

M²Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis

Authors: Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu

Conference: IEEE Transactions on Visualization and Computer Graphics, 2021

Presented by
Rochita Das,
Dept. of Statistics

Motivation of the paper

- Previously for sentiment analysis, **Unimodal sentiment analysis models** are used which are based on a single communication channel (i.e., text or facial expression).
- Communication styles are highly complex and idiosyncratic (a sentence may seem semantically positive, but can be expressed with a sarcastic tone). In such cases, unimodal sentiment analysis is not reliable.
- Currently, deep-learning-based models such **Convolutional Neural Networks** (CNNs) and **Recurrent Neural Networks** (RNNs) are used in multimodal sentiment analysis. However, these models often work like black-boxes, users do not have enough understanding or control over the model.

Proposed Model and its Design

Proposed model

The paper propose M²Lens, a novel **explanatory and interactive visual analytics tool** to help both developers and to better understand and diagnose **Multimodal Models** for sentiment analysis.

Design of the model

To understand users' general needs the design requirements are summarized as follows:

➤ R1: **Show the model performance**

- Q1: What are the overall error distributions for model predictions?
- Q2: What are the instances that are predicted with large/small errors?

Proposed Model and its Design

➤ **R2: Reveal the contributions of modalities to the model predictions**

- Q3: How does each modality influence the model predictions?
- Q4: Which modalities dominate the model predictions? Also, which modalities complement or conflict with each other for model predictions?
- Q5: How do dominant/complementary/conflicting modalities influence the model predictions?

➤ **R3: Identify the influences of multimodal features for the model predictions**

- Q6: What are the feature sets that significantly contribute to positive/ negative sentiment predictions?
- Q7: What features are considered important by the model? Are they plausible for prediction?

Dataset

Data Set

- **CMU-MOSEI:** It consists of 23,454 monologue movie review video clips from 1,000 speakers and 250 topics in YouTube.
- From each video we extract the following information:
 - ✓ Transcripts for language modalities **(l)**
 - ✓ Facial expressions for the visual modalities **(v)**
 - ✓ Voice of speakers as the acoustic modalities **(a)**

Intra and Inter-modal Interactions

When modeling intra- and inter-modal interactions, three typical situations arise:

- One modality is **dominant** for sentiment analysis. For example, people may show agreement by nodding their heads, where the vision modality dominantly indicates their positive attitudes.
- More than one modalities **complement** each other when people are expressing their sentiment. For example, people's positive attitudes in words can be enhanced by a happy tone.
- More than one modalities **conflict** with each other. For example, people may tell sad stories with smiles on their face.

