Unveiling BERT's Ability to Detect Emotions and Sentiments in Spoken Language: An Experimental Analysis

SHEVCHUK Anna*
ENSAE
anna.shevchuk@ensae.fr

TAVERNIER Léo*
ENSAE
leo.tavernier@ensae.fr

Abstract

Owing to the tremendous improvements made in Deep Learning over the past few years, emotion recognition has become a topic of growing interest. In this respect, we have witnessed a huge development of dialogue generation and conversational systems, with ChatGPT as leading figure. Indeed, owing to its wide spectrum of potential applications, such as dialogue generation and conversational systems, emotion recognition is key to generate an appropriate answer and avoid the "generic response problem", with a view to providing the best possible user experience. By virtue of the widespread access to plug-and-play pre-trained NLP models (e.g., BERT, GPT-4, etc.), implementing conversational systems has never seemed so easy and straightforward. However, under closer scrutiny, deploying an efficient end-to-end conversational system encompasses a fair amount of challenges (e.g., fillers, code-switching, communicative intent identification, etc.) that researchers still endeavor to mitigate. In fact, spoken language and oral interactions are usually scrappy, encompassing fillers and less formal, notably from both grammatical and syntactical perspectives.

*In this paper, we precisely aim at assessing whether current state-of-the-art pretained NLP models that have hit the headlines over the past few years for their unprecedented capabilities, can really cope with spoken language when it comes to speech emotion recognition. More specifically, we deploy **BERT** on a dataset of transcribed oral conversations, to evaluate its performances.*[1]

1 Problem Framing

1.1 Related Work

Owing to its potential applications in many challenging tasks, such as dialogue generation, or user behavior understanding, emotion and sentiment recognition ("E/S", henceforth) is a key feature in implementing a conversational system as it enables to generate appropriate responses to users. However, it is no easy task due to the large number of underlying challenges it encompasses, especially when it comes to spoken dialog. The main challenges are related to the less formal feature of spoken interactions, particularly from both a grammatical and syntactical perspective. Moreover, spoken language usually employs words that can either be rare, slang or have limited representation in written language.

While these elements are the most prominent challenges that come to mind when it comes to spoken interactions, E/S recognition from spoken dialog also yields more complex tasks. First, the recognition of the users' emotion is key to accurately capture its communicative intent. Such task requires to comprehend utterances at both self- and discourse-levels, with a view to grasping nuances of the user, as well as to capture patterns over long ranges of the conversation (Chapuis et al., 2020). This is a key undertaking to better seize emotions and guard against the "generic response problem", which accounts for generating an unspecific response that can be an answer to a very large number of user utterances. To exhaustively capture the complex and hierarchical nature of spoken dialogues, implementing a hierarchical pre-training at multiple levels of granularity has proven to outperform the baselines (*Ibid.*).

Besides, traditional models usually model sequential dependencies between utterances (*i.e.,* between two adjacent utterances), therefore solely

[1] https://github.com/a-shevchuk/NLP_intent_classifier_SHEVCHUK_TAVERNIER

apprehending local dependencies. To mitigate this pitfall, (Colombo et al., 2020) proposed to enhance the *seq2seq* model, which rests upon the encoder-decoder framework, with a hierarchical attention mechanism that enabled to capture global dependencies in a dialog as well. This approach achieved outperforming results compared to previous modelling approaches.

Spoken dialog systems also often struggle with code-switching, the fact of associating more than one language within a single utterance or conversation. As it is a common feature of many multilingual communities around the world, an efficient conversational AI cannot escape learning multilingual spoken dialog representations is therefore key to designing an efficient spoken dialog system. Working with both language-specific and multilingual tokenizers, (Chapuis, 2020) developed a new set of loss functions that are inspired by code-switching patterns and designed to better capture the nuances of mixed-language speech. It establishes a milestone towards making spoken dialog systems more effective for multilingual communities.

