Emotion detection on text using BERT language model

Comparing Deep Learning based models

Harsh Yadav
Computer Science
University of Illinois at Chicago
Chicago, IL, USA
hyadav5@uic.edu

Ansul Goenka
Computer Science
University of Illinois at Chicago
Chicago, IL, USA
agoenk2@uic.edu

Aishwarya Sahani
Computer Science
University of Illinois at Chicago
Chicago, IL, USA
asahan2@uic.edu

ABSTRACT

In this project, we attempt to solve the problem of Emotion detection in text based data. Recent developments in neural networks and sequence based models have shown great success in being able to detect emotions like fear, joy, happiness, sadness, etc from text. We compare 4 types of Deep Learning models on this problem- Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated recurrent Unit (GRU), and BERT, and contrast them with the baseline model of logistic regression. We demonstrate the abilities of these models on two datasets of different sample spaces.

KEYWORDS

Emotion Detection, Emotion Classification, EmotionLines, BERT, ISEAR

1 Introduction

Philosophers would argue that the perception of emotion is what separates humans from other non-sentient species. Emotion is arguably the most important factor in the decision making process of humans, While other animals can also express emotion in certain ways, humans have taken this to a new level with language. Written text gives us the power of expressing our feelings without the constraints of space and time. Keats died 200 years ago but his poetry can still send us in deep thought. As such, trying to make computers to understand emotion has been the key focus of Artificial Intelligence since its birth.

Applications of emotion detection are far and wide. It enables us to develop advanced chatbots and personal assistants which can learn the mental state of the user from their utterances. This can help in creating a more realistic experience for the user where the AI is able to respond like a person who is aware of the emotions in a dialogue, and not simply give the same lifeless responses to queries.

Another use for emotion detection could be in a system which does mental health checks and counselling, where we can help more people with mental health issues. We can also employ emotion detection in system which looks at the reactions of people on social media and figures out the state of discourse on any topic or event

The techniques of Natural Language Processing (NLP) enable us to build models which are able to learn the intricacies of human language. Still, emotion detection remains a fairly difficult task for modern AI systems.

In this project we are concerned with developing a neural network based model which is able to correctly identify the emotion being expressed in a sentence or a dialogue. To that end, our plan is five-fold:
(1): Find one or more datasets which work for out problem of emotion detection
(2): Establish a baseline model
(3): Employ various deep learning based models on the dataset, which are expected to give better accuracy.
(4): Evaluate the performance of these models, and compare and contrast them with each other.
(5): Conclude what the best model is.

For our work, we identified two datasets: ELD and ISEAR. Both of these contain text samples annotated with the corresponding emotion being expressed in that piece of text. We then trained RNN classifiers like Vanilla RNN, LSTM & GRU on the datasets. After establishing this baseline, we incorporated the BERT model for the classification tasks on multiple datasets.

After following this plan, we found that the BERT language model significantly outperformed the baseline traditional deep learning recurrent neural network models.

2 Related work

There is a lot of research done for Emotion Recognition but not a lot of them utilize the modern state-of-the-art language models like BERT, ELMO, GPT-2. In the paper, Emotion Detection from Text via Ensemble Classification Using Word Embeddings, they utilize pre-trained, dense word embedding representations for classification of emotions. For the datasets, we explored multiple papers like EmotionLines[18] & ISEARS[19] where they are describing their dataset. Also, using dialogues from one of the most popular & loved series like ‘F.R.I.E.N.D.S.’ makes the dataset more engrossing & makes it interesting for the users to work on it.[18] Also, we believe that since the dialogues are taken from a TV series, the content would be more scripted, well-structured, dramatic and less natural than an actual human conversation. The annotations used for emotions are “anger, disgust, fear, joy, neutral, sadness, and surprise”. The ISEAR dataset explores the emotional profiles of people from different cultures. The pre-trained BERT language model is used to create state-of-the-art models by adding an additional output layer for a wide range of natural processing tasks, such as question answering and language inference.[22] We incorporated this model to classification of emotions tasks to see how it deals with the problem and we were impressed by the results.