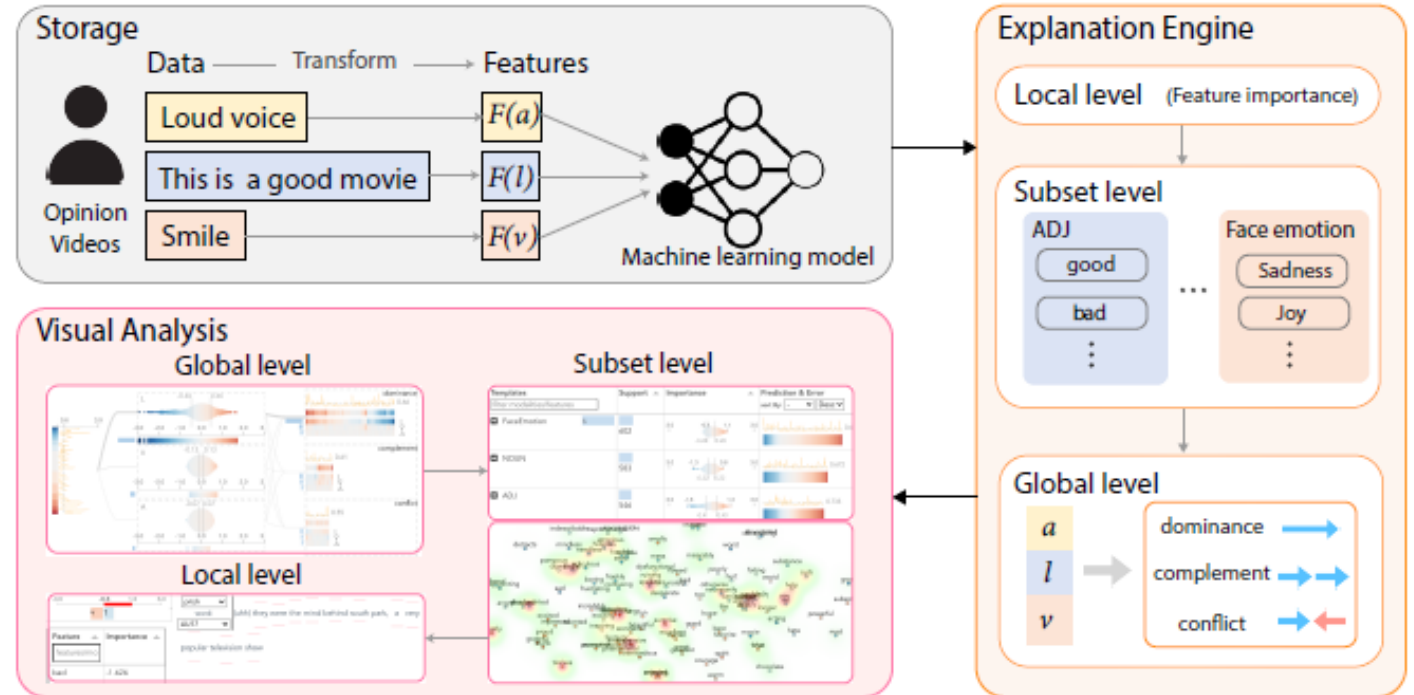
Feature Templates

For different modalities, we can consider different feature extraction techniques.

- **Language:** part of speech (POS)(e.g., noun, adjective, verb) -- by **Glove embeddings** technique each word is transformed to a 300-dimension vector
- **Audio:** pitch, amplitude, glottal/voice quality, and phase -- by **COVAREP** technique features are extracted to 74 dimensions vector
- **Vision:** face parts (i.e., brow, eye, nose, lip, and chin), head movement, and face emotions -- by **Facial Action Coding System (FACS)** it is encoded to 35 facial action units.

System Overview

- The storage module saves users' model and data with processed features.
- Then, the explanation engine inputs the features into the model and generates multi-level explanations of model behaviors.
- The visual analysis module enables interactive exploration of the explanations.



Global Level Explanations

- The influence of the interactions on the model output is based on the importance of each modality (I_l, I_a, I_v) which is the summation of the importance of all its features.
- Then, we extract and summarize the interactions (L) with strong influences for all the predictions.
- The thresholds for our rules are determined by maximizing the distances between the interaction types while minimizing the average influences of interactions that do not belong to dominance, complement, or conflict (i.e., others)

$$\arg \max_{\{Th_{sig}, Th_{dom}, Th_{confl}\}} \frac{1}{|L|^2} \sum_i^L \sum_j^L dist(L_i, L_j) - \bar{L}_{others} \quad (1)$$

where L_i ($i \in \{dominance, conflict, complement, others\}$) is the interaction types output by Algorithm 1 for all the instances, $dist$ is the Euclidean distance between the average influences of L_i and L_j .

Algorithm 1 Rules for extracting important relationships of modalities.

Input: $\{I_l, I_a, I_v\}; Th_{sig}, Th_{dom}, Th_{confl} (\in (0, 1))$;

Output: Label for the interaction types, l ;

```

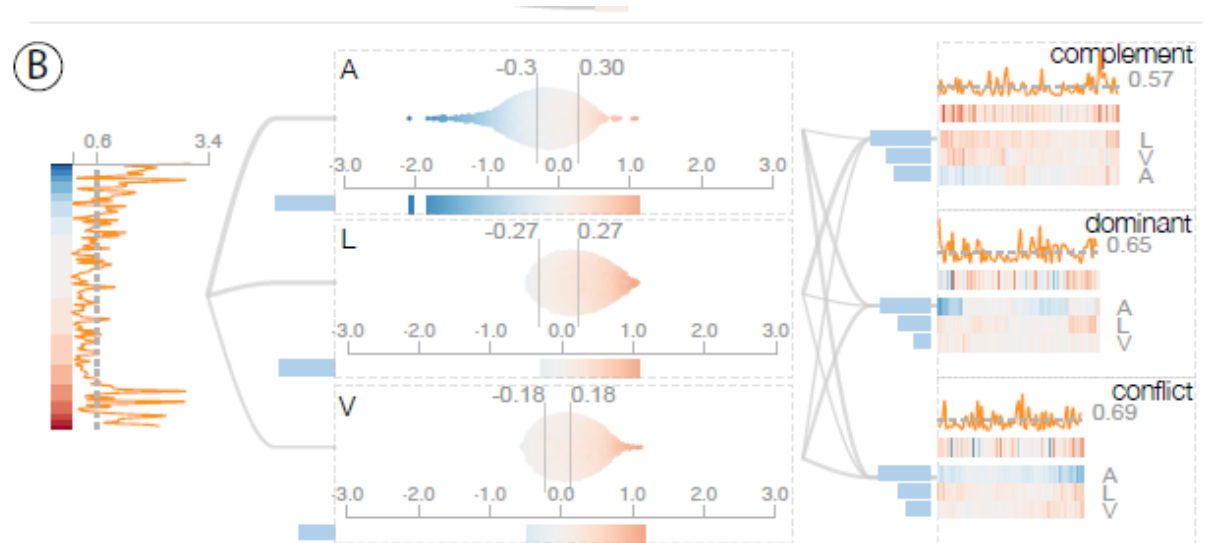
1: if  $\forall i \in \{l, a, v\}, |I_i| > Th_{sig}$  then
2:   /* important interactions */
3:   if  $\exists i, j \in \{l, a, v\}, I_i \cdot \sum I_j > 0, \frac{|I_i|}{\|I\|} \geq Th_{dom}$  then
4:      $l = dominance$ ;
5:   else if  $\exists i, j \in \{l, a, v\}, I_i \cdot I_j < 0, \sum \frac{I_i}{\|I\|} \leq Th_{confl}$  then
6:      $l = conflict$ ;
7:   else if  $\exists i, j \in \{l, a, v\}, I_i \cdot I_j > 0$  then
8:      $l = complement$ ;
9:   else
10:     $l = others$ ;
11: else
12:    $l = others$ ;
```