In addition, fillers, that can be considered as onomatopoeia placed within a spoken dialogue in a episodic fashion, are also a hot topic in NLP modelling. Indeed, fillers are usually considered as an ugly duckling, as they are deemed to convey poor – or even zero – information and are consequently discarded during the data pre-processing phase. Furthermore, as most NLP models are pre-trained on written text, they usually yield poor representations of spontaneous speech words (Barriere et al., 2017). A large body of literature has nonetheless emphasized that fillers play a pivotal role in spoken language (Clark, 2002), and convey key information on the speaker's utterance structure (Dinkar et al., 2020), stance (Grezause, 2017) and commitment to a statement (Smith and Clark, 1993). By mean of deep contextualized embeddings, (Dinkar et al., 2020) demonstrated that modelling fillers can improve the accuracy of spoken languages, both when modelling language and on a downstream task. Most-advanced conversational AI should therefore leverage on modelling fillers to capture important context features and valuable information to improve their performances. These modelling approaches are the current SOTA ones to be implemented when it comes to improving spoken dialog systems. Within the framework of this paper, we will endeavor to evaluate whether SOTA **BERT** is suitable for E/S recognition on spoken language, notably encompassing fillers, slang as well as colloquial utterance structures.

1.2 Data Presentation

Trough this paper, we aim at assessing how well current SOTA NLP framework – **BERT** in this specific instance – perform on spoken language. Indeed, as previously emphasized, spoken language is emancipated from many coercive rules that written languages are usually subjected to. Furthermore, emotions are intrinsic to humans and are associated with an individual's mental state, thoughts and feelings (Poria et al., 2019). As a consequence, even the best performing NLP models find it hard to decipher speakers' emotions, as they are mostly trained on written corpora. With a view to evaluating **BERT** performance on oral conversations, we relied on the *Multimodal EmotionLines Dataset*[2] (henceforth referred to as "MELD"), which contains about 13,000 utterances from 1,433 dialogue transcripts from the TV-series *Friends*. MELD dataset encompasses multiple speakers within each dialogues, and each utterance in a dialogue has been labeled with an emotion[3] as well as a sentiment (*i.e.,* Positive, Negative, Neutral).

Our dataset can be represented as follows: $D = \{C_1, C_2, ..., C_{|D|}\}$, where C_i accounts for conversation i. Each C_i comprises a certain number of utterances u, defined as follows: $u = \{u_1, u_2, ..., u_{|C_i|}\}$, with $Y = \{Y_1, Y_2, ..., Y_7\}$, the corresponding set of emotion labels and $Y^{'} = \{Y_1^{'}, Y_2^{'}, Y_3^{'}\}$, the corresponding set of associated sentiments. As a result, each utterance u_j is associated with both a unique emotion y_k and sentiment $y_{l'}$ labels. Finally, each utterance u_j is composed of a sequence of words or tokens, such that: $u_j = \{w_1^j, w_2^j, ..., w_{|u_j|}^j\}$.

Throughout this paper, our approach is therefore to operate a classification task, aiming at affiliating each utterance with its associated emotion and sentiment at utterance level.

[2] This dataset is available on the following GitHub repository: https://github.com/declare-lab/MELD

[3] As per Paul Ekman (1984), there exists six cardinal emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. These emotions are the ones used to label each utterance of our dataset

1.3 Overview of the Dataset and its Main Features

As we were limited by our computational power (*i.e.,* limited access to GPU), we have relied on a subset of the MELD dataset, that contains 9,989 utterances.

