Emotion Analysis on OpenSubtitles

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

***Abstract*—The availability of large amounts of conversational data and the recent progress on neural approaches to conversational bot are leading to a resurgent interest in developing intelligent open-domain dialogue systems. 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. Televisual subtitles are naturally a good source for developing conversation corpora. Currently the OpenSubtitles dataset is the biggest open-domain resource in this domain. However, subtitle files usually lack clear scene markers, making it difficult to extract self-contained dialogues used for training multi-turn dialogue models. Lison and Meena (2016) [1] have presented a data-driven approach to the segmentation of subtitles into dialogue turns. This paper manually segments the OpenSubtitle dataset into dialogue turns and create a speaker-aligned dataset of 35,000 conversations. On this novel dataset, the research uses a pre-trained BERT model to label the dialogues with emotions. Finally, the present compares our results by reproducing the analysis of Lison and Meena, matching the dialogues with a cleaned subset and applying the same emotion classifier.**

I. Introduction

L_M

Building intelligent open-domain dialogue systems able to converse with humans coherently and engagingly has been a long-standing goal of artificial intelligence [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 used for a chat-oriented dialogue system must cover a wide variety of domains, while being 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 has been developed such as OpenSubtitles [5], Cornell Movie-Dialogue Corpus [17], Movie-DiC [18] and Movie-Triples [19].

This paper investigates the use of user-contributed movie subtitles as a source of emotion analysis. This study is based on the OpenSubtitles corpus (Tiedemann et al 2016 [5]) and restores 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: Section 2 briefly introduces the main parameter of the emotion analysis tools used. While the OpenSubtitles dataset is presented in section 3 through the drawback of the initial block structure and introduction of a dialogue-based version of the data set. However, this section depicts the dataset lacking a valid turn segmentation. Lison and Meena (2016) [1] have tried to address this problem by publishing an automatic turn segmentation of the dataset. Section 4 reproduces their analysis and briefly discuss the heuristic before introducing our manual segmentation in the experiments and results. The findings show that our heuristic is speaker-based. A comparison of the properties of our dialogues with the one from the automatic subset is then studied. Finally, an emotional classification of the subsets will take place before a juxtaposition of 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 texts. The spectrum of sentiment analysis techniques ranges from identifying polarity (positive or negative) to a complex computational treatment of subjectivity, opinion and sentiment. Two broad approaches for calculating the sentiment of a text document exist: rule-based and machine-learning based.

In the following part we will present an overview of Vader, a rule-based approaches that computes the strength of the sentiment expressed in texts, as well as 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 training 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 (e.g. "but") and

preceding Trigram. When a sentence is being rated these keywords are identified and can shift or impact the rating.

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

B. EmoBERT

EmoBERT is a BERT transformer. It is based on a single-sentence emotion classifier [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]. The model was fine-tuned on situation descriptions given in the Empathetic Dialogues dataset [11] tagged with 32 emotions and listener utterances tagged with 8 response intents plus neutral. The training, validation, and test sets comprised respectively 25,023, 3,544 and 3,225 sentences, 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 of 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 previous research started from two fundamental presuppositions: Emotions are discrete and fundamentally different constructs and the characterization 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.

This paper works with the wheel of emotions defined by Robert Plutchik [14] as 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.

EmoBert 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 movies dialogues datasets with our dataset.

