

Aspect-based Automated Evaluation of Dialogues

Arash Shabanpour^{a,b}, Ziwei Hou^a, Akbar Husnoo^a, Khanh Linh Nguyen^b, John Yearwood^a and Nayyar Zaidi^{a,b,*}

^a*School of Information Technology, Faculty of Sci Eng & Built Env, Deakin University, Burwood VIC 3220, Australia*

^b*DataScienceWorks Research Lab, Melbourne 3000, Australia*

ARTICLE INFO

Keywords:

Aspect-based Learning
Fusion models
BERT model
Dialogue Evaluation
Criteria Enforcement
End-to-end Deep Learning Framework


ABSTRACT

Evaluating human dialogue is a complex task, as our conversation are never structured. There are, however, cases where there is some structure in our conversation, e.g., in a typical call center, dialogue between a call center agent and customer revolves around certain topics of conversation. These dialogues can be evaluated based on some pre-specified criteria as well as sub-criteria. This evaluation is typically done manually, which can be time-consuming, motivating the need for an automated system that employs Artificial Intelligence (AI) algorithms to evaluate dialogues efficiently. In this paper, we have proposed a novel dialogue-evaluation framework that leverages recent advancements in deep learning research. The contributions of this work are two fold. Firstly, we introduce a straightforward end-to-end framework – CallAI, for evaluating dialogues in any domain, based on some predefined hierarchical criteria. Secondly, we present a novel algorithm – TAABLM, utilizing a novel combination of aspect-based learning along with traditional TF-IDF features for text. We show in this paper, that TAABLM outperforms conventional baselines such as BERT, LSTM, etc, delivering improved performance in automated dialogue evaluation, whereas CallAI offers a simple yet elegant framework for an AI-based solution to hierarchical dialogue evaluation. We demonstrate the efficacy of our proposed framework and proposed algorithm on three datasets, where we quantify performance in terms of an aggregated dialogue-score as well as in terms of either accuracy or AuROC metrics.

1. Introduction

Evaluating human dialogues poses a significant challenge due to their inherent lack of structure. Our conversations are not solely composed of words; they encompass a range of elements such as tone, emotions, sarcasm, and contextual nuances [20]. However, in certain situations, dialogues tend to exhibit more structure, precision, and systematic characteristics, allowing for objective evaluation of their quality. For instance, call-center interactions often involve time-sensitive and structured discussions between two parties with a specific objective. Other examples include conversations between doctors and patients, interactions between customers and clerks, and exchanges between lawyers and clients, etc. In these scenarios, an additional human element is introduced into the evaluation loop to assess the dialogue's quality. Within context of a call-center environment, the evaluation of an agent's performance relies on predefined standards of communication between agents and customers. Key criteria may include assessing whether the agent greeted the customer appropriately, if the problem was effectively resolved, and did the customer express satisfaction after the call. These factors play a crucial role in determining the agent's performance and the overall quality of the customer interaction. Typically, once the call (dialogue) is finished, a designated individual, often the agent's supervisor, will review the recorded conversation. The purpose of this review is to assess the quality of the call and provide an evaluation score that reflects the agent's performance. Naturally, the supervisor conducts the evaluation based on the predefined standards and criteria established for such assessments. Figure 1a illustrates this standard process of evaluating calls in the agent-customer relationship. It is important to note that this manual evaluation is not only laborious and time-consuming but also expensive, thus motivates the need to automate of the evaluation process. Fortunately, advancements in Natural Language Processing (NLP) and the advent of deep learning techniques [18, 13] have opened up new possibilities for applying these methodologies to dialogue evaluation. This research addresses this need by systematically investigating the application of state-of-the-art deep learning trends in dialogue evaluation. By leveraging cutting-edge techniques, this work aims to bridge the existing gap and contribute to the advancement of automated dialogue evaluation.

*Corresponding author

 nayyar.zaidi@deakin.edu.au (N. Zaidi)

ORCID(s): 0000-0003-4024-2517 (N. Zaidi)

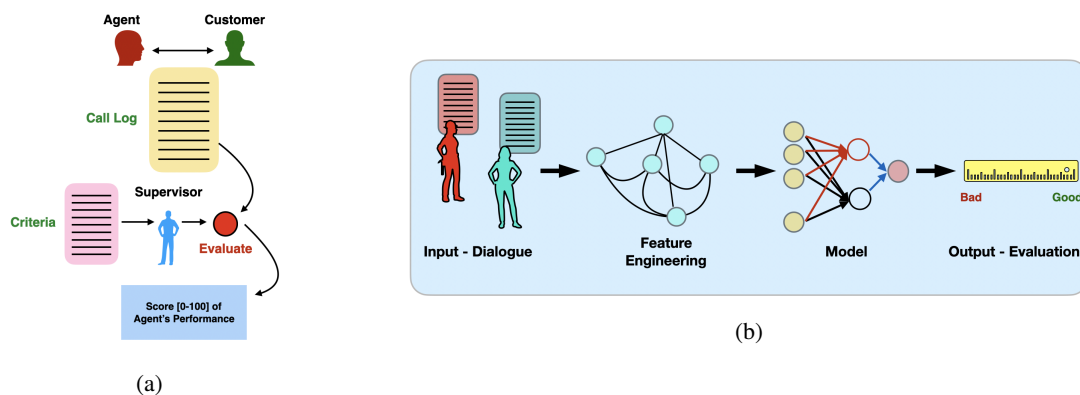


Figure 1: (Left) An illustration of a typical dialog evaluation process, involving three parties (Agent, Customer and Supervisor). (Right) A dialogue evaluation pipeline with deep learning.

A dialogue is constituted of various sentences – each of these sentences must be processed either separately or sequentially to build an effective dialogue evaluation system. It is important to note that dialogue evaluation can be nested (or hierarchical). That is, there can be multiple main criteria – and each of them can have multiple sub-criteria, and so on. For example, a sentence in a dialogue – “*I will look into your record now, can you state your full name for security purpose?*” – can contribute towards fulfilling the Engagement criterion, while also ticks the box for some sub-criterion about customer authentication. How can we process a sentence to determine if it fulfills multiple criterion? Well, recently it has been shown that a sentence can be processed in terms of various aspects. This is motivated from sentiment analysis¹. For example, a review about a restaurant may have positive sentiment about restaurant’s food, but negative sentiment about restaurant’s staff. Here, food and staff are an example of aspects that might exist in a sentence (review). We claim that aspects can play an important role in building an effective dialogue evaluation system as well, as they can contribute towards determining the fulfillment of various sub-criteria in the sentence. For example, the sentence in a dialogue – “*After investigation, I found the problem is with the payment system you are using. I will fix this issue by updating your credit card. Thank you for your patience*”, has at least two aspects – found the problem as well as patience. The first aspect can contribute towards Problem Investigation criterion, while later can contribute towards Engagement criterion. We believe that aspects offers a natural solution to address multiple criterion nature of dialogue evaluation. In this work, We will propose an algorithm for dialogue evaluation based on aspects and we will compare it with other state-of-the-art methods such as LSTM and BERT. We will demonstrate our proposed model’s efficacy empirically. We argue to be the first work to apply aspect-based learning for dialogue evaluation problem.

At the heart, dialogue evaluation is a regression or classification problem. That is, given a sequence of sentences (comprising a dialogue), the goal is to assign a number which quantify the quality of the dialogue. There are at least two problems here:

- Dialogue evaluation can be hierarchical. That is, a score of 80 can be associated with a dialogue, but it can encompass multiple criterion of 30 in engagement, 40 in resolution and 10 in closure. Furthermore, the score of 30 in engagement category could be due to a score of 15 for greetings sub-category and 15 for good-bye, etc. This hierarchical nature of criterion is hard to be enforced in general regression or classification settings.
- The second problem stems from the representation of words and sentences in dialogues. Typically, words are represented by pre-trained word embeddings such as GloVe embeddings – however, given the specialized nature of various dialogue (e.g., in medical domain, law domain, etc.), a better representation of words is required.

To address the first problem of hierarchical dialogue evaluation, we have introduced an end-to-end framework of dialogue evaluation. To the best of our knowledge, all existing works on dialogue evaluation do not consider a nested

¹There has been a lot of interest in *aspects* in NLP research with several applications in sentiment analysis [7, 26], summarization [9, 15], translation [2], etc. E.g., “*New York is a great city, but its people are the best*” – reveals two aspects ‘city’ and ‘people’. There is a positive sentiment for the city but even more positive about people aspect.

criteria. The problem of hierarchical dialogue evaluation is challenging as the criteria can shift from one problem to another problem. We endeavour to be the first work that has presented a simple yet effective strategy for hierarchical dialogue evaluation. By doing so, we highlight and enunciate various elements of the framework such as different levels of models, criteria, evaluation, etc, and test its efficacy on on a simple contrived dataset to demonstrate its effectiveness.

Note that we have constrained ourselves to call-centers dialogues, however, the evaluation framework proposed in the paper is applicable to similar controlled environment dialogues.

