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

Anna Shevchuk*
ENSAE, Institut Polytechnique de Paris
anna.shevchuk@ensae.fr

Léo Tavernier*
ENSAE, Institut Polytechnique de Paris
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 Chat-GPT 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. As a matter of fact, spoken language and oral interactions are usually scrappy as well as less formal, notably from both grammatical and syntactical perspectives.*

*Throughout this paper, we precisely aim at assessing whether current state-of-the-art pre-trained NLP models, that have recently been hitting the headlines due to their unprecedented capabilities, can efficiently cope with spoken language. More specifically, we deploy **BERT** on a dataset of transcribed oral conversations, to evaluate its performances.*[1]

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

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 when it comes to implementing a conversational system, as it enables to generate appropriate responses to users. However, it is no easy task due to the myriad of challenges it encompasses, especially when it comes to spoken dialogue. 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 dealing with spoken interactions, E/S recognition from spoken dialogue also yields more complex tasks. First, the recognition of users' emotions is key to accurately capture their communicative intent. Such a task requires to comprehend utterances at both self- and discourse-levels, with a view to grasping the user's nuances and capturing patterns over longer ranges of the conversation (Colombo* et al., 2020; Garcia* et al., 2019). 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 wide range of user utterances (Colombo* et al., 2019; Colombo, 2021). 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 (Chapuis* et al., 2020; Colombo et al., 2021).

Besides, traditional approaches usually model sequential dependencies between utterances (*i.e.,* between two adjacent utterances), therefore solely 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 dialogue as well. This approach achieved outperforming results compared to previous modelling approaches.

More often than not, spoken dialogue systems also struggle with code-switching, which accounts for the fact of combining 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 fend off learning multilingual spoken dialogue representations. Working with both language-specific and multilingual tokenizers, (Chapuis* et al., 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 speeches. It established a milestone towards making spoken dialogue 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 have usually been 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 corpora, they usually yield poor representations of spontaneous speech words (Dinkar* et al., 2020). A large body of literature has nonetheless emphasized that fillers play a pivotal role in spoken language (Dinkar* et al., 2020), and convey key information on the speaker's utterance structure, stance (Grezause, 2017) and commitment to a statement (Smith and Clark, 1993). By dint of deep contextualized embeddings, it was 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 dialogue systems. Within the framework of this paper, we endeavor to evaluate whether SOTA **BERT** (Devlin et al., 2018) is suitable for E/S recognition on spoken language, which notably encompasses fillers, slang and colloquial utterance structures.

1.2 Data Presentation

Trough this paper, we aim at assessing how well current SOTA NLP frameworks – **BERT** in this specific instance – can perform on spoken language. As previously emphasized, spoken language is indeed 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"). Several alternatives exist to MELD (Li et al., 2017; Colombo, 2021; Shriberg et al., 2004; Mckeown et al., 2013; Busso et al., 2008), but MELD is the largest one. It contains about 13,000 utterances from 1,433 dialogue transcripts from the *Friends* TV-series. The MELD dataset encompasses multiple speakers within each dialogue, and each dialogue utterance 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 conversation 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 sentiment labels. Thus, each utterance u_j is associated with both a unique emotion label y_k and sentiment label y_l'. Finally, each utterance u_j is composed of a sequence of words or tokens, such as: $u_j = \{w_1^j, w_2^j, ..., w_{|u_j|}^j\}$.

[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, Surprise and Fear. In conjunction with these six emotions, a seventh one – Neutral – is also used to label each utterance within the MELD dataset.

Throughout this paper, our approach therefore consists in accomplishing a classification task, which aims to affiliate each utterance with its associated emotion and sentiment.

