Emotion Analysis on OpenSubtitle

Stanislas Furrer*,
School of Engineering (STI)
Ecole polytechnique federale de Lausanne (EPFL)
Lausanne, Switzerland
Email: *stanislas.furrer@epfl.ch,

***Abstract*—There is a resurgent interest in developing intelligent open-domain dialog systems due to the availability of large amounts of conversational data and the recent progress on neural approaches to conversational bot. Building open domain conversational systems that allow users to have engaging conversations on topics of their choice is a challenging task especially for multi-turn settings. Movie and TV subtitles are naturally a good source for developing conversation corpora. Currently the biggest corpus is the Open Subtitles dataset. However, subtitle files usually lack clear scene markers, making it difficult to extract self-contained dialogs used for training multi-turn dialog models. Lison and Meena (2016) [1] has present a data-driven approach to the segmentation of subtitles into dialogue turns. In this paper, we manually segmented the Open Subtitle data set into dialogue turns and create a speaker-aligned Dataset of 35k conversation. On this novel Dataset we use a pre-trained BERT model to label the dialogues with emotions. Finally, We will compare our result by reproducing the paper of Lison and Meena, matching the dialogs with our clean subset and applying the same emotion classifier.**

I. Introduction

Building intelligent open-domain dialog systems that can converse with humans coherently and engagingly has been along-standing goal of artificial intelligence (AI) [15]. A dialogue system requires a large amount of data to learn meaningful features and response generation strategies for building an intelligent conversational agent. Unlike traditional task-oriented bots which are concentrated on a specific domain or area of knowledge, the training Dataset for a chat-oriented dialogue system must cover a wide variety of domains, as well as be able to provide a fair representation of world-knowledge semantics and pragmatics [16]. Movie and TV subtitles are naturally a good source for developing such conversation corpora. In the recent years, some valuable movie subtitle open-domain resource as been developed such as OpenSubtitles [5], Cornell Movie-Dialogue Corpus [17], Movie-DiC [18] and Movie-Triples [19].

In this paper we investigate the use of user-contributed movie subtitles as a source of emotion analysis. We base our study on the OpenSubtitles corpus (Tiedemann et al 2016 [5]) and restore a reliable turn segmentation for a subset of dialogues on which we apply our emotional classifier.

The remainder of this paper is organized as follows : In section 2 we briefly introduce the main parameter of the emotion analysis tools we will use. The OpenSubtitles data set is presented in section 3. We highlight the drawback of the initial block structure and introduce a dialogue-based version of the data set. However, the data set is lacking of a valid turn segmentation. Lison and Meena (2016) [1] has tried to address this problem by publishing an Automatic turn segmentation of the data set. In section 4 we reproduce their paper and briefly discuss about the heuristic. Then we introduce our manual segmentation in the experiments and results. We will show that our heuristic is speaker based. Then we will evaluate the properties of our dialogues with the one from the automatic subset. Finally, we will emotionally classified the subsets and compare the results.

II. Emotion Analysis Tools

Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, attitudes, and emotions via the computational treatment of subjectivity in text. The spectrum of sentiment analysis techniques ranges from identifying polarity (positive or negative) to a complex computational treatment of subjectivity, opinion and sentiment. There are two broad approaches for calculating the sentiment of a text document : rule-based and machine-learning based.

In the following part we will present an overview of Vader, a rule-based approaches that compute the strength of the sentiment expressed in text, and EmoBert a sophisticated emotion classifier developed in the HCI Laboratory at EPFL.

A. Vader

Vader (Valence Aware Dictionary and Sentiment Reasoner) is a higher performing lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Hutto and Gilbert [7] present it as a sentiment intensity polarizer. Vader takes a sentence as input and provides an overall polarity score of the sentence. It uses a lexicon driven approach as well as additional heuristics for rating the input. Since VADER is not a machine learning approach it does not suffer from a speed-performance trade off due to the train on the data.

To build their model the author has gather lexical features of established sentiment lexicons like LIWC, ANEW and GI and include heuristics that can shift or boost the sentiment of a sentence. These heuristics include punctuation marks, capitalization, booster words (negative and positive, e.g.words like "amazingly"), contrasting conjunctions like "but" and preceding Trigram. When a sentence is being rated these keywords are identified and can shift or impact the rating.

Vader as shown great result on social media style text, yet readily generalizes to multiple domains.

B. EmoBert

EmoBERT is a BERT transformer based single-sentence emotion classifier built as suggested in [9]. It consists of both a representation network and a classification network. During training, the representation network was first initialized with weights from the pre-trained language model, RoBERTA [10], and the model was fine-tuned on situation descriptions given in the EmpatheticDialogues dataset [11] tagged with 32 emotions and listener utterances tagged with 8 response intents plus neutral. The training, validation, and test sets comprised of 25,023, 3,544 and 3,225 sentences respectively, which spanned more or less equally across all the emotion and intent categories. The top-1 accuracy of the classifier with altogether 41 different labels over the test set was 65.88%.