To address the second problem (of representation) – we have made use of aspect learning. Additionally, since many sentences in typical dialogue are short², on these short texts, we argue that the presence and absence of some words can help in determining the label of a sentence. For example, mere presence of words Good Morning, Would you like to hear can reveal that the sentence belongs to either Greetings or Engagement category. Any effective dialogue evaluation engine must exploit the presence or absence of these words to determine if some criteria is being enforced or not. Therefore, we have incorporated the use of TF-IDF along with aspect learning to find a better representation of each sentence in a dialogue. In fact, the integration of aspect-based features along with TF-IDF features is novel, and we present this language model denoted as TAABLM – *TF-IDF Augmented Aspect-based Language Model*, as the second contribution of this work. Note, much of the existing related works utilizing deep learning methods relies on state-of-the-art (SOTA) models such as BERT. It will be interesting to see how our proposed algorithm performs as compared to existing SOTA models.

We will discuss that our proposed hierarchical dialogue evaluation model Call-AI, and our proposed language model TAABLM are supervised in nature. This can be seen as the main limitation of the proposed work. One must have access to supervised data, i.e., dialogues where criterion and sub-criterion is demonstrated to retrain our model. This limitation can not be considered as a drawback, as many tasks in NLP such as sentiment analysis, document classification are supervised in nature. For CallAI, it is crucial and can limit its application as finding dialogue data with nested level of criteria is hard to find. In fact, we could not find a single data source in public domain. We will discuss in this paper that we have curated a simple dataset of 150 dialogue, which will be published along side this paper – where multiple criterion of evaluation has been enforced.

The main motivation of this work has been to develop and demonstrate a working AI framework for hierarchical dialogue evaluation and study and investigate and design effective techniques that underlay these dialogue evaluations. The contributions of this work are as follows:

- First, we have proposed a novel framework CallAI – for evaluating and scoring dialogues in a controlled environment. We discuss how multiple criteria and sub-criteria can be incorporated by implementing various layers of models. The main motivation for proposing such framework is to demonstrate a mechanism in which various layered machine learning models can be used to address the need of hierarchical dialogue evaluation.
- Second, we have proposed a language model that is based on the fusion of aspect-based learning and traditional TF-IDF features. We claim to be the first work that studies the augmentation of TF-IDF features with aspect-based learning. Of course, the proposed model plays a pivotal role in determining the effectiveness of our CallAI framework, however, it is important to note that any other model, e.g., BERT or RNN can be used in CallAI.
- We have gathered and curated a dataset of 150 sentences, to test our proposed CallAI framework. We have also generated a dataset of size over 6000 through ChatGPT by providing it examples of our curated small dataset. To test the efficacy of our proposed language model, we have compared it with other SOTA models such as BERT, LSTM, etc. on three dialogue evaluation datasets.

This paper is organized as follows: we will start by discussing the preliminaries and the related work in the following section. Later, we will present our proposed framework. This will be followed by our evaluation of the efficacy of our proposed framework. Lastly, we conclude this work with pointers to future works.

2. Preliminaries

Let us discuss some preliminaries in this section. We will start by discussing the nature of dialogues and the need for evaluation. Later we will delve into the study of typical deep learning pipelines for dialogue evaluation.

²E.g., “Good Morning, how can I help you, sir?”, “Would you like to hear about our newly launched product?”.

2.1. Dialogue Evaluation and Nested Criteria

Typical dialogues among humans are hard to evaluate as there is usually more than one context in which we interact with others. In certain environments, the dialogue can be between two parties and the context of the discussion can be limited. It is in these scenarios, one can impose some evaluation criteria to determine the quality of the dialogue. Of course, determining the quality of such dialogue will benefit the business [19]. Therefore, the business actually derives the criteria of call or dialogue’s quality. Since the criteria are different among businesses (use-cases), any AI model that is built should take into account that use-case-specific criteria. For example, in one use case, a criterion of ‘Customer Greetings’ can account for 50% of the overall score, whereas in other cases, it could be just 5%. When building an AI model, one can not rely on a one-size-fits-all solution, and models need to be trained separately for each business. More importantly, they need to be re-trained as criteria changes.

As mentioned earlier in Section 1, criteria can be nested or hierarchical in nature. E.g., there can be main criterion but this can be broken down into multiple sub-criteria. Developing an end-to-end AI framework for such hierarchical criteria is not trivial. An example of hierarchical criteria for dialogue evaluation in a call center is depicted in Table 1. where evaluation is divided into five categories– Greetings, Account Verification, Engagement, Problem Resolution, Closure. Now the presence or absence of any of these category can constitute a first level of criteria. However, each of the categories can have an associated weight. There can be certain metrics that define the quality of that class. E.g., ‘Introduce himself’ or ‘Ask the name of the person calling’, etc. This constitute a second level criteria. It is important to note that, given a problem and a domain, several nested sub-criteria can exist. We have limited our study in this work to either the first or the second level of nesting. Needless to say that evaluation of hierarchical criteria can be quite tedious, leaving the job of human evaluator extremely challenging.

2.2. Typical Deep Learning Pipeline

Much of the success of deep learning has been on the datasets where data features have a certain structure, e.g., images or text data³ – we will call this ‘Structured Data’. It is on datasets like these, deep learning can engineer features by the use of either convolution, recurrence or attention layers. We have mentioned before that dialogues in constrained environments do have some structure, which makes them amenable to the application of deep learning. A typical pipeline could be pre-processing of the input data, followed by the application of the deep learning model – which produces an output ‘evaluation’ (this simple pipeline is shown in Figure 1b). This pipeline has at least two issues:

- First, there is no way to incorporate criteria in the evaluation yet alone hierarchical criteria.
- Second, feature engineering is neither defined nor trivial for dialogue data. Note, the architecture shows ‘feature engineering’ and ‘Model’, as two separate entities. In practice, a trained model will have an internal feature engineering mechanism.

In this work, we will address these two issues and introduce a framework CaLLAI to incorporate criteria, and a language model TAABLM for effective feature engineering based on TF-IDF and aspect-based learning.

We will discuss various evaluation pipelines in Section 3, but before we move forward, let us in the following discuss three models that are typically used in dialogue evaluation frameworks:

- Simplest approach is that one can train a typical ANN model to classify sentences (where each word is represented as an embedding vector) to either belong to any of the categories or sub-categories.
- One drawback of using ANN models is that it fails to take into account the inter-relationship among the words in a sentence or in a document. Instead of using an ANN, one can instead use Recurrent Neural Networks (RNN) or its variant LSTM [12] models. RNN is an Artificial Neural Network with the ability of modeling sequence data. It processes a sequence of inputs (such as words in a sentence) by iterating through the elements of the sequence and maintaining the information about the elements it has seen in the sequence. Long Short Term Memory (LSTM) model are variants of RNN models and generally address the vanishing gradients problem of RNN, in pursuit of training a deep model on long sequences.
- BERT (Bi-directional Encoder Representations from Transformers) models have presented SOTA results in many NLP tasks such as sentiment analysis, question answering, and others [6]. It is inspired by *Transformer* model.

³Note that in an image, each feature (pixel) is correlated with its neighbour features, similarly in a sentence, there is a correlation among words.

| Class | Scores | Fields | Positive Examples |
|----------------------|--------|---|---|
| Greetings | 10 | Thank the customer for calling. | Hello. Thank you for calling [COMPANY NAME]. My name is [CALL AGENT NAME]. May I have your name? |
| | | Introduce himself. | |
| | | Ask about the reason for the call. | Hello [CUSTOMER NAME]. What can I help you with today? |
| | | Ask the name of the person calling | |
| Account Verification | 20 | Ask accurate and clear details from the user. | For verification purposes, may I have your name and address please? |
| | | Ask for permission through the use of words like "Can", "May" or "Please". | Can I please have your payment ID? |
| Engagement | 15 | Not allowed to use swear words. | I am so sorry for any inconvenience caused by this [CUSTOMER'S PROBLEM]. |
| | | Be impartial to all genders. | |
| | | Be professional and courteous in all dealings. | Thank you for your patience. I will be here to make sure everything is right. |
| Problem Resolution | 50 | Try their best to help the customer and/or resolve the problem on the spot. | Since you did pay for your connection, what I can do is to open a ticket or so that the other department can immediately complete your payment as soon as possible. |
| Closure | 5 | Offer further assistance. | Have I addressed all your problems today? |
| | | Thank the customer for calling. | |
| | | Greet farewell to the customers | I am really glad to help. Thank you for calling Rex Global and have a wonderful evening. |

Table 1: Example of hierarchical criterion for evaluating dialogues in a Call Center.

Note, transformer is designed based on an encoder-decoder architecture, with an attention mechanism that can capture the contextual relationship among the words in sentence [27]. BERT is a multi-layer bidirectional transformer constituting only an encoder. The input to BERT model are individual word embeddings. Its key advantage over other language models includes a) a novel technique named masked-language modelling and b) next sentence prediction which allows contextual learning. In masked-language modelling, 15% of the words in each sentence are masked before feeding into the BERT model and the model should predict the value of these masked words based on the context of the other non-masked words in the sentence. In the sentence prediction step, the BERT model receives the pairs of sentences as its input and needs to learn if the second sentence in the pair is the subsequent sentence of the first sentence or not. In the training phase, 50% of the pairs includes 2 subsequent sentences and the other 50% consists of 2 random sentences. Masked-language modelling and next sentence prediction are trained together which generates a language model that can be fine-tuned for specific tasks. For example, by adding a classification layer on top of the model we can retrain the model for any tasks. It is worth mentioning that an advantage of BERT over RNN model is that, BERT is a non-directional model which means that it read the whole input (sequence of words) at once instead of reading it sequentially (from right to left or left to right).