3 Problem Statement

In this paper, we are concerned with applying deep learning based models to the problem of emotion detection in text. The model should take as input a piece of text such as a sentence, a dialogue, a tweet, etc., and output the corresponding emotion which is being expressed in that piece text.

We want to compare the performance of different types of models on 2 datasets. A simple RNN based model as the baseline, and more advanced models based on LSTM, GRU, and finally BERT.

4 Technical approach

We train four different types of models for our analysis, which are enumerated below:

(1) An RNN based model (baseline)
(2) A LSTM based model
(3) A GRU based model
(4) BERT model (final)

These models are detailed below.

4.1 Baseline (RNN)

The baseline we have chosen for this project is a vanilla Recurrent neural network (RNN) based neural network model implementation. RNNs are the basic component of a network which can handle sequential data. At each timestep, an RNN takes an input and a hidden state, and produces the next hidden state, and an output.

$$\mathbf{h}_t = \mathbf{f}\,(\mathbf{W}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{b})$$

Here, $\mathbf{W}$ is the weight matrix for the input and $\mathbf{b}$ is the bias vector. $\mathbf{h}_t$ and $\mathbf{h}_{t-1}$ are the hidden states at timestep t and t-1. $\mathbf{U}$ is the weight matrix for the previous hidden state. $\mathbf{f}$ is the activation function.

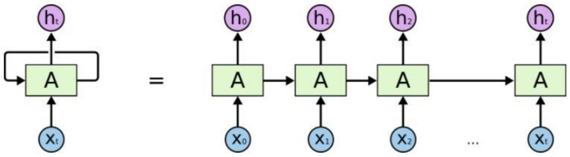


Figure 1: **RNN units**
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

In our model based on the RNN, we first have an embedding layer which calculates the embeddings from the input text data. These embeddings are then fed as sequential input to two RNN layers of dimensions 256 and 128 respectively. These recurrent layers are followed by a Dense layer of size 64, and finally an output layer of predictions for each emotion. The size of the output layer is dependent on the dataset used, and is detailed in a later section.

RNNs are supposed to generate meaningful representations inside their hidden states and therefore should be able to tackle problems of sequential text based data. But in reality, they suffer from many problems and do not perform well with large sequences.

This is mainly due to the problems of vanishing and exploding gradients in RNNs. Due to this, the RNN is unable to look at the inputs which occurred at many steps behind the current position in the sequence. Hence, they tend to lose or ‘forget’ important information, such as context, which is necessary for making good predictions on text based data. Consequently, we see that RNNs fail to perform well on many NLP tasks.

To tackle these drawbacks of the RNN, two modifications of the RNN have been proposed, the LSTM and the GRU.

4.2 Long Short term Memory (LSTM)

To remove the problem of vanishing gradients and the loss of memory in the RNNs, LSTMs were proposed in 1997.
In addition to the hidden state of the RNN, LSTMs added 3 additional gates: the forget gate, the input gate, and the output gate. These gates are dynamic and control how much information if passed through each of these gates, thus giving the LSTM the ability to erase, write, and read information from the cell.

$$\mathbf{i_t = \sigma\ (W^i x_t + U^i h_{t-1} + b^i)}$$
$$\mathbf{f_t = \sigma\ \ W^f x_t + U^f h_{t-1} + b^f}$$
$$\mathbf{o_t = \sigma\ (W^o x_t + U^o h_{t-1} + b^o\ )}$$
$$\mathbf{g_t} = \mathrm{relu}\ \mathbf{(W^g x_t + U^g h_{t-1} + b^g\ )}$$
$$\mathbf{c_t = f_t \,.\, c_{t-1} + i_t \,.\, g_t}$$
$$\mathbf{h_t = o_t \,.\, tanh(c_t)}$$

Here, i_t , f_t, and o_t are input, forget, and output gates respectively. g_t is a new memory cell, c_t is the hidden state from the RNN.