C. Plutchik's Wheel of Emotions

Defining axes of polarity is not a hard task, typically one has negativity, positivity and a notion of neutrality or objectivity in between. For emotions however, defining a complete and clear set of emotions is much more difficult. When classifying emotions, the research done started from two fundamental presuppositions: that emotions are discrete and fundamentally different constructs and that they can be characterized on a dimensional basis in groupings. Though several researchers attempted at defining standards in this field (Parrott, 2001 [13]; Plutchik, 1980 [14]) there is still no consensus on a basic set of emotions that is generally accepted and could be objectively verified.

In this paper we works with the wheel of emotions defined by Robert Plutchik [14] because it defines only eight basic emotions that are assumed to be complete in the sense that any expressed emotion is related or subsumed by one of the eight. Furthermore, Plutchik defines eight human feelings that are derivatives of combinations of two basic emotions. This in fact means that we can get sixteen dimensions of emotions and feelings. In our work we will ignore the class emotion awe and add a neutral class. The table I illustrate the mapping of the 41 Emobert category onto the sixteen Plutchick labels:

Emo Bert Labels	Plutchick Labels
Nostalgic, sentimental, sad, Lonely, disappointed, devastated	Sadness
Guilty	Remorse
Disgusted	Disgust
Furious, Angry, Annoyed	Anger
Jealous	Aggressiveness
Prepared, hopeful, anticipating	Anticipation
Proud	Optimism
Excited, joyful, content	Joy
Caring	Love
Grateful, confident, trusting	Trust
Faithful	Submission
Terrified, Afraid, Anxious, Apprehensive	Fear
Impressed, surprised	Surprise
Ashamed, Embarrassed	Disapproval
Agreeing, acknowledging, encouraging, consoling, sympathizing, suggesting, questioning, wishing, neutral	Neutral

Table I. Mapping of Emo Bert labels onto the sixteen Plutchick emotions and feelings. We remove the Plutchik feeling awe and add the neutral emotion.

Dataset	# Dialogues	Description
OpenSubtitle [5]	8.8M	Movie subtitles which are not speaker-aligned
Movie-Triples [9]	245k	Dialogues of three turns between two interlocutors.
Movie-DiC [18]	132k	American movie scripts
Cornell Movie-Dialogue [17]	220k	Conversation from the movie scripts.
Our dataset	35k	Movie subtitles with a speaker alignment

Table II. A comparison of existing movie dialogue datasets with our dataset.

```
row 1 | 00:00:02,025 | 00:00:04,823 |
So , let 's take a look at what 's going on around the country .

row 2 | 00:00:04,857 | 00:00:06,241 |
You know what ?

row 3 |00:00:06,275 | 00:00:09,527 |
Just bear with me while I take off these annoying pants .
```

Fig. 1. OpenSubtitle samples of consecutive row

III. Datasets

Movie and TV subtitles constitute a prime resource for many purposes such as machine translation [2], cross-lingual studies [3] but also monolingual tasks such multitask learning to improve natural language understanding [4]. Although they transcribe scripted interactions, subtitles do cover a large variety of dialogue phenomena, including e.g. the widespread use of colloquial language, multiple speaker styles, and the presence of complex conversational structures. Various movie-subtitles datasets has been present over the past year. A comparison of our novel dataset with the existing movie dialogue datasets is shown in Table II.

A. The Original OpenSubtitles

The OpenSubtitles database (Tiedemann et al 2016 [5]) provides a large collection of users contributed subtitles in various languages for movies and TV programs. The data base contain more than 3million subtitles in over 60 languages. An augmented version of the original dataset has been released in 2018 with almost 5 million subtitles. In this paper we will extract and work with the English Subtitle from the 2018 OpenSubtitles database. The raw dataset is structured in row which are short text segments associated with a start and end time. These blocks are expected to obey specific time and space constraints: at most 40-50 characters per line, a maximum of two lines and an on-screen display between 1 and 6 seconds [6]. The Fig.1 illustrate a sequence of 3 row.

The consecutive block in figure 1 correspond to a single dialog between two personas. In fact, the constrained applied to the blocks are too restrictive to build a meaningful emotion

analysis. Consequently, we employed an easy pre-processing rules to build a dialog oriented data set. We use the associate start and end time of each row to suggest the following rule: *We expand the dialog until the time between the end of the current sentence and the start of the previous one is higher than 5 seconds.* In the case where the gap time is missing the row is added to the dialog.

Once we had performed our simple process, the original OpenSubtitle data set contains almost 120 million row and 8.8 million dialogues.