3. Related Work

In the past decades, there are some related works in the context of dialogue evaluation from the literature. Paprzycki et al. [22] were among the first few researchers who initiated research on automated call-centre agent's performance evaluation using well-known data mining approaches. Their research focused on the building of six different models using machine learning techniques⁴. The models were trained and then evaluated on an actual dataset consisting of one year records of five call centres of a big insurance company. The authors claimed that CART produced the best results with an accuracy of 89.48% in determining customer satisfaction. Note that in order to evaluate each call, the authors hand-crafted 11 different features which were further fed to the aforementioned classification model. Similarly, Zweig et al. [30] proposed an automated system for evaluating call centre agents' performance based on a set of 31 questions from the IBM's North American Help Desk evaluation form. Utilizing a partial score-based approach

⁴Multi-layer Perceptrons (MLP), Linear Neural Networks (LNN), Probabilistic Neural Networks (PNN), Classification and Regression Trees (CART), Support Vector Machines (SVM) and Hybrid Decision Tree.

and a maximum entropy classifier, the probability of a call being classified as bad or good was determined based on a set of user-derived features, e.g., ‘maximum silence length’, ‘the occurrence of selected n-gram word sequences’, etc. The authors evaluated their system on 195 calls from one of the IBM call centers and achieved a precision of 60%. Kim [17] proposed an online call monitoring approach to detect bad calls before it is too late to drive customer retention. To do so, the author employed a class conditional N-gram based language model and an adaptive BoosTexter classifier [24]. Using an undisclosed dataset with 10,000 call logs, the author trained the models and evaluated the models using 1000 logs. After testing the models, the author concluded that the N-gram based language model achieved a higher accuracy of 83.0% as opposed to 73.3% achieved by BoosTexter classifier. Our work differs from these works as we do not use hand-crafted features. Additionally, unlike our framework, these works can not incorporate any criteria or sub-criteria in their evaluation.

Similarly, Ezzat et al. [8] proposed an approach to perform sentiment analysis on call centre conversations using text classification techniques. Textual features were extracted, pre-processed and selected. Using Chi-Square keyword extractor, the relevant keywords are extracted and fed to four machine learning classifiers⁵. After evaluation and comparison of their classification models on an artificially generated dataset, the authors concluded that SVM with key graph extraction produced the highest accuracy. Note that authors extracts textual features identified using bag-of-words method while feature selection method is applied to reduce dimensionality. Karakus and Aydin [16] developed a novel system by leveraging big data technologies and distributed NLP methods for automated call centre performance evaluation based on six pre-defined metrics. After the conversion of calls⁶ to text, various similarity-based approaches were used. The authors highlighted that Cosine and Jaccard similarities achieved higher performances. This work differs from our works as the main focus of their work is scalability and not performance. These lines of work relies on traditional machine learning models and text mining strategies unlike the use of deep learning approach approach in our work.

Roy et al. [23] and Mariappan et al. [21] put forwarded a system named – QART^{RT}, an end-to-end real-time quality assurance for automated evaluation of call centre agent performance. The proposed system consists of four main components. *Customer behavior* component is responsible for emotion recognition and categorization, *conversational characteristics* component is responsible for automatic detection of deviations from standard operating procedures employing various deep learning models, *dialogue summarization* is responsible for easing context-at-a-glance by supervisors through the use of information templates, and lastly, an *interactive visualization* component is used for visualization of evaluations at real-time. Overall, the authors concluded that their approach drastically reduced the tedious manual process and achieved an overall better accuracy in most components as opposed to manual evaluation. A similar work is that of Abhinav et al. [1], where authors proposed a deep learning-based approach to evaluate call centre agents and provide feedback to improve call quality in real time. Their model consisted of an embedding layer, one CNN, and two LSTM layers and five dense layers to evaluate a call based on ten pre-defined organizational compliance metrics. To evaluate their proposed method, the authors used 300 calls from a call centre and claimed that their approach recorded an F1-score of 80% and above for almost all predefined metrics. This work in particular is the closest to our work, as the authors also employ deep learning methods for call scoring. We will compare the performance of our model with that of language model of QART, and nominate this as state-of-the-art work in this area.

We mentioned aspect-based learning briefly in Section 1, and will delve into its details in Section 4, but let us first briefly discuss some works related to aspects and dialogue systems. Most notable work is that of Guhr et al. [10], in which a sentiment classification model based on aspect learning is built to rate conversation and evaluate users’ reactions to changes in the behaviour of an (automated) dialog engine. Javdan et al. [14] considered aspects detecting sarcasm in dialogues. The paper proposed a combination of BERT pre-trained model with aspect-based sentiment analysis models. Song et al. [25] focuses on the sentiment and mention extractions in dialogues. The authors used BERT to extract the sentiment expressions and mentions of particular words in dialogue. Note, all these studies have focused on sentiment classifications on dialogue datasets, whereas, we focus on dialogue evaluation in any domain utilizing a pre-defined criteria. However, we share similarity with these works as we utilize aspect-based learning to extract better word-level and sentence-level representations.

4. Proposed Framework and Language Model

Let us discuss our proposed framework – CallAI and proposed language model– TAABLM in this section.

⁵SVM, Naive-Bayes, Decision tree and the K-Nearest Neighbor classifiers

⁶Gathered from an undisclosed Turkish call center.

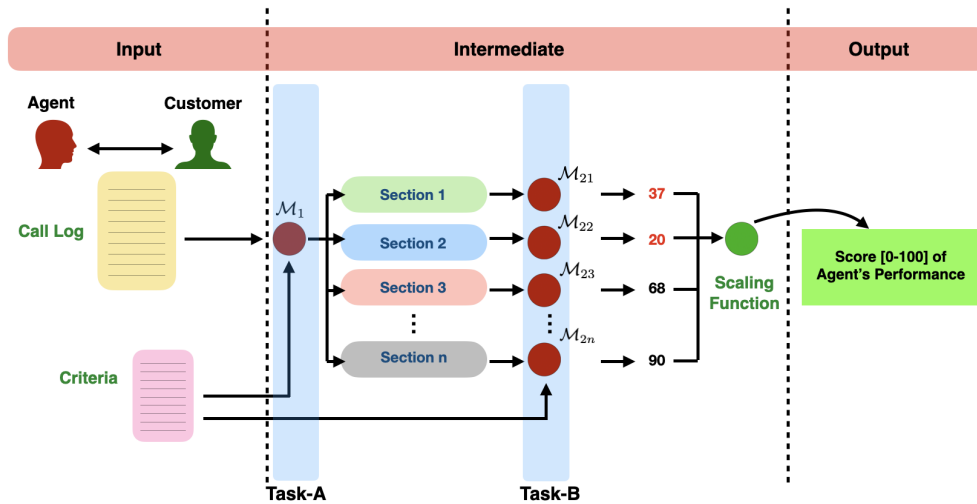


Figure 2: Our Proposed Dialog Evaluation Framework. Red circles depict trained language models. Section 1, ...Section n represents sentences to be input to the model. Models can either perform Task-A (dialogue classification), i.e., determine existence of a category in input or Task-B (Criteria Enforcement), i.e., determine a score associated with that criteria or sub-criteria.

4.1. CallAI

CallAI is based on a layered approach to accommodate criteria and sub-criteria for dialogue evaluation. It is composed of three layers, namely: a) Input Layer, b) Intermediate Layer, and c) Output Layer, as illustrated in Figure 2. The intermediate layer can be further nested (divided) to incorporate a nested criteria. Let us discuss these three layers in the following.

4.1.1. Input Layer

The input level of CallAI includes application of diarization model [28] and is followed by speech-to-text. The output of input level is a text file clearly labelling dialogue sentences belonging to either customer (person-A) or agent (person-B). Note, diarization and speech-to-text models are not re-trained, and standard off-the-shelf models are used. It is important to note, that the output of diarization and speech-to-text models should be checked for sanity, and any issues as a result should be addressed. For example, typical issues include incorrect assignment of ‘customer’ and ‘agent’ to their dialogue in the call. There will be some cases where dialogues are already present in form of text. In such cases, diarization and speech-to-text will not be needed. However, even in this case, input layer will do some dataset specific pre-processing of input text sentences.

4.1.2. Intermediate Layer

The input to intermediate layer is the output of input layer (i.e., text containing labelled dialogue between two parties). The role of intermediate layer of CallAI is to incorporate the specified criteria or sub-criteria. It does this by fundamentally achieving two tasks:

- *dialogue classification* (we will refer to this task as task-A) – determines the existence of a sentences that pertains to a particular criterion for evaluation (note, this is effectively a multi-class text classification task), and
- *criteria enforcement* (we will refer to this task as task-B) – if the sentences exists in the dialogue, how can a score be assigned, again based on a specified criteria.