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 relied on a subset of the MELD dataset, which encompasses 9,989 utterances.
We first conducted a data-processing phase, which consisted in cleaning and reorganizing our dataset for efficiency purpose. To this end, we undertook a certain number of pre-processing steps. First and foremost, we considered that removing stop-words[4] was a key pre-processing step, as their removal does not jeopardize the overall meaning of the utterance. However, we did not consider 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 retained only utterances that comprised at least three words. Indeed, we assumed that utterances with fewer words convey poor information when it comes to detecting a speaker's emotion and also 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 that encompasses 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
Word 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 the data based on emotion labels (*cf.* Figure 1), as well as the frequency of the ten most recurrent words (*cf.* Figure 2). We notice that our data are highly skewed towards the Neutral and Surprise classes.

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

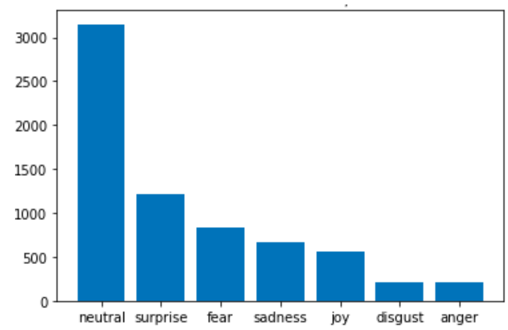

Figure 1: Emotions Frequency in MELD Dataset

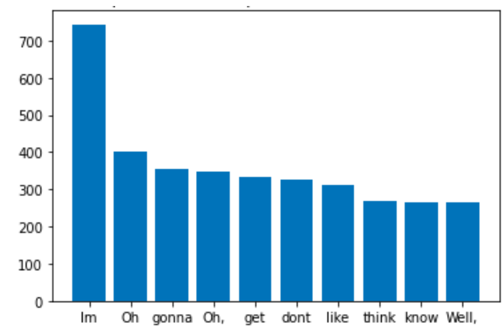

Figure 2: Ten Most Frequent Words in the Corpus

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

We also computed the correlation matrix between all numerical variables: Emotion, Sentiment, Word Count and *Polarity Score* (*cf.* Figure 8 in Appendix section). It is worth emphasizing that the *Polarity Score* is positively correlated with sentiments, which appears congruent as it is

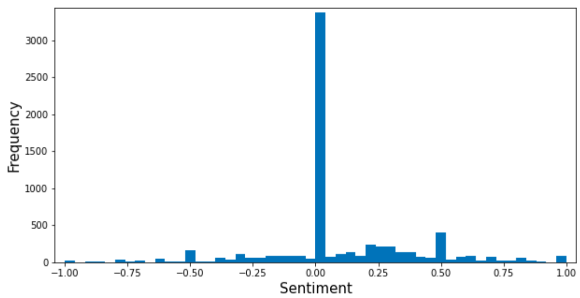

Figure 3: Polarity Score over the MELD Dataset

precisely what such a score endeavors to capture. Moreover, emotions and sentiments are positively correlated as well, which also appears consistent as these concepts are relatively entangled. On the contrary, the word count is sightly negatively correlated with 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`, which generates a vector of size 256. For classification purpose, we also converted both emotion and sentiment labels to numerical representations. Thus, emotions can now be represented by the following set $Y = \{0, ..., 6\}$ and sentiment by the following one: $Y^{'} = \{0, 1, 3\}$. Our dataset can henceforth be represented as follows: $\mathcal{D} = \{(X_1, Y_1, Y_1^{'}), (X_2, Y_2, Y_2^{'}), ..., (X_N, Y_N, Y_N^{'})\}$, where N accounts for the total number of samples and $X_i \in \mathbb{R}^{256}$. 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 relied on the `BertForSequenceClassification` model, 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** comprises 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.

Learning Rate	Optimizer Epsilon	Epochs	Batch Size
2e-5	1e-8	10	16

Table 2: Hyperparameters setting

We also considered the AdamW optimizer (*cf.* subsection 2.3), widely used in Deep Learning implementations, which combines computational efficiency, momentum and weight decay regularization.

2.3 AdamW Optimizer

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:

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

where η accounts for the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ represent 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 MELD dataset, on both emotions and sentiments.

