worry	26
Happiness	5	Embarrassment	1
Pleasure	7	Guilt	8
Love	28	Disappointment	9
Awe	12	Envy	2
Relief	7	Disgust	8
Surprise	2	Contempt	15
Nostalgia	9	Irritation	15
Neutral	28	Other	4

Table 3: Number of dialogs validated by human raters in each emotion category

5 Results and Analysis

To evaluate the performance of multi-emotion classifiers, we use micro and macro F1-scores to both assess how they perform on average and to take into consideration class imbalances. Furthermore, we also use area under curve (AUC), which takes into consideration the imbalances in the number of instances for each class.

Based on both micro-averaged and macro-averaged F1-scores shown in table 4, the approach that yields the best performance is the rule-based GALC extended with PMI (**micro-f1=34.0%, macro-f1=25.4%, AUC=60.2%**). This reflects a significant increase over the **baseline Initial GALC-R** of **9.2%, 8.4% and 12.9%** in micro-f1, macro-f1 and AUC respectively. Higher recall of GALC-Plus 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. This also proves the importance of the use of rule-based syntactic approach and natural language processing techniques for cleaning and analyzing the affective weight of certain parts of the sentence, which enriches the methodology and can create a difference into any weakly supervised approach. This also shows the role of applying pre-processing to get rid of noise in the data and extract only relevant features.

As for **voting method**, table 5 shows that on a fine-grained level, it outperforms other approaches in classifying positive and negative emotions with average f1-scores of 27.4% and 25.0% respectively, which is an **increase of 0.1% for positive emotions** and a **slight improvement of 3.0%** over GALC-Plus PMI. However, against our expectations, it ranks second with a macro-f1 score of 21.2%, it doesn't improve the overall performance and this can be due to its aim to increase the recall of emotional instances, which affects neutrality accuracy. In the third place 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 voting model and GALC-Extended with PMI.

After trying different semi-supervised algorithms with different parameters and using cross-validation over different subsets of training and testing datasets, we observe that although some algorithms did better in training and testing dataset, they generalized poorly even with the use of domain adaptation. As shown in table 6, with random forest we get a high f1-score of 70% on both training and testing datasets, however on the validation dataset, it doesn't perform as good reaching only a score of 4.9% which tells us that this algorithm over-fitted on validation dataset although it performs evenly good on testing dataset chosen randomly. On the other hand, Naïve 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 use to compare with our other approaches.

This semi-supervised approach with 10 iterations used in domain adaptation achieves on average the lowest performance before the baseline with an AUC of 54.5%, 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. However, for some emotion classes, it still has the best f1-score for e.g. relief (22.2%) and envy (28.6%) for which all other classifiers gave a f1-score of 0.0% as shown in table 6. 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 a better approach that takes into consideration the expertise of each classifier to assign weights in a more elaborate voting model.

Methodology	**Micro**	**Macro**			**AUC**	**Rank**
	F1	**P**	**R**	**F1**		
Initial *GALC-R*	24.8	23.3	20.3	17.0	47.3	5
GALC-Plus PMI	**34.0**	**24.7**	**26.3**	**25.4**	**60.2**	**1**
GALC-Plus Word2Vec	32.0	21.8	25.6	20.6	57.4	3
Semi-Super NB	12.4	13.3	14.4	10.4	54.5	4
Voting Model	33.2	22.8	26.4	21.2	56.5	2

Table 4: Performance Comparison of Different Methodologies

Positive	**GALC**	**PMI**	**W2V**	**NB**	**VM**	**Negative**	**GALC**	**PMI**	**W2V**	**NB**	**VM**
Involvement	33.3	**56.2**	54.5	6.6	**56.2**	*Pity*	0.0	22.2	**25.0**	0.0	**25.0**
Amusement	30.8	**61.5**	**61.5**	21.4	59.9	*Sadness*	16.7	**34.5**	27.6	0.0	**34.5**
Pride	0.0	0.0	0.0	0.0	0.0	*Worry*	44.0	54.5	53.7	23.5	**56.2**
Happiness	18.2	32.0	**34.8**	0.0	32.0	*Embarrassment*	33.3	33.3	**40.0**	0.0	33.3
Pleasure	16.7	18.2	17.4	**22.2**	17.4	*Guilt*	40.0	30.8	**46.2**	19.0	**46.2**
Love	52.2	53.7	52.9	26.6	**55.1**	*Disappointment*	0.0	0.0	0.0	0.0	0.0
Awe	**41.2**	33.3	25.0	8.3	35.3	*Envy*	0.0	0.0	0.0	**28.6**	0.0
Relief	0.0	0.0	0.0	**22.2**	0.0	*Disgust*	22.2	**30.8**	26.7	12.2	26.7
Surprise	**33.3**	0.0	0.0	28.6	0.0	*Contempt*	0.0	0.0	0.0	0.0	0.0
Nostalgia	0.0	**18.2**	17.4	0.0	**18.2**	*Irritation*	16.0	13.8	12.9	**20.7**	13.8
Average	22.6	27.3	26.4	13.6	**27.4**	*Average*	17.2	22.0	23.2	10.4	**25.0**

