Multimodal Extraction and Recognition of Arabic Implicit Discourse Relations

Ahmed Ruby[1] **Christian Hardmeier**[2] **Sara Stymne**[1]

[1]Uppsala University, Department of Linguistics and Philology
[2]IT University of Copenhagen, Department of Computer Science
{ahmed.ruby, sara.stymne}@lingfil.uu.se, chrha@itu.dk

Abstract

Most research on implicit discourse relation identification has focused on written language, however, it is also crucial to understand these relations in spoken discourse. We introduce a novel method for implicit discourse relation identification across both text and speech, that allows us to extract examples of semantically equivalent pairs of implicit and explicit discourse markers, based on aligning speech+transcripts with subtitles in another language variant. We apply our method to Egyptian Arabic, resulting in a novel high-quality dataset of spoken implicit discourse relations. We present a comprehensive approach to modeling implicit discourse relation classification using audio and text data with a range of different models. We find that text-based models outperform audio-based models, but combining text and audio features can lead to enhanced performance.

1 Introduction

Understanding discourse relations in spoken language is crucial for effective communication and significantly enhances various language technology applications such as emotion recognition, text-to-speech (TTS) systems, and dialogue systems (Kharitonov et al., 2022; Kakouros et al., 2023; Ma et al., 2019). However, to the best of our knowledge, there are no datasets of labeled spoken data available that captures implicit discourse relations in a way that can be directly used to train and enhance these systems. This study addresses this gap by proposing a method for their automatic identification, that we use to construct a dataset of spoken discourse relations.

Identifying implicit discourse relations (IDR) or connectives in written text poses significant challenges. Research has shown that prosody or acoustic audio features can aid in their identification or characterization (Murray et al., 2006; Kleinhans et al., 2017; Ruby et al., 2024). Given this insight, our study focuses specifically on exploring implicit discourse relations in speech, investigating how auditory cues can help understand and identify complex relations, comparing the performance with models trained on text, or speech+text.

We noted that for TEDx talks, subtitles often maintain a formal style despite the informal language used by speakers. This motivated us to investigate the discrepancies between these subtitles and the spoken content. We found many instances where connectives explicitly stated in Modern Standard Arabic (MSA) subtitles were only implicitly present in the Egyptian Arabic (EGY) speech, as shown in Table 1. This phenomenon, known as explicitation in translation studies (Blum-Kulka, 2000), refers to the tendency of translators to make implicit elements explicit. It is frequently observed in translation and has even been proposed as a universal property of the translation process. Moreover, it has been claimed that explicitation can be used to obtain annotated data (Shi et al., 2017, 2019). We use this insight to design a method for automatically extracting and labeling EGY implicit discourse connection, based on the link to an explicit MSA connective. We use our proposed method to construct a dataset of EGY spoken implicit discourse relations from TEDx talks, paired with MSA subtitles and EGY transcripts. A manual verification of the quality of the corpus shows that the automatically created corpus is of very high quality, with no errors regarding discourse relation labels, and very few errors regarding the identification of the span of discourse arguments. This indicates that our proposed method can be used to create high-quality resources labeled with implicit discourse relations in speech.

We present experiments on IDR where we compare the usefulness of audio, text, and their combination. We present two sets of experiments: simpler models, where we avoid the bias of varying

Proceedings of the 31st International Conference on Computational Linguistics, pages 5415–5429
January 19–24, 2025.

Transcription	
Original Transcr. (MSA)	**Automatic Transcr. (EGY)**
00:08:37.666 –> 00:08:40.446	00:08:37.666 –> 00:08:40.446
وأخذها الميكانيكي كي يصلحها.	وخدها الميكانيكي يصلحها.
mechanic took it in order to repair it.	mechanic took it, repairing it.

Table 1: An example of an implicit connective in a TED Talk by an Egyptian speaker, explicitly stated in the subtitles provided by TED and remaining implicit in the automatic transcription generated by Whisper.

pre-training, and advanced deep learning models. We find that text-based models outperform audio-based models, but that combining the two can lead to further improvements. Even our pre-trained models struggle with IDR, though, with our highest accuracy being 0.45, showing that the IDR task still is challenging even for strong pre-trained models.

Our main contributions are:

1. We propose a novel method for automatically identifying implicit discourse relations in speech and text.

2. We introduce a novel high-quality dataset[1] of spoken implicit discourse relations in Egyptian Arabic with corresponding explicit connectives in MSA, which, to our knowledge, is the first of its kind.

3. We present experiments on IDR using audio and/or text data, using a range of different models.

2 Related Work

Most of the existing datasets for discourse relation identification are text-based. Notable text-based datasets include the Penn Discourse Treebank (PDTB) (Prasad et al., 2008b, 2019), the Rhetorical Structure Theory Discourse Treebank (RST) (Carlson et al., 2002), the Discourse Graphbank (Wolf et al., 2005), and the TED Multilingual Discourse Bank (TED-MDB) (Zeyrek et al., 2019). These datasets have been extensively used to train and evaluate models for discourse relation identification in written texts.

Since discourse relation identification, particularly for implicit relations, remains a significant challenge even with advancements in large language models, creating additional annotated datasets could be beneficial. However, this process is costly and requires expert annotations. Therefore, some researchers have investigated methods to automatically construct datasets. Shi et al. (2017) proposed a method to augment training data for English implicit discourse relation classification by leveraging explicitation in English-French parallel corpora. They back-translated the French target text into English and then applied a discourse parser to identify cases where connectives appeared in the back-translated text but not in the original English source, signaling implicit relations. Building on this, Shi et al. (2019) expanded the approach by incorporating translations from multiple languages such as, French, German, and Czech, to improve the quality and reliability of the additional training data. However, this method relies heavily on the quality of machine translation and discourse parsers, which can introduce errors, particularly with ambiguous connectives. Additionally, the effectiveness of the approach may be limited, when back-translating from a more verbose language into a less verbose language, connectives that are presented in the target language may not appear in the back-translated version. In a related effort Ma et al. (2019) propose a method to extract implicit discourse relation pairs from an English dialogue dataset. This is achieved by converting explicit relation pairs into implicit pairs by dropping the connectives. By leveraging unique dialogue features, their method significantly enhances the performance on IDR. However, the linguistic behavior of explicit relations may differ from implicit ones, introducing additional complexities. From a different angle, Omura et al. (2024) introduced a method for generating written synthetic data for IDR using a large language model. This technique significantly enhanced performance, particularly improving the recognition of infrequent discourse relations. However, synthetic data may lack diversity and authenticity in the training examples.

