# EmotionX-Area66: Predicting Emotions in Dialogues using Hierarchical Attention Network with Sequence Labeling

**Rohit Saxena**
TCS Research
rohit.saxena2@tcs.com

**Savita Bhat**
TCS Research
savita.bhat@tcs.com

**Niranjan Pedanekar**
TCS Research
n.pedanekar@tcs.com

## Abstract

This paper presents our system submitted to the EmotionX challenge. It is an emotion detection task on dialogues in the EmotionLines dataset. We formulate this as a hierarchical network where network learns data representation at both utterance level and dialogue level. Our model is inspired by Hierarchical Attention network (HAN) and uses pre-trained word embeddings as features. We formulate emotion detection in dialogues as a sequence labeling problem to capture the dependencies among labels. We report the performance accuracy for four emotions (*anger, joy, neutral* and *sadness*). The model achieved unweighted accuracy of 55.38% on *Friends* test dataset and 56.73% on *EmotionPush* test dataset. We report an improvement of 22.51% in *Friends* dataset and 36.04% in *EmotionPush* dataset over baseline results.

## 1 Introduction

Emotion detection and classification constitutes a significant part of research in the area of natural language processing (NLP). The research aims to detect presence of an emotion in a text snippet and correctly categorize the same. The emotions are typically classified using categories proposed by (Ekman et al., 1987), namely *anger, disgust, fear, joy, sadness, surprise*. Significant amount of research has been dedicated to emotion classification in variety of texts like news and news headlines (Strapparava and Mihalcea, 2008; Staiano and Guerini, 2014), blogposts (Mishne, 2005), fiction (Mohammad, 2012b).

With the advent of social media and dialogue systems like personal assistants and chatbots, emotion analysis of short texts has garnered a lot of attention. Short texts are defined as small text chunks in the form of tweets, messenger conversations, social network posts, conversational dialogues etc. Unlike large documents, these texts have unique set of characteristics such as informal language, incomplete sentences, use of emoticons. Different approaches for emotion detection in short texts are proposed in (Krcadinac et al., 2013) for instant messages, (Mohammad, 2012a) and (Wang et al., 2012) for *Twitter* and (Preotiuc-Pietro et al., 2016) for status updates in *Facebook*.

| **Speaker** | **Utterance** | **Emotion** |
|---|---|---|
| Joey | Whoa-whoa, Treeger made you cry? | surprise |
| Rachel | Yes! And he said really mean things that were only partly true. | sadness |
| Joey | I'm gonna go down there and teach that guy a lesson. | anger |
| Monica | Joey, please don't do that. I think it's best that we just forget about it. | fear |
| Rachel | That's easy for you to say, you weren't almost just killed. | anger |
| Joey | All right that's it, school is in session! | neutral |

Table 1: Example of a dialogue from *Friends* dataset

Conversational short texts consist of dialogues between two or more entities. A dialogue naturally has a hierarchical structure, with words contributing to an utterance and a set of utterances contributing to a dialogue (Kumar et al., 2017). Table 1 shows an example of a dialogue which consists of 6 utterances with corresponding speakers

*Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* , pages 50–55,
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

