SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi and Puneet Agrawal
Microsoft, India
{anchatte, kedharn, mejoshi, punagr}@microsoft.com

Abstract

In this paper, we present the SemEval-2019 Task 3 - EmoContext: Contextual Emotion Detection in Text. Lack of facial expressions and voice modulations make detecting emotions in text a challenging problem. For instance, as humans, on reading "*Why don't you ever text me!*" we can either interpret it as a sad or angry emotion and the same ambiguity exists for machines. However, the context of dialogue can prove helpful in detection of the emotion. In this task, given a textual dialogue i.e. an utterance along with two previous turns of context, the goal was to infer the underlying emotion of the utterance by choosing from four emotion classes - *Happy*, *Sad*, *Angry* and *Others*. To facilitate the participation in this task, textual dialogues from user interaction with a conversational agent were taken and annotated for emotion classes after several data processing steps. A training data set of 30160 dialogues, and two evaluation data sets, Test1 and Test2, containing 2755 and 5509 dialogues respectively were released to the participants. A total of 311 teams made submissions to this task. The final leader-board was evaluated on Test2 data set, and the highest ranked submission achieved 79.59 micro-averaged F1 score. Our analysis of systems submitted to the task indicate that Bi-directional LSTM was the most common choice of neural architecture used, and most of the systems had the best performance for the *Sad* emotion class, and the worst for the *Happy* emotion class.

1 Introduction

Emotions are basic human traits and have been studied by researchers in the fields of psychology, sociology, medicine, computer science etc. for several years. Some of the prominent work in understanding and categorizing emotions include Ekman's six class categorization (Ekman, 1992) and Plutchik's "Wheel of Emotion" (Plutchik and Kellerman, 1986) which suggested eight primary bipolar emotions . In recent times, several Artificial Intelligence (AI) agents like Siri, Cortana, Alexa have emerged and they primarily focus on providing users with assistance on specific tasks such as booking tickets or scheduling meetings etc. However, we believe that for machines and humans to develop a deeper partnership, an Intelligence Quotient (IQ) is not enough. These agents need to also possess an Emotional Quotient (EQ). Social conversational agents like Mitsuku[1] or Ruuh [2] (Damani et al., 2018) are experimental agents designed to have human-like persona, and possess a deeper sense of EQ; understanding and expressing emotions is an inherent aspect of these agents.

Detecting emotions in textual dialogues is a challenging problem in absence of facial expressions and voice modulations. Moreover, we observed that context of ongoing dialogue can completely change the emotion for an utterance as compared to perceived emotion when the utterance is evaluated standalone. Table 1 presents few such examples. Note that, in the first example *"I started crying"* will be perceived as '*Sad*' by a majority, however considering it in context, it turns out to be a '*Happy*' emotion. Similarly, in the second example, the last turn *"Try to do that once"* is very likely to be perceived as '*Others*', however again, a majority will judge it as '*Angry*' with the given context.

Naturally, considering context to estimate emotion of a text utterance becomes even more important for aforementioned scenarios of digital assistants and conversational agents, because of their text-based conversational interface. This task was

[1] www.pandorabots.com/mitsuku
[2] www.ruuh.ai

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 39–48
Minneapolis, Minnesota, USA, June 6–7, 2019.

User Turn-1	Conversational Agent Turn-1	User Turn-2	True Class
I just qualified for the Nabard internship	WOOT! Thats great news. Congratulations!	I started crying	Happy
How dare you to slap my child	If you spoil my car, I will do that to you too	Just try to do that once	Angry
I was hurt by u more	You didn't mean it.	say u love me	Sad

Table 1: Examples showing influence of context in determining emotion of last utterance.

designed to invite research interest in the area of emotion detection in text. More details about the task can be found on our web page[3]. The evaluation data set served as a benchmark to compare various techniques and the task received attention from a wide range of researchers from industry as well as academia. We believe continued interest in this field will be beneficial towards making the AI-agents more human-like.