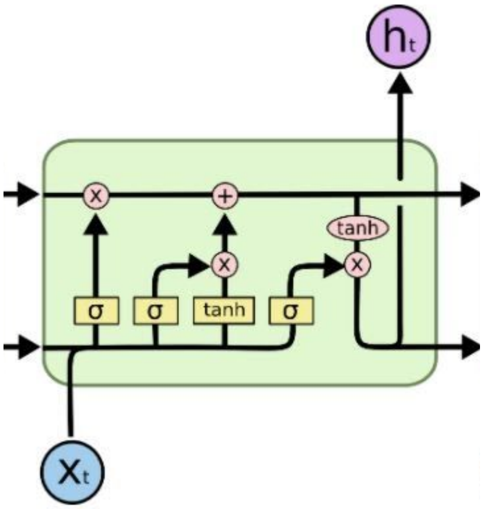


Figure 2: **LSTM cell**
Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

These modifications to the vanilla RNN structure enable the LSTM to keep memory for longer periods of time, which ultimately results in better performance on the data.

In our model based on the LSTM, we first have an embedding layer. These embeddings are then fed as sequential input to an LSTM layer of dimension 128. The recurrent layer is followed by an output layer of predictions for each emotion.

4.3 Gated Recurrent Unit (GRU)

GRUs are an additional modification of the LSTM model. They have only one hidden state, and only two gates - an update gate, and a reset gate.

$$\mathbf{r_t = \sigma\ (W^r x_t + U^{\,r} h_{t-1} + b^{\ r}\ )}$$
$$\mathbf{z_t = \sigma\ (W^z x_t + U^{\,z} h_{t-1} + b^{\ z}\ )}$$
$$\mathbf{h\tilde{}_{\ t} = tanh\ (\ W x_t + r_t * U^{\ h\tilde{}}\ h_{t-1} + b^{\ h\tilde{}}\ )}$$
$$\mathbf{h_t = z_t * h_{t-1} + (1 - z_t) * h\tilde{}}$$

Here, r_t and z_t are the reset and the update gates respectively.
GRU also has the ability to remember information for long periods of time, but it is computationally more efficient than LSTM.

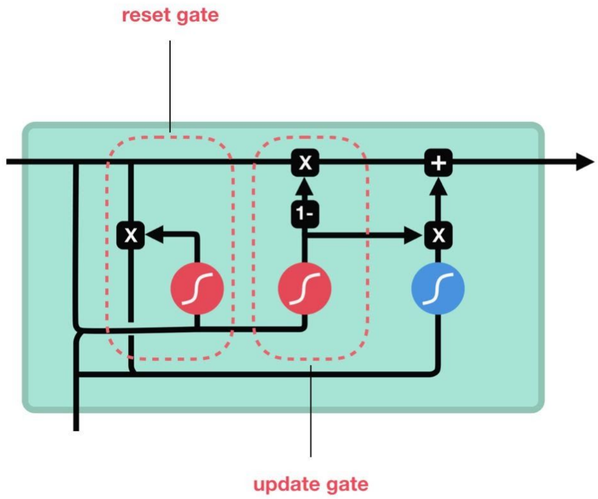


Figure 3: **GRU cell**
Source:https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

In our model based on the GRU, we first have an embedding layer. These embeddings are then fed as sequential input to an GRU layer of dimension 128. The recurrent layer is followed by an output layer of predictions for each emotion.

For each of the models detailed above, an activation of ReLu is applied on all layers except the output layer, where a softmax activation is applied.
We have also used Dropout on all layers with a dropout rate of 30%.

4.4 BERT

The Bidirectional Encoder Representations from Transformers, or BERT, is a more advanced deep learning model for sequential input. It is based on the transformer architecture.