It can be seen from Figure 2 that intermediate layer trains several models. For instance, \mathcal{M}_1 is trained to produce a Boolean value, highlighting if certain sections exists in the dialogue related to a specific criteria. As we discussed earlier in Section 1 that in one of our dataset, criteria incorporate existence of following categories: Greetings, Account Verification, Engagement, Problem Resolution, and Closure. Here, \mathcal{M}_1 will determine if these five categorize exist, and if they do, it outputs sentences associated with these categories, as shown in the form of Section

1, Section 2, ..., Section n in Figure 2. These section of dialogues are used as input to another set of models to perform task-B. Note, in Figure 2, they are denoted as $\mathcal{M}_{21}, \dots, \mathcal{M}_{2n}$. Given the form of criteria or sub-criteria, the output of these models can take two form:

- perform task-A again – that is given Section n of the text – determine the existence of some sub-criteria, this will be followed by task-B, or
- assign a score based on how well a criteria has been satisfied. We call this the terminal case, as the output of these models is combined and is fed to the output layer.

The terminal case is shown in Figure 2, where the models $\mathcal{M}_{21}, \dots, \mathcal{M}_{2n}$ actually produce a scalar score, which is passed through a criteria dependent *Scaling Function*. For example, *Greeting* criterion can contribute to 20% of the final score, and some sub-criterion with-in *Greeting* class can contribute to 50% of that 20%.

A nested intermediate layer is shown in Figure 3. It can be seen that all models determines the existence of a category (related to criteria) – i.e., they perform task-A. If sentences exists, they are passed to second layer of models, which again determines if a sentences related to sub-criteria exists, and if they do – they are passed to the models which performs task-B, i.e., determine how well the criteria is satisfied. Note, this nested approach is scalable as long as enough examples are provided for a criterion or sub-criterion, the models can learn to either do dialogue evaluation or criteria enforcement tasks, – therefore, the nested intermediate layer can easily accommodate any level of hierarchical criteria. We will discuss the complexity of underlying language model in the later section.

4.1.3. Output Layer

The output layer in CallAI takes care of presentation and visualization of the results. The layer can take an input from any of the model in the intermediate layer, and presents the results in a form needed for an individual scenario.

4.1.4. Evaluation Metric

CallAI is a novel and unique framework that has the capability to incorporate both criteria and associated sub-criteria. As we discussed earlier, evaluating such framework requires datasets that are not publicly available. It is important to note that, the goal of such framework is to output one final score – we call it the **Dialogue Score** –which is weighted sum of various sub-criteria from any dialogue. One measure of evaluation of this framework is how much the predicted dialogue score deviates from original dialogue score. We will use this metric while evaluating our proposed framework on our hand curated dataset. Note, the framework relies on models to either perform task-A or task-B (as discussed earlier). In the experimental section, we will consider various state-of-the-art language model and compare their performance with our proposed language model (TAABLM) under the umbrella of CallAI.

4.1.5. Limitations of CallAI

The existence of supervised (labelled) data is one of the limitation of CallAI model. The dialogues must be first annotated, and all the section in the data must be attributed to a criteria or sub-criteria (task-A), as well as conformance to the criteria as well (task-B). We conjecture that such limitation should not considered a short-coming of our proposed framework, this is because organizations already have large quantities of available data that can be annotated to train CallAI framework.

4.2. TAABLM – TF-IDF fused Aspect-based Model

Now that we have discussed a framework for dialogue evaluation based on nested criteria – CallAI, let us discuss various language models that can be used in the framework. We discussed that underlying models in CallAI performs two tasks, that is:

1. determines the category or label of each section of the dialogue – *dialogue evaluation* (task-A).
2. determines the quality of the dialogue, in other words, it determines conformance of that section of dialogue to the criteria – *criteria enforcement* (task-B).

For both tasks, one can use any of the language models that we discussed in Section 2.2, however, in this work, we have proposed a novel model – TAABLM, which combines typical TF-IDF representation with that of aspect learning (Figure 4). The main motivation for such model stems from two observations. First, aspect-based models have been show to be quite effective in a wide range of sentiment analysis tasks. Second, by their definition, they are useful in scenarios where a context specific representation of words or sentences is required. In our case where a model is

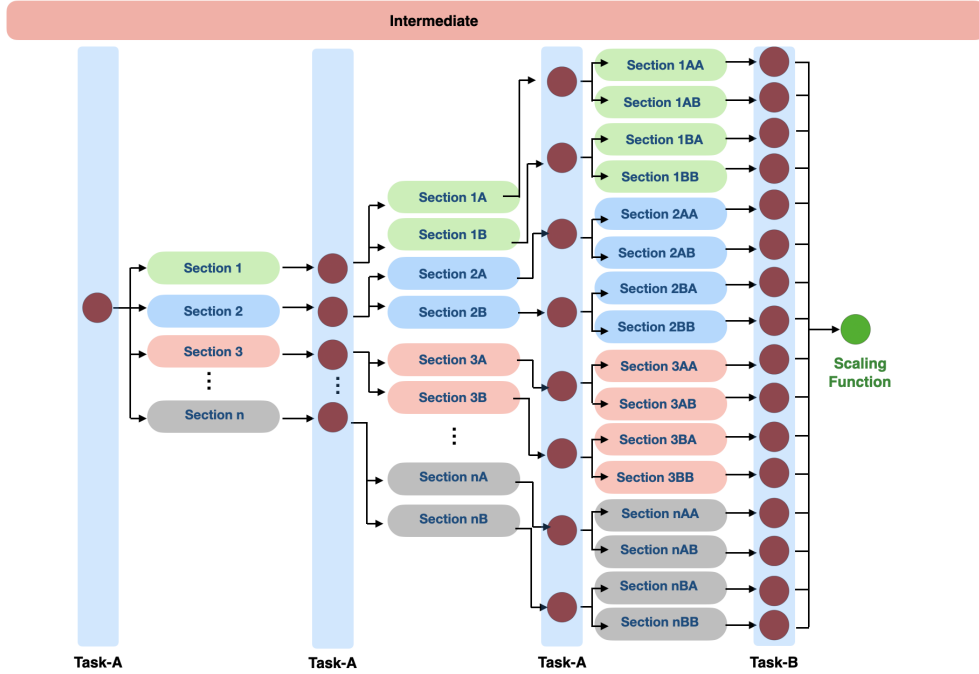


Figure 3: Nested Intermediate Layer.

expected to do both dialogue evaluation, and criteria enforcement – aspect-based learning has its merits. Additionally, we conjecture that in nested criteria based dialogue evaluation, aspect-based learning can play a natural role. We will discuss in the following that TAABLM is based on the fusion of TF-IDF features along with aspect-based learning. Though, aspect-based learning has been combined with models such as BERT in the recent past, combination of TF-IDF with aspect-based learning has not been explored. The motivation for the use of TF-IDF is because of one property of sentence in dialogue text. Our initial investigation showed that most sentences in dialogues were extremely short. Simple features based on TF-IDF has been shown to be quite effective on shorter texts. We studied the performance of CallAI model separately with TF-IDF features and then with aspect-based learning – later studied the combination of these two models, leading to our proposed formulation TAABLM language model.

Let us start by discussing how aspects are learned in aspect-based learning. Well, the determination of aspects is based on attention mechanism [11]. Note, the goal of aspect-based learning is to capture the relevance of each word with some aspect in the sentence. Specifically, for each word w_i in the sentence, a positive weight a_i is learned which can be interpreted as the degree of the relevancy of the word to an aspect, as:

$$a_i = \frac{\exp s_i}{\sum_{j=1}^n \exp s_j}. \quad (1)$$

Here n denotes the number of words in our sentence and s_i is defined as:

$$s_i = e_{w_i}^T \cdot M \cdot y_s. \quad (2)$$

Here $e_{w_i} \in \mathbb{R}^d$ denotes the embedding vector of the word i , and M is the *Aspect Matrix* – i.e., we have potentially d number of aspects. Note, $M \in \mathbb{R}^{d \times d}$ constitutes a set of learnable parameter of the attention layer, where d is the size of word embedding e_{w_i} . The matrix M can be interpreted as a matrix that captures the relevance of each word with each aspect. The global context of the sentence s is denoted as y_s , and is defined as:

$$y_s = \frac{1}{n} \sum_{i=1}^n e_{w_i}. \quad (3)$$

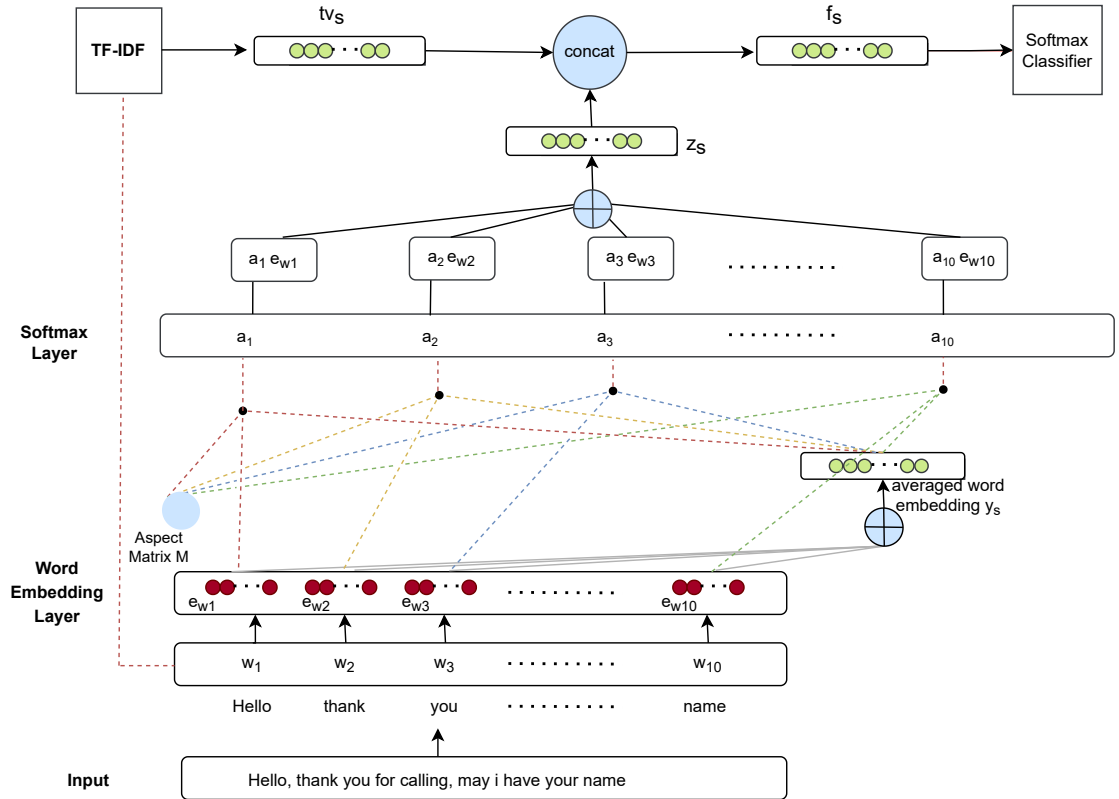


Figure 4: Our proposed aspect-based text classification with sentence embeddings.

Aspect-based sentence embedding is obtained by weighted averaging of the word embeddings as follows:

$$z_s = \sum_{i=1}^n a_i e_{w_i}. \quad (4)$$

Our proposed model – TAABLM, also incorporates TF-IDF representation of sentences as well – which is represented by the symbol tv_s . TAABLM concatenates aspect-based and TF-IDF based sentence embeddings to obtain a final representation f_s as depicted in the following equation:

$$f_s = \text{Concatenate}(z_s, tv_s). \quad (5)$$

Note, $f_s \in R^m$, where m is the size of final sentence embedding. The final sentence embedding – f_s , is fed to a Softmax layer to either train the *dialogue classification* task or *criteria enforcement* task. The Softmax layer has the trainable weights: $W \in R^{m \times k}$, where k is the number of classes, and can be written as:

$$\hat{Y} = \text{Softmax}(W^T f_s). \quad (6)$$

Once \hat{Y} is obtained, we used the cross-entropy loss as the objective function, and optimize it by back-propagating the error through a standard deep learning optimization procedure to update the learnable weights of our model, that is: W and M .

4.2.1. On Computational Efficiency

Aspect-based learning is based on attention mechanism – in attention based algorithms, complexity is dependent on the number of words in input sentences. We mentioned earlier that since the number of words in a standard sentence of

a dialogue are not many – complexity component of aspect-based learning is not huge. TF-IDF, on the other hand, can be computed quickly, making it efficient in real-time dialogue systems, especially when dealing with large volumes of data. Moreover, TF-IDF uses sparse representations of vectors that can reduce memory consumption and computation, especially for large dialogue systems involving a large number of conversations.

5. Empirical Evaluation

In this section, we will empirically evaluate our proposed CALLAI framework and TAABLM language model. We will start by discussing the datasets that we have used to train the model, followed by discussion of our experimental settings. Later we will discuss the evaluation of our language model for *dialogue classification* and *criteria enforcement*.

5.1. Dataset

For evaluation of our proposed framework, we have used three datasets:

- Call Centre data (CCD),
- Banking data (BD),
- Call Centre data (Generated) (CCD-CGPT).

In the following we will discuss these datasets along with the annotation process we undertook.

5.1.1. Call Center Data

The call center data as denoted as CCD is a collection of dialogues between agent and customer of a call centre, and is curated from available resources on the internet. The dataset is disclosed as the supplementary material for this paper. It consists of over 150 call.

The conversations were automatically transcribed to text format and customer and agent were identified using the diarization model [28], as mentioned earlier. The output was manually revised for ensuring consistency. Punctuation and grammar issues were also addressed in each dialogue. An example of dialogue can be seen in Figure 5 after a proper transcription. On average, each dialogue in this dataset has around 25 samples (sentences).

Annotation: Dataset is annotated based on the criteria of Figure 1. That is, five classes are identified *Greetings*, *Account Verification*, *Engagement*, *Problem Resolution*, *Closure*, and underlying sub-criteria is established for each class. The average number of sentences in each category is shown in Figure 6. Note, most existing work in dialogue evaluation does not publish the criteria – by curating this dataset, we have endeavoured to create a dataset for research community, along with formulation of a simple criteria and a sub-criteria. The dataset is annotated by the authors – the annotation involved:

- identifying various categories in each call for *dialogue classification*, and
- identifying the quality for each category – in our case, we have constrained to binary classification, i.e., identifying if adheres to the criteria or not – *criteria enforcement*.

5.1.2. Banking Data

This dataset is used to detect intent in fine-grained single-domain text. It is composed of online banking queries annotated with their corresponding intents. Due to lack of available datasets in public domains, we have used this dataset to test the efficacy of our proposed framework and language model. It comprises 13,083 customer service queries labeled with 77 intents, and can be downloaded from: <https://huggingface.co/datasets/banking77>. Let us see a few example in this dataset, along with their labels:

```
{'label': 11, # "card_arrival" intent
  'text': "I am still waiting on my card?"}
{'label': 13, # "card_linking" intent
  'text': "Okay, I found my card, can I put it back in the app?"}
{'label': 32, # "exchange_rate" intent
  'text': "What is my money worth in other countries?"}
{'label': 24, # "country_support" intent
  'text': "Are cards available in the EU?"}
```

It is noteworthy that most of the sentences are short, also, presence of some words can play key role in determining the category of each sentence, motivating the need for TF-IDF and aspect-based learning.

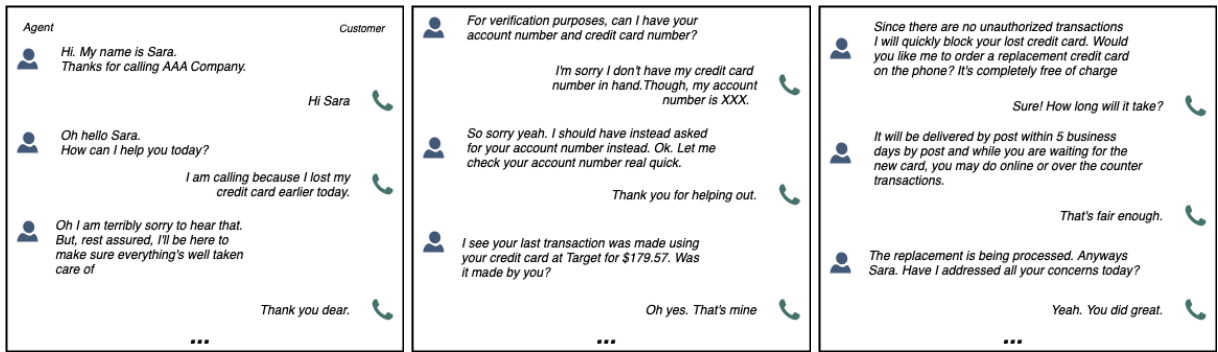


Figure 5: A sample of a call-log from CCD dataset after transcription.

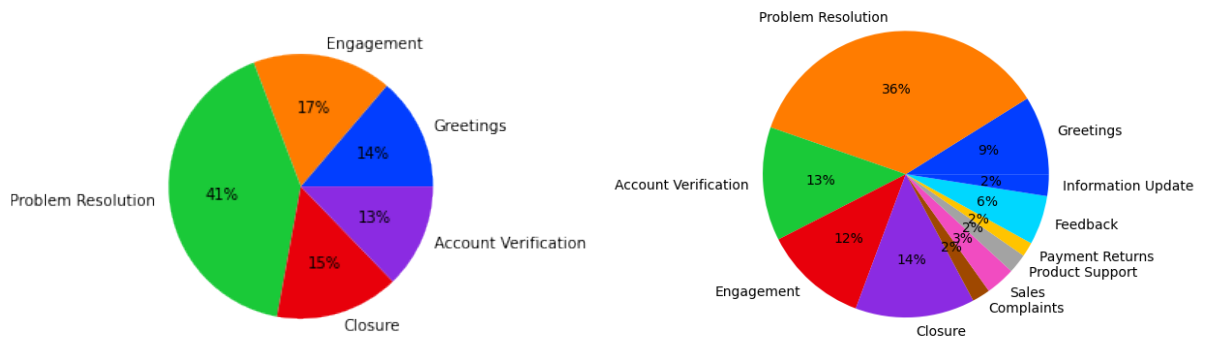


Figure 6: (Left) Class distribution in CCD dataset, (Right) Class distribution in CCD-CGPT dataset.