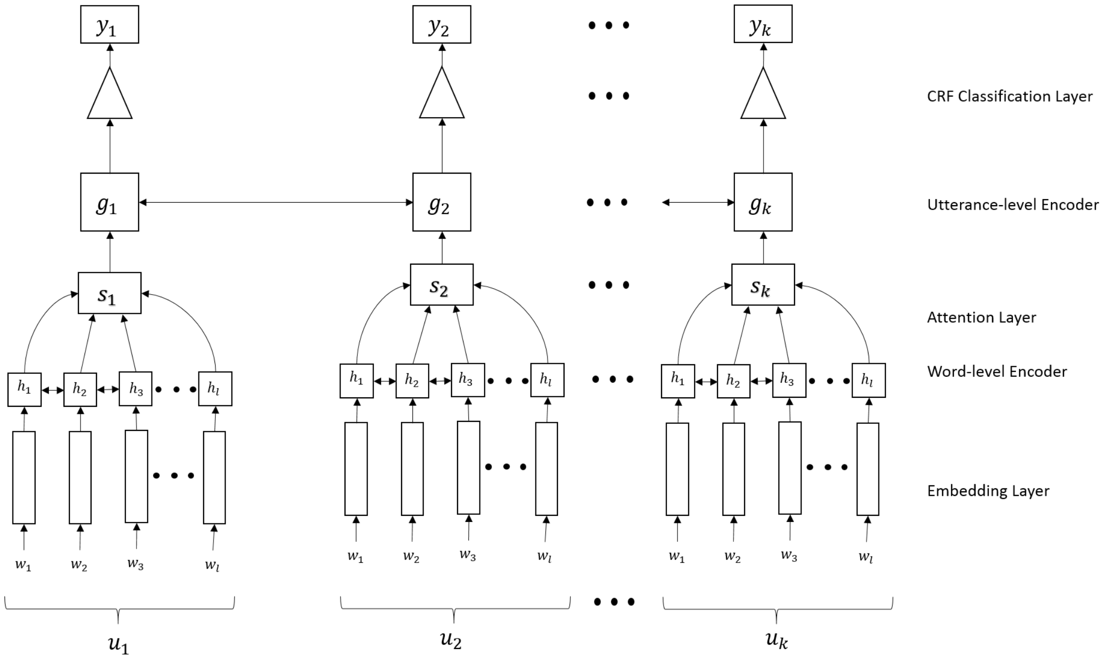

Figure 1: An illustration of proposed Hierarchical Attention Network

and emotions. In these dialogues, context builds as the dialogue progresses. There is a dependency between consecutive utterances and hence the classification of such utterances can be treated as a sequence labeling problem. In particular, (Stolke et al., 2000; Venkataraman et al., 2003) and (Kim et al., 2010; Chen and Eugenio, 2013; Kumar et al., 2017) have captured dependencies in utterances for dialogue act classification using Hidden Markov Model (HMM) and Conditional Random Field (CRF) respectively. Also, several ways of incorporating such context information in artificial neural networks have been proposed in (Liu, 2017).

The EmotionX shared task consists of detecting emotions for each utterance from EmotionLines dataset. The dataset (Chen et al., 2018) contains dialogues collected from *Friends* TV show scripts and private *Facebook* messenger chats. Each of the utterances has been annotated for one of the eight emotions viz. six basic emotions proposed by (Ekman et al., 1987) and two other emotions viz. *neutral, non-neutral*. The shared task focuses on detecting only four of these eight emotions, namely *joy, sadness, anger* and *neutral*. In this paper, we present our approach to detect emotions in utterances. Inspired by (Kumar et al., 2017), we use Hierarchical Attention Network (HAN) to build context both at utterance and dialogue level. We treat emotion detection at utterance level as a sequence labeling problem and use a linear chain CRF as a classifier.

## 2 Proposed Model

The dataset for the task consists of dialogues, each dialogue ($D^i$) consists of sequence of utterances denoted as $D^i = (u_1, u_2, \dots u_n)$, where $n$ is the number of utterances in a given dialogue. Each utterance $u_j$ is associated with a target emotion label $y_j \in \mathcal{Y}$. To build context within a dialogue, we consider a moving context window $N_k$ of length $k$ and combine all the utterances within the window with their target labels to create multiple sets of context utterances. These sets of utterances are given as input to our model.

The model consists of HAN (Yang et al., 2016), where the first part is a word-level encoder with the attention layer, encoding each word in an utterance. The second part is an utterance-level encoder, encoding each utterance in the dialogue. The HAN is combined with a linear chain CRF classification layer for detecting emotions. The utterance level emotion detection is treated as a sequence labeling problem based on the fact that the emotion in an utterance depends on emotions of previous utterances. An illustration of complete model comprising of embedding layer, word level encoder, attention layer , utterance level encoder with final layer of CRF classification is depicted in Figure 1.

## 3 Model Desscription

**Embedding Layer:** A context window $N_k$ consists of $k$ utterances each having $l$ number of words. Each word $w_{ij}$ in an utterance $u_j$, where $j \in [1, k]$ , is embedded to a low-dimensional vector space $R^d$ using an embedding layer ($f_{embed}$) of size $d$. It projects the word into representative word vector $x_{ij}$. We initialize the weights of the embedding layer with pre-trained GloVe embeddings[1].

$$x_{ij} = f_{embed}(w_{ij})$$

**Word-level Encoder:** We use a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014) as the word-level encoder in the hierarchical network to summarize information from both directions for words. The bidirectional GRU contains the forward GRU which reads the utterance $u_j$ from $w_{1j}$ to $w_{lj}$ and a backward GRU which reads from $w_{lj}$ to $w_{1j}$:

$$\overrightarrow{h}_{ij} = \overrightarrow{GRU}(x_{ij}), i \in [1, l]$$

$$\overleftarrow{h}_{ij} = \overleftarrow{GRU}(x_{ij}), i \in [l, 1]$$

The forward hidden state $\overrightarrow{h}_{ij}$ and backward hidden state $\overleftarrow{h}_{ij}$ are concatenated to obtain word encoded representation $h_{ij}$.

