LinguisTech at SemEval-2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversation

Mihaela Alexandru
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
alexandrumihaela227@gmail.com

Călina-Georgiana Ciocoiu
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
calinaciocoiu@gmail.com

Ioana Măniga
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
ioana.mna@gmail.com

Octavian Ungureanu
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
tavi2105@gmail.com

Daniela Gîfu
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
Institute of Computer Science, Romanian Academy - Iasi Branch
daniela.gifu@iit.academiaromana-is.ro

Diana Trandăbăț
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
dtrandabat@info.uaic.ro

Abstract

The "Emotion Discovery and Reasoning Its Flip in Conversation" task at the SemEval 2024 competition focuses on the automatic recognition of emotion flips, triggered within multi-party textual conversations. This paper proposes a novel approach that draws a parallel between a mixed strategy and a comparative strategy, contrasting a Rule-Based Function with Named Entity Recognition (NER)—an approach that shows promise in understanding speaker-specific emotional dynamics. Furthermore, this method surpasses the performance of both DistilBERT and RoBERTa models, demonstrating competitive effectiveness in detecting emotion flips triggered in multi-party textual conversations, achieving a 70% F1-score. This system was ranked 6th in the SemEval 2024 competition for Subtask 3.

1 Introduction

The field of emotion analysis continues to be rich with surprises (Kumar et al., 2022), especially within the context of conversations. For this competition, we have implemented a competitive method for Subtask 3 (Kumar et al., 2024). Uncovering the reasons (triggers) behind a speaker's emotional shift during a conversation—taking the example of "Friends," an American television sitcom—presents a unique challenge, especially in the realm of response generation (Gifu and Cioca, 2013). With the rising popularity of chatbots (Ouatu et al., 2020), it appears that emotions are the critical link missing between establishing trust and simulating genuine connections (Madasu et al., 2023). Furthermore, the detection of emotions and their triggers (Cristea et al., 2015) could play a significant role

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 412–419
June 20-21, 2024 ©2024 Association for Computational Linguistics

in new digital marketing strategies, enhancing user feedback, and analyzing overall customer centricity.

This raises a pertinent question: *Is AI capable enough to identify emotions and their triggers with high accuracy within code-mixed dialogues?*

The remainder of this paper is organized as follows: Section 2 briefly reviews studies related to emotion recognition (Kumar et. al., 2023) and the concept of an emotion flip in conversations. Section 3 describes the system developed to detect the specific emotional dynamics that occur during a conversation. Section 4 outlines the experimental setups. Section 5 discusses the results of the experiments conducted, and Section 6 presents the conclusions.

2 Background

Recent research in dialogue emotion detection has witnessed significant advancements. The literature suggests that the challenge of recognizing emotions in conversations can be tackled from various perspectives. For instance, a notable approach involves the use of models based on transformers, as well as iterative emotion interaction networks.

The most prevalent method for emotional discovery and analysis in recent years involves employing various transformers. Variants of BERT have been frequently utilized, whether they are pre-trained or not. Some of the notable examples include mBERT (De Bruyne et al., 2022), LFTW-RoBERTa, YT-Bert, MNLI-BART-large, MNLI-RoBERTa (Bulla et al., 2023), and EmoRoBERTa (Bayram & Benhiba, 2022), among others. Additionally, a study by Li et al. (2020) introduced HiTrans, an innovative model specifically designed to discern emotions within multi-speaker conversations. A team of researchers (Kumar et al., 2023) has presented a pioneering approach that focuses on identifying the triggers behind emotion shifts in conversations. Using BERT as a foundation, their findings indicate that TGIF (a novel neural architecture) more effectively addresses the increase in instigator labels compared to existing baselines. Some studies concentrate on the application of zero-shot models to emotion classification and hate speech detection (Bulla et. al., 2023), while others adopt a modified approach, developing a semi-zero-shot model. This variation aims to investigate and determine whether significant challenges and differences exist in emotion detection across various language families (De Bruyne et al., 2022). Interestingly, the F1-scores for all transformer types employed in zero-shot scenarios are reported to be similar across both studies.