BERT has broken many records in recent times and has proved to be the state-of-the-art model in many applications of NLP. It also has models which were pre-trained on very large datasets.

The BERT model contains 12 encoder layers, 768 hidden units, 12 attention heads. Each layer of BERT applies self-attention, and passes the results to a feed-forward network.

After the processing from BERT, we add a Dense layer and output layer. These final layers are used for fine-tuning the weights of the model so that BERT performs well on our specific task.

5 Experimental Setup

5.1 Dataset

We have used EmotionLines [18] and ISEAR (International Survey on Emotion Antecedents and Reactions) [19] as the dataset for this classification problem. EmotionLines contains dialogues from Friends (TV series), which is categorised into 7 different categories.

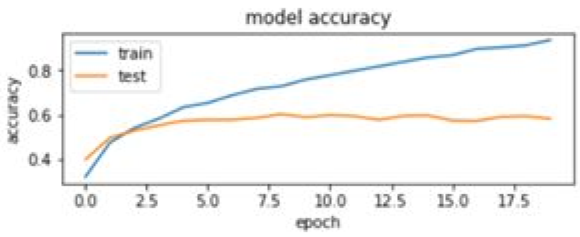
Datasets	Dialogues
Training Set	9989
Testing Set	1109
Validation Set	2610

Table 1: **Dataset sizes for EmotionLines**

The ISEAR dataset consists of 7,666 sentences [19], with regard to our experiments.For building the ISEAR, 1,096 participants who have different cultural backgrounds completed questionnaires about experiences and reactions for seven emotions including anger, disgust, fear, joy, sadness, shame and guilt.

5.2 Experiments

We have tried RNN, LSTM, GRU and BERT on EmotionLines and ISEAR dataset. We tried a different number of hidden units for all of them and found 64 hidden units gave us the best results for EmotionLines dataset and 128 to be the best for ISEAR dataset. We found that a dropout of 30% gave us the best regularization. We tried with different numbers of epochs and realised that 10 epochs gave us sufficient results for EmotionLines and 20 epochs for ISEAR, after that our model (RNN, LSTM, GRU) started to overfit and for BERT we got sufficient result with 10 and 5 epochs respectively.



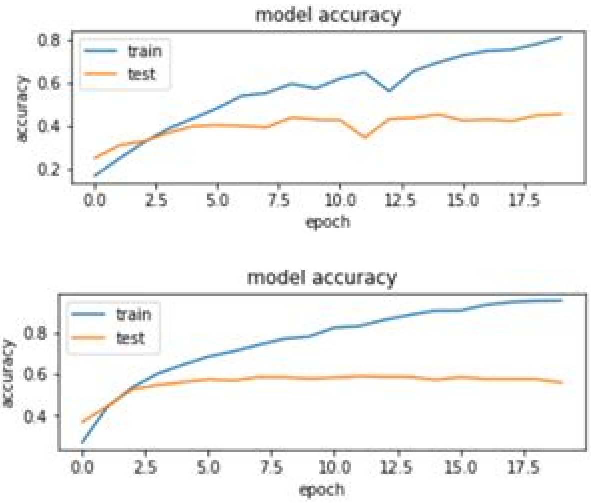


Figure 4: **Line graph showing the change the training and validation accuracy of LSTM, RNN and GRU models for ISER dataset**