**Attention Layer:** The intuition for using an attention layer is that a few words in an utterance are more important in identifying an emotion. Moreover, the informativeness of words is context dependent i.e. same set of words contribute differently in different context. We augment the Word-level Encoder with a deep self-attention mechanism (Bahdanau et al., 2014; Baziotis et al., 2017) to obtain a more accurate estimation of the importance of each word. The attention mechanism assigns a weight $\alpha_{ij}$ to each word representation. Formally:

$$r_{ij} = tanh(Wh_{ij} + b)$$

$$\alpha_{ij} = \frac{exp(r_{ij})}{\sum_{i=1}^{l} exp(r_{ij})}$$

$$s_j = \sum \alpha_{ij} h_{ij}$$

where $s_j$ is the utterance representation.

**Utterance-Level Encoder:** Similar to Word Level Encoder, the set of utterance representations $s_j$ is passed to a bidirectional GRU to obtain the final representation $g_j$ at utterance level. These representations are passed to CRF classification layer.

**Linear Chain CRF:** Bidirectional encoder captures dependencies among utterances. To model the dependency among labels, the final utterance representations are passed to the linear chain CRF classifier layer. CRFs are undirected graphical models that predict the optimal label sequence given an observed sequence. For a given context window $N_k$, the probability of predicting sequence of emotion labels for a set of utterance representations $\overline{g}$ and corresponding emotion label set $\overline{y}$ is

$$P(\overline{y}|\overline{g}; w) = \frac{exp(\sum_j w_j F_j(\overline{g}, \overline{y}))}{\sum_{y\prime \in Y} exp(\sum_j w_j F_j(\overline{g}, \overline{y\prime}))}$$

where $w_j$ is the set of parameters corresponding to CRF layer and $F_j(\overline{g}, \overline{y})$ is the feature function (Maskey, Spring 2010).

## 4 Data Preparation

The dataset consists of two sets, viz. 1) dialogues collected from *Friends* TV show script and 2) *Facebook* messenger private chats. Both these datasets have characteristics of *short texts.* We describe our preprocessing strategies for these datasets below.

### 4.1 Pre-processing

***EmotionPush***: These are informal chats between two individuals. This data has typical characteristics of short texts. It contains incomplete sentences, informal language, use of emoticons, excessive use of punctuations like *'?'* and *'!'*. As a part of preprocessing, we convert all the emoticons to appropriate emotion word. We also replace all occurrences of date and time with named entities **'DATE'** and **'TIME'**. We convert all contracted forms like ***'can't'***,***'haven't'*** to appropriate expanded forms like ***'can not'*** and ***'have not'***. The dataset contains named entities such as **'PERSON_354'**, **'ORGANIZATION_78'** and

[1] https://nlp.stanford.edu/projects/glove/

**'LOCATION_8'**. These entities are important to build the context but they do not appear in word embeddings. We convert all these named entities to pseudo entities which are present in word embeddings but not present in the *EmotionPush* dataset vocabulary.

| **Accuracy (%)** | **EmotionPush** | **Friends** |
|---|---|---|
| *Unweighted* | 56.73 | 55.38 |
| *neutral* | 88.2 | 73.5 |
| *anger* | 21.6 | 39.8 |
| *joy* | 63.1 | 57.6 |
| *sadness* | 54 | 50.6 |

Table 2: Final results on Test Sets.

| **Emotion** | **Precision (%)** | **Recall (%)** | **F1 (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| *anger* | 31 | 44 | 36 | 44 |
| *joy* | 59 | 64 | 61 | 64 |
| *neutral* | 82 | 85 | 84 | 85 |
| *sadness* | 30 | 61 | 40 | 61 |

Table 3: Experimental results on *EmotionPush* Development Set.

| **Emotion** | **Precision (%)** | **Recall (%)** | **F1 (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| *anger* | 36 | 34 | 35 | 34 |
| *joy* | 47 | 67 | 55 | 67 |
| *neutral* | 67 | 78 | 72 | 78 |
| *sadness* | 25 | 47 | 33 | 47 |

Table 4: Experimental results on *Friends* Development Set.

***Friends - TV Show* scripts**: This dataset contains scene snippets having interaction between two or more speakers. Some of the utterances are incomplete and some have excessive use of punctuations. Unlike *EmotionPush* dataset, there are no emoticons and tagged named entities in this data. We convert the contracted forms as mentioned above and remove extra punctuations. In this dataset, speaker and words uttered by the speaker play an important role in building the context. To incorporate this, we concatenate speaker information to every utterance.

## 5 Experiments and Results

The EmotionX challenge consists of detecting emotions for each utterance from EmotionLines dataset. Each of the utterances has been annotated for one of the eight emotions, *anger, sadness, joy, fear, disgust, surprise, neutral and non-neutral*. Even though the shared task consists of detection of only four emotions, viz. *joy, sadness, anger* and *neutral*, we consider all emotions in our model. We train the model separately for each dataset. We use pre-trained 100-dimensional GloVe-Tweet embedding for both datasets. These embeddings are used to initialize weights of the embedding layer.