For audio-based discourse relation identification, some researchers have focused exclusively on using prosodic features. Murray et al. (2006) constructed a small dataset from the ICSI Meetings corpus, which consisted of manually labeled examples. In a related effort, Kleinhans et al. (2017) explored the correlation between the discourse structure of spoken monologues and their prosody by predicting discourse relations using various prosodic attributes. They used automatic annotation with a discourse parser to generate training data from TED.

More specifically related to this work, Al-Saif (2012) and Al-Saif and Markert (2010) conducted an annotation study for discourse relations in Ara-

[1] https://github.com/UppsalaNLP/Spoken-ImplicitDR

bic, focusing on explicit discourse connectives. This scheme was used to create the Leeds Arabic Discourse Treebank, a discourse corpus for Arabic[2].

To our knowledge, there is no existing Arabic dataset that includes labels for implicit discourse relations. This lack of resources highlights a research gap in developing models specifically designed to identify discourse relations in spoken language. Notably, we have proposed a method to construct a discourse relation dataset from TED Talks, which we consider to be the first of its kind.

3 Dataset Construction

We collected 87 TEDx Talks[3] presented in Egyptian Arabic by Egyptian speakers, each accompanied by subtitles in VTT format. These subtitles were created by TED's volunteer transcribers in accordance with TED guidelines,[4] which specifically recommend using language that is universally accessible and understood across all dialects. Given that Egyptian Arabic is the language used in these talks, the transcribers produced the transcriptions in MSA, as it is the standard form recognized across all Arabic dialects. Figure 1 presents a detailed flowchart summarizing the process of creating the spoken implicit discourse relation dataset.

3.1 Egyptian Arabic Speech recognition

To obtain aligned segments in both MSA and EGY, timestamps were extracted from the MSA subtitles to guide the segmentation of the audio files. Following this, the audio segments were transcribed [5] using Whisper-large-v3[6] (Radford et al., 2022), a model for speech-to-text conversion across multiple languages and dialects. This process allowed for the exact alignment of the audio with its corresponding transcriptions in both MSA and automated EGY subtitles.

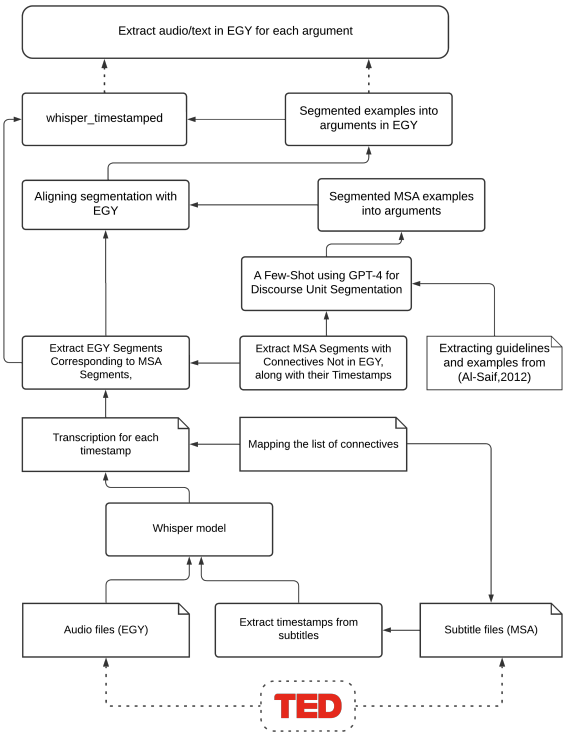

Figure 1: The process of creating spoken implicit discourse relations dataset

The transcription was performed using specific parameters to optimize performance: `beam_size=5` and `best_of=5` for improving the selection of most likely transcription paths, and a `temperature` value of 0.2, as recommended.[7]

3.2 Creating a List of connectives

To accurately map the automatically extracted connectives to their corresponding discourse relations, we focused only on unambiguous connectives that could be clearly associated with a single relation type. We identified instances where explicit connectives in the provided MSA subtitles were absent in the EGY automatic transcription by compiling two lists of connectives—one from the comprehensive lexicons available at Connective Lex[8] for MSA and another custom list that we created for EGY. By using the Connective Lex, we were able to exclude ambiguous connectives and focus on those that correspond to a single relational type. Connective Lex shows the potential relations for each connective, which allowed us to identify and use only those connectives with clear and

[2]The authors have confirmed that this dataset is no longer available.

[3]We used all Egyptian Arabic TED Talks that had MSA subtitles available at the time of our study, but selected only those that were transcribed by at least two volunteers—one serving as a transcriber and another as a reviewer—ensuring that the resulting subtitles provide a reliable source for identifying connectives absent in spoken audio. Moreover, more talks are continually added to TEDx, and could be used to expand our dataset in the future.

[4]https://www.ted.com/participate/translate/guidelines

[5]Since subtitles are only available for MSA, automatic transcription was necessary to generate subtitles for the EGY segments.

[6]https://github.com/openai/whisper

[7]https://github.com/linto-ai/whisper-timestamped

[8]http://connective-lex.info/

unambiguous relational types, such as لكن indicating *COMPARISON:Concession:Arg2-as-denier*. These lists are in Appendix A. With these refined lists, we analyzed both transcriptions, synchronized by matching timestamps, to accurately detect missing connectives.

3.3 Discourse Unit Segmentation

Discourse Units (DUs) are distinct segments of text or speech that contribute to building a representation of discourse. They can be clauses, sentences, or dialogue turns. Defining the boundaries of DUs is generally dependent on theoretical frameworks, as each theory specifies its own guidelines for segmentation and the scale of the units (Keskes, 2015). For instance, the definition of discourse units in Rhetorical Structure Theory (RST) slightly varies from the definition of arguments in the PDTB annotation (Al-Saif, 2012).

Since there are no resources available for Arabic discourse analysis, we extracted the annotation guidelines, principles, and examples from the thesis of Al-Saif (2012) to facilitate a few-shot learning approach with large language models (LLMs). This thesis presents the Leeds Arabic Discourse Treebank (LADTB), which follows the same principles as the English Penn Discourse Treebank (PDTB) but adapts and expands the annotations to account for the specific linguistic properties of Arabic. Using these guidelines, we direct LLMs to effectively identify only the elementary discourse units (EDUs) within Arabic texts.

We use the `gpt-4-turbo-2024-04-09` version of OpenAI's GPT-4 (OpenAI, 2023) model to segment discourse in MSA texts, based on the presence of connectives. This approach uses a few-shot learning setup where the model is instructed to identify and segment texts into distinct discourse units based on predefined guidelines concerning the function of connectives. The instructions and a small evaluation can be found in Appendix B.