IV. OpenSubtitles with automatically segmented turns

The original Opensubtitle is lacking of a valid turn segmentation. This factor prevent any meaningful emotion analysis. To adress this issue Lison and Meena (2016) [1] has publish an Automatic turn segmentation of the data set. They use various linguistic marker to the detection of turn boundaries. They extract features such as timing gap, punctuation, bigram between row and length from the training set and run a classifier. The classifier take a pair of two consecutive sentences and determines whether they are part of the same turn or not. We reproduce their paper and get a classifier accuracy of 76.69 %, which is in fact close to the 78% that they claim. For the following part of the paper, the automatic segmented data set will be refer as *Automatic data set*.

V. Experiments and Results

The motivation of this paper is to build from the Opensubtitle (In dialog form) a high quality emotionally labelled subset. In this work, We will start by analysing the original OpenSubtitle and outline the critical characteristic of the initial dialogues. The key observation highlight the importance of Speaker information. It will suggest our manual rule-based segmentation. Once our subset build, we extract the corresponding id in the automatic data and compare the structure of the subsets. Finally, we will apply the Emotion Analysis Tools describe above and compare the result.

A. Manual segmentation

Movie and TV subtitles contain large amounts of conversational material, but lack an explicit turn structure. Most of the subtitle do not provide any information about who is speaking at a given time. It makes hard to extract seflcontained dialogues for training multi-turn dialog models. Without speakers information, it's almost impossible for the machine to guess how many speaker are interacting and how long is their dialog. Their is a high risk to mix dialog without considering speaker information. Therefore, we will built a novel data set with dialog that explicitly contain row with personas and their speech. Furthermore, only the dialog that contain at least two distinct speaker will be consider.

Initially, the English subtitles from Open-Subtitles contains almost 120 million row and 8.8 million dialog. In average a dialog is composed of 13.56 sentence with an average of word per sentence of 7.35. More in details, 90% of the dialog include up to 20 sentences and 99.4 % of the sentences has up to 29 token. In Table III we illustrate sample of dialog from the original dialogues-based data set. A deep analyse of the dialog structure within the corpus will drive our rule based algorithm.

Dialog id	Raw Text
1	This is an emotional time for all of us .
1	I 'm not being emotional .
1	I 'm ... I 'm an orphan !
1	I 'm a jobless and homeless orphan .
2	BRIAN :
2	Hey , are you okay ?
2	BRIAN :
2	Feel you guilty ?
2	PETER :
2	What am I doing wrong , Brian ?
3	Guest <NUM >:
3	Hey , are you okay ?
3	Guest <NUM >:
3	Feel you guilty
3	Guest <NUM >:
3	What am I doing wrong , Brian ?

Table III. sample of dialogues from the original Open-Subtitles 2018. The first dialog don't contain any speaker information. Nevertheless, The second dialog as the properties to be turn segmented as in Table IV since speaker information are include. Finally, The third one contain a typography mistake. In fact, All the numbers of the corpus were replaced by <NUM >when initially process. This mistake can yield to confusion during segmentation. It has to be ignore.

The first dialog don't contain any speaker information. Obviously their is two personas. In order to build a relevant turn segmentation the three last row from the first dialog should be merge since it belongs to the same speaker. Unfortunately, due to the lack of speaker information it's almost impossible to build a model that can produce the expected result. By contrast, the second dialog we have speaker information. In this particular case with rule based algorithm it's possible to get a multi turn segmentation as in Table IV. On the other hand, in the third dialog of Table III occurs a mistake of typography. In fact, in the full corpus, all the numbers have been replaced by <NUM >. The form of the third dialog yield to confusion even for humans interpretation.

Dialog id	Speaker	Cleaned Text
2	BRIAN	Hey , are you okay ? Feel you guilty ?
2	PETER	What am I doing wrong , Brian ?

Table IV. Expected turn segmentation result. The dialog contain two distinct persona with alternate speech

Our target consist in the transformation of the full data set into a multi turn form as in Table IV. Each dialog has at least two distinct speaker and each row self contain the speech of each character. To give an illustration, in the dialog 2 from the Table III the same speaker show up consecutively in two row. It has to be merge. The main process of the manual turn segmentation of a dialog are summarized as follow:

1) We assume that a Speaker is always followed by the special character ":" and it name is only one words. We

validate our assumption by considering all the sentences that contain one special character ":". In 98.13 % of the cases their is only one words before the special character. The side effect of this assumption will discarded many outliers and the ambiguity produce by <NUM >.

2) If it doesn't exceed a threshold size, the text between two occurrence of speaker is merge and belongs to the first speaker otherwise the speaker is discarded. We set the treshold at 31 row to reach 80% of the cases and discarded dialog which should be outlier because too long.
3) When one Speaker appears consecutively twice or more, its text is merge.
4) Once all the above process have been applied, only dialog that contain at least two distinct speaker with non empty text are kept.

The resulting subset contain 35k dialog with 195k sentences. Table V show a random example from the subset.