2 Related Work

Researchers have achieved good results on image based emotion recognition (Wang et al., 2018), (Zhang et al., 2016) as well as voice based emotion recognition (Pierre-Yves, 2003). Techniques have been proposed to detect emotions in spoken dialog systems (Liscombe et al., 2005). However, classifying textual dialogues based on emotions is relatively new research area. Emotion-detection algorithms for text can be largely bucketized into following two categories:

(a) *Hand-crafted Feature Engineering Based Approaches:* - Many methods exploit the usage of keywords in a sentence with explicit emotional/affect value (Balahur et al., 2011), (Strapparava and Mihalcea, 2008), (Sykora et al., 2013). To that end, several lexical resources have been created, such as WordNet-Affect (Strapparava et al., 2004) and SentiWordNet (Esuli and Sebastiani, 2007). Part-of-Speech taggers like the Stanford POS tagger are also used to exploit the structure of keywords in a sentence. These pattern/dictionary based approaches, although attaining high precision scores, suffer from low recall.

Hasan et al. (2014), Purver and Battersby (2012), Suttles and Ide (2013) and Wang et al. (2012) have also harnessed cues from emoticons and hashtags. Other methods rely on extracting statistical features such as presence of frequent n-grams, negation, punctuation, emoticons, hashtags to form representations of sentences which are then used as input by classifiers such as Decision Trees, SVMs among others to predict the output (Alm et al., 2005), (Balabantaray et al., 2012), (Davidov et al., 2010), (Kunneman et al., 2014), (Yan and Turtle, 2016). However, all of these methods require extensive feature engineering and they often do not achieve high recall due to diverse ways of representing emotions. For example, the following utterance, *"Trust me! I am never gonna order again"*, contains no affective words despite conveying an emotion of anger or frustration perhaps.

(b) *Deep Learning Based Approaches:* - Deep Neural networks have enjoyed considerable success in varied tasks in text, speech and image domains. Variations of Recurrent Neural Networks, such as Long Short Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and Bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997) have been effective in modeling sequential information. Also, Convolutional Neural Networks (CNN) (Krizhevsky et al., 2012) have been a popular choice in the image domain. Their introduction to the text domain has proven their ability to decipher abstract concepts from raw signals (Kim, 2014).

Recently, approaches which employ Deep Learning for emotion detection in text have been proposed. Zahiri and Choi (2017) predicts emotion in a TV show transcript. Abdul-Mageed and Ungar (2017) and Köper et al. (2017) tries to understand emotions of tweets. Li et al. (2017) learns to detect emotions on user comments in Chinese language. Felbo et al. (2017) learns representation based on emoticons, and uses it for emotion detection. A further detailed analysis of various approaches have been provided by Chatterjee et al. (2019). It is worth noting that textual dialogues are informal and laden with misspellings which pose serious challenges for automatic emotion detection approaches. Prior to this task, to the best of our knowledge, the methods proposed by Mundra

[3]Task webpage: humanizing-ai.com/emocontext.html

et al. (2017) and Chatterjee et al. (2019) are some of the few methods that tackled the problem of emotion detection in English textual dialogues.

3 Task Details

Problem Definition: *In a textual dialogue, given an utterance along with its two previous turns of context, classify the emotion of the utterance as one of the following classes: Happy, Sad, Angry or Others.*

The motivation for restricting the number of emotion classes stems from the popularity of these emotions in conversational data. The task proceeded in two phases. A training corpus, Train, of 30160 dialogues was provided at the beginning of Phase 1. The evaluation in this phase was done on an evaluation data set, Test1, comprising of 2755 dialogues. The labels for Test1 were made public five weeks before the end of Phase 1, allowing participants time and data to improve their models. The final evaluation was carried out in Phase 2 on a evaluation data set, Test2, which comprised of 5509 dialogues. It is important to note that while the maximum number of submissions a participant could make in Phase 1 was 20 per day, it was reduced to 10 per day during Phase 2.

4 Data Collection

A data set of textual dialogues was released to facilitate participation in this task. Several data processing steps were performed to create the final set of textual dialogues which are further explained in this section.

4.1 Dialogue Collection and Processing

A dialogue mined from the user's interaction with agent is defined as a tuple of 3 values - User Turn-1 (Utterance of the user), Conversational Agent Turn-1 (Response by the agent), User Turn-2 (User utterance as response to agent).
To begin with, user interactions with the agent over a period of one year were considered and over 2 million dialogues were randomly sampled. These dialogues further went through the processing and data cleaning as described in further subsections.