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 (e.g. multitask learning to improve natural language understanding) [4]. Although they transcribe scripted interactions, subtitles do cover a large variety of dialogue phenomena, including non-exhaustively the widespread use of colloquial language, multiple speaker styles, and the presence of complex conversational structures. Various movie-subtitles datasets have been presented over the past years. A comparison of our novel dataset with the existing movie dialogue datasets is depicted in Table II.

```
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

A. The Original OpenSubtitles

The OpenSubtitles database (Tiedemann et al 2016 [5]) provides a large collection of users contributed subtitles in various languages for televisuals. The data base contains more than 3,000,000 subtitles in over 60 languages. An augmented version of the original dataset has been released in 2018 with almost 5 million subtitles. This paper extracts and works 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 corresponds to a single dialogue between two personas. In fact, the constraints applied to the blocks are too restrictive to build a meaningful emotion

analysis. Consequently, this analysis employs an easy pre-processing rule to build a dialogue oriented data set. The analysis uses 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 dialogue.

Once we had performed our simple process, the original OpenSubtitle dataset contains almost 120 million rows and 8.8 million dialogues.

IV. OpenSubtitles with automatically segmented turns

The original OpenSubtitle is lacking of a valid turn segmentation. This factor prevents any meaningful emotion analysis. To address this issue Lison and Meena (2016) [1] have published an automatic turn segmentation of the data set. The scholars 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 takes a pair of two consecutive sentences and determines whether they are part of the same turn or not. This analysis reproduces their study and get a classifier accuracy of 76.69 %, which corroborates their 78% that they claim. For the following part of the paper, the automatic segmented data set will be referred to as: *Automatic dataset*.

V. Experiments and Results

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

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 hardens the extraction of self-contained 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. There is a high risk of mixing dialogue without considering speaker information. Therefore, the code builds a novel data set with dialogues that explicitly contain row with personas and their speech. Furthermore, only the dialogues that contain at least two distinct speaker are considered.