3.4 Discourse Unit Alignment

In order to align the text segmented units in MSA with their equivalents in EGY, we evaluated two methods for alignment: Awesome-Align tool (Dou and Neubig, 2021) and GPT-4 API (OpenAI, 2023). Awesome-Align tool uses the start and end boundaries of the MSA segments for precise alignment, while GPT-4 employs prompted text segmentation, guided explicitly by the MSA segmented units. After annotating 50 Egyptian Arabic examples with discourse boundaries, we compared the performance of both approaches (see Table 7). GPT-4 demonstrated superior effectiveness, particularly in handling complex alignment challenges involving significant word order differences. Consequently, we selected GPT-4 for our alignment tasks. The instructions can be found in Appendix C.

Data	Precision	Recall	F_1-score
Awesome-Align	0.88	0.95	0.91
GPT-4	0.90	0.98	0.94

Table 2: Evaluating alignment tools on discourse unit alignment in MSA and EGY.

To extract the audio for each argument, we used the initial timestamps provided for each instance to accurately extract the corresponding audio segments. We employed the whisper_timestamped[9] (Radford et al., 2022; Giorgino, 2009) tool to transcribe segments, which generated detailed timestamps for each word. We then aligned the start and end of the EGY text segments with these corresponding transcribed segments that included precise word-level timestamps. This alignment process enabled us to accurately determine the start and end times for each argument within the original audio file, thus facilitating the precise extraction of audio for each defined argument.

3.5 Defining Categories of IDR

Based on the connectives from examples we extracted from the MSA subtitles, we choose the categories of discourse relations—cause-effect, contrast, elaboration, and temporal sequence. These categories align closely with the fundamental roles of structuring and organizing discourse in a meaningful and coherent manner, taking into account prosodic structure. They are well-established in various linguistic theories and frameworks, such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Discourse TreeBank (PDTB) (Prasad et al., 2008a), making them reliable and widely accepted in the field.

The selected connectives for these categories in our dataset include:

- **Cause-effect:** 'because / *Li-Anna*', 'in order to / *Kay*', 'therefore / *Ldhā* | *Idhn*', 'because of / *bi-Sabab*', 'from what, which / *Mimmā*'
- **Contrast:** 'but / *Lākin*', 'but, rather /*Bal* |

[9] https://github.com/linto-ai/whisper-timestamped

Relation type	Instances
'cause-effect'	216
'contrast'	212
'temporal-sequence'	135
'elaboration'	197
Total	760

Table 3: Statistics of the Spoken Implicit Discourse Relations Dataset

Innamā', 'although / *Bālrghm min*', 'while / *Baynamā*'

- **Elaboration:** 'or / *Aw* | Am', 'where, in that / *áaythu*', 'in addition / *Bāláāfh*', 'for example/ textitAlá sabīl al-mithāl', 'and this / *wa-Hādhā*'

- **Temporal sequence:** 'then / *Thumma*', 'after / *Bad*', 'when / *Indamā*', 'during / *Athnā*'

This categorization considers the primary roles of these connectives in spoken Arabic, making it easier to understand their use and impact on effective communication. We present the size of the processed dataset in Table 3.

3.6 Quality control

To assess the quality of our dataset as well as the feasibility of our proposed method for extraction, all discourse relations and segmentations for the entire dataset were manually reviewed by a native Egyptian Arabic speaker. The review process found that out of 760 instances, all instances had the correct IDR label, whereas only 19 required segmentation corrections. These errors typically involved cases where GPT-4 failed to follow guidelines regarding the order of arguments, especially with the pattern <DC+Arg2, Arg1>. Another issue was that the heads of relative clauses (e.g., الذي) were not always excluded from Arg1 when the arguments were within a relative clause. Additionally, a few instances were not correctly extracted from the MSA text, resulting in incomplete segments. Furthermore, 11 instances with no discourse function were removed. These included cases where adjectives or prepositional phrases did not act as separate discourse arguments. While GPT-4 correctly identified some of these cases as single arguments, it struggled with others, particularly when faced with similar connectives, leading to incorrect segmentation.

4 Experimental Setup

The main purpose of our experiments is to explore the usefulness of text versus speech features, including prosodic features. State-of-the-art methods use large pre-trained models that have been trained on data of varying size, from different languages and language variants, and from different domains. This makes it hard to compare results from such models, since they differ both in what they are pre-trained on, and in model architecture, as well as whether they are text or speech-based. We thus also train a number of simpler models only on our own dataset, and not any additional data. Using such methods allows us to make a more fair comparison between text and speech, and to combine text and speech features in the same model. We formulate implicit discourse relation classification as a multi-class classification task.

4.1 Models for IDR

Simpler Models For audio, we start with simpler models that use traditional signal processing techniques—prosodic features and Mel Frequency Cepstral Coefficients (MFCC)—paired with conventional classifiers (Davis and Mermelstein, 1980; Nilu Singh, 2012). For text, we employ TF-IDF (Term Frequency-Inverse Document Frequency) vectorization combined with a classifier to assess the capability of the simpler model in handling this challenging task. Bridging text and audio, we experiment with integrating prosodic features and TF-IDF, as well as MFCC and TF-IDF, with classifiers. This integration aims to explore the impact of audio-textual features on enhancing the model's ability to identify complex discourse patterns.

Advanced Models Recognizing the limitations of simpler models in capturing nuanced features, we then transition to more advanced models, leveraging their advanced representation learning capabilities. For audio, we fine-tune cutting-edge acoustic models wav2vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021) via S3PRL, and Whisper (Radford et al., 2022), which are all multilingual. For text, we fine-tune the XML-R (Conneau et al., 2020), a multilingual model, as well as Arabic-only models such as AraBERT (Antoun et al., 2020), CAMelBERT (Inoue et al., 2021). Additionally, we use GPT-4 (OpenAI, 2023) for few-shot learning. Bridging text and audio, we experiment with two approaches: concatenating the AraBERT output with the processed prosodic features, and concate-

Dataset	Relations	Talks
Train	418	43
Validation	114	17
Test	228	27

Table 4: Dataset Distribution with discourse relations and number of talks

nating the AraBERT output with Wav2vec speech features. In both cases, we feed the combined representations into a fully connected layer for classification. Due to the random initialization and stochastic training of these models, we run the models 3 times with different random seeds and report the average score of the 3 runs.

4.2 Data

We divided the dataset into three distinct sets — 30% for testing, 15% for validation, and 55% for training — with no overlap in talks between the sets, summarized in Table 4. This split was carefully designed to maintain a consistent class distribution across all sets.

4.3 Audio Classification of IDR

Prosodic & MFCC Features: We used the Librosa library to extract two types of audio features—prosodic features and MFCC coefficients—for classifying implicit discourse relations. We extracted the mean and standard deviation of pitch and energy as prosodic features and the first 13 MFCC coefficients from each audio segment. We then concatenated the features from both segments into a single feature vector. All features were normalized using z-score normalization to ensure a mean of 0 and a standard deviation of 1.