4.1.1 Offensive filtering

All the dialogues were passed through a filtering layer to remove offensive and sensitive content

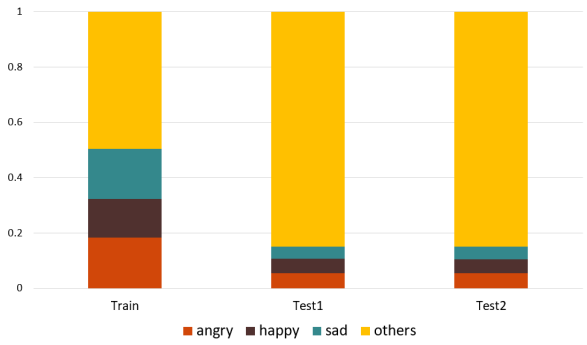


Figure 1: Comparison of class distribution in Training vs Evaluation data sets.

Emotion	Happy	Sad	Angry	Others	#
Train	4243	5463	5506	14948	30160
Test1	142	125	150	2338	2755
Test2	284	250	298	4677	5509

Table 2: Emotion label count across classes in Train, Test1 and Test2 data sets.

such as adult information, politically sensitive topics, or ethnic-religious content, or other potentially contentious material, such as inappropriate references to violence, crime and illegal substances etc. Several lexicons and human judgments were used to achieve this filtering.

4.1.2 PII filtering

Personally Identifiable Information (PII) identifies the unique identity of a given user. This includes personal data like names, phone numbers, email Ids, among others. Dialogues containing any PII content were removed using hand crafted rules and via human judgments.

4.1.3 Language filtering

Given that the agent was available for users across geographies, the dialogues contained multiple languages and users employed code-mixed language as well. We used language detectors as well as user modeling to identify the language in the dialogues and filter non-English dialogues from the data set.

4.2 Training Data Set Creation

In the collected textual dialogues the emotion classes were not frequently expressed and hence directly annotating a random sample of textual dialogues results in very low volume of textual dialogues with emotion class. This problem was tackled by Gupta et al. (2017) and we used similar heuristics and strategies to ensure a higher ratio of

textual dialogues with emotion classes. This exercise was primarily conducted to reduce the cost of human judgments and is further explained below. We started with a small set (approximately 300) of annotated dialogues per emotion class obtained by showing a randomly selected sample to human judges. Using a variation of the model described by Palangi et al. (2016), we created embedding for these annotated dialogues. Potentially similar dialogues were further identified from the entire pool of dialogues using a threshold-based cosine similarity and these dialogues form our candidate set for each emotion class. Various heuristics like presence of opposite emoticons (example ":'(" in a potential candidate set for *Happy* emotion class), sentiment analysis, length of utterances etc. are used to further prune the candidate set in certain cases. The candidate set is then shown to human judges to determine if they belong to an emotion class. Using this method, we cut down the amount of human judgments required by five times as compared to showing a random sample of dialogues and then choosing dialogues with emotion class from them.

Data belonging to class "*Others*" is collected by randomly selecting dialogues from our pool of dialogues and were human labelled to discard any dialogues with emotion class such as *Happy*, *Sad* or *Angry*.

Figure 1 shows the distribution of different classes in training data set.

4.3 Evaluation Data Set Creation

Unlike training data set where we intentionally over sampled dialogues from emotion classes to help participants with a larger volume of data with emotion classes, we maintained the natural distribution of emotion classes in evaluation data sets. We randomly sampled and annotated two evaluation sets, Test1 and Test2, of size 2755 and 5509 respectively. Detailed distribution of emotion classes in these sets is described in Table 2.

4.4 Emotion Class Labeling

For this specific task of emotion class labelling, 50 human judges were trained. Given a dialogue, i.e an utterance with two previous turns as context, a judge was asked to annotate the utterance as belonging to one of the following four classes: *Happy*, *Angry*, *Sad* or *Others*. All dialogues were judged by 7 human judges and a majority consensus was taken as the final class label. Fleiss' Kappa score (Shrout and Fleiss, 1979) of 0.58 was observed on training data set and of 0.59 on evaluation data set. Such a Kappa score indicates the existence of multiple perspectives about the underlying emotion of a conversation.

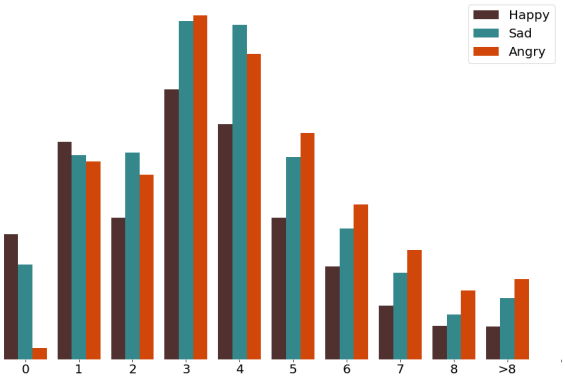


Figure 2: Comparison of word count of utterances per emotion class. Emoticons were removed for this calculation, as a result of which the leftmost bin of 0 word count can be seen as well.