In the experiments dedicated to the KET model (Zhong et al., 2019), several key findings were highlighted: notably, the KET model demonstrated superior performance, surpassing existing state-of-the-art models in various datasets as measured by F1 score. This underscores its effectiveness in detecting emotions within textual conversations. Additionally, there is research (Lu et al., 2020) exploring non-transformer-based solutions, such as the innovative Iterative Emotion Interaction Network. This approach specifically addresses the challenge of the absence of gold-standard emotion labels during inference, offering a novel solution to a prevalent issue in emotion detection.

Additional research (Zhu et al., 2021) explores the use of baselines such as DialogueGCN and KET, but it is COSMIC that emerges as the superior model among these baselines. This advancement began with the development of a topic-augmented language model (LM), which includes a dedicated layer for detecting topics. These collective efforts significantly push the boundaries of dialogue emotion detection forward by incorporating a blend of knowledge, contextual insight, and cutting-edge neural architectures.

The third subtask of SemEval-2024 Task 10, titled 'Emotion Discovery and Reasoning its Flip in Conversation' (EDiReF), is dedicated to exploring the point in a dialogue at which the last emotion flip occurs For the Emotion Flip Reasoning subtask, Task 10 of SemEval-2024 provides three types of datasets: training, validation, and testing, detailed in the table below:

Training Dataset	**Validation Dataset**	**Testing Dataset**
400 entries	426 entries	1002 entries
13500 dialogue lines	3522 dialogue lines	8642 dialogue lines

Table 1: Task Dataset Statistics

The datasets contain dialogues extracted from different episodes of the 'Friends' series, stored in a JSON array. Each entry comprises the following fields:

- episode: the name of the episode (e.g. "episode": "utterance_0");
- speakers: a list of speakers in order of their participation in the conversation (e.g. "Chandler", "The Interviewer", "Chandler", "The Interviewer", "Chandler");
- emotions: a list of emotions in order (e.g. "neutral", "neutral", "neutral", "neutral", "surprise",);
- utterances: the list of utterances from the dialogue in sequential order (e.g. "also I was the point person on my company's transition from the KL-5 to GR-6 system.", "You must've had your hands full.", "That I did. That I did.", "So let's talk a little bit about your duties.", "My duties? All right.");
- triggers: a list of triggers in sequential order. This field is the output of our models and represents a list of ‘0.0s’ and only one value of ‘1.0’, indicating the trigger in that conversation.

Before proceeding further, we conducted a thorough examination of the training dataset for our subtask to gain insights into the appearance of triggers and the functioning of the Emotion Flip Reasoning (EFR) system. Our analysis revealed that all triggers are associated with the same (last) emotion flip in the dialogue. Additionally, we observed that triggers can manifest in any utterance within the same segment of the conversation where the emotion change occurs. To achieve this understanding, we initially examined the speakers, emotions, and triggers. Subsequently, we delved into the utterances, particularly focusing on cases where triggers were less clear. As observed in numerous papers, the implementation of models often revolves around transformers, with BERT being a prominent choice. This observation significantly influenced our approach, leading us to adopt a strategy centered on utilizing the DistilBERT transformer. DistilBERT, developed to reduce the size and enhance the computational efficiency of BERT while preserving a substantial portion of its functionality (Sanh et al., 2019), emerged as a key component of our investigation. Additionally, we incorporated the RoBERTa transformer into our architecture's model, reflecting our commitment to leveraging state-of-the-art techniques. This initiative can be seen in the baseline part of our architecture model.

3 System Overview

Our objective is to enhance emotion recognition technology by investigating the underlying reasons for sudden emotional changes. Specifically, our research concentrates on emotional flips, which denote abrupt shifts in emotions during conversation—an aspect often overlooked in existing studies. Despite the progress achieved by previous methods, recognizing emotions in conversation remains challenging due to the nuanced conveyance of emotions and the varying significance of utterances, influenced by the specific topics discussed and implicit understandings shared among participants.

Upon analyzing the dataset, we identified seven distinct emotion labels: neutral, joy, surprise, anger, sadness, fear, and disgust, with varying frequencies. Dialogues in the dataset involve a range of one to eight participants, with dialogues between two speakers being the most common.