User Interface: Summary View

The **Summary View** presents an overview of the intra- and inter-modal interactions that are learned by the selected model in the User Panel.

In the parent node, a **barcode chart** and a **line chart** show the distributions of the ground truths and model prediction errors.

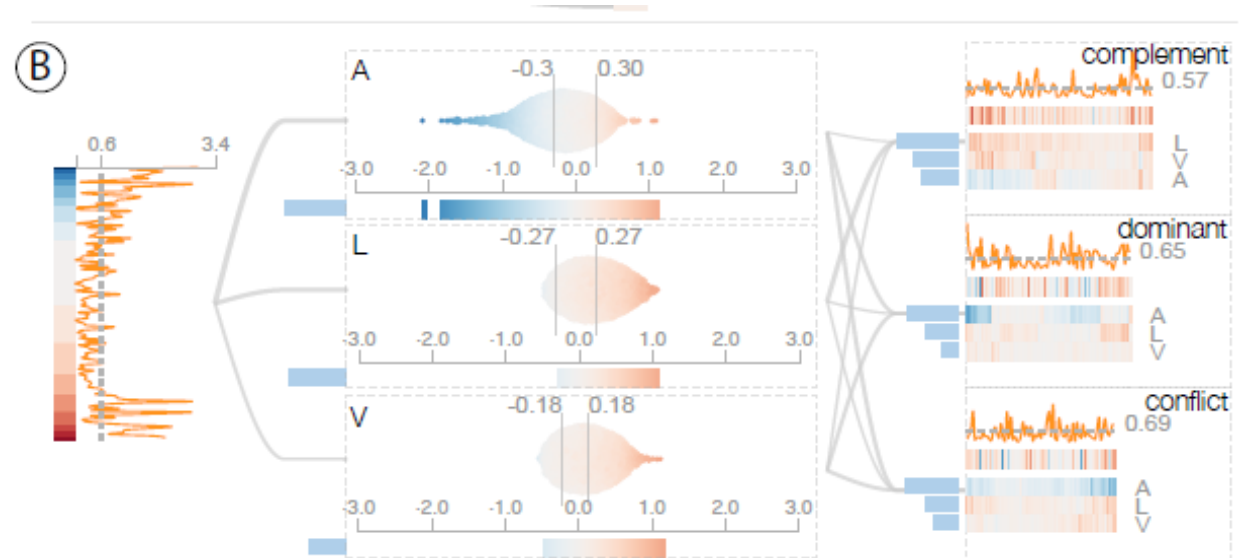
- the vertical height of the barcode represents the total number of instances
- the color displays the sentiment
- the horizontal position of the line chart suggests the absolute error
- mean error is represented as a dashed line



User Interface: Summary View

The second layer presents the importance of individual modalities in **bee swarm** plots .