Table 5: F1 scores per emotion category for each methodology

	Training (Labelled Only)				**Testing**				**Validation**			
	P	R	F1	AUC	P	R	F1	AUC	P	R	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**

Table 6: Summary of Performance of Machine Learning Algorithms

6 Conclusion and Future Work

This paper explores fine-grained emotion recognition and devises techniques for its application to the domain of human dialogues. We refined the affective representation of subtitles through the use of various techniques such as POS tagging, Named entity recognition and lemmatization. We combined both lexicon-based, syntactic rule based, semantic analysis using both count-based (PMI) and predictive word embedding (word2vec) and semi-supervised machine learning approach to leverage the capability of our model to detect emotions with better recall and precision. We managed to improve the accuracy of our predictions with an **average macro-f1 of 34.0%**, which is an **improvement of 8.4%** over our baseline method applied on the same dataset. We got the best results using GALC-Extended PMI on average. The reason why the combination of rule-based and lexicon-based with PMI performs the best is that people not only use explicit keywords to express an emotion but use implicit expressions that can only be understood by referring back to the context. However, higher results on fine-grained recall of individual emotions using **voting model**. Therefore, we conclude that all tried methodologies can work in a collaborative manner to increase the recall of certain emotions as each one has strength to detect certain emotions that others sometimes cannot detect in a satisfying way. This motivates us to consider this kind of aggregation models and to investigate other possibilities to improve them in future work.

More future work can also be done to analyze syntactic dependencies, as there exist other types of modifiers such as intensifiers, diminishers, modality and conditionality worth investigating. We could also customize more hand-coded rules depending on the emotion category. Through crowdsourcing, we gained better understanding of the challenges and venues to be looked at when dealing with such fine-grained emotion recognition task. One of those challenges is the existence of multiple dominant emotions, which may 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 family. If we can detect fine-grained emotions with a high accuracy, 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 response generation techniques. It remains to think of how to embed emotions into current chatbots to constrain responses.

7 References

Agrawal, A.; and An, A. 2012. Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT),* 346–353.

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 238–247.

Burget, R.; Karasek, J.; and Smekal, Z. 2011. Recognition of emotions in Czech newspaper headlines. *Radioengineering* 20(1):39-47.

Burkhardt, F.; Ballegooy, M. V.; Engelbrecht, K. P.; Polzehl, T.; and Stegmann, J. 2009. Emotion Detection in Dialog Systems: Applications, Strategies and Challenges. Deutsche Telekom Laboratories. In the *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, 684–689.

Chelba, C.; Mikolov, T.; Schuster, M.; Ge, Q.; Brants, T.; Koehn, P.; Robinson, T. 2013. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. arXiv:1312.3005.

Danescu-Niculescu-Mizil, C. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (ACL)*.

Ghazi, D.; Inkpen, D.; and Szpakowicz, S. 2010. Hierarchical versus flat classification of emotions in text. In *Proceedings of NAACL HLT workshop on computational approaches to analysis and generation of emotion in text*, 140-146. Association for Computational Linguistics.

Holzapfel, H.; Fuegen, C.; Denecke, M.; and Waibel, A. 2002. Integrating emotional cues into a framework for dialogue management. In *Proceedings of Fourth IEEE International Conference on Multimodal Interfaces*, 141-146.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop*. arXiv:1301.3781.

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

Shaheen, S.; El-Hajj, W.; Hajj, H.; and Elbassuoni, S. 2014. Emotion recognition from text based on automatically generated rules. In *Proceedings of IEEE International Conference on Data Mining Workshop (ICDMW)*, 383–392.

Sintsova, V.; Musat, C.; and Pu, P. 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*, 12-20. Association for Computational Linguistics.

Sintsova, V.; Musat, C.; and Pu, P. 2014. Semi-Supervised Method for Multi-Category Emotion Recognition in Tweets. In *IEEE International Conference on Data Mining Workshop*, 393-402.

Strapparava, C.; and Mihalcea, R. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing (ACM)*, 1556-1560.

Strapparava, C.; and Valitutti, A. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC),* 1083-1086.

Takayuki, H.; Kaji, N.; Yoshinaga, N.; and Toyoda, M. 2013. Predicting and Eliciting Addressee's Emotion in Online Dialogue. In *Proceedings of the 51st Annual Meeting of Association for Computational Linguistics (ACL)*, 964–72.

Yang, H.; Willis, A.; Roeck, A.; Nuseibeh, B. 2012. A hybrid model for automatic emotion recognition in suicide notes. In *Biomedical informatics insights* 5(1): 17-30.

Yan, J. L. S.; and Turtle, H. R. 2016. Exploring Fine-Grained Emotion Detection in Tweets. In *Proceedings of NAACL-HLT*, 73-80. Association for Computational Linguistics.

Zhou, H.; Huang, M.; Zhang,T.; Zhu, X.; and Bing, L. 2017. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. arXiv:1704.01074.