5.1.3. Call Center Data (Generated)

The dataset is created by us by synthesizing and generating content using ChatGPT with appropriate prompts. The main motivation of creating this dataset is to address small data size of CCD. The objective here is to replicate CCD while creating more data, as well as introducing as many categories as possible. The dataset is annotated by our team. The prompts used for this purpose are structured as follows:

I am doing the problem called Aspect-based sentiment analysis for position caller

Take a look at these examples:

```
Example 1:[{'Person': 'A',
  'Category': 'Greetings',
  'Quality': 'Positive',
  'Text': 'Thank you for calling Triple-A bank.'},
{'Person': 'A',
  'Category': 'Greetings',
  'Quality': 'Positive',
  'Text': 'My name is Arianna.'},
{'Person': 'A',
  'Category': 'Greetings',
  'Quality': 'Positive',
  'Text': 'May I have your name?'},
{'Person': 'C', 'Category': nan, 'Quality': nan, 'Text': 'James breed; '}]
```

```
Example 2: [{'Person': 'A',
  'Category': 'Greetings',
  'Quality': nan,
  'Text': 'How can I make you smile today?'},
```

```
{'Person': 'C',
 'Category': nan,
 'Quality': nan,
 'Text': 'Why do I have fees on my bank account? '}]
```

```
Example 3: [{'Person': 'A',
 'Category': 'Problem Resolution',
 'Quality': 'Positive',
 'Text': "I'll see what I can do Mr. Breed."},
 {'Person': 'A',
 'Category': 'Problem Resolution',
 'Quality': 'Positive',
 'Text': 'As I have checked you were charged some fees because you were not able to pay
 on time so that means we really cannot do anything about it anymore.'}]
```

```
Example 4: [{'Person': 'A',
 'Category': 'Engagement',
 'Quality': 'Positive',
 'Text': "I truly understand what you're going through Mr. Breed."}]
```

```
Example 5: [{'Person': 'A',
 'Category': 'Account Verification',
 'Quality': 'Positive',
 'Text': 'To put your account, can I please have your account number or debit card number
?'}],
```

```
Example 6: [{'Person': 'A',
 'Category': 'Closure',
 'Quality': 'Positive',
 'Text': 'I have already submitted a request and you will receive an email or text
message after three days for the decision.'},
 {'Person': 'C',
 'Category': nan,
 'Quality': nan,
 'Text': 'Thanks for trying.'}]
```

With:

```
Person: C - customer, A: assistance
Category: Greetings, Account Verification, Problem Resolution, Engagement Closure and nan
if you dont know type of category
Quality: positive, negative, nan if neutral
Text: The text the person said
```

Give me some random full conversations with a variety of topics such as: banking, sales, product support, payment returns, complaints, problem solving, information updates, feedback that ends in about 10 to 30 sentences.

Includes one or a combination of 2, 3, 4, 5 of the sections Category: Greetings, Account Verification, Problem Resolution, Engagement Closure, Nan and Quality: positive, negative, nan. Just return the result as a python list of dict-like in the example.

Note, we tried various prompts and choose this prompt after experimenting and interacting with ChatGPT. The dataset generated was proof-read by our team, and corrected for any obvious flaws. The dataset has over 6000 dialogues. The average number of sentences in each category is shown in Figure 6.

5.1.4. Evaluation Metric

When testing binary classification performance of our proposed language model TAABLM, we will make use of AUC (criteria enforcement task), while accuracy will be used to test the performance of of multi-class problem (dialogue classification task).

5.1.5. Pre-processing Steps

Most of the pre-processing is needed for CCD data, as it is obtained in audio form, and requires speech-to-text and diarization, followed by manual annotation and sanctity checks. BD and CCD-CGPT datasets are already in text

formats. For all datasets, we removed all the punctuation, converted all texts to lowercase, and turned the texts into space-separated sequences of words.

5.2. Experimental Settings

To train various language models, we initialize the word embedding matrix with word vectors trained by Word2vec and set the embedding size to 300. We also set the word embedding layer trainable during the training process, other learnable parameters are initialized randomly. The model is trained with Adam optimizer with initial learning rate of 0.001. The batch size is set to 8 and early stopping is used. The batch size parameter is set after doing some preliminary experiments with varying batch sizes. The number of aspects are set equal to 100. Again, this is configured after doing some preliminary experiments.

Our empirical evaluation constitutes comparing following language models:

1. **ANN** – Typical ANN model (5 hidden layers with 10 nodes each (ReLU activations)).
2. **QART** – One CNN layer, two LSTM layers (10 nodes each) and five dense layers (10 nodes each) – with ReLU activations.
3. **BERT** – BERT model. The model is illustrated in Figure 8 in Appendix.
4. **TAABLM** – Our proposed language model as illustrated in Figure 4.
5. **ASPECTS** – Our proposed model but utilizing only aspects, i.e., excluding the TF-IDF features. In this formulation, Equation 5 is replaced as:

$$f_s = z_s. \quad (7)$$

6. **TF-IDF** – Language model that is trained on only TF-IDF. In this formulation, Equation 5 is replaced as:

$$f_s = tv_s. \quad (8)$$

We use 80 : 20 split of data for training and test, and repeated the experiments 5 times, and report the average results.

5.3. Dialogue Classification

We compare the performance of different language models in terms of Accuracy and AUC evaluated under our proposed CallAI framework as shown in Table 2 on CCD dataset. It can be seen that TAABLM model outperforms all other five models, reaching an accuracy of 81.55%. It is surprising to see that the simple TF-IDF model performs better than BERT, which is usually the state-of-the-art model in various NLP tasks. One reason of this is that BERT does require large quantities of training data to train its parameters. Nonetheless, it is encouraging to see an improvement of 6% with our proposed model against BERT. Similarity QART model’s performance is not good (even worst than standard ANN – this could be due to over-fitting of the model on this dataset. It is important to note, that, we use the term QART for representing the model only – however, as we discussed in Section 3, QART entails a complete platform of dialogue evaluation, and its model is just one component. The individual confusion matrices of the two models (TAABLM and BERT) is shown in Figure 7. Again, it can be seen from the confusion matrices that our proposed model does a much better job of classifying than state-of-the-art model BERT.

The comparison of the performance of different language models in terms of accuracy evaluated under our proposed CallAI framework on BD and CCD-CGPT datasets is shown in Table 3 and 4 respectively. It is encouraging to see that our proposed language model TAABLM outperforms all other baselines, including state-of-the-art BERT model.

5.4. Criteria Enforcement

Criteria enforcement results for six language models are shown in Table 2, where we present the average accuracy and AUC for each model. It can be seen that our proposed model TAABLM has similar performance in terms of the accuracy as compared to state-of-the-art BERT model, with TAABLM reaching an accuracy of 96.83% as compared to 97% of BERT model. Similar pattern can be seen in terms of AUC, with TAABLM reaching an AUC of 93.52% as compared to 93.38% for BERT. It is important to note that since there is class imbalance, AUC results are relevant here. Getting a performance that is as good as state-of-the-art BERT is extremely encouraging.

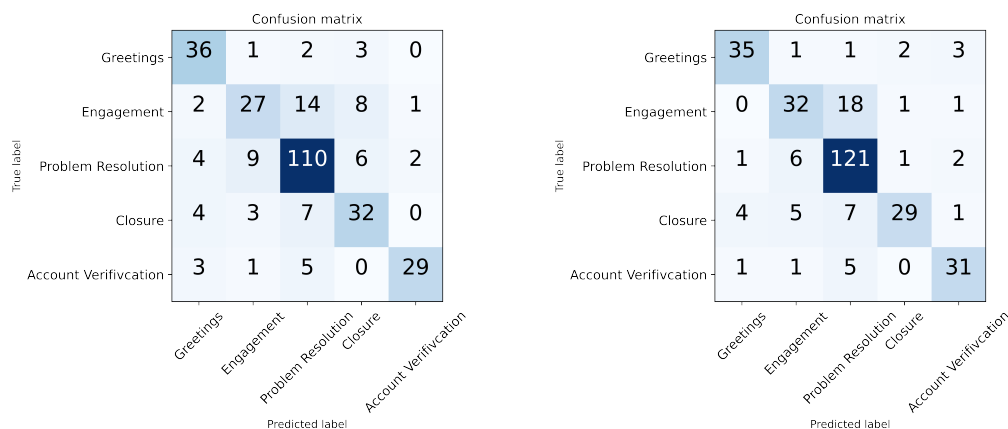


Figure 7: Confusion matrices for Dialogue evaluation task – (Left) BERT, (Right) TAABLM.