The primary focus of this paper is to identify speaker-specific emotional dynamics occurring during conversation. Our approach utilizes two transformer-based baselines, RoBERTa and DistilBERT. Additionally, we compare their performance with a mixed and comparative method employing rule-based and Named Entity Recognition (NER) techniques.

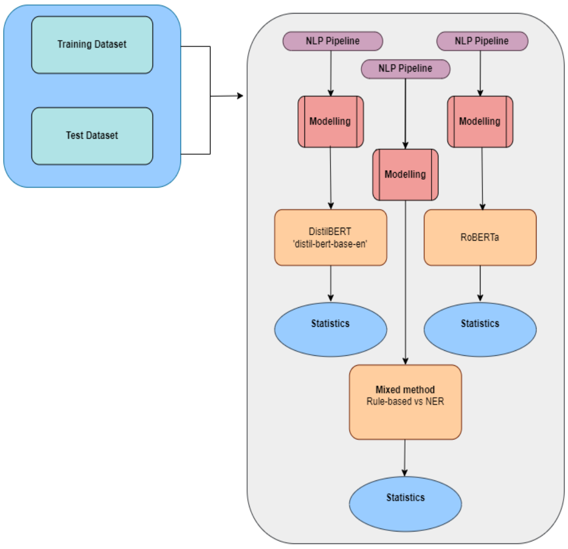

Figure 1: The LinguisTech system architecture

The first transformer baseline we utilized was a pre-trained RoBERTa model (TFRobertaModel) based on the BERT-base architecture. This model is described by: 2-layer, 768-hidden, 12-heads, 125M parameters. As for the parameters, we configured the model with the following settings:

- metrics=['acc', f1_m, precision_m, recall_m]
- loss='sparse_categorical_crossentropy'
- optimizer=tf.keras.optimizers.Adam(lr=1e-5)

In addition, we employed 'relu' and 'softmax' as activation functions. We segmented each conversation into utterances, and for each utterance, the training data is structured as a dictionary containing the following fields:

- utterance – the current utterance
- emotion - the current emotion
- context – containing arrays with: all emotions in that dialog, all speakers, all utterances

In the pre-processing phase for the RoBERTa baseline, we pursued several approaches and actions:

- Extracted all replicas from the context and applied tokenization, lemmatization, stopword removal, etc.
- Extracted emotions from contexts.
- Extracted emotions and utterances from context.
- Extracted emotions, utterances, and speakers from context.
- Retained the context along with the following: id, list of utterances, list of emotions, list of speakers.
- Retained the context along with individual replicas, list of utterances, list of emotions, list of speakers.
- Maintained the original context while eliminating the first half, followed by attempting to remove the first half of the context and combining speakers, emotions, speakers, and emotions.

As for the second baseline model, we chose the DistilBertClassifier from the keras_nlp framework. We utilized the 'distil_bert_base_en' preset, which is a 6-layer DistilBERT model maintaining case sensitivity. This model comprises 65.19 million parameters and was trained on English Wikipedia + BooksCorpus using BERT as the teacher model. For parameters, we configured the model with the following settings:

- loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True)
- optimizer=keras.optimizers.Adam(5e-5)
- jit_compile=True
- metrics=['accuracy', f1_m, precision_m, recall_m], where f1, precision and recall are functions defined by us with the traditional method.

In the preprocessing phase for the DistilBERT baseline, we divided each conversation into utterances. For each utterance, the training data is structured as a dictionary containing the following fields:

- entry_index - the index of the utterance in conversation
- entry – a string representing the intervention of index entry_index, formed from entry_index - speaker - utterance - emotion
- context – a string formed by concatenating the entire conversation, every dialogue line being formed with this rule: speaker: utterance – emotion

After preprocessing, we applied a DictVectorizer from sklearn to convert the data into a numerical format. Additionally, we performed feature selection by selecting the 100 best features using SelectKBest (also from sklearn), with the chi-square test as the scoring function.

Examples of preprocessed data objects for RoBERTa and DistilBERT can be observed in the first and second annexes, respectively.