Turn	Speaker	Cleaned Text
1	RACHEL	Stepping off the last step , I want you to drift into the water .
2	DEBS	This isn't going to work .
3	LAUREN	If we've a chance , even a slim one , of Mum being less full-on and off our backs for the next 30 years ...
4	DEBS	Please let it work.

Table V. Random dialog from the manual segmented data set. The structure is optimal for a sentimental analysis

B. A statistics comparisons

The dialog id in our subset hasn't been change. We can thus get an equivalent subset from the automatic data set by matching the index. In other words, the automatically segmented data set is reduced to 35k dialog.

The 35k dialog from the automatic segmented data set contain 4.28 (835k sentences) times more sentences than the manual one (195k sentences). In fact, the heuristic is very different between both method of segmentation. The automatic segmentation depend mainly on the timestamp between sentences. In their paper Lison and Meena has mention that the human annotators made little use of the timing information and in the heuristic it state that in absence of a time gap the two sentences are part of the same visual block, which often indicates a continued turn. As a result, It produce many inconsistency in the size of a dialog. The automatic heuristic is timestamp based as opposed to speaker based heuristic from the manual segmentation.

The various segmentation errors from the automatic segmentation are reflect in the cumulative function of the numbers of row in a dialog (Figure 2). The variance in the numbers of row in a dialog is prominent with the automatic subset. On the contrary, the manual subset mainly contain dialog smaller than 20 row (97% of the dialog). Those consideration are corroborate with the box plot illustrate in figure 3.

To conclude our dialog properties analysis, we compute the distribution of the mean number of token in a turn in a

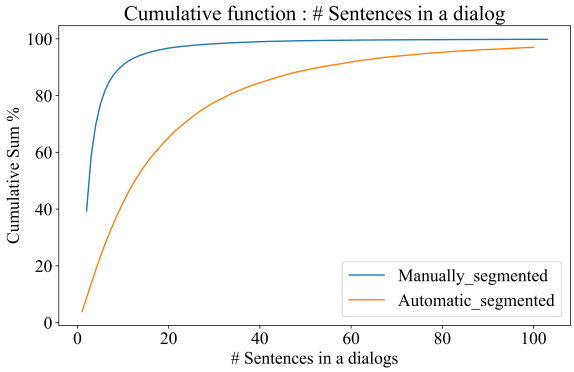

Fig. 2. The automatic segmentation has a higher variance and median number of row per dialog than the manual. This can be explained by the difference in the segmentation heuristic

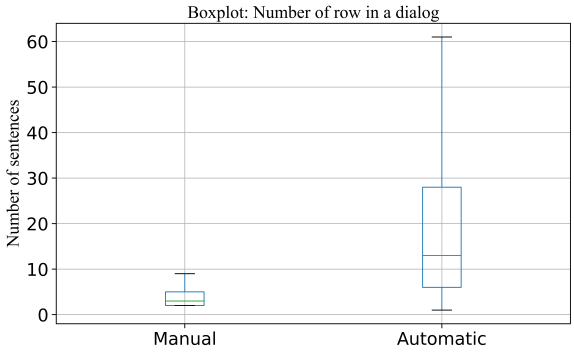

Fig. 3. The variance and the median measure corroborate the observation in figure 2. It illustrate the main weakness of the timestamp based heuristic

dialog. The figure 4 and 5 highlight an other properties of the heuristics. In the manual segmentation the row between two speaker has been merge when they don't exceed a threshold. On the contrary, the automatic heuristic is more sophisticated and tend to merge less the row. To summarize, the dialog structure found in the manual set seems to be more adapted for an emotional analyse. The number of turn inside a dialog is more or less consistent within the all corpus for manual. On the contrary, the weakness in the heuristic of the automatic segmentation is reflected by a higher median and variance of turn number. Additionally, the mean length of each turn has more variation in the manual set than in the automatic. In fact, it's a nice result as it illustrate the variance of typical human dialogues. Finally, in appendix we discuss about the joint distribution of the dialogue's properties. This study reveal independence between dialog properties and hence support our analyse.

C. Emotion Analysis

The simplicity of Vader carries several advantages. First, it is both quick and computationally economical without sacrificing accuracy. It does not require training data, yet it performs well in diverse domains. A corpus that takes a fraction of a second to analyze with VADER can take hours

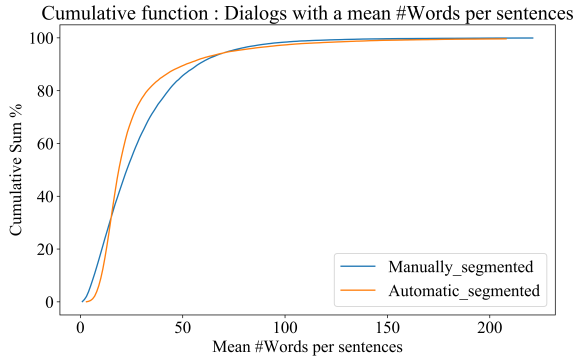

Fig. 4. As opposite to the analyse of the number of turn in a dialog, the mean length of a turn within a dialog has higher median and variance in the manual set. Nevertheless, the cumulative function have more similarities than in 2.