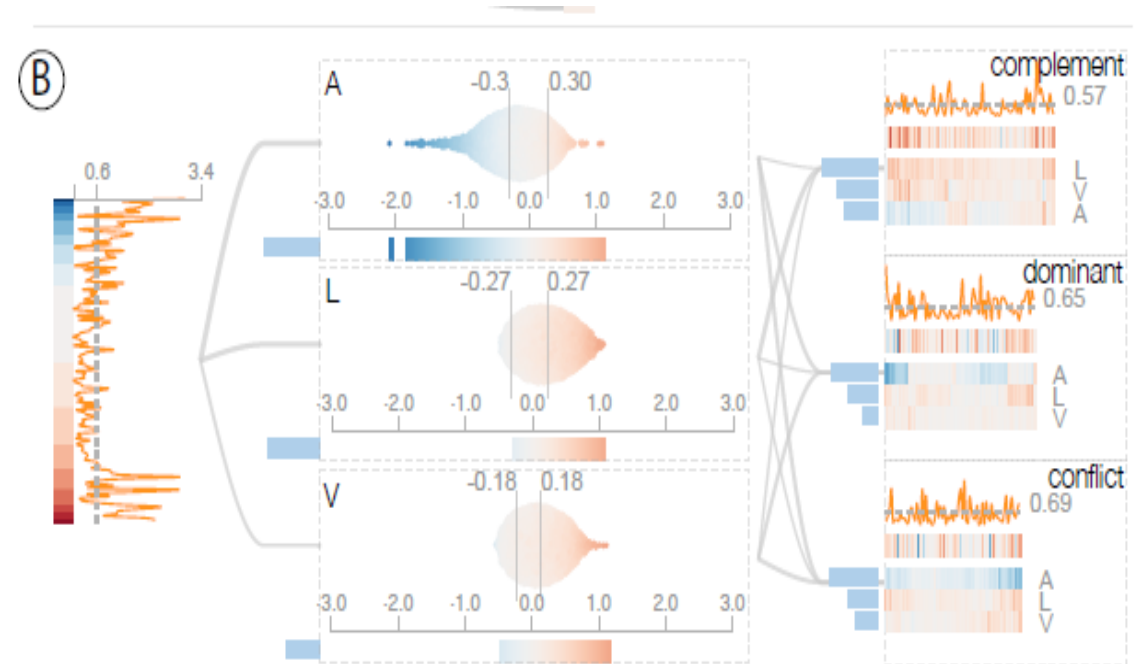
- For each node in the layer, a blue bar is put to the left, whose horizontal length summarizes the total influences of the modality
- the dots in the bee swarm plot and the barcode below demonstrate the distribution of the influences of that modality
- The color of the dots encode the importance values
- two gray lines indicate the magnitude of mean absolute importance



User Interface: Summary View

The last layer summarizes the information about the three types of **Interactions**.

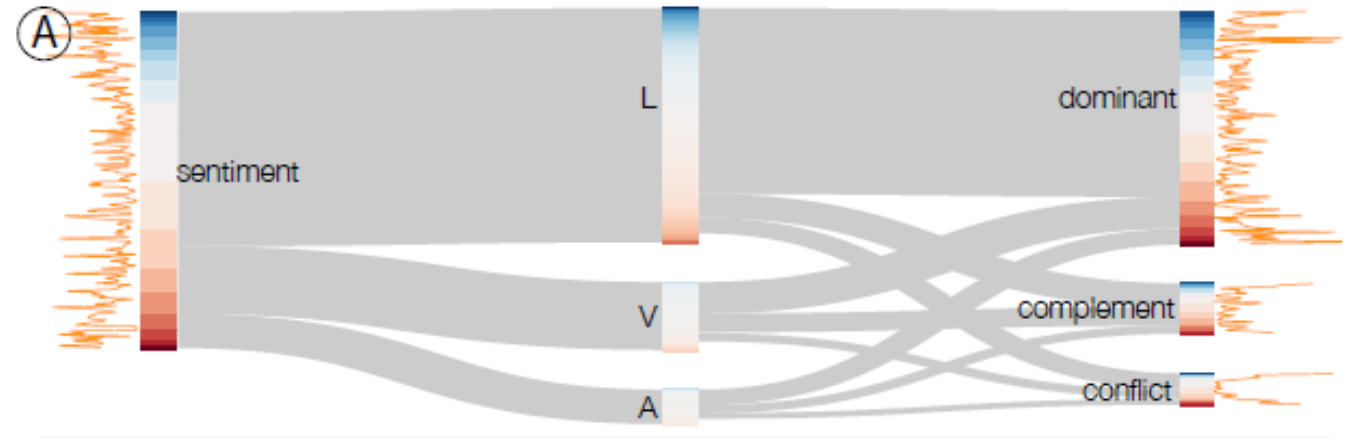
- a line chart and a barcode chart at the top summarize the error and prediction patterns.
- three barcode charts present the distribution of importance of all three modalities.
- The color represents the importance values.



User Interface: Summary View

The **Sankey diagram** reveals the intra- and inter-modal interactions and their importance to the predictions.