4 Experimental Setup

Based on the results obtained from implementing the two transformers, RoBERTa and DistilBERT, we observed outcomes that did not meet our expectations. Consequently, we initiated an experimental investigation aimed at combining and comparing two alternative methods to achievestyle improved performance. These methods include a rule-based function constructed from observations on the dataset, as well as a Named-Entity Recognition (NER) Model.

Our initial observation revealed that triggers are generally present in the second part of the conversation. To validate our hypothesis, we calculated the instances where this statement holds true, as well as the percentage of cases where it does not. The results are as follows:

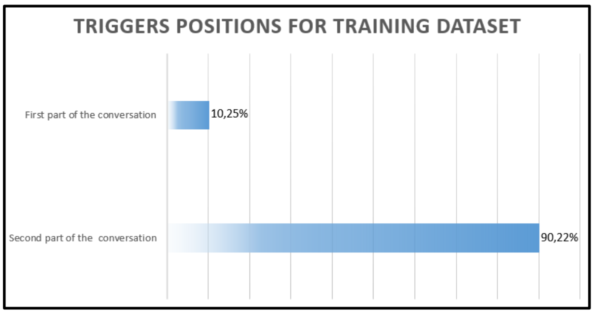

Figure 2: Trigger positions for training dataset in first/second part of conversation

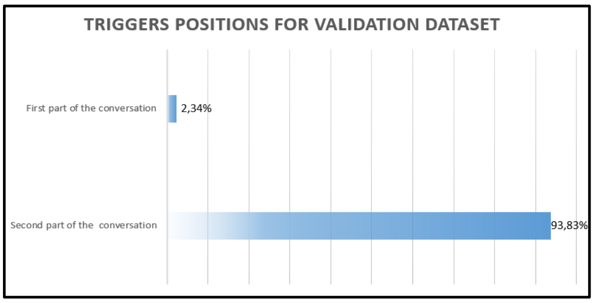

Figure 3: Trigger positions for validation dataset in first/second part of conversation

Having said that, the first rule we applied focused solely on the second part of each conversation.

The second rule is based on the emotion flips observed for each speaker. Whenever a change in emotion occurs between two consecutive interventions by a speaker, we designate the utterance preceding the second intervention as a trigger.

	Speaker	Utterance	Emotion	Trigger
1	Chandler	Hey, Mon.	Neutral	0
2	Monica	Hey-hey-hey. You wanna hear something that sucks.	Neutral	0
3	Chandler	Do I ever.	Joy	0
4	Monica	Chris says they’re closing down the bar.	Sadness	0
5	Chandler	No way!	Surprise	1
6	Monica	Yeah, apparently, they’re turning it into some kind of coffee place.	Neutral	0

Table 2: Dialogue example for the second rule detected

For the NER method, we utilized TFAutoModelForTokenClassification from python library transformers library with the 'bert-base-cased' preset.

As for the parameters, we configured the model with the following settings:

- optimizer=tensorflow.keras.optimizers.Adam(learning_rate=2e-5)
- epochs = 3 (the best score was obtained on running with 3 epochs)
- metrics: 'precision', 'recall', 'f1', 'accuracy'
- tensorflow.keras.callbacks.EarlyStoppig(monitor='val_loss', patience=3)

From the dataset, we only used emotions and triggers from every conversation. Because the model solves a tagging problem, we arranged the attributes in two separate lists, so that there is a 1-1 correspondence between their elements. We also renamed the triggers into labels: 0.0 = ’no’ and 1.0 = ‘yes’. An example of preprocessed data objects for NER can be observed in the third annexe.

{

"tokens": ["neutral", "neutral", "neutral", "neutral", "surprise"],

"labels": ["no", "no", "no", "yes", "no"]

}

After that, we applied tokenization with AutoTokenizer from transformers.

We also concatenate the train and validation dataset and applied a random split on the result, with the pivot value of 80% of the dataset length, so that we use 80% for training and 20% for validation.