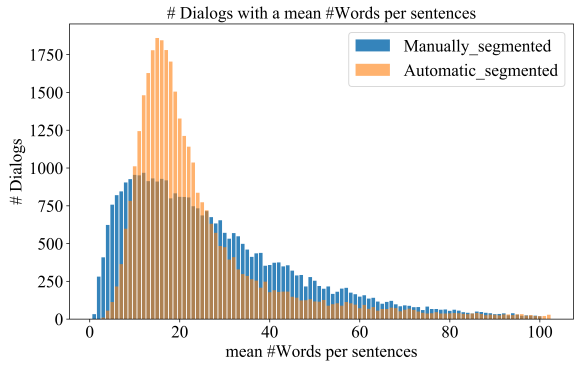

Fig. 5. The mean length of a turn is more compact within the automatic corpus.

when using more complex models like SVM [12] (if training is required) or tens of minutes if the model has been previously trained. Consequently we applied Vader to get the emotion intensity overview from the subsets. Each turn of a single dialog is independently process by Vader and the mean score is attribute to the dialog. It is mentioned in paper of Hutto et al [7] to set a standardized thresholds as follow:

1) Positive sentiment: $Vaderscore > 0.05$
2) Negative sentiment: $Vaderscore < -0.05$
3) Neutral sentiment : $Vaderscore \in [-0.05, 0.05]$

Figure 6 illustrate the Vader emotion score distribution of both manual and automatic subsets. The plot deduce a trend of both subset to be emotionally positive. Further, it makes in evidences that the manual set spread over a wider range of intensity than the automatic one. In other words, the manual subset trend to be more emotionally colored. It's supported by Table VI.

The intensity polarizer Vader predict that 81.34 % of the manual and 78.48 % of the automatic subset are emotionally colored. Nonetheless, the measure in Table VI aren't supported by any quality criteria since it's purely unsupervised. In their paper Hutto and Gilbert [7] specify that the performance of Vader depend on the intrinsic properties of the data set. It work

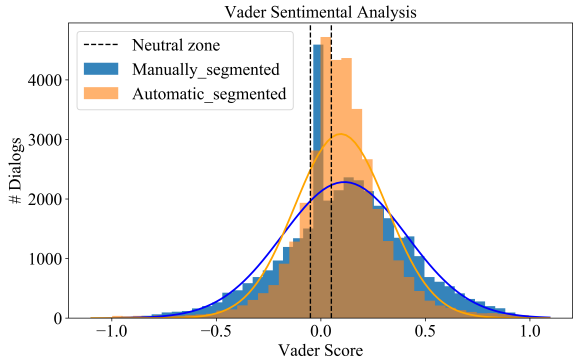

Fig. 6. The standardized thresholds used by Hutto et al [7] is illustrate with the black dashed line. All dialog that have a Vader score outside the neutral zone are emotionally colored. We observe that both data set trend to be positive. Additionally, the automatic data set tend to be more neutral than the manual one.

with a human accuracy for classifying tweets [7] but has poor performance for some other data set [8]. However, it constitute a good baseline to binary emotion classification.

Subset Type	Negative	Neutral	Positive
Manual	24.45 %	18.66 %	56.88 %
Automatic	18.25 %	21.52 %	60.22 %

Table VI. Percentage of dialog in the subset classified as positive, negative or neutral by Vader.

Next we applied our trained Emo BERT classifier. Rather than scoring a dialog with an emotion intensity, Emo BERT compute the probability of belonging to each of the 41 class. Each turn of a dialog is independently classified by EmoBert and the average among the turn is attribute to the dialog. Afterward, we map the Emo Bert labels into the Plutchik category according to I. For the following analyse, we only keep the prominent Plutchik category of each dialog

The results differ significantly from the one with Vader. 42.69 % of the manual and only 20% of the automatic data set are emotionally colored. The figure 7 highlight the variety of emotion in both corpus. The plot makes in evidence that the manual subset is more distributed and emotionally colored than the automatic. Surprisingly, their is a clear distinction between the number of emotion classified as basics or as feelings according to the Plutchik wheel. From all the emotionally dialogues in the manual subset 81 % are classified as a basics Plutchik emotion (joy, anticipation, joy, trust, fear, suprise, sadness, disgust, anger) and 19 % as a human feeling (optimism, love, submission, disapproval, remorse, aggressiveness, disgust). This trend is similar for the automatic subset (76 % basics and 24% feelings).

Table VII illustrate some nice examples of dialogues with emotions.