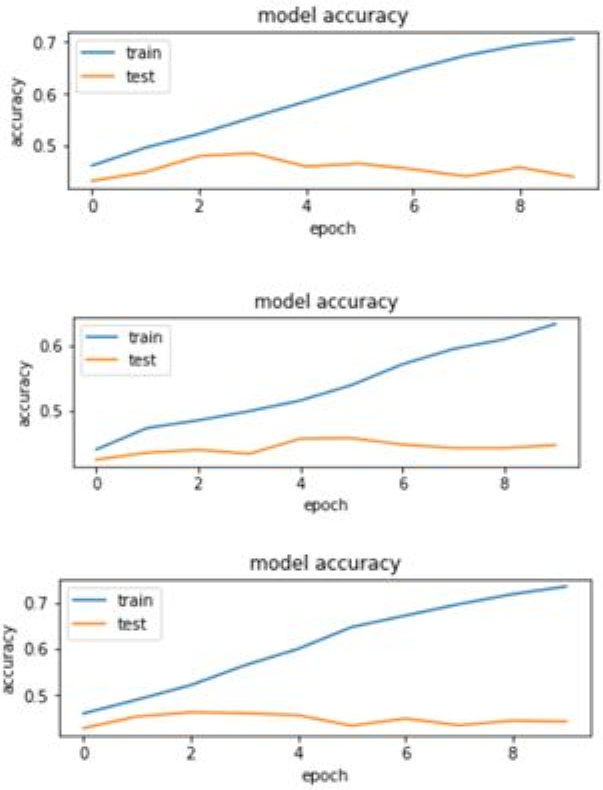


Figure 5: **Line graph showing the change the training and validation accuracy of LSTM, RNN and GRU models for EmotionLines dataset**

The learning rate for RNN, GRU and LSTM models were 0.01 and for BERT we tried with 0.0005 and 0.00003 and found that 0.00003 as the initial learning rate gave us the best results. We realised that adam performed better than

SGD and RMS Prop for RNN, LSTM and GRU models. We got better results with categorical cross entropy as our loss function than the mean square. For BERT, we used adamw as the optimizer and categorical cross entropy as the loss function. We tried relu, tanh and sigmoid as our activation function but got better results with relu as the activation function.

6 Result

6.1 Performance Metrics

For classification tasks, we generally use accuracy, precision, recall and f1 score. We have compared results of RNN, GRU and LSTM models with our fine tuned BERT model for both of our datasets.

Model	Accuracy	Precision	Recall	F1
RNN	49.12	26.19	23.40	22.87
LSTM	48.47	32.16	26.70	27.71
GRU	47.66	31.95	26.82	27.89
BERT	61.57	37.11	34.10	33.73

Table 2: **Table describes the comparison of the BERT model with the baseline models on the EmotionLines dataset.**

For the EmotionLines dataset, our baseline model of RNN performs better than the LSTM and GRU models, but BERT seems to perform significantly better than the other approaches. BERT has produced state-of-the-art results in many natural language processing tasks. We are exploiting this modern architecture for classification of emotions. As expected, the pretrained model does the trick. Although, the reasons for state-of-the-art performances are not yet well understood.

In EmotionLines [18], the state-of-the-art accuracy was 63.9% and with our fine-tuned BERT model with 10 epoch we achieved an accuracy 61.57%.

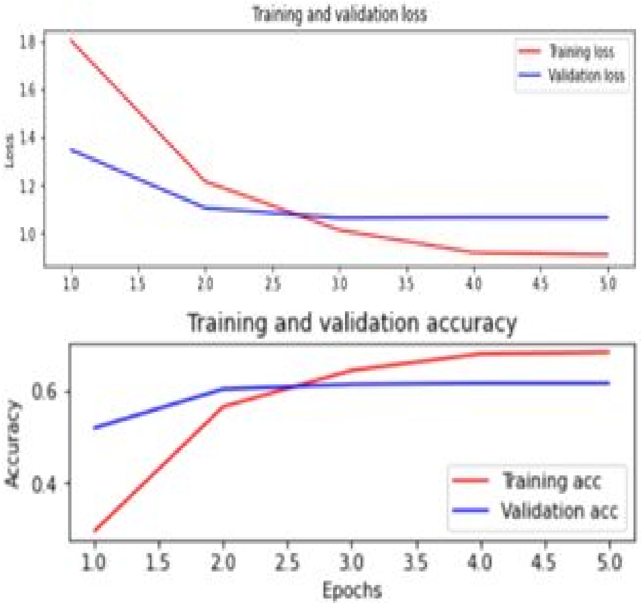


Figure 6: **Line graph showing the change the training and validation accuracy of our fine tuned BERT model for ISER dataset**