We applied the same machine learning models separately to both feature sets:

- **Logistic Regression Model**: We optimized the model using grid search with 5-fold cross-validation to tune hyperparameters such as C values ranging from 0.01 to 100 and solvers (`lbfgs`, `liblinear`), with L2 regularization and a maximum of 1,000 iterations. Additionally, we applied SMOTE to address class imbalance in the training data.

- **Neural Network**: The model was trained with a learning rate of 0.001, class-weighted cross-entropy loss, the Adam optimizer, 50 epochs, two hidden layers of size 256 neurons each, and dropout (0.5) for regularization.

- **XGBoost Model**: The model was configured with the `multi:softmax` objective, used class-weighted sample weights to handle class imbalance, and used `mlogloss` as the evaluation metric.

Pre-trained Speech Models: We leveraged two self-supervised learning (SSL) pre-trained models from the S3PRL toolkit[10], specifically cvhubert (Chen et al., 2023; Hsu et al., 2021) and wav2vec2_large_lv60_cv_swbd_fsh (Baevski et al., 2020), for fine-tuning on the classification of implicit discourse relations. Both models are multilingual and were chosen because they were trained on CommonVoice, which includes Arabic data. We fine-tuned the last ten layers of both the wav2vec2 and cvhubert models and used their extracted features as input to a custom neural network classifier. To address class imbalance, we computed class weights for use in the Cross-Entropy Loss function. The training involved using the Adam optimizer with a learning rate of 1e-5 for the pre-trained model parameters and 0.0001 for the custom layers. We trained the models for up to 50 epochs with a batch size of 8, incorporating early stopping after 5 epochs of no improvement in validation loss to select the best model. We also employed a Cosine Annealing Learning Rate Scheduler (T_max = 10) to adjust learning rates dynamically and applied a weight decay of 1e-5 for regularization.

Whisper Features: We used the Whisper large-v3 model to extract audio features for implicit discourse relation classification. The extracted features were aggregated using the mean and standard deviation of the encoder outputs to create a fixed-length representation. These features were then fed into a custom neural network classifier consisting of three hidden layers with 512 units each, using ReLU activation functions and dropout rates of 0.5 to prevent overfitting. The model was trained for 10 epochs with a learning rate of 0.0005 and a weight decay of 1e-5 using the Adam optimizer. A batch size of 32 was used during training. Early stopping was implemented based on validation loss to select the best model. Class imbalance was addressed using the CrossEntropyLoss function.

4.4 Text Classification of IDR

TF-IDF: We used TF-IDF to vectorize text data for implicit discourse relations classification. The dataset was preprocessed by combining the text

[10] https://github.com/s3prl/s3prl

of Arg 1 and Arg 2. These combined texts were then vectorized using a TF-IDF vectorizer with a maximum of 5000 features, considering both unigrams and bigrams. We trained three models—a Logistic Regression, a Neural Network, and an XGBoost classifier—separately on the TF-IDF features to classify implicit discourse relations, using the same hyperparameters as those applied in the audio feature experiments.

Pre-trained Textual Models : We used three pre-trained text-based models—AraBERT (bert-base-arabertv02) (Antoun et al., 2020), CAMeLBERT (bert-base-arabic-camelbert-da) (Inoue et al., 2021), and XLM-R (xlm-roberta-base) (Conneau et al., 2020)— loaded from the Hugging Face model hub to classify implicit discourse relations in Arabic text. Each model was fine-tuned using a consistent set of hyperparameters: the Adam optimizer with a learning rate of 1e-5, a batch size of 16, and training for 15 epochs. We employed the Cross-Entropy Loss function to guide the optimization process. For each model, we concatenated the texts from both segments and tokenized them using the respective tokenizer, with a maximum sequence length of 512 tokens. The output from the [CLS] token, obtained as bert_outputs.pooler_output, represents the entire input sequence. This output was then processed through a dropout layer followed by a fully connected layer added atop the pre-trained model for classification. To select the best model, we used early stopping based on validation loss, saving the model with the lowest validation loss for test set evaluation.

GPT4 (Few-Shot Learning): We used the GPT-4 (OpenAI, 2023) API with few-shot learning to classify implicit discourse relations by providing simple instructions and 20 examples from the training data, then evaluated the model on the test data.

4.5 Audio-Text Classification of IDR

MFCC + TF-IDF: We combined MFCCs audio features with TF-IDF text features to enhance the classification of implicit discourse relations. An XGBoost model was trained with these combined features, using the same hyperparameters as those applied in the individual audio (MFCC) and text (TF-IDF) experiments.

Prosodic + TF-IDF: We integrated prosodic features with TF-IDF text features to enhance the classification. Two models—a Logistic Regression model and a Neural Network—were trained separately on these combined features, using the same hyperparameters as those applied in the individual audio (Prosodic) and text (TF-IDF) experiments.

BERT + Audio Features: We enhanced AraBERT's [CLS] token output with audio features in two experiments. In the first experiment, we concatenated the pooled output from the [CLS] token, which represents the input sequence, with prosodic features like pitch mean, Last F0 max, and F0 interquartile range. In the second experiment, we combined AraBERT's [CLS] token text features with standardized Wav2Vec2 speech features. For both experiments, the combined feature vectors were passed through a fully connected layer for classification, using the same hyperparameters as the individual text model.

5 Results and Discussion

Table 5 shows the results of the simpler models. Overall, models using text performed better than audio-only models. For the audio-only models, Prosodic+NN (prosodic features with neural network) performed better overall compared to MFCC+XGB and Prosodic+XGB (XGBoost-based models), with the highest accuracy. However, Prosodic+LogReg (prosodic features with logistic regression) had a slightly higher F1 score than Prosodic+NN. Prosodic+LogReg achieved the best performance in Contrast and matched MFCC+XGB in Elaboration. Prosodic+NN excelled in Cause-Effect but struggled in Temporal Sequence, similar to the other models.

For text-only models, TF-IDF+LogReg achieved the highest overall performance, with the best precision, recall, F1 score, and accuracy. TF-IDF+NN closely followed, showing high precision, recall, and a solid F1 score. TF-IDF+XGB demonstrated moderate performance in these metrics. In terms of classes, TF-IDF+NN excelled in both Contrast and Cause-Effect. TF-IDF+LogReg outperformed the others in Contrast and showed strong results in Temporal Sequence.