| Category | QART | ANN | BERT | ASPECT | TF-IDF | TAABLM |
|--|------|------|-------|--------|--------|--------------|
| Dialogue Evaluation – Accuracy | | | | | | |
| | 51 | 59 | 75.73 | 80.57 | 80.91 | 81.55 |
| Criteria Enforcement – Accuracy | | | | | | |
| Greetings | 81.7 | 85.5 | 97.62 | 97.62 | 92.86 | 97.62 |
| Account Verification | 75.3 | 82.3 | 100 | 100 | 97.44 | 100 |
| Problem Resolution | 87.1 | 89.3 | 100 | 100 | 98.48 | 100 |
| Engagement | 84.6 | 64.9 | 94.23 | 94.23 | 92.31 | 94.23 |
| Closure | 75.5 | 80.3 | 93.48 | 92 | 91.30 | 92.31 |
| Average | 81.1 | 81.3 | 97.0 | 96.81 | 94.44 | 96.83 |
| Criteria Enforcement – AUC | | | | | | |
| Greetings | 65.9 | 65.8 | 95 | 95 | 85 | 95 |
| Account Verification | 51.7 | 63.7 | 100 | 100 | 93.75 | 100 |
| Problem Resolution | 52.4 | 50.1 | 100 | 100 | 96.8 | 100 |
| Engagement | 50.6 | 49.6 | 88 | 83.8 | 80 | 88.8 |
| Closure | 53.6 | 56.6 | 83.8 | 82.2 | 80 | 83.8 |
| Average | 54.4 | 56.0 | 93.38 | 92.2 | 87.11 | 93.52 |

Table 2: Comparison of the performance of different language models in terms of Accuracy and AUC evaluated under our proposed Ca11AI framework on CCD dataset.

| QART | ANN | BERT | ASPECT | TF-IDF | TAABLM |
|---------------------------------------|-----|-------|--------|--------|--------------|
| Dialogue Evaluation – Accuracy | | | | | |
| 78.71 | 70 | 86.79 | 88.34 | 86.85 | 89.97 |

Table 3: Comparison of the performance of different language models in terms of Accuracy evaluated under our proposed Ca11AI framework on BD dataset.

5.5. Call Scoring

In this section, we will demonstrate the effectiveness of our framework by formulating a call scoring and evaluation end-to-end system. This is to highlight the efficacy of our proposed Ca11AI framework. We are interested to determine if our proposed framework (powered by TAABLM language model) can achieve the job for what it is designed for – i.e., read a dialogue and then produce a dialogue score based on some criteria. To perform this experiment, we selected 15 dialogues from CCD dataset (note that these dialogues are not included in the training). Our framework first applies TAABLM model for dialogue classification to these dialogues, and based on its prediction, applies the model for criteria enforcement task. A scaling function converts the output of criteria enforcement task to a scalar number. The output of the model for all categories is averaged to produce a final score. We threshold the probability score of our

| QART | ANN | BERT | ASPECT | TF-IDF | TAABLM |
|---------------------------------------|-------|-------|--------|--------|--------|
| Dialogue Evaluation – Accuracy | | | | | |
| 52 | 56.23 | 75.67 | 76.80 | 76.27 | 76.94 |

Table 4: Comparison of the performance of different language models in terms of Accuracy evaluated under our proposed CallAI framework on CCD-CGPT dataset.

model to covert its output into a score.

This is followed by the application of averaging to produce a final dialogue score that is presented as ‘Predicted Score’ in Table 5. We also present the ‘True Score’ as well as the absolute difference between the two scores (labelled as ‘Error’). We define the average error as ϵ defined as: $\epsilon = \frac{1}{N} \sum_{i=1}^N \text{Error}_i$. Note that a smaller value of ϵ is desirable. It can be seen that on these randomly selected $N = 15$ calls, our proposed CallAI framework achieves an average error (ϵ) of 20.6. Also, CallAI works much better for calls that have intermediate true scores, i.e., between 85 and 60, where ϵ is equal to 12. For calls that have an associated high true score of 100, our model has an ϵ of 25.6, and for calls with a low true score that is < 50 – CallAI achieves an ϵ of 27.3. We see these results as extremely encouraging and step in the right direction. Nonetheless, it demonstrates that our proposed CallAI framework can achieve the task of call dialogue scoring effectively.

5.6. Ablation Study

In the ablation studies, we would like to study the impact of adding TF-IDF features to standard BERT model, as well as combining our model with that of BERT model. We will compare the performance of TAABLM with following two language models:

1. **BERT + TF-IDF** – We augmented traditional BERT model with TF-IDF sentence representation features.
2. **BERT + ASPECTS + TF-IDF** – We combined our proposed TAABLM model with BERT model as shown in Figure 9 in Appendix.

It can be seen from Table 6, that BERT + TF-IDF leads to an improvement of 1.63% over BERT for dialogue classification task. This is consistent with comparison of ASPECT vs. TAABLM – demonstrating the usefulness of TF-IDF for dialogue classification task. As we discussed in Section 1, the presence of TF-IDF features can greatly facilitate state-of-the-art models on datasets comprising of short texts, and these results support our claim. In terms of criteria enforcement task, both BERT and BERT + TF-IDF lead to similar accuracy and AUC. Additionally, Table 6 shows BERT + ASPECTS + TF-IDF results for dialogue classification. It can be seen that the resulting model has better performance than BERT + TF-IDF but worst results than TAABLM for dialogue classification. Note that in BERT + ASPECTS + TF-IDF model, we used word embeddings computed by BERT model (instead of static word embeddings from Glove used in TAABLM model). We conjecture that due to short text and limited data size, BERT embeddings are not as representative as Glove embeddings are. This is one reason why BERT + ASPECT + TF-IDF does not perform better than TAABLM model, however, it does warrant further study as well as analysis on larger training data.

In Table 7, we report results of TAABLM on two datasets – BD and CCD-CGPT with varying training data. It is encouraging to see that the performance of our model improves with more training data. Nonetheless, with extremely small training data (i.e., only 1000 examples), TAABLM reach an accuracy of over 70%, which is extremely encouraging.

6. Conclusion

In this work, we proposed an end-to-end framework – CallAI, that measures the quality of a dialogue between a customer and an agent based on some pre-defined criteria. We employed a novel fusion of aspect-based learning and TF-IDF features – TAABLM and evaluated it for dialogue classification and criteria enforcement under CallAI framework. On a standard dataset, our proposed model – TAABLM resulted in better performance than state-of-the-art model such as BERT. Our ablation studies demonstrated the importance of simple TF-IDF features for short texts. Furthermore, we conducted call scoring using our proposed end-to-end system, and demonstrated we can achieve this task quite effectively. The scarcity of datasets has been one of the biggest challenges for this work. We believe, that with more data and better annotation, the performance of our framework can be further improved.

One future line of research that we have undertaken is the incorporation of customer’s answers in the evaluation process and building the model to incorporate these answers. Note, this will involve formulating dialogues as an MDP

| | True Score | Predicted Score | Error |
|--------------------------------|------------|-----------------|-------|
| Call-1 | 100 | 84 | 16 |
| Call-2 | 100 | 72 | 28 |
| Call-3 | 100 | 70 | 30 |
| Call-3 | 100 | 75 | 25 |
| Call-5 | 100 | 71 | 29 |
| Average Score of Calls [1-5] | 100 | 74.4 | 25.6 |
| Call-6 | 85 | 80 | 5 |
| Call-7 | 80 | 61 | 19 |
| Call-8 | 75 | 90 | 15 |
| Call-9 | 73.5 | 48 | 25.5 |
| Call-10 | 65 | 67.5 | 2.5 |
| Call-11 | 60 | 65 | 5 |
| Average Score of Calls [6-11] | 73 | 68.6 | 12 |
| Call-12 | 50 | 45 | 5 |
| Call-13 | 45 | 69.5 | 24.5 |
| Call-14 | 45 | 100 | 55 |
| Call-15 | 45 | 70 | 25 |
| Average Score of Calls [11-15] | 46.2 | 71 | 27.3 |
| Average Score of Calls [1-15] | 71.2 | 74.9 | 20.6 |

Table 5: Comparison of True and Predicted call scores.

| Category | BERT + TF-IDF | BERT + ASPECT + TF-IDF | TAABLM |
|---|---------------|------------------------|--------------|
| dialogue classification – Accuracy | | | |
| | 77.35 | 78.96 | 81.55 |
| Criteria Enforcement – Accuracy | | | |
| Greetings | 97.62 | 97.6 | 97.62 |
| Account Verification | 100 | 100 | 100 |
| Problem Resolution | 100 | 100 | 100 |
| Engagement | 94.23 | 94.23 | 94.23 |
| Closure | 93.48 | 93.48 | 92.31 |
| Average | 97 | 97 | 96.83 |
| Criteria Enforcement – AUC | | | |
| Greetings | 95 | 95 | 95 |
| Account Verification | 100 | 100 | 100 |
| Problem Resolution | 100 | 100 | 100 |
| Engagement | 88 | 88 | 88.8 |
| Closure | 83.8 | 83.8 | 83.8 |
| Average | 93.38 | 93.38 | 93.52 |

Table 6: Comparison of the performance of different language models in terms of Accuracy and AUC evaluated under our proposed CallAI framework on CCD dataset.

| training size | TAABLM | |
|---------------|--------|----------|
| | BD | CCD-CGPT |
| 1000 | 71 | 73.58 |
| 5000 | 86.51 | 76.94 |
| 10000 | 89.97 | NA |

Table 7: Comparative analysis of TAABLM’s performance in terms of accuracy on two datasets BD and CCD-CGPT with varying training data.

and we endeavour to solve it via reinforcement learning algorithms. Note, this line of research has been explored before [29, 5, 3, 4, 7], we are working on integrating CallAI and TAABLM under the framework of reinforcement learning. Secondly, we are also utilizing document summarization methods as well incorporation of hand-crafted features and sentiment analysis based features to further improve TAABLM model.