We first operated a data-processing phase, consisting in cleaning and reorganizing our dataset for efficiency purpose. As a result, we deployed a certain number of pre-processing steps. First, we considered that removing stopwords[4] was a key pre-processing step to undertake, as their removal does not jeopardize the overall meaning of the utterance. However, we do not considered removing punctuation marks, as we assume they convey much information about the underlying emotion of the speaker, especially in a transcribed version of an oral conversation. In addition, we also removed dashes and some types of single quotes, and eventually kept only utterances that comprise at least three words, assuming that utterances with fewer words convey poor information when it comes to detecting a speaker's emotion and add noise to the data, thus downgrading the model's ability to accurately disentangle the underlying pattern of the data. Finally, we ended up with a dataset encompassing a total of 6,843 utterances, thus removing 3,146 utterances in total. The main features of our sanitized dataset are presented hereinbelow:

	neutral	surprise	fear	sadness	joy	disgust	anger
Max	30	17	17	40	23	19	22
75th	8	7	9	8	8	9	9
Median	5	5	6	6	6	7	6
Avg	6	5	6	6	6	7	6
25th	4	3	4	4	4	4	4
Min	3	3	3	3	3	3	3
StDev	3	2	3	3	3	3	3
Words count	19725	3743	1476	3813	7845	1477	5530
# Utt.	3149	660	217	560	1217	206	834

Table 1: Descriptive Statistics of MELD Dataset

We also provide the distribution of data based on emotion labels (*cf.* Figure 1), as well as the frequency of the ten most frequent words in the corpus (*cf.* Figure 2). We notice that our data are highly skewed towards the Neutral and Surprise classes.

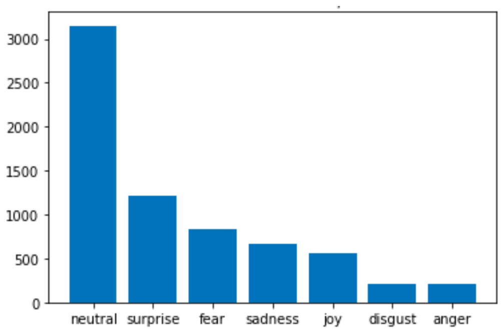


Figure 1: Emotion Frequency in MELD Dataset

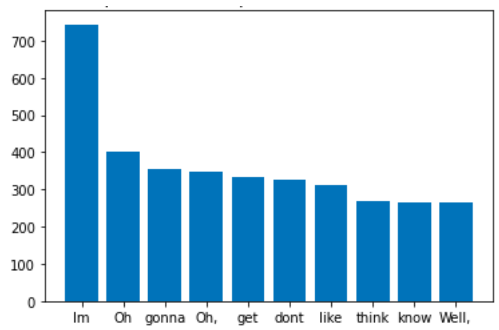


Figure 2: Ten Most Frequent Words in the Corpus

We also considered that scrutinizing the sentiment distribution, by way of the *Polarity Score*, could be useful to better grasp the underlying pattern or our dataset. We can observe that the calculated *Polarity Score* is consistent with the overall distribution of emotions and sentiments within our dataset, as the distribution is highly leptokurtic, with most of its mass centered around zero, which precisely corresponds to the over-represented Neutral emotion/sentiment.

We also computed the correlation matrix (*cf.* Figure 8 in Appendix section) between all numerical variables: Emotion, Sentiment, Words Count and *Polarity Score*. It is worth observing that the *Polarity Score* is positively correlated with sentiments, which appears congruent as it is precisely what such a score endeavors to capture. In addition, emotions and sentiments are also positively correlated, which also appears consistent, as these concepts are relatively entangled. On the contrary,

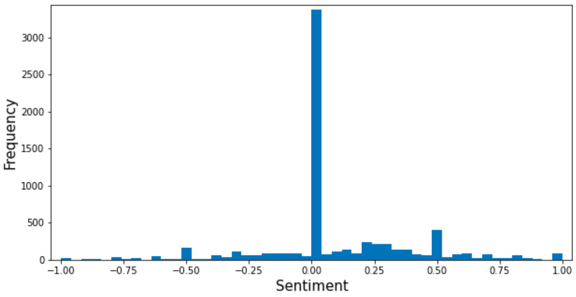


Figure 3: Polarity Score over the MELD Dataset

[4] A set of commonly used words in any language, usually considered as weak information providers.

the words count is sightly negatively correlated with the both emotions and sentiments. In Appendix, we also provide a word cloud depicting the most frequent 1-grams (*i.e.,* words) in our corpus, where the size of the word is relative to its frequency (*cf.* Figure 9 in Appendix section).