When combining audio and text features, Prosodic+TF-IDF+LogReg achieved the highest overall performance, with the best precision, recall, F1 score, and accuracy. Prosodic+TF-IDF+NN followed, showing strong recall and competitive accuracy. MFCC+TF-IDF+XGB demonstrated moderate performance across these metrics. In terms of classes, Prosodic+TF-IDF+LogReg excelled in Contrast and Elaboration, while Prosodic+TF-IDF+NN performed best

Data	Models	P	R	F1	Acc.	Cause-E	Contrast	TempSeq	Elaboration
Audio	MFCC+XGB.	0.24	0.24	0.24	0.25	0.27	0.21	0.15	0.32
	Prosodic+XGB.	0.24	0.24	0.24	0.26	0.37	0.23	0.10	0.24
	Prosodic+LogReg	0.30	0.28	0.28	0.29	0.23	0.37	0.19	0.32
	Prosodic+NN.	0.28	0.27	0.26	0.31	0.39	0.35	0.05	0.22
Text	TF-IDF+XGB.	0.30	0.31	0.30	0.32	0.34	0.41	0.16	0.30
	TF-IDF+NN.	0.39	0.39	0.38	0.40	0.43	0.49	0.29	0.31
	TF-IDF+LogReg.	0.42	**0.43**	0.40	0.42	0.37	**0.53**	0.33	0.38
Both	MFCC + TF-IDF+ XGB.	0.32	0.31	0.31	0.33	0.36	0.35	0.15	0.38
	Prosodic + TF-IDF+ LogReg.	**0.43**	**0.43**	**0.42**	**0.43**	0.39	0.49	**0.34**	**0.45**
	Prosodic + TF-IDF+ NN.	0.37	0.38	0.36	0.38	**0.46**	0.43	0.27	0.29

Table 5: Results of simpler models on the test set for the classification of IDR in Text, Audio, and combined Text-Audio (in macro), and F1 score of relations.

Data	Models	P	R	F1	Acc.	Cause-E	Contrast	TempSeq	Elaboration
Audio	HuBERT	0.29	0.25	0.22	0.27	0.29	0.34	0.10	0.17
	Wav2vec2	0.30	0.30	0.23	0.24	0.22	0.18	0.32	0.18
	Whisper+NN	0.32	0.31	0.31	0.33	0.37	0.34	0.19	0.34
Text	AraBERT	0.42	0.43	0.41	0.43	0.41	0.48	0.35	0.42
	XLM-R	0.43	0.39	0.38	0.40	0.36	0.46	0.32	0.39
	CAMeLBERT	0.42	0.41	0.40	0.42	0.43	0.49	0.30	0.39
	GPT4	0.41	0.41	0.41	0.42	**0.47**	0.48	0.33	0.35
Both	AraBERT+Pitch mean	**0.46**	**0.44**	**0.43**	**0.45**	0.45	**0.50**	0.34	0.43
	AraBERT+LastF0max	0.43	0.41	0.40	0.43	0.43	0.47	0.26	**0.44**
	AraBERT+LastF0max+F0IQR	0.45	**0.44**	0.42	0.44	0.42	0.53	**0.38**	0.36
	AraBERT + Wav2vec2	0.43	0.41	0.41	0.43	0.45	0.47	0.29	**0.44**

Table 6: Results of advanced models on the test set for the classification of IDR in Text, Audio (in macro), and F1 score of relations.

in Cause-Effect. MFCC+TF-IDF+XGB showed strength in Elaboration but had lower performance in Temporal Sequence. Confusion matrices for logistic regression models are in Appendix D. The confusion matrices show that the text-only model performs relatively well for 'Contrast' and 'Elaboration.' The prosody-only model struggles across categories, while the combined model provides more balanced results, improving 'Cause-Effect' and 'Contrast'. However, 'Temporal-Sequence' remains challenging for all models.

Comparing models using the same classifier but different feature sets, it is evident that combining features generally leads to improved performance. For example, MFCC+TF-IDF+XGB showed enhanced accuracy compared to either MFCC+XGB or TF-IDF+XGB alone. Prosodic+TF-IDF+LogReg achieved the best overall performance, surpassing both Prosodic+LogReg and TF-IDF+LogReg, particularly excelling Contrast and Elaboration. However, this trend does not hold for all combinations. TF-IDF+NN outperformed Prosodic+TF-IDF+NN overall, but Prosodic+TF-IDF+NN achieved the best performance Cause-Effect class. This suggests that while combining features from both audio and text data often leads to better results, this is not always the case, as seen with the NN-based models. The choice of features and classifiers needs careful consideration to avoid potential drops.

Table 6 presents the results of the advanced models. Again, audio-only models have the lowest scores. For audio-only models, Whisper+NN achieved the best overall performance with the highest F1 score and accuracy. In terms of classes, HuBERT performed best in Contrast but struggled in Temporal Sequence and Elaboration. Wav2vec2 excelled in Temporal Sequence, while Whisper+NN showed strong performance in both Contrast and Elaboration. For text-only models, AraBERT had the highest overall accuracy and F1 score, with balanced precision and recall. In terms of classes, it excelled in Temporal Sequence and Elaboration, while CAMeLBERT performed best in Contrast. GPT-4 showed consistent performance across all metrics, with strong results in both Cause-Effect and Contrast but slightly lower performance in Elaboration.

When combining text and audio features, AraBERT+Pitch mean had the highest accuracy and F1 score. In terms of classes, it excelled in Contrast. AraBERT+LastF0max showed strong results in Elaboration. AraBERT+LastF0max+F0IQR was most effective in Temporal Sequence, while

AraBERT+Wav2vec2 showed balanced performance across classes, particularly in Elaboration. Confusion matrices for AraBERT and Wav2vec2 models are in Appendix D. The confusion matrices show that AraBERT + LastF0Max+IQR achieves the best performance for 'Contrast'. Prosodic-enhanced models outperform text-only AraBERT for 'Contrast' and 'Elaboration' and slightly improve 'Cause-Effect'. AraBERT + Wav2Vec2 excels in 'Cause-Effect', highlighting the value of detailed audio features for capturing causal relations. However, all models continue to struggle with 'Temporal-Sequence', which remains the most challenging category. These findings suggest that combining text and prosodic features enhances overall performance balance.

Across all models, the scores for TempSeq relations are consistently low, highlighting a common challenge in discourse relation classification, which likely stems from their imbalanced representation in the dataset. This issue is evident not only in our dataset but also in established datasets such as the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008b, 2019; Wang et al., 2023) where these relations are relatively imbalanced compared to other relation types. This likely limits the models' ability to effectively learn and accurately predict these relations. Additionally, audio-level features are not particularly useful on their own because they lack the semantic and syntactic information needed to fully capture discourse relations. However, they can be helpful when combined with text.

IDR is a challenging task, with a highest accuracy of 0.45. While the advanced models performed somewhat better than the simple models, the difference is relatively small, showing that this task is still challenging, even with large pre-trained models. For both simple and advanced models, text-based models performed better than audio-based models, but combining text and audio led to some further improvements.