Initially, the English subtitles from OpenSubtitles contain almost 120 million row and 8.8 million dialogues. In average a dialogue is composed of 13.56 sentence with an average of 7.35 words per sentence. More in details, 90% of the dialog includes up to 20 sentences and 99.4 % of the sentences have up to 29 tokens. In Table III we illustrate sample of dialogue from the original dialogues-based data set. A deep analyse of the dialogue 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. samples of dialogues from the original OpenSubtitles 2018. The first dialogue doesn't contain any speaker information. Nevertheless, The second dialogue has the properties to be turn segmented as in Table IV since speaker information are included. Finally, The third one contains a typography mistake. In fact, All the numbers of the corpus were replaced by <NUM >when initially process. This mistake can lead to confusion during segmentation. It has to be ignored.

The first dialogue doesn't contain any speaker information. Obviously there are two personas. In order to build a relevant turn segmentation the three last rows from the first dialogue should be merged since they belong to the same speaker. Unfortunately, due to the lack of speaker information it is almost impossible to build a model that can produce the expected results. By contrast, the second dialogue with speaker information, with rule based algorithm it's possible to get a multi turn segmentation as in Table IV. On the other hand, in the third dialogue of Table III occurs a typography mistake. In fact, in the full corpus, all the numbers have been replaced by <NUM >. The form of the third dialog yields to confusion even for human 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 dialogue contains two distinct personas with alternate speech

The analysis target consists in the transformation of the full data set into a multi turn form as in Table IV. Each dialogue has at least two distinct speakers and each row self contains the speech of each character. To give an illustration, in the dialog 2 from the Table III the same speaker shows up in two rows consecutively and require a merger. The main process of the manual turn segmentation of a dialogue are summarized as follow:

1) We assume that a Speaker is always followed by the special character ":" and his name is only one word. We validate our assumption by considering all the sentences that contain one special character ":". In 98.13 % of the cases there is only one word before the special character. The side effect of this assumption will discard many outliers and the ambiguity produced by <NUM >.
2) If it doesn't exceed a threshold size, the text between two speaker's occurrences is merged and belongs to the first speaker otherwise the speaker is discarded. We set the threshold at 31 rows to reach 80% of the cases and discarded dialog which should be outliered for being too long.
3) When one Speaker appears consecutively twice or more, his text is merged.
4) Once all the above process have been applied, only dialog that contains at least two distinct speakers with non empty texts are kept.

The resulting subset contains 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 dialogue from the manual segmented dataset. The structure is optimal for a sentimental analysis

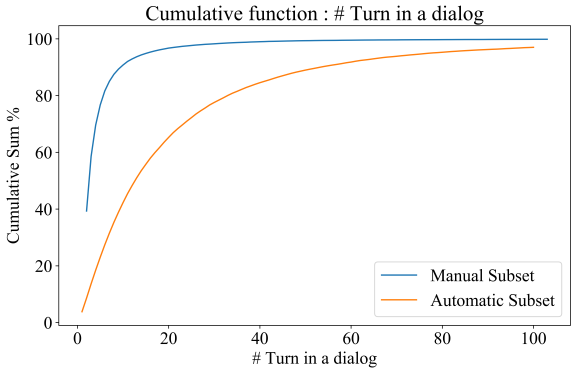

Fig. 2. The automatic segmentation has a higher variance and median number of row per dialog than the manual. This can be explained by the difference in the segmentation heuristic

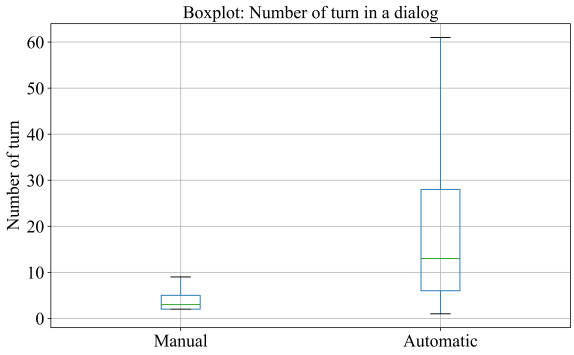