2 Experiment Protocol

2.1 Data Encoding

Our work aims at evaluating whether the state-of-the-art **BERT** framework performs well on oral speech. To achieve this goal, we first converted all utterances into word embeddings (*i.e.,* numerical vectors), relying on the uncased `BertTokenizer` (*i.e.,* insensitive to case) available in `Python`. For classification purpose, we also converted both emotion and sentiment labels to numerical representations. Thus, emotions are now represented by $Y = \{0, ..., 6\}$, while sentiment by $Y^{'} = \{0, 1, 3\}$. Our dataset can now be represented as $\mathcal{D} = \{(X_1, Y_1, Y_1^{'}), (X_2, Y_2, Y_2^{'}), ..., (X_N, Y_N, Y_N^{'})\}$, where N is its size and $X_i \in R^{256}$. We then converted our features and labels into tensors by using the `PyTorch` library, which is a data structure commonly used by Deep Learning algorithms.

2.2 Model

We rely on the pre-trained `BertForSequenceClassification`, a bidirectional transformer pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the *Toronto BookCorpus* and Wikipedia. **BERT** encompasses one embedding layer, twelve transformers and one output layer. As our problem encompasses seven classes for emotion recognition and three classes for sentiment analysis, we consequently fine-tuned **BERT** to match our classification problem.

2.3 Hyperparameters tuning

Table 2 summarizes the value of hyperparameters that we chose prior to deploying the model. We also consider the AdamW optimizer, widely used in Deep Learning implementation, that combines computational efficiency, momentum and weight decay regularization (see next subsection for further details).

Learning Rate	Optimizer Epsilon	Epochs	Batch Size
2e-5	1e-8	5	16

Table 2: Hyperparameters setting

2.4 AdamW Optimizer

We use the AdamW optimizer to implement the model previously exposed, as it is largely used in the existing literature. AdamW is a stochastic optimization method (Loshchilov and Hutter, 2017) that modifies the typical implementation of weight decay in Adam, by decoupling weight decay from the gradient update.

To see this, L_2 regularization in Adam is usually implemented with the below modification, where f is the objective function, θ_t the parameters at time t and w_t the rate of the weight decay at t:

$$g_t = \nabla f(\theta_t) + w_t\theta_t,$$

while AdamW adjusts the weight decay term to appear in the gradient update:

$$\theta_{t+1,i} = \theta_{t,i} - \eta\Big(\frac{1}{\sqrt{\hat{v}_t + \epsilon}} \cdot \hat{m}_t + w_{t,i}\theta_{t,i}\Big), \forall t,$$

with η the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ the first and second moments estimated at time t respectively.

3 Results

In this section, we present and compare the results obtained by applying the model previously exposed to MLDS emotions versus sentiments.

Regarding emotions, the training loss behaves well and decreases with the epochs (Figure 4). After the testing phase, we obtain a model accuracy of 0.57, *i.e.,* the model predicts the right label *ca.* half of the time only, making it more efficient than a random classifier on very meager scale. A confusion matrix is displayed in Figure 5, as well as a classification report in Figure 10 in Appendix – quite intuitively, the best predictions come out for the neutral emotion as it is overrepresented in the dataset compared to the other ones.

In comparison, when applying the model to sentiments, the training loss after five epochs is *ca.* 2 times smaller (Figure 6), while the accuracy is of 0.64 (see Figure 7 for confusion matrix, and Figure 11 in Appendix for classification report). As for the emotions, the model better predicts the neutral sentiment.