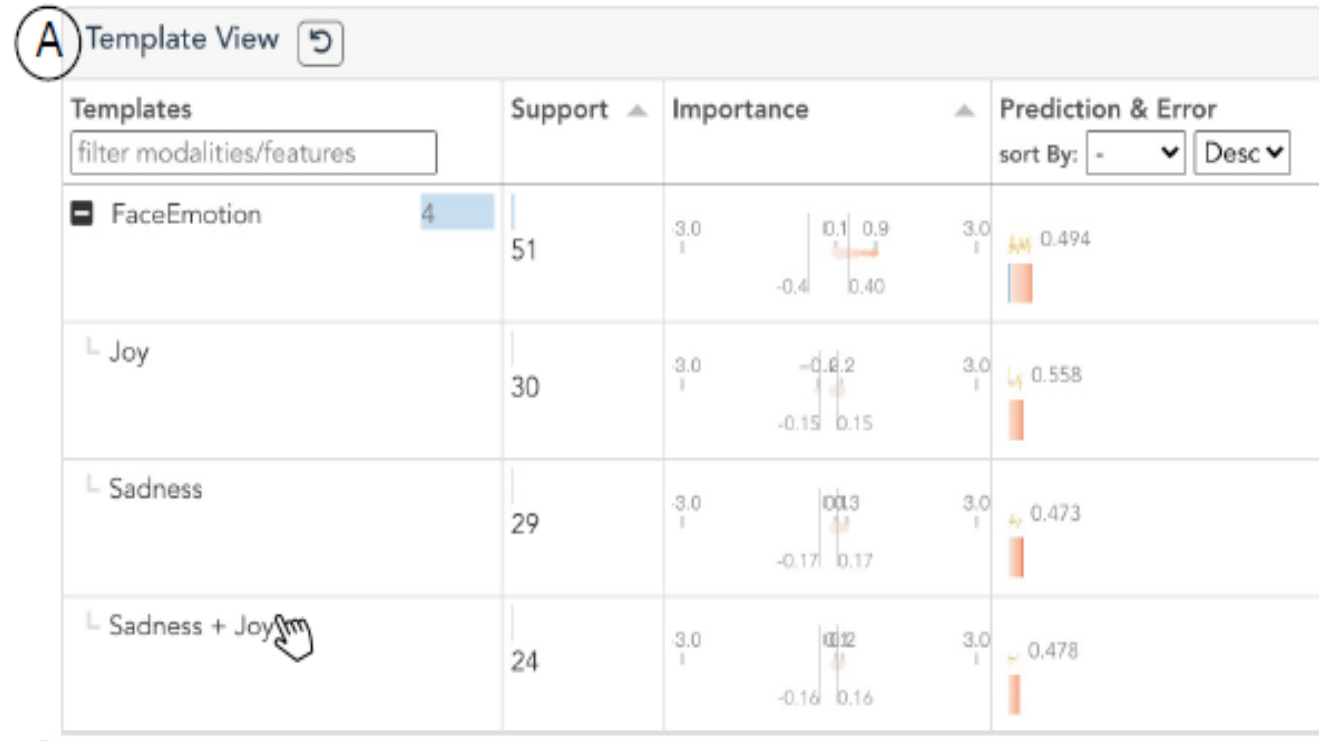
- It consists of three parts, the ground truth at the left, the influences of individual modality at the center, and the inter-modal interactions at the right.
- The width of a flow is proportional to the importance of the target node of the flow.
- The barcode chart of each node displays the importance distribution.
- The orange lines of the nodes show the error distribution.



User Interface: Template View

The Template View has four columns -template types, support, importance, predictions and errors .

- The first column records the names of feature sets. A green bar is placed to the right denoting the number of children for the feature set. Users can collapse the corresponding row by clicking the '+'.
- The second column displays the frequency for the templates. The distribution of the importance and prediction information is visualized in the third and fourth columns.
- Users can sort the templates according to their support, importance, and errors. In such a way, they can prioritize their efforts in diagnosing the complex model behavior.