As regards emotions, the training loss behaves in a satisfying way, decreasing as the number of epochs increases (Figure 4). After the testing phase, we obtained an accuracy of 0.52, *i.e.,* the model predicts the right label *ca.* half of the time only, outperforming 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 accuracy is better the neutral emotion as it is over-represented within the dataset.

Comparatively, when applying the model to sentiments, the training loss after ten epochs reaches 0.04 (Figure 6), with an accuracy of 0.61 (see Figure 7 for confusion matrix, and Figure 11 in Appendix for classification report). As for emotions, the model better predicts the neutral sentiment.

A comparative table is displayed in Table 3, which evidences that the model applied to sentiments is

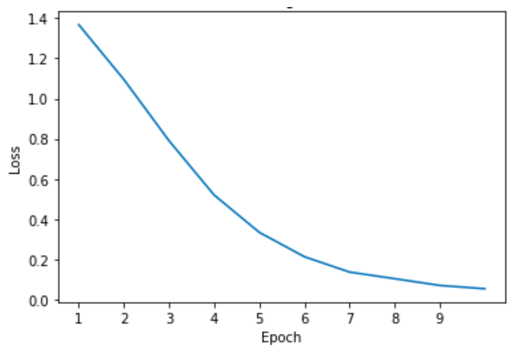

Figure 4: Training Loss on Emotions

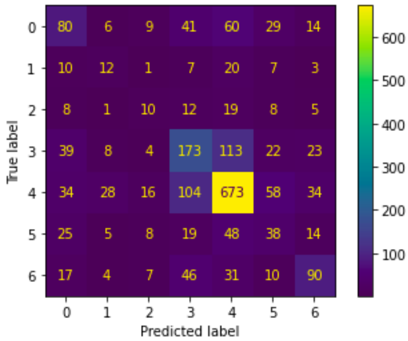

Figure 5: Confusion Matrix on Emotions
Anger: 0 – Disgust: 1 – Fear: 2 – Joy: 3 – Neutral: 4 – Sadness: 5 – Surprise: 6

relatively more efficient in view of both the accuracy and the loss, yet in a marginal fashion. It is a result that one could have easily inferred, as the dataset contains only three sentiments but seven emotions, making the classification task harder with emotions. In absolute terms, the model produces contrasting results whatever the classification output, which is partly attributable to the fact that **BERT** was pre-trained on written corpora, which strongly contrasts with oral utterances used within the framework of this paper. A few improvement areas are discussed in the next section.

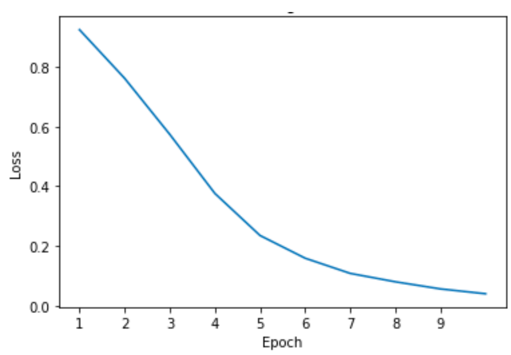

Figure 6: Training Loss on Sentiments

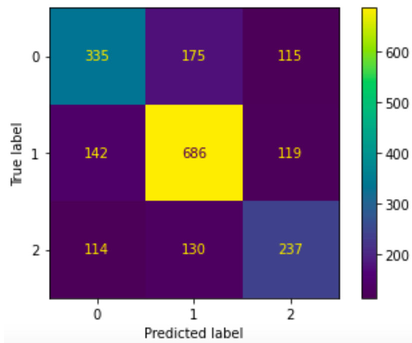

Figure 7: Confusion Matrix on Sentiments
Negative: 0 – Neutral: 1 – Positive: 2

	Training Loss after ten Epochs	Model Accuracy
Emotions	0.053	0.524
Sentiments	0.039	0.613

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. As regards emotion recognition, we managed to outperform a random classifier in a very marginal fashion. This is not much of a surprise, as **BERT** was trained on written corpora, which significantly differ from spoken language. However, such a low accuracy score raises certain questions. Utterances are obviously associated with their own emotions, but when considering a dialogue, context is key. Indeed, a dialogue or a conversation has an overall pattern, which entails 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 could 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 is 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 we would naturally consider would be to increase the size of the dataset to assess whether it would yield better performances. Fine-tuning **BERT** (*e.g.,* changing the loss function, etc.) also appears as a key procedure for performance enhancement.