A comparative table is displayed in Table 3, which shows that the model applied to sentiments is relatively more efficient in view of both the accuracy

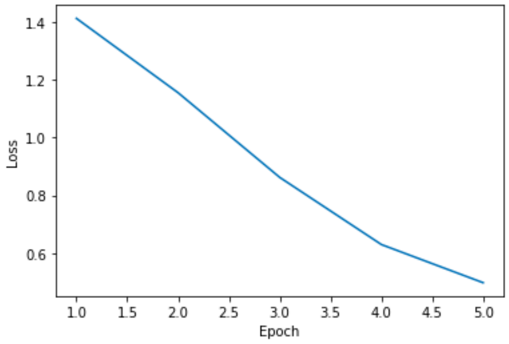


Figure 4: Training Loss on Emotions

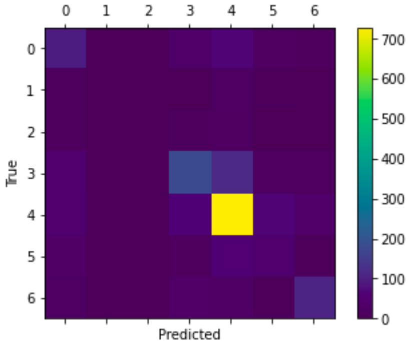


Figure 5: Emotions – Confusion Matrix
Anger: 0 – *Disgust*: 1 – *Fear*: 2 – *Joy*: 3 – *Neutral*: 4 – *Sadness*: 5 – *Surprise*: 6

and the loss. It is a result that one could easily anticipate as the dataset contains only three sentiments but seven emotions, which makes the classification task considerably more complicated. In absolute terms, the model produces contrasting results, whatever we apply it on sentiments or emotions, which is partly attributable to the fact that **BERT** was pre-trained on written corpora as opposed to oral speech that we use here. A few improvement areas are discussed in the next section.

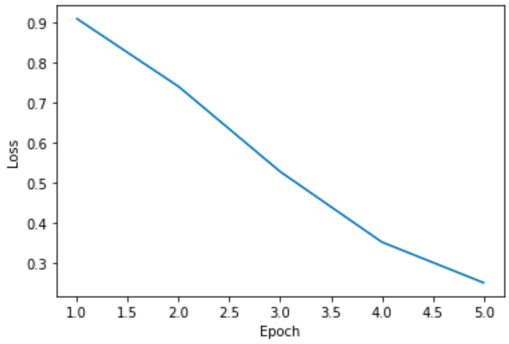


Figure 6: Training Loss on Sentiments

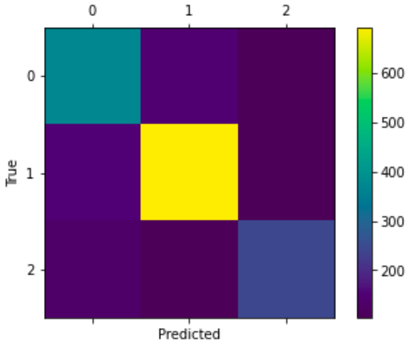


Figure 7: Sentiments – Confusion Matrix
Negative: 0 – *Neutral*: 1 – *Positive*: 2

	Training Loss after five Epochs	**Model Accuracy**
Emotions	0.497	0.566
Sentiments	0.271	0.636

Table 3: Emotions and Sentiments Training Loss and Model Accuracy comparison

4 Discussion

4.1 Results Discussion

In spite of **BERT** tremendous capabilities in NLP, our approach yielded mixed results in terms of accuracy standpoint. As regards emotion recognition, we managed to beat a random classifier by only a tiny margin. It is not much of a surprise, as **BERT** was trained on written corpora, which significantly differ from spoken languages. However, such a low accuracy score raises certain questions. Utterances are obviously associated with their own emotions, but when considering a dialog, context is key. Indeed, a dialogue or a conversation have an overall pattern, meaning that only a restricted sample of emotions can materialize. As a matter of fact, it is very unlikely that an interlocutor experiences all types of emotions in a single conversation, except in some specific cases, where many (or even all) different emotions cohabit. As a result, we assume that a given conversation possesses an overall pattern that, when taken into account, could precisely help a model to better decipher emotions at utterance-level. In this respect, modelling approaches that take global dependencies have precisely achieved the best results so far (Colombo et al., 2020; Bothe et al., 2018). In further work, it would consequently be a very interesting topic to scrutinize and implement.