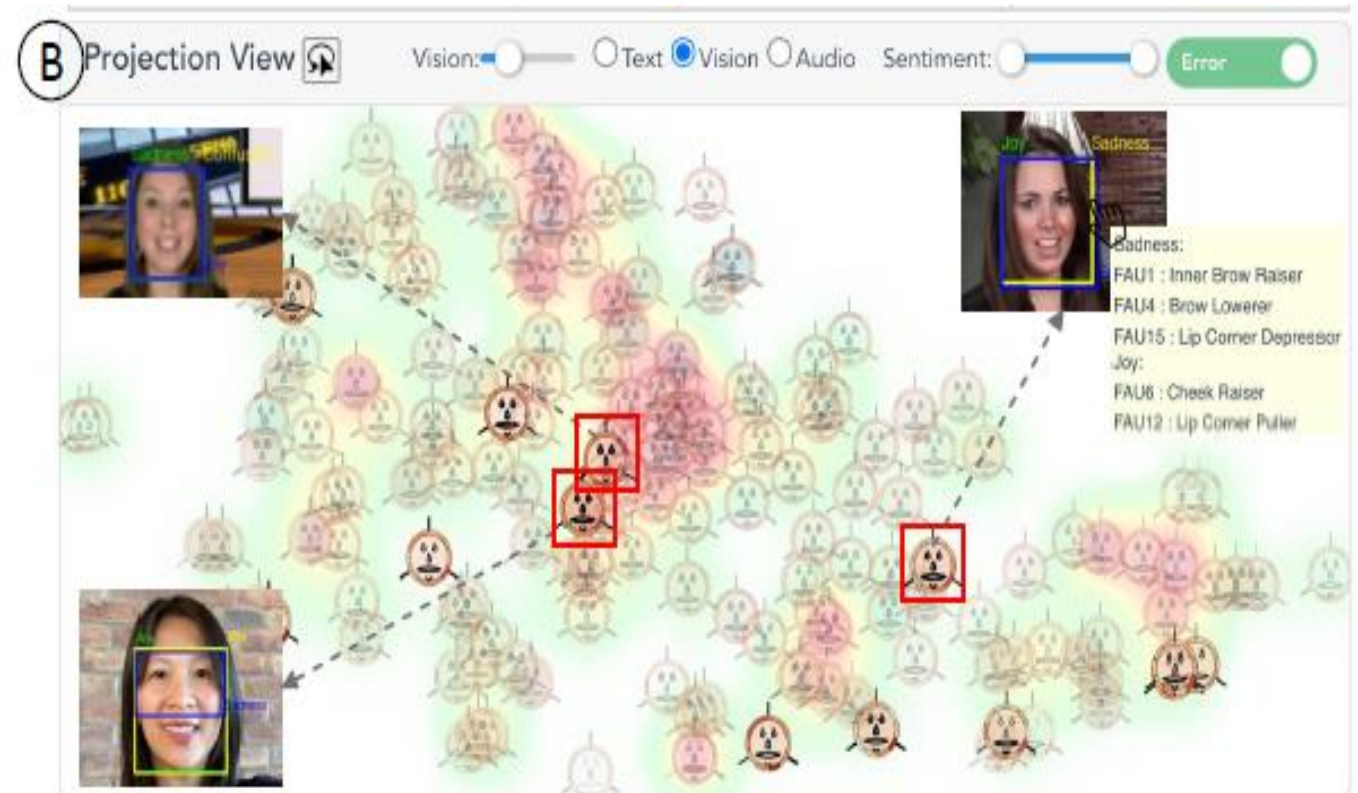
The screenshot shows the 'Template View' interface. It has a header bar with a logo, the title 'Template View', and a refresh icon. Below the header is a table with four main columns: 'Templates', 'Support', 'Importance', and 'Prediction & Error'. The 'Templates' column has a search box labeled 'filter modalities/features'. The 'Support' column shows the frequency of each template. The 'Importance' column shows a distribution plot for each template. The 'Prediction & Error' column shows a prediction error bar and a numerical value. The table lists four templates: 'FaceEmotion', 'Joy', 'Sadness', and 'Sadness + Joy'. A hand cursor is pointing at the 'Sadness + Joy' row.

Templates	Support	Importance	Prediction & Error
FaceEmotion	51		0.494
Joy	30		0.558
Sadness	29		0.473
Sadness + Joy	24		0.478

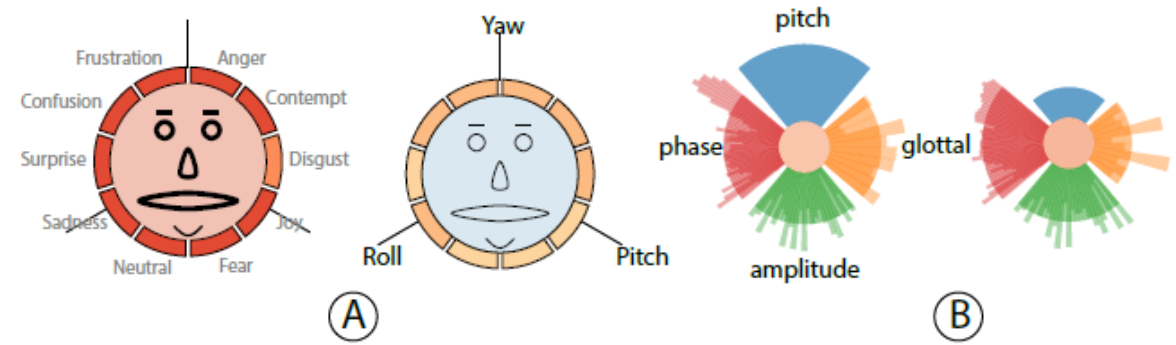
User Interface: Projection View

To summarize the feature sets, we project the high-dimensional features onto a 2D plane using **t-SNE**. Thus, instances with similar features will be placed close to each other.