5 Data Analysis

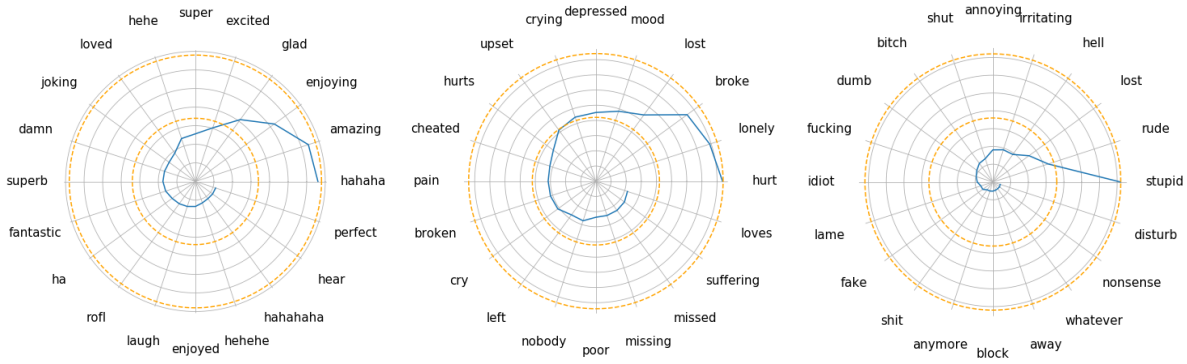
In this section we analyze the utterance in the dialogue that was judged by human judges for emotion classes.

5.1 Word Count

Figure 2 shows the distribution of the word count of utterances per emotion class. We observed that users tend to repeat emoticons several times. Hence emoticons were removed from utterances for this calculation, as a result of which the utterances which had only emoticons are clubbed in the leftmost bin with utterance of length 0. It can be observed that happiness is often expressed through emoticons and hence happy emotion class has highest count under the bin of 0 word count. Also, happiness is often expressed in fewer words as compared to other emotions can be observed from the graph. Another point to note is that angry emotion class is often expressed using more words as compared to other emotion classes.

5.2 Top Unigrams

Figure 3 shows the most frequent unigrams per emotion class in our data set. Note that emoticons are not considered as unigrams for this analysis. The length of the radius in the spiral graph denotes the frequency of the unigram in all the utterances belonging to that particular emotion class. In order



(a) Happy (b) Sad (c) Angry

Figure 3: Most frequent unigrams per emotion class in our data set. The length of the radius in the spiral graph denotes the frequency of the unigram in all the utterances for a emotion class. Only those unigrams which are not in the top 500 list of most frequent unigrams of the "Others" class have been considered.

Happy	😂	😁	😀	😄	😆
Sad	😭	😞	😢	💔	😔
Angry	😡	😠	😤	😑	😤

Table 3: Top five emoticons per emotion class.

to avoid neutral words like "my", "what", "sure" from showing up in the analysis, we consider only those unigrams which are not in the top 500 list of most frequent unigrams of the "Others" class.

5.3 Top Emoticons

Emoticons are frequently used in textual dialogues, as was observed by Gupta et al. (2017), who found 21% of textual dialogues to contain emoticons. Table 3 shows the top emoticons observed in utterances per emotion class. While most emoticons align with our expectations of the most frequent emoticons, it is interesting to note the frequent use of broken-heart emoticon to express sad emotion.

6 Evaluation Metric

Evaluation was carried out using the micro-averaged F1 score ($F1_\mu$) for the three emotion classes - *Happy*, *Sad* and *Angry* on the submissions made with predicted class of each sample in the evaluation data set. To be precise, we define the metric as following:

$$P_\mu = \frac{\Sigma TP_i}{\Sigma(TP_i + FP_i)} \forall i \epsilon \{Happy, Sad, Angry\}$$

$$R_\mu = \frac{\Sigma TP_i}{\Sigma(TP_i + FN_i)} \forall i \epsilon \{Happy, Sad, Angry\}$$

$$F1_\mu = 2 \cdot \frac{P_\mu \cdot R_\mu}{P_\mu + R_\mu}$$

where TP_i is the number of samples of class i which are correctly predicted, FN_i and FP_i are the counts of *Type-I and Type-II errors* [4] respectively for the samples of class i.