6 Conclusion and Future Work

We introduce a novel method for automatically extracting and classifying implicit discourse relations in speech and text, by aligning speech and speech transcripts in one language or language variant, with subtitles in another language, to extract examples of semantically equivalent pairs of implicit and explicit discourse markers. We apply the method to create a dataset for Egyptian Arabic, the first of its kind, addressing the identification of implicit discourse relations in spoken and written Egyptian Arabic, facilitating further advancements in this area. We manually verified this corpus and found that it was 100% accurate for labeling discourse relations, and over 97% accurate in identifying the span of discourse relations. Given that TED recommends its transcribers use language that is universally accessible and understood across all dialects, we anticipate that other pairs of languages will exhibit the same phenomenon, which suggests that our method can potentially be effectively used also for other language pairs.

We propose a comprehensive approach to modeling implicit discourse relations using both audio and text data, conducting two sets of experiments. We find that the IDR task is challenging, even for advanced models. While advanced models achieve the overall best performance, simpler models are not far behind. Text-based models outperform audio-based ones; however, integrating text and audio features can lead to further performance gains. In future work, we aim to explore the integration of text and audio features in more detail. Additionally, we want to explore our corpus creation method for other language pairs, including not only language variants but also more distantly related language pairs.

Limitations

Our work is limited to the language pair of Egyptian Arabic and Modern Standard Arabic (MSA), two variants of the same language, which are relatively similar. We do not explore whether the proposed method for corpus creation also works for more distantly related or unrelated languages. We also do not cover all the types of implicit discourse relations, only those that are explicitated with an unambiguous connective, leaving an investigation of the properties of the data set for future work. For the experiments in this work, we focus on relatively simple methods for combining text and audio, since our main goal is to explore the usefulness of text versus speech for classification. There has been research on implicit discourse classification proposing stronger methods (Wang et al., 2023). Using such methods would likely lead to stronger results also on our corpus, than those reported in the paper.

References

Amal Al-Saif. 2012. *Human and automatic annotation of discourse relations for Arabic*. Ph.d. thesis, University of Leeds.

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2046–2053, Valletta.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Shoshana Blum-Kulka. 2000. Shifts of cohesion and coherence in translation. In Lawrence Venuti, editor, *The Translation Studies Reader*, pages 298–312. Routledge, London & New York. First published 1986 in J. House & S. Blum-Kulka (Eds.), *Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition* (pp. 17-35). Tübingen: Narr.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST discourse treebank LDC2002T07. Web Download. Philadelphia: Linguistic Data Consortium.

William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe. 2023. Joint prediction and denoising for large-scale multilingual self-supervised learning. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sofoklis Kakouros, Themos Stafylakis, Ladislav Mošner, and Lukáš Burget. 2023. Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Iskandar Keskes. 2015. *Discourse Analysis of Arabic Documents and Application to Automatic Summarization*. Ph.D. thesis, Université de Toulouse.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Janine Kleinhans, Mireia Farrús, Agustín Gravano, Juan Manuel Pérez, Catherine Lai, and Leo Wanner. 2017. Using prosody to classify discourse relations. In *Proc. Interspeech 2017*, pages 3201–3205.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019. Implicit discourse relation identification for open-domain dialogues. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 666–672, Florence, Italy. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Gabriel Murray, Maite Taboada, and Steve Renals. 2006. Prosodic correlates of rhetorical relations. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 1–7, New York City, New York. Association for Computational Linguistics.

Raj Shree Nilu Singh, R. A. Khan. 2012. MFCC and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54(1):9–13.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad et al. 2008b. Penn Discourse Treebank version 2.0 LDC2008T05. Web Download. Philadelphia: Linguistic Data Consortium.

Rashmi Prasad et al. 2019. Penn Discourse Treebank version 3.0 LDC2019T05. Web Download. Philadelphia: Linguistic Data Consortium.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Ahmed Ruby, Christian Hardmeier, and Sara Stymne. 2024. Investigating the role of prosody in disambiguating implicit discourse relations in Egyptian Arabic. In *Speech Prosody 2024*, pages 926–930.

Wei Shi, Frances Yung, and Vera Demberg. 2019. Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 12–21, Minneapolis, MN. Association for Computational Linguistics.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chenxu Wang, Ping Jian, and Mu Huang. 2023. Prompt-based logical semantics enhancement for implicit discourse relation recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 687–699, Singapore. Association for Computational Linguistics.

Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. 2005. Discourse Graphbank LDC2005T08. Web Download. Philadelphia: Linguistic Data Consortium.

Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

A Lists of Connectives

Table 7 lists Arabic connectives in MSA and their Egyptian Arabic counterparts. These connectives were expanded in our system to include potential clitics, such as proclitics and enclitics, allowing for more accurate alignment between the connectives in MSA and EGY.

B Argument Segmentation

B.1 GPT4 Instructions

Evaluate the role of the connective in each sentence:

- Segment the text if the connective introduces an independent clause or a contrasting idea (acting as a discourse marker).
- If the connective merely connects items or extends the same thought without introducing an independent idea, do not segment.

Instructions for Output Formatting

- Start each output by echoing the full input.
- Start each argument with a bracketed number like '[1]', '[2]' for order.
- Place each argument on a new line.

Guidelines on Connectives

- Consider the following list of potential discourse connectives and their paired forms for segmentation:

و بسبب, بسبب, وإنما, لأنها, و, بل, و مع ذلك,
وهي, ام, ثم, لأن, ولأنني, لأنه, الا بعد, بعدما,
ب, ف, ل, مما, وبالتالي, في حال, بعد ذلك,
منذ, ورغم ذلك, إذن, إذاً, وحتى, وقد, اذ, إذا,
ففي النهاية, عندها, عندما, وبعد, من, وأن,
و من خلال, وكانت, وعندها, إذ, وكان, وفي,
وأثناء, أحيانًا, فيما, فقد, حيث, و من ثم,
حين, فلقد, وفي مرحلة أخرى, ولما, حينذاك,