4.2 Extension: Broadening the Perspective

First, it worth emphasizing that data quantity is of paramount importance when training models on complex patterns, especially when relying on neural network architectures. However, we were highly constrained by our computational power, with limited access to GPU. Thus, the first development that would naturally considered would be to increase the dataset size to assess whether it yields better performances. Fine-tuning **BERT** (*e.g.,* changing the loss function, etc.) also appears as a key upstream improvement procedure.

Secondly, even if our losses demonstrated a good overall behavior during training, *i.e.,* converging towards zero, it would be insightful to augment the number of epochs and subsequently evaluate the generalization capacity of a better trained model on unseen data. This would also require higher computational capacities.

Third, implementing our approach on other datasets could also help understand whether the results are specific to our data, or if the model better performs on other datasets, all the more so as our dataset is highly skewed towards one specific emotion (*i.e.,* Neutral). It would also be interesting to deploy the same methodology on Dialog Act (”D/A” henceforth) classification, or even trying to predict both simultaneously, to appraise whether D/A can provide information on E/S, the other way round.

Furthermore, implementing approaches that take global dependencies into account appear key in this type of classification assignments, as both the underlying content and the overall context of a conversation convey, without a shadow of a doubt, key information on the emotions it encompasses.

5 Conclusion

To summarize, our approach aimed at assessing whether SOTA **BERT** could perform well on oral utterances, while having been trained on written corpora. Indeed, the main challenge is that oral interactions, precisely owing to their oral and informal dimension, can strongly differ from written language. As a matter of fact, spoken language usually pays no heed to formalism, employing words that can either be rare, slang or have limited representation in written language. Moreover, spoken language can also heavily rely on redundant words, such as ”like” in English or ”du coup” in French, that more formal sentence structures dot not manipulate that much. Besides, written language abides by a stranglehold of – usually – coercive rules (*i.e.,* grammatical, syntactic, etc.), from which oral exchanges usually get away. In addition, code-switching and fillers make it all the more harder, even for the best NLP models, to properly decipher the speaker’s intent. At utterance level, relying on the MELD dataset, we managed to achieve a 57% accuracy on emotion during the validation phase, making it more efficient than a random classifier on a very marginal basis. When it comes to sentiment classification, we however achieved a 64% accuracy on validation data, which is rather encouraging. From an accuracy standpoint, higher accuracy results as regards sentiment classification appears congruent with the fact that this an easier pattern to identify, as it only encompasses three classes. Managing to incorporate code-switching, fillers by way of deep neural contextualized embeddings and especially global dependencies, would arguably yield significantly higher performances, and be consistent with current SOTA approaches in this respect.

References

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38.

Treeb Clark. 2002. Using uh and um in spontaneous speaking. *https://doi.org/10.1016/S0010-0277(02)00017-3*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Essid Barriere et al., Clavel. 2017. Opinion dynamics modeling for movie review transcripts classification with hidden conditional random fields. *arXiv:1806.07787v1*.

Esther Le Grezause. 2017. Um and uh, and the expression of stance in conversational speech. *HAL Id: tel-02069026*.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. *CoRR*, abs/1805.06280.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances.

Labeau Clavel Chapuis, Colombo. 2020. Code-switched inspired losses for generic spoken dialog representations. *arXiv:2108.12465*.

Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *CoRR*, abs/2002.08801.

Matteo Manica Matthieu Labeau Chloe Clavel Chapuis et al., Pierre Colombo. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv:2009.11152*.

Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. pages 7985–7993.