Our final metric $F1_\mu$ is calculated as the harmonic mean of P_μ and R_μ.

7 Baseline Model

To encourage and assist participants in making their first submission, we provided a starter kit, which consisted of scripts for training a naive baseline model. The script also enabled participants to cross-validate their model and create a submission file. This section explains the baseline model in detail.

7.1 Data Processing

Minimal data pre-processing steps were provided. These included replacing certain repeated punctuation marks with their single instances, lower casing, removing extra space and tokenization. For example, "I am so happy!!" was converted to "i am so happy !".

7.2 Model Architecture

We modeled the task of detecting emotions as a multi-class classification problem where given a dialogue, the model outputs probabilities of it belonging to four output classes - *Happy*, *Sad*, *Angry* and *Others*. The three turns are concatenated using a special <*eos*> token. The concatenated input is passed into a pre-trained word embedding

[4] http://en.wikipedia.org/wiki/Type_I_and_type_II_errors

Team	GloVe	Word2Vec	NTUA-SLP	BERT	ELMO	ULMFit	Others
NELEC							✓
SymantoResearch	✓			✓			✓
ANA	✓			✓	✓		
CAiRE_HKUST	✓			✓	✓		✓
SNU_IDS		✓			✓		✓
THU-HCSI		✓	✓				
Figure Eight			✓	✓		✓	✓
YUN-HPCC	✓				✓		
LIRMM-Advanse						✓	
MILAB	✓						
PKUSE	✓						
THU_NGN	✓	✓					✓

Table 4: Input representations used by top systems.

layer, which projects the words into continuous vector representations. We used 100 dimensional GloVe embeddings (Pennington et al., 2014) for this purpose. The embeddings are processed by an LSTM layer, which produces a 128 dimensional representation of the sentence. This representation is then mapped to a 4 dimensional output vector which outputs probabilities per emotion class using a fully connected neural network. The architecture of the model was kept deliberately simple and was intended to serve as a starting point for participants. The baseline model achieved a $F1_{\mu}$ score of 0.5861 on the final leader board and most teams were able to beat the baseline model. Further details on the model and its comparison with other systems can be seen in Table 5.

8 Systems and Results

As mentioned earlier in section 3, the task was conducted in two phases. The first phase saw a participation from 311 teams and 164 teams participated in the second phase. In this section, we briefly describe the top systems [5], followed by observations across systems regarding the techniques used and their performance across different emotion classes.

[5]The top 2 systems - *Leo1020* and *Mfzszgs* did not submit system description papers, and hence have been omitted from discussion in this Section.

8.1 Top Systems

Due to the overwhelming number of participants, we cannot describe all systems. We describe the main features of the top few systems ranked according to their final performance.

- **NELEC** uses a combination of lexical features such as word and character grams, along with additional signals like emotional intensity, valence-arousal-dominance scores. In addition, they use adult, offensive and sentiment classifiers' scores from neural models. Using these features, the authors trained a Light-GBM tree (Ke et al., 2017), which achieves better performance than their deep-learning based architecture.

- **SymantoResearch** explores different deep-learning based architectures, some of them employing multi-task learning to better classify *Others* class vs. emotion classes. By ensembling such architectures with fine-tuned BERT (Devlin et al., 2018) and USE (Cer et al., 2018) models, the authors are able to distinguish three emotions (*Sad*, *Happy*, *Angry*) and separate them from the rest (*Others*) more accurately.

- **ANA** uses an ensemble of fine tuned BERT model and Hierarchical LSTMs, where the semantic and emotional content of text is encoded via GloVe, ELMo (Peters et al., 2018)