We expand our analyse in the appendix by plotting the prominent EmoBert class and studying the relation between the Vader score and the Plutchik classification.

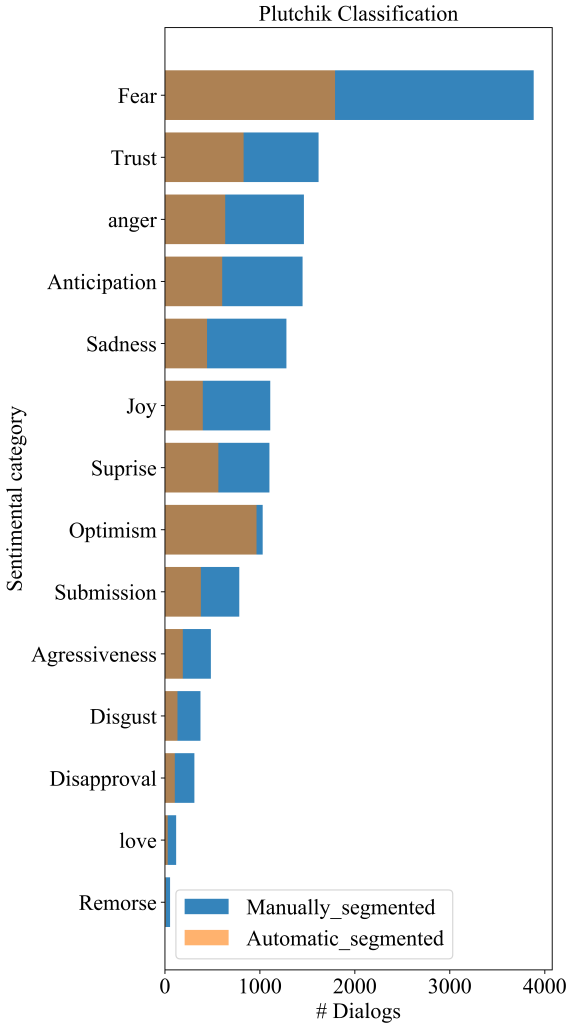

Fig. 7. The manual subset set has twice more emotionally colored dialog comparing to the automatic subset (15'064 and 7'070 dialog respectively).

Labels	Turn	Text
Fear	Daniel	South side , he 's on his way out .
Fear	Rebecca	Stay close , something 's up . This guy looks scared.
Trust	Voiceover	Let 's go get married , shall we ?
Trust	Billy	You got the ring mom , let 's go .
Trust	Mother	I 've got it .
Sadness	FIONA	The doctors say it 's terminal
Sadness	CORDELIA	Do me a favor . Die before Thanksgiving , so none of us have to suffer through that mess of raisins and Styrofoam you call stuffing
Disapproval	NYHOLM	Goes to character .
Disapproval	ALICIA	My son was pulled over once , the prosecutor dropped the charges .
Disapproval	NYHOLM	Excuse me , Your Honor . I 'm questioning a witness , not his mother .

Table VII. Example of nice emotionally colored dialogues

VI. Conclusion

Movie and TV subtitles are a valuable source for natural language task and are frequently used. The Opensubtitles database provide the largest collection of users contributed subtitles. However, the initial block structure lack considerably of meaningful sense for emotion analysis. Therefore, We started to build a dialog structured following an easy rule. This data set structure will be used as the baseline for the rest of the paper. Nevertheless, their is a deficit of a valid turn segmentation. This factor prevent any meaningful emotion analysis. To address this issue Lison and Meena (2016) [1] has publish an Automatic turn segmentation of the data set. Their key features to the detection of the turn boundary is the time gap and various linguistic marker. We successfully reproduce their paper by getting an classifier accuracy of 76.69 %. In parallel, we develop a manual rule-based algorithm that extract self-contained dialogues for multi-turn dialog model. Our heuristic segments the dialog based on the speaker information. A dialog has at least two distinct character with their speech as a turn. We end up with a subset of 35k dialog with a nice dialog structure.

We then compare the dialog structure between the manual and the automatic subset. This study reveal the weakness of the automatic segmentation. The human annotators made little use of the timing information and in the automatic heuristic it state that in absence of a time gap the two sentences are part of the same visual block, which often indicates a continued turn. It result in a very high variety of turn's number in the dialog that implies wrong turn structure. When it's question of the mean length of a turn in a dialog it has been shown that the manual has a higher variety of result than the automatic subset. In the manual segmentation the row between two speaker has been merge when they don't exceed a threshold. On the contrary, the automatic heuristic is more sophisticated and tend to merge less the row. It could produce a loss in the real size of the turn.