MSA	EGY	English
لكن	بس / إنما / لكن	but
و مع ذلك	بس / إنما / و مع ذلك	nevertheless / however
إنما / بل	بس / إنما / لكن / بل	rather
رغم ذلك / بالرغم / مع ان	بس / إنما / لكن / رغم ذلك	although
وليس	إنما	not but
في المقابل	في المقابل / قصاد كده	in contrast
بسبب	علشان / عشان	because of
لأن	علشان / عشان كده	because
وبالتالي	وبكده / ف /فعشان كده	therefore
ومن ثم	ف / عشان كده	thus
ل	علشان / ف / ل / عشان كده	to
كي / لكي	علشان / عشان كده	in order to
لذا / لذلك	علشان / ف / عشان كده	so
إذن	إذن / ف	then
وبما ان	بما إن / علشان / عشان كده	since
نتيجة لذلك	علشان / ف / عشان كده	as a result
إذا	لو / أما لو	if
في الواقع	في الواقع / في الحقيقة / الحقيقة	in fact
أي	أى	any
أو / أم	أو / ولا	or
هو	هو	he/it
هي	هي	she/it
وقد	وقد / أما / و	might/has/already
بحيث	بحيث / اللي / ان	so that
ثم	ثم / بعدين / بعد كده	then
منذ	منذ / من	since
قبل	قبل / قبل كده	before
بعد	بعد / بعد كده	after
بعدما	بعديها / بعد كده / بعد ما	after
عندما	عندما / لما	when
كان / كنا	كان / كنا	was/were
بينما	في حين / بينمت / في نفس الوقت	while
خلال	في / طول / وقت	during
حين	لما / وقت ما / ساعة	when
لو	لو	if
كذلك	كمان / برضه	also
فهذا	فده / فكده	so this
وبهذا	وكده / وبكده	and thus
إنه / إنها	هي / دي	it is
خصوصا	بالذات / خصوصا / خاصة	especially
في النهاية	بالنهاية / في الاخر / في النهاية	in the end
في البداية	بالنهاية / في الأول / في النهاية	at the beginning
على سبيل المثال / مثلا	مثلا / يعني / زي	for example
بالإضافة إلى	وكمان / و / بالإضافة	in addition to
أما	أما / بالنسبة ل	as for
هم	هم / دول	they

Table 7: List of Arabic connectives in MSA and EGY.

بينما, لكن, اذا, في البداية, في أي, حينما,
وليست, وليس, ف, كأن, ولكن, مع ان, حتى و,
لعله, وهم, وكنا, او, في, ليس, والآن, بالرغم,
فهو, بالإضافة إلى, وكلاهما, على سبيل المثال,
خلال, كما, وقد, فهم, الا, وكذلك, وهناك, وهو,
رغم ان, لذا, لكنها, بعد, عند, بحيث, إلا بعد,
ب, أو, إما, في الواقع, حينئذ

- Do not include the discourse connectives in segmented arguments.
- Arabic clitic connectives that have a discourse function like ف, ب, ل, and و should not be included within segmented arguments.
- Do not modify the original phrasing of arguments; extract them exactly as they appear in the input.

Structural Rules

- Use canonical forms for ordering components based on connective presence:
  - Simple Connectives:
    - Linear Order (Normal): <Arg1 + DC + Arg2> - The first argument followed by the connective and then the second argument.
    - Reverse Order: <DC+Arg2, Arg1> - The connective (e.g., بعد) comes first, followed by the second argument and then the first argument.
  - Paired connectives are connectives which consist of non-adjacent lexical items, i.e. they have two parts DCP1 and DCP2. For paired connectives, only one order is possible: <DCP1+Arg2+DCP2+Arg1>.
- Exclude relative clause heads (e.g., الذي, التي, الذين) from the first argument if contained within a relative clause. Both arguments are in a relative clause.
- Two independent clauses or sentences can be joined by a coordinating conjunction such as أو, و, لكن. These conjunctions indicate discourse relations.
- Prepositional clitic discourse connectives such as ل و ب are usually attached to المصدر nouns.

Data	Precision	Recall	F_1-score
86 examples	0.75	0.70	0.72
50 examples	0.88	0.88	0.88

Table 8: Evaluating the GPT-4 Model on Discourse Segmentation Tasks for Modern Standard Arabic Texts.

- The preposition عند is rarely used to signal a discourse relation, but it is a discourse connective when followed by <al-maSdar> noun.
- Include all obligatory complements in VP and NP arguments by expanding the boundary of the argument to cover tokens in their trees.
- The clause involving verb ellipsis is usually considered as the second argument.

Special Considerations

- Ensure that the use of connectives as discourse connectors is distinct from their use in grammatical structures (e.g., simple conjunctions of verbs or nouns should not be misconstrued as discourse-level segmentation cues). (e.g., إلى المكتبة ثم إلى المدرسة, يلعبون ويصرخون).

B.2 Evaluation

To evaluate the model's capability to perform discourse unit segmentation, we used 86 examples from Al-Saif (2012) for testing. Additionally, we annotated 50 MSA examples with discourse boundaries, following the Leeds Arabic Discourse Treebank (LADTB) guidelines to further validate the model. The results are presented in Table 8. The performance on the 50 examples are notably high, with precision, recall, and F_1-score all at 0.88. This is likely due to the presence of unambiguous connectives in these examples, making the segmentation task easier. In contrast, the 86 examples dataset includes a broader range of cases extracted from the annotation guidelines, which likely contain more ambiguous and challenging instances, resulting in slightly lower performance metrics.

C GPT4-instructions for alignment

prompt = f""" Segment the following EGY sentence based on the MSA segments. There should be exactly two segments corresponding to the two MSA segments. Each segment should correspond to the MSA segments in order. Output the segmented EGY sentence in the following format:

[1] EGY segment corresponding to the first MSA segment

[2] EGY segment corresponding to the second MSA segment

Only output the segments in the specified format. Do not include any explanations or additional information.

MSA Segments:
{ ' | '.join(msa_segments) }

EGY Sentence:
{ egy_sentence }

Output: """

D Confusion Matrices

We show confusion matrices for the logistic regression classification, with text, audio or combined features with the logistic regression in Figures 2–4 We show the confusion matrices for the AraBERT-based classification models, with only text and combined audio features with AraBERT, in Figures 5–9

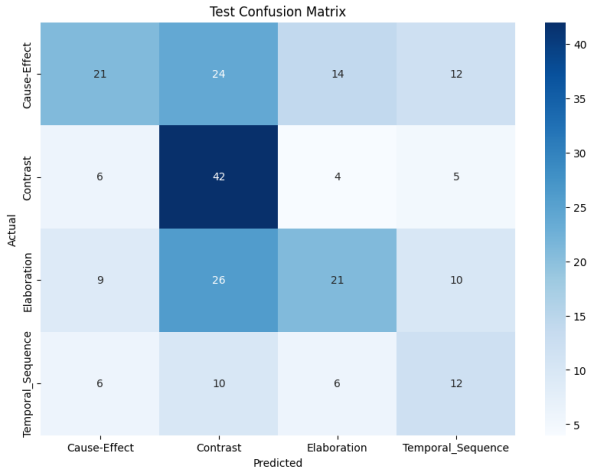

Figure 2: TF-IDF + LogReg

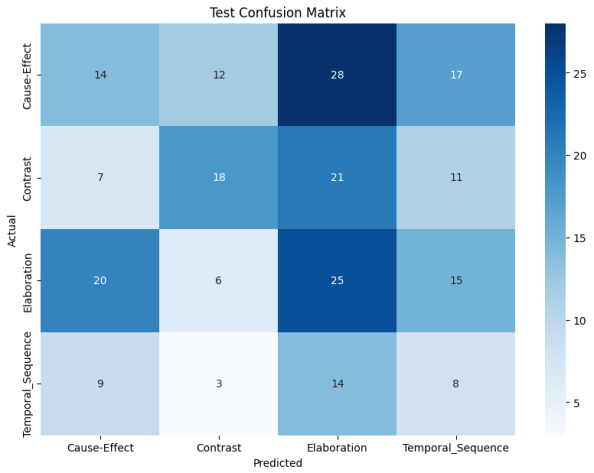

Figure 3: Prosodic + LogReg.