- Given textual, acoustic, and visual features are heterogeneous, we design three different **glyphs** to encode the feature sets of the instances.
- Moreover, to help diagnose the model behavior (e.g., errors), a **heatmap** is added as the background to display the distribution of prediction errors.



User Interface: Projection View

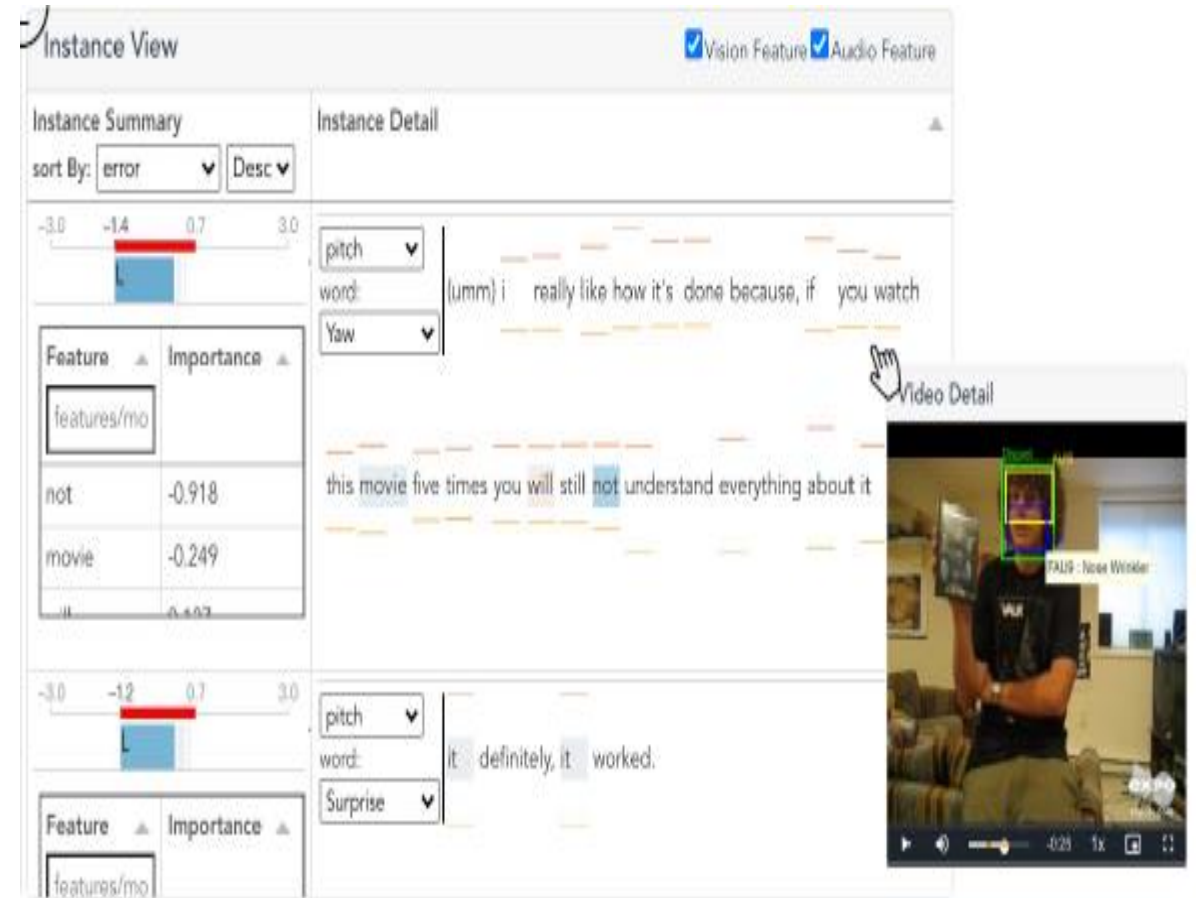


- **Language:** As words already carry semantic meanings, we use them to represent the textual features. In addition, we add a circle for each word, whose color encodes the sentiment prediction.
- **Vision:** The glyph designs for facial features are inspired by Chernoff face and in addition we add three sticks around the face to indicate the head movement in the yaw, pitch, and roll axis, respectively.
- **Audio:** To understand acoustic features, we group them into higher-level classes. Each colored sector represents the features of a class, where the radius relates to feature values. The sectors at the front summarized the average values of normalized features, while the small ones at the back display detailed feature values of the classes. Additionally, the inner circle color shows the sentiment prediction.