Finally, we conduct an emotion analysis of the subsets. The intensity polarizer Vader predict that 81.34 % of the manual and 78.48 % of the automatic subset are emotionally colored. The EmoBert classifier shows significantly different result ; 42.69 % of the manual and only 20 % of the automatic subset are emotionally colored. We demonstrate that the manual subset has more variety of emotions

The diversity of emotion in the corpus as well as the structure of the dialog are keys element for training a social bot. We conclude that our 35k dialogues subset has a high variety of emotions and reveals appropriate structure to train a social bot

VII. Acknowledgements

I would like to say a big thank you to everyone who helped me toward my goal. The biggest nod of appreciation goes to my mentors, Svikhnushina Ekaterina and Kalpani Anuradha, who faithfully monitored my progress each week. Finally, thank you to Dr. Pearl Pu for arranging the project, enabling me to learn so much and a neat project to pursue.

REFERENCES

[1] P. Lison and R. Meena, "Automatic turn segmentation for Movie & TV subtitles," 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, 2016, pp. 245-252, doi: 10.1109/SLT.2016.7846272.

[2] Volk, M., Sennrich, R., Hardmeier, C., and Tidstr om, F. (2010). Machine trans-lation of TV subtitles for large scale production. In Proceedings of the Second Joint EM+/CNGL Workshop on "Bringing MT to the User: Research on Integrating MT in the Translation Industry", pages 53–62, Denver.

[3] Lavecchia, Caroline Smaïli, Kamel Langlois, David. (2007). Building a bilingual dictionary from movie subtitles based on inter-lingual triggers.

[4] S. Constantin, J. Niehues, A. Waibel (2019) Multi-task learning to improve natural language understanding, arXiv:1812.06876

[5] Lison, P. and Tiedemann, J. (2016). Opensubtitles 2016: Extracting large parallel corpora from movie and tv subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016), Portoroz, Slovenia.

[6] Aziz,W.,de Sousa,S.C.M.,and Specia,L. (2012). Cross lingual sentence compression for subtitles. In 16th Annual Conference of the European Association for Machine Translation (EAMT 2012),pages103–110,Trento, Italy.

[7] Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[8] Ribeiro, F.N., Araújo, M., Gonçalves, P. et al. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci. 5, 23 (2016). https://doi.org/10.1140/epjds/s13688-016-0085-1

[9] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[11] Rashkin, H., Smith, E.M., Li, M. and Boureau, Y.L., 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. arXiv preprint arXiv:1811.00207.

[12] Hsu, C.-W., Chang, C.-C., Lin, C.-J. et al. (2003). A practical guide to support vector classification.

[13] W.G. Parrott. 2001. Emotions in Social Psychology. Psychology Press, Philadelphia.

[14] R. Plutchik, 1980. A general psycho evolutionary theory of emotion, pages 3–33. Academic press, New York.

[15] M. Huang, X. Zhu, J. Gao, (2019) Challenges in Building Intelligent Open-domain Dialog Systems, Microsoft Research, arXiv preprint arXiv:1905.05709

[16] Bunt H (ed) (2000) Abduction, belief, and context in dialogue: studies in computational pragmatics. J. Benjamins.

[17] Danescu-Niculescu-Mizil, C.; Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, Portland, OR, USA, 23 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011.

[18] Banchs, R.E. Movie-DiC: A movie dialogue corpus for research and development. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), Jeju Island, Korea, 8–14 July 2012.

[19] Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.C.; Pineau, J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI, Phoenix, AZ, USA, 12–17 February 2016.

[20]

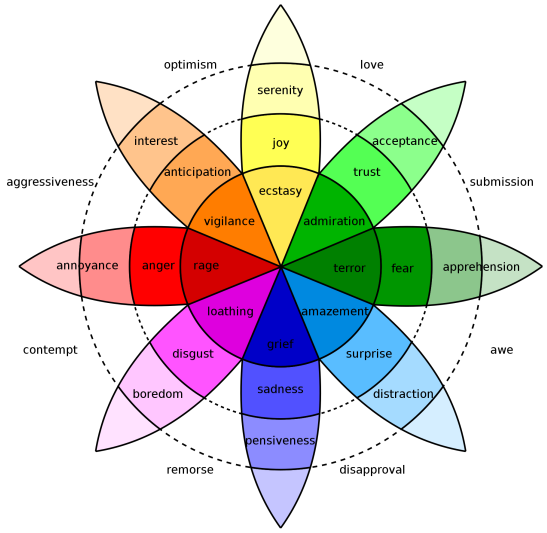

Fig. 8. Plutchik's wheel of emotions (Plutchik, 1980)

VIII. APPENDIX

A. The plutchik's wheel

Plutchik created the wheel of emotions, which illustrates the various relationships among the emotions (Figure 8). The eight basics Plutchik emotions are joy, anticipation, joy, trust,fear, suprise, sadness, disgust and anger. Furthermore, Plutchik defines eight human feelings that are derivatives of combinations of two basic emotions : optimism, love, submission, disapproval, remorse, aggressiveness, disgus and awe.