Appendix

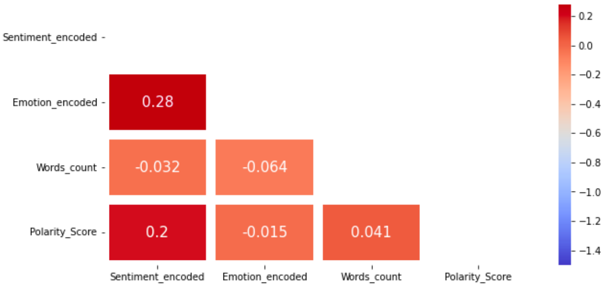


Figure 8: Correlation Matrix between Numerical Variables

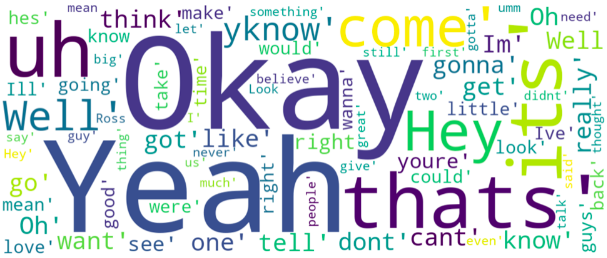
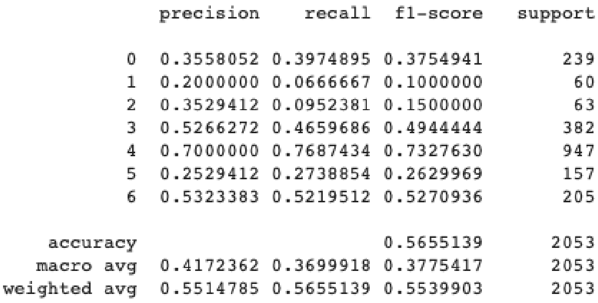


Figure 9: World Cloud on sanitized MELD Dataset

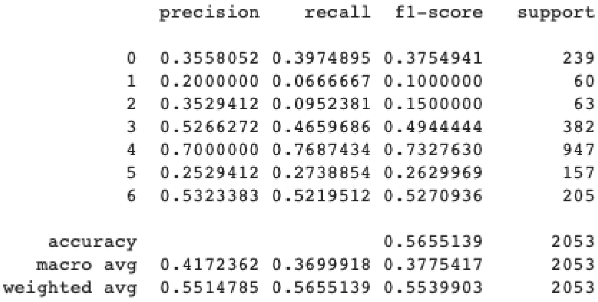


```
              precision    recall  f1-score   support

           0  0.3558052 0.3974895 0.3754941       239
           1  0.2000000 0.0666667 0.1000000        60
           2  0.3529412 0.0952381 0.1500000        63
           3  0.5266272 0.4659686 0.4944444       382
           4  0.7000000 0.7687434 0.7327630       947
           5  0.2529412 0.2738854 0.2629969       157
           6  0.5323383 0.5219512 0.5270936       205

    accuracy                      0.5655139      2053
   macro avg  0.4172362 0.3699918 0.3775417      2053
weighted avg  0.5514785 0.5655139 0.5539903      2053
```

Figure 10: Emotions – Classification Report *Anger*: 0 – *Disgust*: 1 – *Fear*: 2 – *Joy*: 3 – *Neutral*: 4 – *Sadness*: 5 – *Surprise*: 6



```
              precision    recall  f1-score   support

           0  0.3558052 0.3974895 0.3754941       239
           1  0.2000000 0.0666667 0.1000000        60
           2  0.3529412 0.0952381 0.1500000        63
           3  0.5266272 0.4659686 0.4944444       382
           4  0.7000000 0.7687434 0.7327630       947
           5  0.2529412 0.2738854 0.2629969       157
           6  0.5323383 0.5219512 0.5270936       205

    accuracy                      0.5655139      2053
   macro avg  0.4172362 0.3699918 0.3775417      2053
weighted avg  0.5514785 0.5655139 0.5539903      2053
```

Figure 11: Sentiments – Classification Report *Negative*: 0 – *Neutral*: 1 – *Positive*: 2