Model	Accuracy	Precision	Recall	F1
RNN	48.67	49.24	48.94	48.60
LSTM	58.98	58.83	58.96	58.68
GRU	56.58	57.96	56.64	57.02
BERT	62.03	62.38	62.20	62.00

Table 3: **Table describes the comparison of the BERT model with the baseline models on the ISEAR dataset.**

For the ISEAR (International Survey on Emotion Antecedents and Reactions), our baselines doesn't tend to perform well, it produces an accuracy of 48% but as this dataset is more balanced than the EmotionLines dataset, we get comparatively better precision, recall and f1 score for all our models. LSTM performs better than GRU as well as RNN model and BERT delivers the best performance even for this dataset.

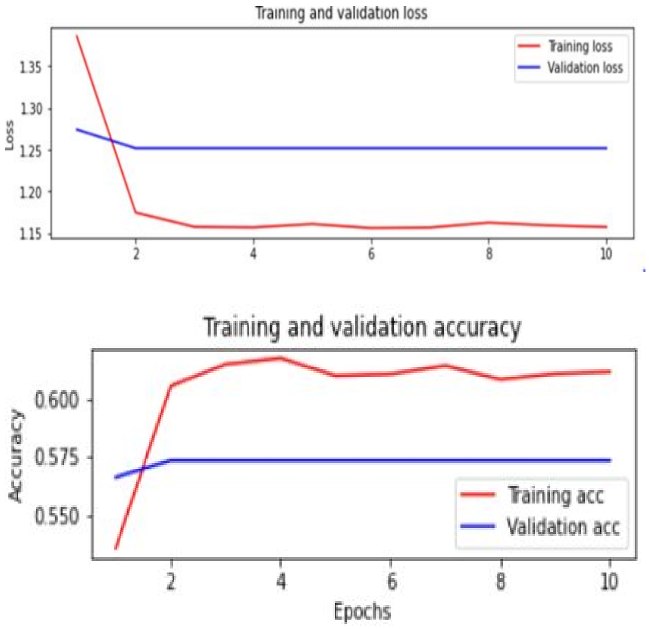


Figure 7: **Line graph showing the change the training and validation accuracy of our fine tuned BERT model for EmotionLines dataset**

7 Conclusion & Future Work

We have successfully created a model to classify the text based emotions into various categories. The model takes advantage of the BERT model to preprocess and encode the data. Also, the model is very fast in terms of training from large datasets. The simplicity, performance and the results achieved by the model makes it competitive with the baselines. Emotion recognition is a challenging task and it gets more challenging as the number of emotions / labels increase. This model satisfactorily classifies emotions for multiple classes upto 8.

The future work can be to incorporate multimodal datasets viz. MELD[21] where multiple modes like audio & visual cues are used along with textual dialogues to classify the emotion. The additional modes would give us more information about the emotion of the utterance. Also, we can fine-tune the BERT model to generate class labels more specific to the dataset. A simple BERT model works wonders, perhaps, a more complex & layered model will perform better. We have just used 2 different datasets, adding more and multi-faceted datasets would really test the model and we could fine-tune the model based on various datasets. Perhaps, if we add more dialogues from other sitcoms as well, we will have a complete, humongous and a utilitarian dataset to completely train & test the emotion classification system.

8 Team Work division

Ansul: Worked on constructing and training models of LSTM and BERT.
Harsh: Worked on dataset research and preprocessing, and RNN model
Aishwarya: Worked on constructing BERT and GRU models, and evaluation metrics.

Everyone worked on the initial research of related papers, as well as the experimentation with different settings of the models.