Team Name	Angry			Happy			Sad			$F1_\mu$
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	
Leo1020	0.7723	0.8423	**0.8058**	**0.804**	0.7077	**0.7528**	0.8494	0.812	0.8303	**0.7959**
Mfzszgs	0.759	0.8456	0.8	0.7769	0.7113	0.7426	**0.8595**	0.832	**0.8455**	0.7947
NELEC	0.747	0.8322	0.7873	0.7632	0.7148	0.7382	0.7938	0.816	0.8047	0.7765
SymantoResearch	**0.7807**	0.7886	0.7846	0.738	0.7042	0.7207	0.8193	0.816	0.8176	0.7731
ANA	0.7198	0.8188	0.7661	0.7698	0.6831	0.7239	0.8458	0.812	0.8286	0.7709
CAiRE_HKUST	0.6997	0.8289	0.7588	0.7301	**0.743**	0.7365	0.7774	**0.852**	0.813	0.7677
SNUIDS	0.7405	0.7852	0.7622	0.772	0.6796	0.7228	0.8135	0.82	0.8167	0.7661
THU-HCSI	0.7155	0.8356	0.7709	0.7702	0.6725	0.718	0.796	0.796	0.796	0.7616
Figure Eight	0.6954	**0.8658**	0.7713	0.7055	0.7254	0.7153	0.7695	0.828	0.7977	0.7608
YUN-HPCC	0.7198	0.8188	0.7661	0.7169	0.6866	0.7014	0.8016	0.824	0.8126	0.7588
LIRMM-Advanse	0.7229	0.8054	0.7619	0.7256	0.7077	0.7166	0.8291	0.776	0.8017	0.7582
MILAB	0.7295	0.8054	0.7656	0.7481	0.7007	0.7236	0.7652	0.808	0.786	0.7581
Huxiao	0.7362	0.8054	0.7692	0.7403	0.6725	0.7048	0.7757	0.816	0.7953	0.7564
PKUSE	0.745	0.755	0.75	0.7351	0.6937	0.7138	0.8056	0.812	0.8088	0.7557
THU_NGN	0.7329	0.7919	0.7613	0.7452	0.6796	0.7109	0.8117	0.776	0.7935	0.7542
Baseline	0.4777	0.7867	0.5945	0.5123	0.5845	0.5461	0.5163	0.7600	0.6149	0.5861

Table 5: Performance comparison of top 15 teams on leaderboard.

and DeepMoji (Felbo et al., 2017) embeddings, following which a contextual LSTM encodes the entire dialogue for prediction.

- **CAiRE_HKUST** experiments with combinations of feature based models and end-to-end neural models. The feature based models use various pre-trained word embeddings and emotional embeddings, combining them with Logistic Regression and XGBoost (Chen and Guestrin, 2016). For the end-to-end neural models, the authors found the performance of hierarchical models, which take sequential nature of dialogue into account, to be better.

- **SNU_IDS** proposes several methods for alleviating the problems caused by difference in class distributions between training data and test data. The authors also present a semi-hierarchical neural architecture combining character and word embeddings that effectively encodes an utterance in context of the previous utterances.

- **THU-HCSI** is composed of three CNN-based neural network models trained for different base tasks - four-emotion classification, *Angry-Happy-Sad* classification and *Others-or-not* classification respectively. The authors use multiple steps of voting to combine the predictions of these base classifiers, resulting in a more accurate and robust model performance.

- **Figure Eight** uses an ensemble of transfer learning models for capturing the representations of the utterances. Using sophisticated fine-tuning techniques described in ULMFiT (Howard and Ruder, 2018), the authors observe that transfer learning using pre-trained language models outperforms models trained from scratch.

8.2 Miscellaneous Observations

From the system description papers of the top 15 teams, we observed that BiLSTMs/LSTMs were the most frequently used neural models. GRU (Chung et al., 2014) and CNN models were used by a few teams, and some variations of attention mechanism were employed by most of the teams to enhance performance of their models. Transfer learning using BERT, ELMo, ULMFit was a popular choice among top teams, and almost all the teams used an ensemble of their best models to create the final model.

	$\mathbf{F1}_\mu$
Max	0.7959
Min	0.0143
Mean	0.6599
Median	0.694
1st Quartile	0.637
3rd Quartile	0.7317
Std. Dev.	0.1264

Table 6: Performance statistics of all participants.

Table 4 shows the embeddings used by the top 5 teams. It can be observed that GloVe was used most frequently. BERT and ELMo were the most popular choice for transfer learning. NTUA-SLP embeddings (Baziotis et al., 2018) were used as well to leverage its affective information. Participant teams tried various ways to encode the emotional content expressed by emoticons, and Deepmoji and Emoji2Vec (Eisner et al., 2016) were utilized in this regard. A good number of teams used the "ekphrasis" package (Baziotis et al., 2017) for tokenization, word normalization and word segmentation.