User Interface: Instance View

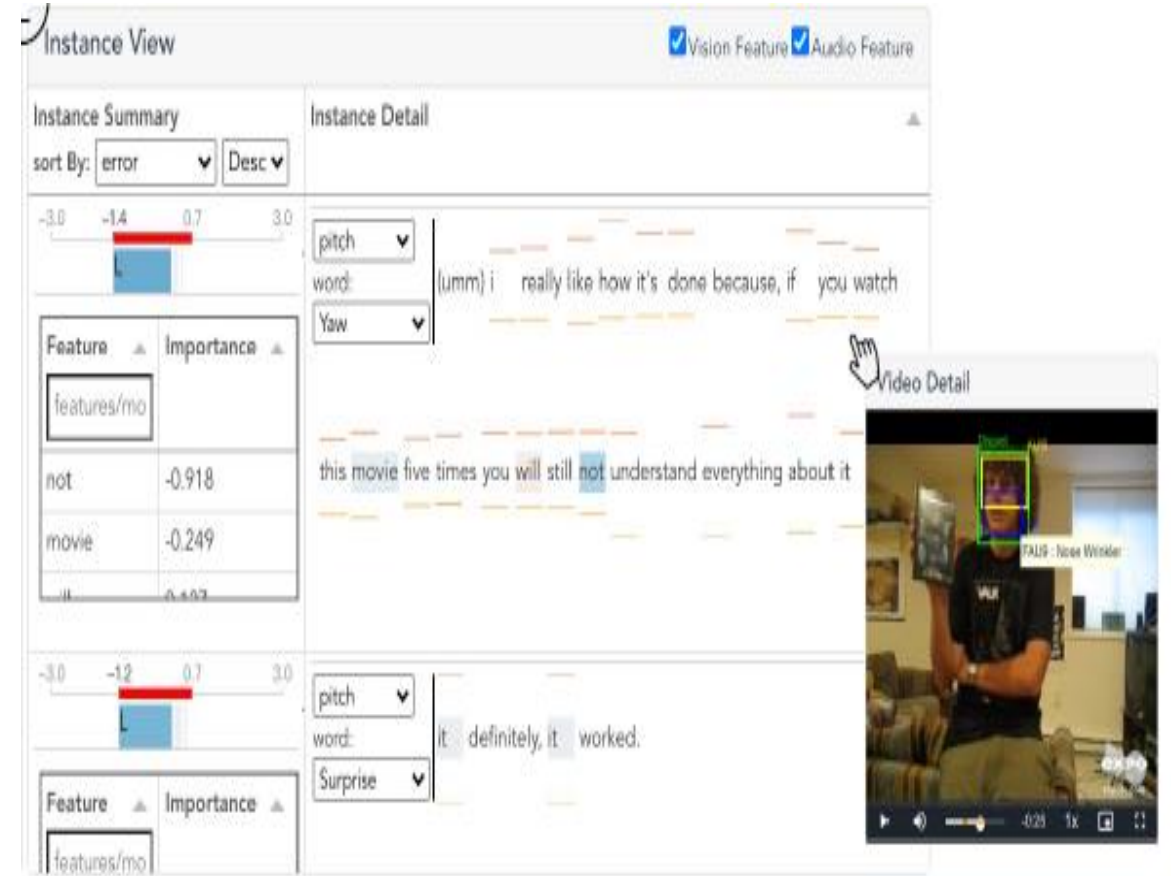
The **Instance View** provides **local explanations** by visualizing the important multimodal features and the context (i.e., transcripts and videos) of individual instances.

- In Left Column, in each row, the horizontal axes demonstrate the sentiment range, where the prediction and ground truth are marked. Between the two values, the thick red line suggests the error.



User Interface: Instance View

- The right column highlights the important features of the spoken words and draw the most important ones using orange lines. The lines above the words correspond to acoustic features, while the lines below represent the visual features.
- It also provides video context. When users click on the rows of the table, the corresponding video clips will pop up and play.



Future Work

- In the future, the system usability can be enhanced by adding functions, such as model comparison, data error correction.
- Also, system can be expanded to other multimodal applications (e.g., emotion recognition).

Thank You