REFERENCES

[1] Jonathan Herzig, Michal Shmueli-Scheuer, and David Konopnicki. 2017. Emotion Detection from Text via Ensemble Classification Using Word Embeddings. In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '17). Association for Computing Machinery, New York, NY, USA, 269–272. DOI:https://doi.org/10.1145/3121050.3121093

[2] Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A Comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention. In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP'19 Adjunct). Association for Computing Machinery, New York, NY, USA, 63–68. DOI:https://doi.org/10.1145/3314183.3324983

[3] Ameeta Agrawal and Aijun An. 2012. Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '12). IEEE Computer Society, USA, 346–353.

[4] Haji Binali and Vidyasagar Potdar. 2012. Emotion detection state of the art. In Proceedings of the CUBE International Information Technology Conference (CUBE '12). Association for Computing Machinery, New York, NY, USA, 501–507. DOI:https://doi.org/10.1145/2381716.2381812

[5] Patel, Manish. “TinySearch - Semantics based Search Engine using Bert Embeddings.” ArXiv abs/1908.02451 (2019): n. pag.

[6] Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In <i>Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10). Association for Computational Linguistics, USA, 45–53.

[7] Jose Maria Garcia-Garcia, Victor M. R. Penichet, and Maria D. Lozano. 2017. Emotion detection: a technology review. In Proceedings of the XVIII International Conference on Human Computer Interaction (Interacción '17). Association for Computing Machinery, New York, NY, USA, Article 8, 1–8. DOI:https://doi.org/10.1145/3123818.3123852

[8] Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2010. Emotion detection in email customer care. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10). Association for Computational Linguistics, USA, 10–16.

[9] Mary Jane C. Samonte, Hector Irvin B. Punzalan, Richard Julian Paul G. Santiago, and Peter Joshua L. Linchangco. 2017. Emotion detection in blog posts using keyword spotting and semantic analysis. In Proceedings of the

3rd International Conference on Communication and Information Processing (ICCIP '17). Association for Computing Machinery, New York, NY, USA, 6–13. DOI:https://doi.org/10.1145/3162957.3162963

[10] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R. Zaiane. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Comput. Surv. 50, 2, Article 25 (June 2017), 33 pages. DOI:https://doi.org/10.1145/3057270

[11] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2010. Hierarchical versus flat classification of emotions in text. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10). Association for Computational Linguistics, USA, 140–146

[12] Jackson Feijó Filho, Thiago Valle, and Wilson Prata. 2014. Exploring non-verbal communications in mobile text chat: emotion-enhanced chat. In Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14). Association for Computing Machinery, New York, NY, USA, 1069–1072. DOI:https://doi.org/10.1145/2639189.2670278

[13] Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10). Association for Computational Linguistics, USA, 26–34.

[14] Jackson Liscombe. 2006. Detecting emotion in speech: experiments in three domains. In Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: doctoral consortium (NAACL-DocConsortium '06). Association for Computational Linguistics, USA, 231–234. DOI:https://doi.org/10.3115/1225797.1225803

[15] 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 (HLT '05). Association for Computational Linguistics, USA, 579–586. DOI:https://doi.org/10.3115/1220575.1220648

[16] Saif M. Mohammad. 2012. #Emotional tweets. In Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval '12). Association for Computational Linguistics, USA, 246–255.

[17] Jordan S. Huffaker, Jonathan K. Kummerfeld, Walter S. Lasecki, and Mark S. Ackerman. 2020. Crowdsourced Detection of Emotionally Manipulative Language. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376375

[18] Chen, Sheng-Yeh, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao Huang and Lun-Wei Ku. “EmotionLines: An Emotion Corpus of Multi-Party Conversations.” ArXiv abs/1802.08379 (2018): n. pag.

[19] Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66(2), 310–328. https://doi.org/10.1037/0022-3514.66.2.310

[20] Kim, S., Valitutti, A., & Calvo, R. (2010). Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. HLT-NAACL 2010.

[21] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ACL.

[22] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.