References

- [1] Abhinav, K., Dubey, A., Jain, S., Arora, V., Puttaveerana, A., Miller, S., 2019. Aqua: Automatic quality analysis of conversational scripts in real-time, in: ICAISC.
- [2] Barnes, J., Lambert, P., Badia, T., 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification., in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1613–1623.
- [3] Chen, L., Cao, J., Liang, W., Wu, J., Ye, Q., 2022. Keywords-enhanced deep reinforcement learning model for travel recommendation. *ACM Trans. Web* doi:10.1145/3570959.
- [4] Chen, L., Cao, J., Tao, H., Wu, J., 2023a. Trip reinforcement recommendation with graph-based representation learning. *ACM Trans. Knowl. Discov. Data* URL: <https://doi.org/10.1145/3564609>, doi:10.1145/3564609.
- [5] Chen, L., Zhu, G., Liang, W., Wang, Y., 2023b. Multi-objective reinforcement learning approach for trip recommendation. *Expert Systems with Applications* 226, 120145. URL: <https://www.sciencedirect.com/science/article/pii/S0957417423006474>, doi:<https://doi.org/10.1016/j.eswa.2023.120145>.
- [6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [7] Do, H.H., Prasad, P., Maag, A., Alsadoon, A., 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications* 118, 272–299.
- [8] Ezzat, S., Gayar, N.E., Ghanem, M., 2012. Sentiment analysis of call centre audio conversations using text classification. *IJCISIM* 4, 619–627.
- [9] Frermann, L., Klementiev, A., 2019. Inducing document structure for aspect-based summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6263–6273.
- [10] Guhr, O., Schumann, A.K., Bahrmann, F., Böhme, H.J., 2020. Training a broad-coverage german sentiment classification model for dialog systems, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 1627–1632.
- [11] He, R., Lee, W.S., Ng, H.T., Dahlmeier, D., 2017. An unsupervised neural attention model for aspect extraction, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 388–397.
- [12] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural computation* 9, 1735–1780.
- [13] Hong, S., Cohn, A., Hogg, D.C., 2022. Using graph representation learning with schema encoders to measure the severity of depressive symptoms, in: ICLR'2022.
- [14] Javdan, S., Minaei-Bidgoli, B., et al., 2020. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection, in: Proceedings of the second workshop on figurative language processing, pp. 67–71.
- [15] Kansal, H., Toshniwal, D., 2014. Aspect based summarization of context dependent opinion words. *Procedia Computer Science* 35, 166–175.
- [16] Karakus, B., Aydin, G., 2016. Call center performance evaluation using big data analytics, in: ISNCC '16, pp. 1–6.
- [17] Kim, W., 2007. Online call quality monitoring for automating agent-based call centers, in: INTERSPEECH '07.
- [18] Kumar, A., Tyagi, V., Das, S., 2021. Deep learning for hate speech detection in social media, in: IEEE GUCON'2021, pp. 1–4.
- [19] Lloyd, A., 2020. Efficiency, productivity and targets: The gap between ideology and reality in the call centre. *Critical Sociology* 46, 83–96.
- [20] Mann, W.C., Moore, J.A., Levin, J.A., 1977. A comprehension model for human dialogue, in: IJCAI'77, p. 77–87.
- [21] Mariappan, R., Peddamuthu, B., Raajaratnam, P.R., Dandapat, S., Pande, N., Roy, S., 2016. Qart: A tool for quality assurance in real-time in contact centers, in: CIKM '16, p. 2493–2496.
- [22] Paprzycki, M., Abraham, A., Guo, R., Mukkamala, S., 2004. Data mining approach for analyzing call center performance. ArXiv cs.AI/0405017.
- [23] Roy, S., Mariappan, R., Dandapat, S., Srivastava, S., Galhotra, S., Peddamuthu, B., 2016. Qart: A system for real-time holistic quality assurance for contact center dialogues, in: AAAI'16, p. 3768–3775.
- [24] Schapire, R.E., Singer, Y., 2000. Boostexter: A boosting-based system for text categorization. *Machine learning* 39, 135–168.
- [25] Song, L., Xin, C., Lai, S., Wang, A., Su, J., Xu, K., 2022. Casa: Conversational aspect sentiment analysis for dialogue understanding. *Journal of Artificial Intelligence Research* 73, 511–533.
- [26] Thet, T.T., Na, J.C., Khoo, C.S., 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science* 36, 823–848.
- [27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in NIPS, pp. 5998–6008.
- [28] Wang, Q., Downey, C., Wan, L., Mansfield, P.A., Moreno, I.L., 2018. Speaker diarization with lstm, in: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE. pp. 5239–5243.
- [29] Zhu, G., Cao, J., Chen, L., Wang, Y., Bu, Z., Yang, S., Wu, J., Wang, Z., 2023. A multi-task graph neural network with variational graph auto-encoders for session-based travel packages recommendation. *ACM Trans. Web* 17. URL: <https://doi.org/10.1145/3577032>, doi:10.1145/3577032.
- [30] Zweig, G., Siohan, O., Saon, G., Ramabhadran, B., Povey, D., Mangu, L., Kingsbury, B., 2006. Automated quality monitoring for call centers using speech and nlp technologies, p. 292–295.

A. Code and Datasets

The datasets used in this work, as well as the code for the models and pipelines, will be published along side the paper.

B. Alternative Architectures

The BERT architecture is shown in Figure 8. We represent, BERT + ASPECT + TF-IDF in Figure 9.

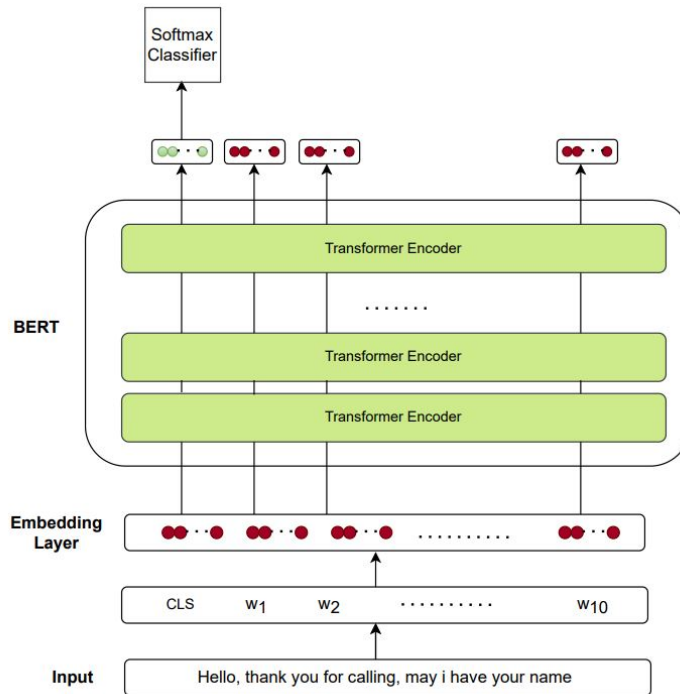


Figure 8: BERT Architecture.

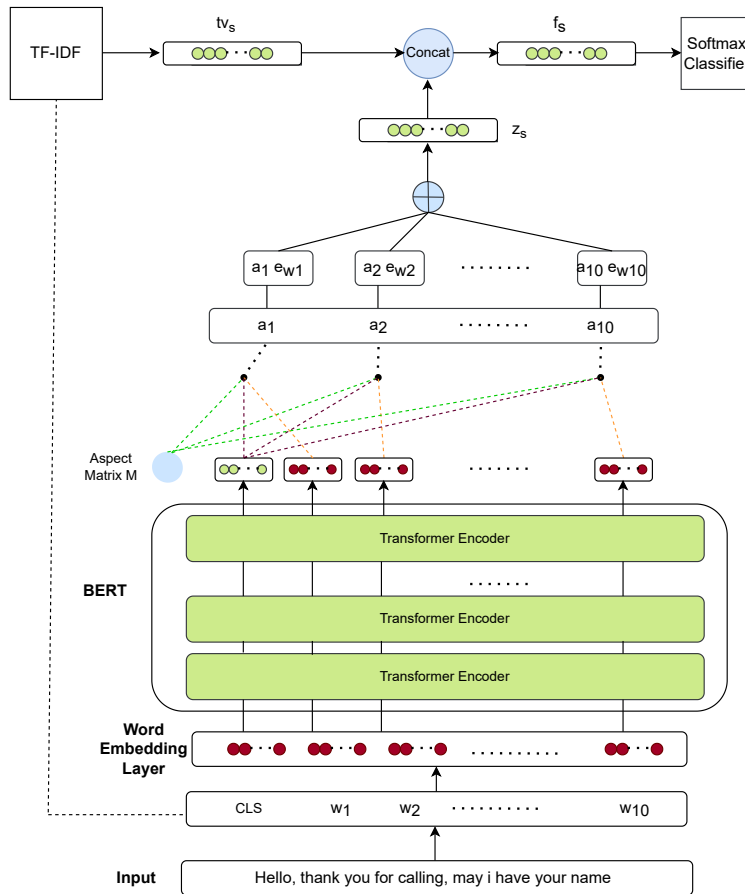


Figure 9: BERT + ASPECT + TF-IDF Architecture.