Fig. 3. The variance and the median measures corroborate the observation in figure 2. It illustrates the main weakness of the timestamp based heuristic

B. A statistics comparisons

The dialogue 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 dialogues.

The 35k dialogues (835k sentences) from the automatic segmented data set contain 4.28 times more sentences than the manual one (195k sentences). In fact, the heuristic is very different between both method of segmentation. The automatic segmentation depends mainly on the timestamp between sentences. In their paper Lison and Meena have mentioned that the human annotators made little use of the timing information and in the heuristic it states 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 produces many inconsistency in the size of a dialogue. The automatic heuristic is timestamp-based as opposed to the speaker-based heuristic from the manual segmentation.

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

To conclude our dialogue properties analysis, we compute the distribution of the mean number of tokens in a turn in a dialogue. The figure 4 and 5 highlights an other properties of the heuristics. In the manual segmentation the row between two speaker has been merged 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 dialogue structure found in the manual set seems to be more adapted for an emotional analysis. On the one hand, the number of turns inside a dialogue is more or less consistent within the all corpus for manual. On the other hand, the weakness in the heuristic of the automatic segmentation is reflected by a higher median and variance of turn numbers. 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 illustrates the variance of typical human dialogues. Finally, in the appendix the discussion while be articulated about the joint distribution of the dialogue's properties. This study reveals independence between dialogue properties and hence supports our analyse.

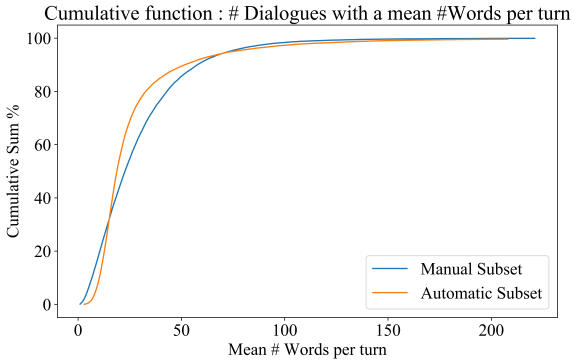

Fig. 4. As opposite to the analyse of the number of turn in a dialog, the mean length of a turn within a dialog has higher median and variance in the manual set. Nevertheless, the cumulative function have more similarities than in 2.

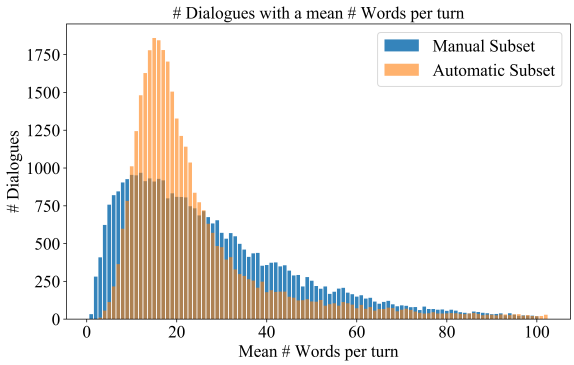

Fig. 5. The mean length of a turn is more compact within the automatic corpus.

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 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 dialogue. 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 deduces a trend of both subsets to be emotionally positive. Further, it highlights that the manual set spreads over a wider range of intensity than the automatic one. In other words, the manual subset trend is more emotionally colored. This is supported by Table VI. The intensity polarizer Vader predicts that 81.34 % of the

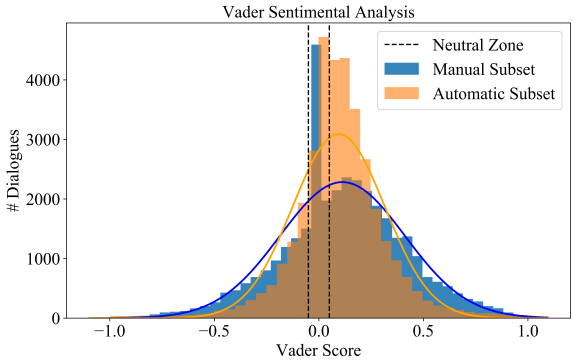

Fig. 6. The standardized thresholds used by Hutto et al [7] is illustrated with the black dashed line. All dialog that have a Vader score outside the neutral zone are emotionally colored. We observe that both data set trend to be positive. Additionally, the automatic data set tend to be more neutral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