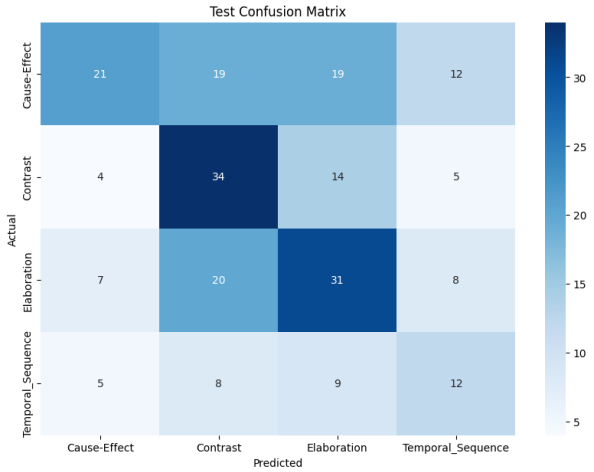

Figure 4: Prosodic + TF-IDF + LogReg.

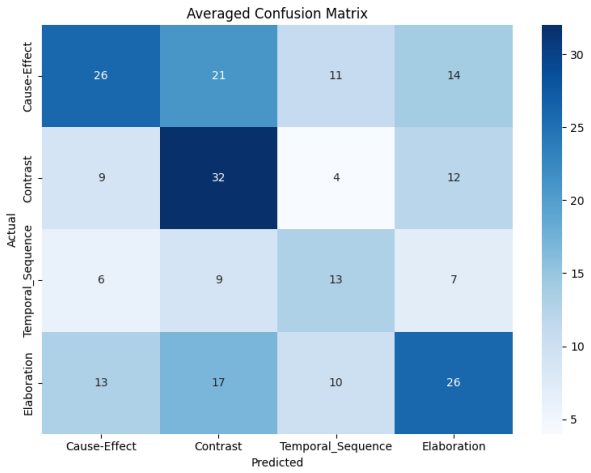

Figure 5: AraBERT

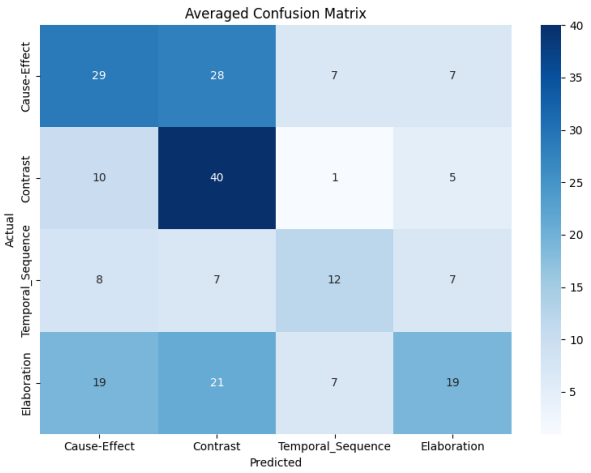

Figure 8: AraBERT+LastF0max+F0IQR

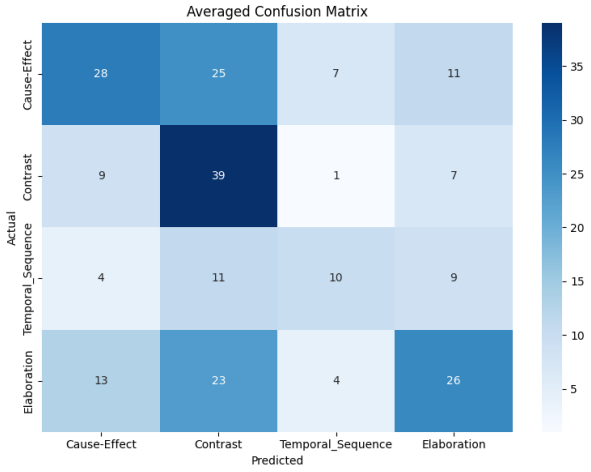

Figure 6: AraBERT+Pitch mean

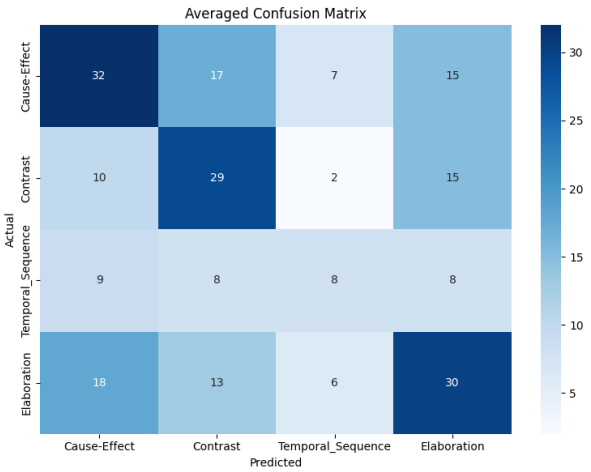

Figure 9: AraBERT+Wav2vec2

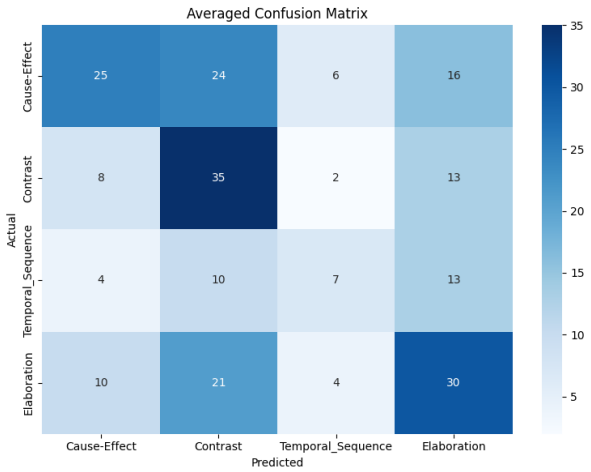

Figure 7: AraBERT+LastF0max