5 Results

Upon comparing the Rule-Based Function and Named-Entity Recognition Methods, we obtained the results (F1 score of the triggers) displayed in the following table:

	Method	Score
1	Rule-based method	0.45
2	NER model with 3 epochs - cased	0.68
3	NER model with 3 epochs with rule-based method (XOR function applied on outputs) cased	0.47
4	NER model with 1 epoch cased	0.67
5	NER model with 5 epochs cased	0.66
6	NER model with 3 epochs uncased	0.70

Table 3: Comparing Scores (Rule-Based Function – NER) methods

From the results, it is evident that the highest F1 score is achieved by submission 2, which utilized the NER model trained over 3 epochs. Interestingly, as the number of epochs exceeded 5, we observed a consistent decrease in the F1 score.

	Method	F1 Score
1	RoBERTa Baseline	0.00
2	DistilBERT Baseline	0.00
3	NER model with 3 epochs - cased	0.68
4	NER model with 3 epochs - uncased	0.70

Table 4: Comparing Scores (Baselines vs NER)

The preceding table showcases the results achieved with the various methods we applied. Notably, the method using NER with 3 epochs outperformed the others, achieving F1 scores between 0.6 and 0.7 (Training/Validation). In comparison, our implementations using baseline methods yielded lower F1 scores: sthe DistilBERT Baseline method obtained a score of 0.1811%, and the RoBERTa Baseline method achieved 0.2452% (Training/Validation). It’s crucial to note that these scores were calculated using our custom-defined F1 scoring function, tailored to the traditional method. Furthermore, a 0.00% score was observed when applying a different F1 scoring approach.

6 Conclusion

In this paper, we demonstrated that employing a Named-Entity Recognition (NER) model trained over 3 epochs for emotion flip detection yields superior results compared to classical approaches such as the RoBERTa and DistilBERT baselines, as well as a rule-based strategy. Our team's mixed and comparative solution outperformed the baseline models in terms of outcomes and provided valuable insights for future research on architecture and model enhancements. Notably, our method, utilizing the NER model trained over 3 epochs, achieved the highest F1 score. However, it is crucial to note that increasing the number of epochs beyond 5 led to a consistent decrease in the F1 score. Our evaluation indicates a significant performance improvement (~60% in F1-score) compared to previous studies.

In this way, we discovered that this is a complex problem, revealing numerous intriguing avenues for further exploration. Nevertheless, it is crucial to consider the potential benefits of incorporating audio and visual support, which could lead to enhanced performance. This insight prompts us to contemplate an exciting investigation for the future.

References

Kumar, S., Shrimal, A., Akhtar, Md S., Chakraborty, T. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. In: Knowledge-Based Systems, Vol. 240, 108112, ISSN 0950-7051 https://doi.org/10.1016/j.knosys.2021.108112.

S. Kumar, S. Dudeja, M. S. Akhtar and T. Chakraborty, "Emotion Flip Reasoning in Multiparty Conversations," in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 3, pp. 1339-1348, March 2024, doi: 10.1109/TAI.2023.3289937.

Kumar, Shivani and S, Ramaneswaran and Akhtar, Md and Chakraborty, Tanmoy 2023. From Multilingual Complexity to Emotional Clarity: Leveraging Commonsense to Unveil Emotions in Code-Mixed Dialogues. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 9638-9652 doi: 10.18653/v1/2023.emnlp-main.598

Kumar, Shivani and Akhtar, Md Shad and Cambria, Erik and Chakraborty, Tanmoy 2024. "SemEval 2024 -- Task 10: Emotion Discovery and Reasoning its Flip in

Conversation (EDiReF)". In: Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics. https://arxiv.org/abs/2402.18944

Gîfu, D. and Cioca, M. 2013. Online Civic Identity. Extraction of Features. In: Procedia - Social and Behavioral Sciences, Vol. 76, University of Piteşti Publishing House 2013, pages 366-371, Elsevier, ISSN 1844-6272, https://doi.org/10.1016/j.sbspro.2013.04.129.

Ouatu, B., Gîfu. D. 2020. Chatbot, the Future of Learning? In: Proceedings of the 5th International Conference on Smart Learning Ecosystems and Regional Development (SLERD 2020), in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education, Springer, pages 263-268. https://api.semanticscholar.org/CorpusID:224946479.