B. Correlation in dialogues properties

The number of row in a dialog of the automatic data has a higher median and variance than the manual. In opposite it has a smaller median and variance of the mean length of a row in a dialog than the manual. We could think that while a dialog get big the mean length of a sentence get small as it can catch noisy data such as single word row. The plot of the joint distribution in figure 9 refute this hypothesis. Their is no correlation between the distribution. They are independent.

C. Emo Bert Prominent Analyse

The Figure 11 illustrate the prominent Emo Bert classification. Only the emotionally colored class as been considered. In our study, The class [agreeing, acknowledging, encouraging, consoling, sympathizing, suggesting, questioning, wishing, neutral] are considered as neutral. The result is similar than the one with the Plutchik class 7

D. Correlation Vader score and Plutchick classification

The rule-based Vader classifier attribute an emotion intensity of a dialog when the prominent Emo Bert attribute a class to the dialog. The figure 11 maps all the Plutchick category to a Vader score. The mapping has been done with the measure

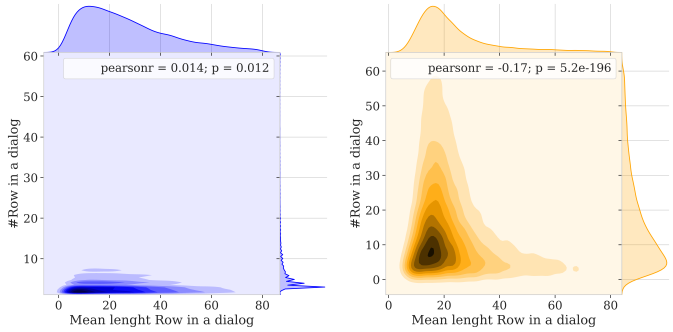

Fig. 9. Joint distribution: Number of Row in dialog with mean length of a Sentences in a dialog. The distribution are not correlated. They are independent.

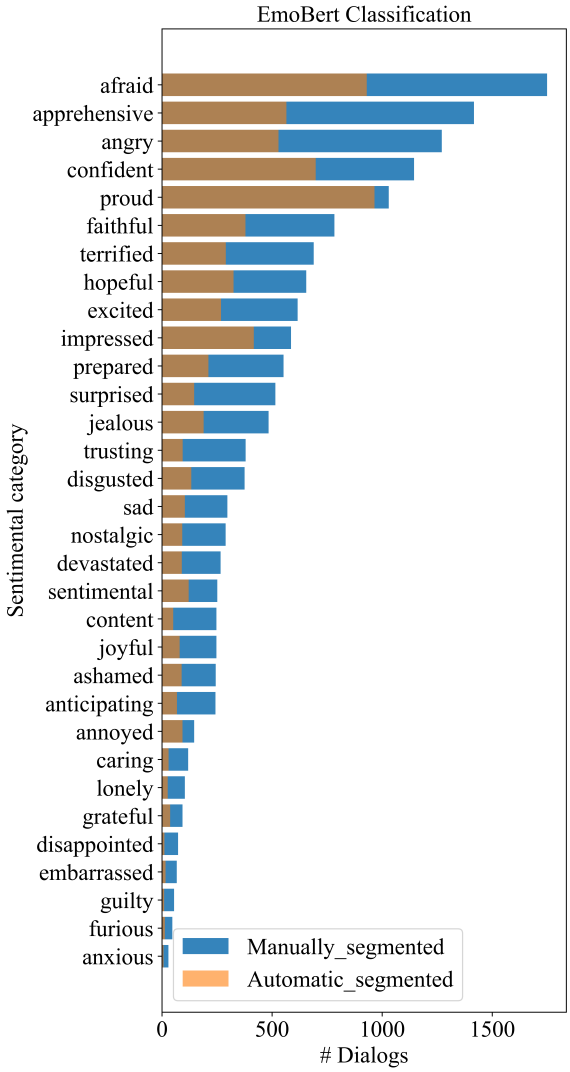

Fig. 10. As the plot 7 of the prominent Plutchik class , the manual data set is more distributed within the emotion class.

of emotions obtain with the manual data set. We observe that each Plutchick category have a big variance in the Vader space.

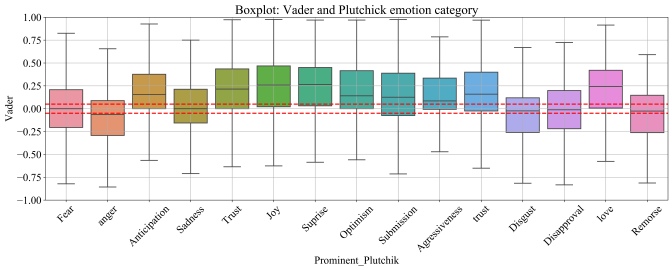

Fig. 11. The mapping of the Plutchik category in the vader Space has been done with the result of our study on the manual segmented data set.