manual and 78.48 % of the automatic subsets 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 depends on the intrinsic properties of the dataset. It works with a human accuracy for classifying tweets [7] but has poor performances for some other datasets [8]. However, it constitutes 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 dialogue in the subset classified as positive, negative or neutral by Vader.

Next, we applied our trained EmoBERT classifier. Rather than scoring a dialogue with an emotion intensity, EmoBERT computes the probability of belonging to each of the 41 class. Each turn of a dialogue is independently classified by EmoBert and the average among the turn is attributed to the dialogue. Afterwards, the mapping of the Emo Bert labels into the Plutchik category takes place according to I. For the following analyse, we only keep the prominent Plutchik category of each dialogue. 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 highlights the variety of emotions in both corpus. The plot highlights that the manual subset is more distributed and emotionally colored than the automatic one. Surprisingly, there is a clear distinction between the number of emotions classified as basics or as feelings according to the Plutchik wheel. From all the emotional dialogues in the manual subset 81% are classified as basics Plutchik emotions (joy, anticipation, joy, trust, fear, suprise, sadness, disgust, anger) and 19% as human feelings (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.

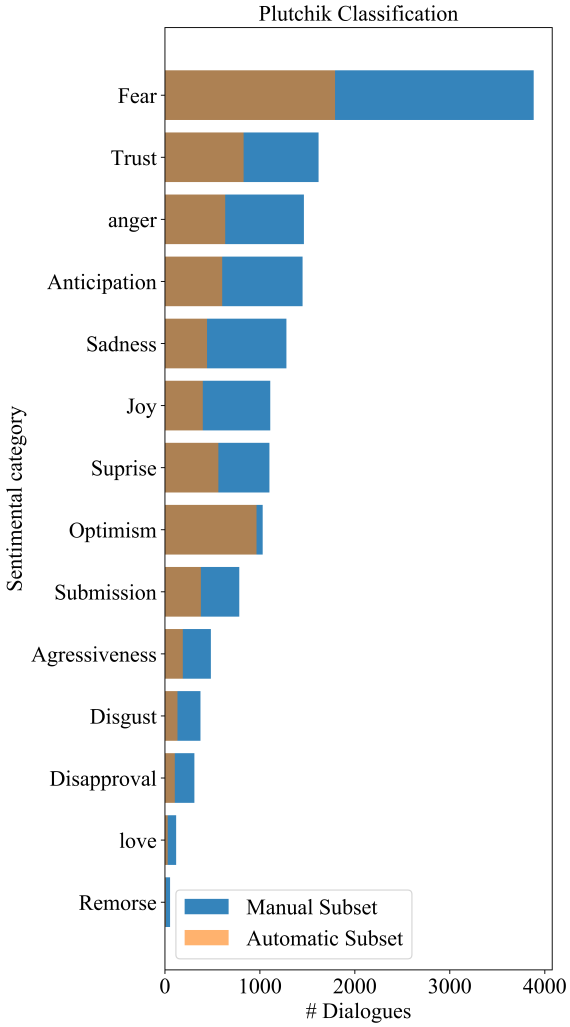

Fig. 7. The manual subset set has twice more emotionally colored dialog comparing to the automatic subset (15'064 and 7'070 dialog respectively).

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

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.

VI. Conclusion

Televisual subtitles are a valuable source for natural language tasks and are frequently used. The OpenSubtitles database provides the largest collection of users contributed subtitles. However, the initial block structure lacks considerably of meaningful sense for emotion analysis. Therefore, this study started to build a dialogue structured following an easy rule. This dataset structure is used as the baseline for the rest of the paper. Nevertheless, there is a deficit of a valid turn segmentation. This factor prevents any meaningful emotion analysis. To address this issue Lison and Meena (2016) [1] have published an Automatic turn segmentation of the dataset. Their key features to the detection of the turn boundary is the time gap and various linguistic marker. We have successfully reproduced the paper by getting an classifier accuracy of 76.69%. In parallel, we have developed a manual rule-based algorithm that extract self-contained dialogues for multi-turn dialog model. Our heuristic segments the dialogue based on the speaker information. A dialogue has at least two distinct characters with their speech as a turn. The paper ends up with a subset of 35k dialogues with a satisfied dialogue structure.