Madasu, A., Firdaus, M., and Ekbal, A. 2023. A Unified Framework for Emotion Identification and Generation in Dialogues. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 73–78, Dubrovnik, Croatia. Association for Computational Linguistics. https://aclanthology.org/2023.eacl-srw.7/.

Cristea, D., Gîfu, D., Colhon, M., Diac, P., Bibiri, A., Mărănduc, C., and Scutelnicu, L.-A. 2015. Chapter - Quo Vadis: A Corpus of Entities and Relations. In: Language, Production, Cognition, and the Lexicon. Text, Speech and Language Technology, Part VI - Language Resouces and Langauge Engineering, Nuria Gala, Reinhard Rapp and Gemma Bel-Enguix (eds.), Vol. 48, New York, USA, pages 505-543. https://doi.org/10.1007/978-3-319-08043-7_28

De Bruyne, L, Singh, P., De Clercq, O., Lefever, E., Hoste, V. 2022. How Language-Dependent is Emotion Detection? Evidence from Multilingual BERT. In Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL), pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://aclanthology.org/2022.mrl-1.7/

Bulla, L., Gangemi, A., Mongiovi', M. 2023. Towards Distribution-shift Robust Text Classification of Emotional Content. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8256–8268, Toronto, Canada. Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.524/.

Bayram, U. and Benhiba, L. 2022. Emotionally-Informed Models for Detecting Moments of Change and Suicide Risk Levels in Longitudinal Social Media Data. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, pages 219–225, Seattle, USA. Association for Computational Linguistics. https://aclanthology.org/2022.clpsych-1.20/.

Li, J., Ji, D., Li, F., Zhang, M., and Liu, Y. 2020. HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.370/.

Kumar, S., Dudeja, S., Akhtar, M. S., and Chakraborty, T. 2023. "Emotion Flip Reasoning in Multiparty Conversations," in IEEE Transactions on Artificial Intelligence, doi: 10.1109/TAI.2023.3289937. https://ieeexplore.ieee.org/document/10164178.

Zhong, P., Wang, D., and Miao, C. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 165–176, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1016/.

Lu, X., Zhao, Y., Wu, Y., Tian, Y., Chen, H., and Qin, B. 2020. An Iterative Emotion Interaction Network for Emotion Recognition in Conversations. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4078–4088, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.360/.

Zhu, L., Pergola, G., Gui, L., Zhou, D., and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1571–1582, Online. Association for Computational Linguistics. https://aclanthology.org/2021.acl-long.125/.

Sanh, V. et al. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108. https://arxiv.org/pdf/1910.01108.pdf

A. Appendices

Data Object After Preprocessing RoBERTa

```
{
    "utterance": "also I was the point person on
my company's transition from the KL-5 to GR-6
system",
    "emotion": "neutral"
    "speakers": [
      "Chandler",
      "The Interviewer",
      "Chandler",
      "The Interviewer",
      "Chandler"
    ],
    "utterances": [
      "also I was the point person on my
company's transition from the KL-5 to GR-6
system.",
      "You must've had your hands full.",
      "That I did. That I did.",
      "So let's talk a little bit about your duties.",
      "My duties?  All right."
    ],
    "emotions": [
      "neutral",
      "neutral",
      "neutral",
      "neutral",
      "surprise"
    ]
  }
```

B. Appendices

Data Object After Preprocessing DistilBERT

```
{
"entry_index": 0,
      "entry": "0 - Chandler - also I was the point
person on my company's transition from the KL-5 to
GR-6 system. - neutral",
    "context":
"Chandler: also I was the point person on my company's
transition from the KL-5 to GR-6 system. – neutral
The Interviewer: You must've had your hands full. –
neutral
Chandler: That I did. That I did. – neutral
The Interviewer: So let's talk a little bit about your
duties. – neutral
Chandler: My duties?  All right. - surprise"
      }
```

C.Appendices

Data Object After Preprocessing NER

```
{
"tokens": [ "neutral", "neutral", "neutral", "neutral",
"surprise"],
"labels": [ "no", "no", "no", "yes", "no"]
}
```