We also consider *word priors* as features. *Word prior* for a *word* is computed as

$$p(w_i|c_j) = \frac{count(w_i, c_j)}{count(c_j)}$$

where *count*($w_i, c_j$) is frequency of *word* $w_i$ in *class* $c_j$ and *count*($c_j$) is total number of words in *class* $c_j$. We determine *word priors* for every word for all 8 emotion classes and concatenate these 8 features to embedding feature vectors.

The hyper-parameters such as *window length for context window*, *learning rate*, *optimizer*, *early stopping* and *dropout* were tuned for performance during experimentation.

Results on both *EmotionPush* and *Friends* test sets are listed in Table 2. We also report model performance on both the development datasets in Table 3 and Table 4. The model achieved improvement of 22.51% in *Friends* dataset and 36.04% in *EmotionPush* dataset over baseline (Chen et al., 2018) results. We report overall unweighted accuracy of 56.73% on *EmotionPush* test dataset and accuracy of 55.38% on *Friends* test dataset.

## 6 Discussion

To understand how the context builds over the dialogues, we performed exploratory analysis on both the datasets. In *Friends* dataset, we found some anomalies which can impact the performance of our system.

1. A few dialogues consist of utterances from different scenes which breaks the continuity of the dialogue.

2. Some utterances have scene descriptions as part of the utterance. For example, in record {"speaker": "Joey", "**utterance**": "**and Phoebe picks up a wooden baseball bat and starts to**

**swing as Chandler and Monica enter.)**", "emotion": "non-neutral"}, utterance is a scene description and not spoken by any speaker.
3. We also found few utterances having no words but only a punctuation (*'.'* or *'!'*) which is attached with an emotion. For example,

a) {"speaker": "Rachel", "**utterance**": "**!**", "emotion": "non-neutral"}

b) {"speaker": "Phoebe", "**utterance**": "**.**", "emotion": "non-neutral"}

We did not find such anomalies in *EmotionPush* dataset.

The word embeddings do not have explicit emotion information for words. To incorporate this, we added *word priors* per class to word vectors and examined their effect on the performance of our model. *Word priors* improve the model performance by 17% in *EmotionPush* dataset and 19% in *Friends* dataset. For example, utterances like "Lol weird" and "I also have no shoes lol" belonging to emotion class '*joy*' were misclassified without using word priors as features. Similarly, utterances such as "Sorry he cannot" and "Sorry about that person_107" belonging to emotion class '*sadness*' were also misclassified.

## 7 Conclusion

In this paper, we present our submission for EmotionX emotion detection challenge. We use Hierarchical Attention Network (HAN) model to learn data representation at both utterance level and dialogue level. Additionally, we formalize the problem as sequence labeling task and use a linear chain Conditional Random Field (CRF) as a classification layer to classify the dialogues in both *Friends* and *EmotionPush* dataset. The model achieved improvement of 22.51% in *Friends* dataset and 36.04% in *EmotionPush* dataset over baseline results. In future, we would like to explore the speaker-listener relation with emotion and lexical features to improve the performance of the system.

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.

Lin Chen and Barbara Di Eugenio. 2013. Multimodality and dialogue act classification in the robohelper project. In *Proceedings of the SIGDIAL*.

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao (Kenneth) Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*.

Kyunghyun Cho, Bart Van Merrinboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Paul Ekman, Wallace V. Friesen, and Maureen O'Sullivan et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717.

Su Nam Kim, Lawrence Cavedon, and Timothy Baldwin. 2010. Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic. 2013. Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing*, 4(3):312–325.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, Sachindra Joshi, and Arun Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. *arXiv preprint*, arXiv:1709.04250:712–717.

Yang Liu. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178.

Sameer Maskey. Spring 2010. Statistical methods for natural language processing. Course Slides- Week 13 - Language Models, Graphical Models.

Gilad Mishne. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, pages 321–327.

Saif M Mohammad. 2012a. #emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255.

Saif M Mohammad. 2012b. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Daniel Preotiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes C Eichstaedt, Margaret Kern, Lyle Ungar, and Elizabeth P Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: A lexicon for emotion analysis from crowdannotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 427–433.

Andreas Stolke, Klaus Ries, and Noah Coccaro. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computation Linguistics*, pages 339–373.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied Computing*, pages 1556–1560.

Anand Venkataraman, Lucianna Ferrer Andreas Stolcke, and Elizabeth Shriberg. 2003. Training a prosody based dialog act tagger from unlabeled data. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter big data for automatic emotion identification. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.