Secondly, although our losses demonstrated a satisfying overall behavior during training, rapidly 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. However, 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 can better perform 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 Dialogue Act (”D/A” henceforth) classification, or even investigate predicting both simultaneously, to appraise whether D/A can provide information on E/S, as well as the other way round.

Furthermore, implementing approaches that take global dependencies into account appear key for this type of classification tasks, 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

In a nutshell, our approach aimed at assessing how well could SOTA **BERT** perform on oral utterances, while having been trained on written corpora. Indeed, the main challenge lies in the fact that oral interactions, precisely owing to their oral and informal dimensions, 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, and 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 are usually emancipated. In addition, code-switching and fillers make it all the more difficult to properly decipher a speaker’s intent, even for the best NLP models. At utterance level, relying on the MELD dataset, we managed to achieve a 52% 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 61% accuracy on validation data, which is rather encouraging. From a sentiment standpoint, better accuracy results appear congruent with the fact that this is an easier pattern to identify, as it only encompasses three classes.

Consequently, managing to incorporate code-switching, fillers by way of deep neural contextualized embeddings, as well as taking global dependencies into account would arguably yield improved 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.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

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

Alexandre Garcia*, Pierre Colombo*, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. *EMNLP 2019*.

Pierre Colombo*, Emile Chapuis*, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. *AAAI 2020*.

Emile Chapuis*, Pierre Colombo*, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *Finding of EMNLP 2020*.

Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Code-switched inspired losses for spoken dialog representations. In *EMNLP 2021*.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Appendix

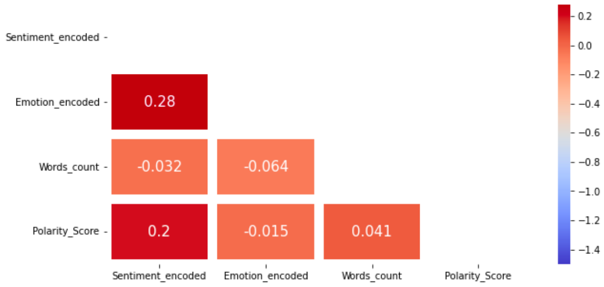

Figure 8: Correlation Matrix between Numerical Variables

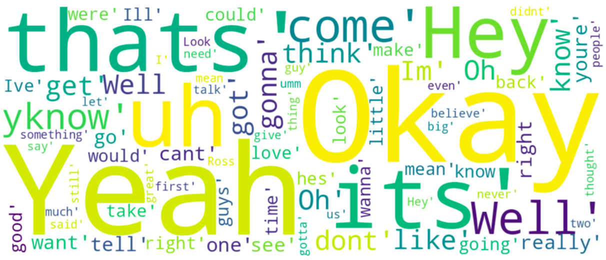

Figure 9: World Cloud on sanitized MELD Dataset

	precision	recall	f1-score	support
0	0.3755869	0.3347280	0.3539823	239
1	0.1875000	0.2000000	0.1935484	60
2	0.1818182	0.1587302	0.1694915	63
3	0.4303483	0.4528796	0.4413265	382
4	0.6981328	0.7106653	0.7043433	947
5	0.2209302	0.2420382	0.2310030	157
6	0.4918033	0.4390244	0.4639175	205
accuracy			0.5241111	2053
macro avg	0.3694457	0.3625808	0.3653732	2053
weighted avg	0.5228935	0.5241111	0.5230702	2053

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

	precision	recall	f1-score	support
0	0.567	0.536	0.551	625
1	0.692	0.724	0.708	947
2	0.503	0.493	0.498	481
accuracy			0.613	2053
macro avg	0.587	0.584	0.586	2053
weighted avg	0.610	0.613	0.611	2053

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