The research then compares the dialogue structure between the manual and the automatic subset. This study reveals the weakness of the automatic segmentation. The human annotators made little use of the timing information while automatic heuristic states that in absence of a time gap the two sentences are part of the same visual block, which often indicates a continued turn. It results in a very high variety of turn's number in the dialogue that implies wrong turn structures. When it is question of the mean length of a turn in a dialogue, 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 tends 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 predicts 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. This paper demonstrates that the manual subset has more variety of emotions.

The diversity of emotion in the corpus as well as the structure of the dialogue are key elements for training a social bot. The conclusion is that our 35k dialogues subset have a high variety of emotions and reveal appropriate structure to train a social bot

VII. Acknowledgements

I would like to warmfully thank everyone who helped me towards my goal. The biggest nod of appreciation goes to my mentors, Svikhnushina Ekaterina and Kalpani Anuradha, who faithfully monitored my progress each week. In addition, I express my gratitude to Dr. Pearl Pu for arranging the project,

enabling me to deepen my knowledge and giving me the wish to further neat this project.

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.

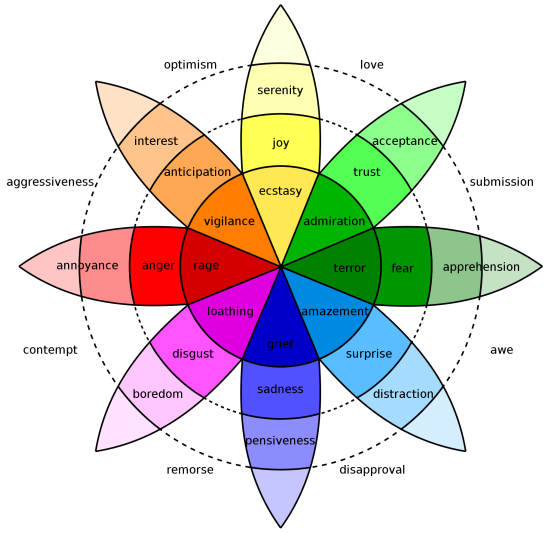

Fig. 8. Plutchik's wheel of emotions (Plutchik, 1980)

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, surprise, 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, disgust and awe.

B. Correlation in dialogues properties

The number of turn in a dialogue 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 turn in a dialog than the manual. We could think that while a dialogue get big the mean length of a sentence get small as it can catch noisy data such as single word turn. The plot of the joint distribution in figure 9 refutes this hypothesis. There is no correlation between the distributions. They are independent.

C. Emo Bert Prominent Analyse

The Figure 11 illustrate the prominent EmoBERT classification. Only the emotionally colored class as been considered. In this 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 attributes an emotion intensity of a dialogue when the prominent EmoBERT attributes 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 of emotions obtain with the manual dataset. We

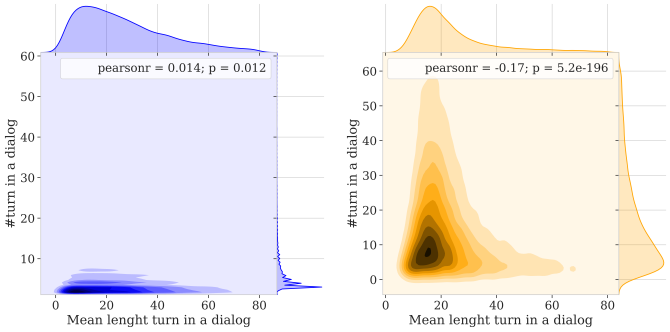

Fig. 9. Joint distribution: Number of Row in dialog with mean length of a Sentences in a dialog. The distribution are not correlated. They are independent.

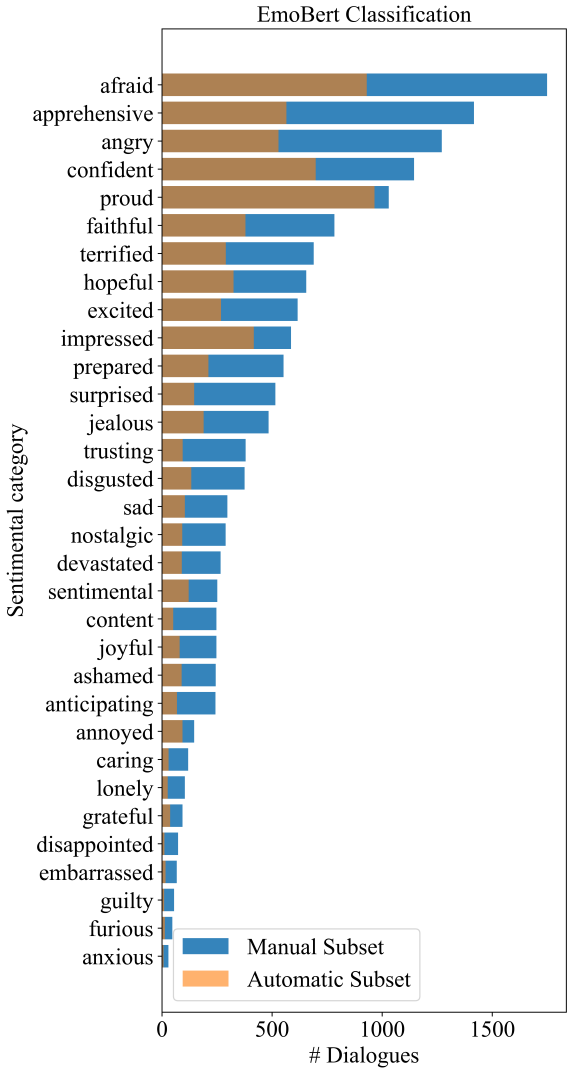

Fig. 10. As the plot 7 of the prominent Plutchik class , the manual subset is more distributed within the emotion class.

observe that each Plutchick category has a big variance in the Vader space.

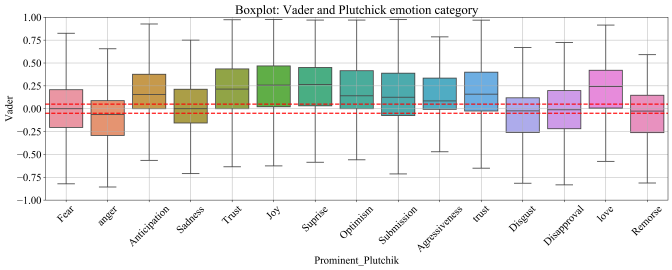

Fig. 11. The mapping of the Plutchik category in the vader Space has been done with the result of our study on the manual segmented dataset.