8.3 Performance across Emotion Classes

Table 5 displays the detailed performance of the top 15[6] participant teams. Upon inspection, it can be observed that the performance of the systems on the *Happy* class was not as good as the other emotion classes for the evaluation set. We believe, this is largely due to the natural ambiguity existing between neutral and happy utterances. For example, a greeting like "*Happy Morning*" can be thought of as expressing a happy emotion by some, while being judged to be neutral by others. We also observed that most systems performed best for the *Sad* emotion class. Table 6 provides some basic statistics on the results obtained by the whole set of participants.

9 Conclusion

A total of 311 teams made submissions to the task. The final leader-board was evaluated on Test2 data set, and the highest ranked submission achieved 79.59 $F1_\mu$ score. Our analysis of systems submitted to the task indicate that Bi-directional LSTM was the most common choice of network architecture used by participants, and most systems had best performance for *Sad* emotion class, and worst for *Happy* emotion class. A large number of teams have participated in the task but only 46 teams submitted their final system description papers; in fact, the top 2 teams in Phase 2 did not submit their system description paper. It was also observed that the ranking of various systems across both the phases varied significantly. In this task, we released the evaluation set without labels to participants, in future tasks it might be useful to also experiment with system submissions such that the entire evaluation set is never seen, with or without labels to the participants during the evaluation phase in a bid to have completely blind evaluation.

[6]Final rankings of all participating systems can be consulted via the CodaLab website of our task: https://competitions.codalab.org/competitions/19790

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume Vol. 1, pages 718–728.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. ACL.

Rakesh C Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, Vol. 4, pages 48–53.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 53–60. ACL.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Sonam Damani, Nitya Raviprakash, Umang Gupta, Ankush Chatterjee, Meghana Joshi, Khyatti Gupta, Kedhar Nath Narahari, Puneet Agrawal, Manoj Kumar Chinnakotla, Sneha Magapu, and Abhishek Mathur. 2018. Ruuh: A deep learning based conversational social agent. *32nd Conference on Neural Information Processing Systems (NIPS 2018), Montral, Canada*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 241–249. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, Vol. 6, pages 169–200.

Andrea Esuli and Fabrizio Sebastiani. 2007. Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*, pages 1–26.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD Workshop on Health Informatics, New York, USA*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, Vol. 9, pages 1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. Ims at emoint-2017: emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *WASSA*, pages 50–57.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Florian Kunneman, Christine Liebrecht, and Antal van den Bosch. 2014. The (un) predictability of emotional hashtags in twitter. *In European Chapter of the Association for Computational Linguistics*, pages 26–34.

Panpan Li, Jun Li, Feiqiang Sun, and Peng Wang. 2017. Short text emotion analysis based on recurrent neural network. In *Proceedings of the 6th International Conference on Information Engineering*. ACM.

Jackson Liscombe, Giuseppe Riccardi, and Dilek Hakkani-Tur. 2005. Using context to improve emotion detection in spoken dialog systems.

Shreshtha Mundra, Anirban Sen, Manjira Sinha, Sandya Mannarswamy, Sandipan Dandapat, and Shourya Roy. 2017. Fine-grained emotion detection in contact center chat utterances. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 337–349. Springer.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol. 24, pages 694–707.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume Vol. 14, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Oudeyer Pierre-Yves. 2003. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1-2):157–183.

Robert Plutchik and Henry Kellerman. 1986. *Emotion: theory, research and experience*. Academic press New York.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. ACL.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, Vol. 45, pages 2673–2681.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, Vol. 86, page 420.

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

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *The 4th International Conference on Language Resources and Evaluation*, volume Vol. 4, pages 1083–1086.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer.

Martin D Sykora, Thomas Jackson, Ann O'Brien, and Suzanne Elayan. 2013. Emotive ontology: Extracting fine-grained emotions from terse, informal messages. *IADIS International Journal on Computer Science and Information Systems*.

Shui-Hua Wang, Preetha Phillips, Zheng-Chao Dong, and Yu-Dong Zhang. 2018. Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm. *Neurocomputing*, 272:668–676.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust, 2012 International Conference on Social Computing*, pages 587–592. IEEE.

Jasy Liew Suet Yan and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *The North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 73–80.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

Yu-Dong Zhang, Zhang-Jing Yang, Hui-Min Lu, Xing-Xing Zhou, Preetha Phillips, Qing-Ming Liu, and Shui-Hua Wang. 2016. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access*, 4:8375–8385.