Multi-Modal Video Ad Classifier

Mohan Bhosale

Detailed Overview and Results

In this project, the primary objective was to classify 150 video advertisements based on 21 binary yes/no questions by utilizing both textual and visual data from the ads. The process began with data preprocessing, where the textual data from ad descriptions and speech, was cleaned and standardized. This involved handling some exceptions like yes and yes,both, the we did tokenization, and normalization. The processed data was then used to train a multi-modal classifier, integrating features from both text and numbers. The classifier used `BERT` for textual data and CNNs for visual data, and was designed to maximize performance metrics such as agreement percentage, F1 score, precision, and recall.

The project utilized various models and libraries, such as DistilBERT for text processing and ResNet3D for video analysis. The data preparation involved merging multiple data sources and normalizing numerical features using `StandardScaler`. The classifier architecture included a fully connected layer to combine textual and numerical features, trained using the AdamW optimizer and evaluated using metrics like F1 score, precision, and recall.

Upon evaluating the trained classifier, we observed that it performed exceptionally well on questions related to explicit features, such as specific product mentions and calls to action, achieving high agreement percentages and F1 scores. The model achieved an **overall F1 score of 0.8088** and **accuracy of 0.8184** on the test data. However, the classifier struggled with more subjective and nuanced questions, such as those related to emotional impact and creativity, resulting in lower scores for those categories. To address all videos, a function was developed which just take the video and does feature extraction, summarization and generates the video description just using video, then feeds this processed data into the trained classifier. This approach achieved an **F1 score of 0.74** on videos, indicating the model's robust performance in most areas but also highlighting the need for further improvements.

Methodology:

Data Loading and Preprocessing:

The process began with loading and preprocessing the data. The `load_and_preprocess_data` function handled data from multiple sources, including sample data, ground truth responses, video text with sentiment scores, video features, and video descriptions. The textual data was cleaned and standardized, which included handling missing values, tokenization, and normalization. This ensured the text data was in a suitable format for feature extraction. Visual data was processed using convolutional neural networks (CNNs) to extract relevant features from video frames. The aggregated ground truth data was merged with other datasets to form a comprehensive dataset containing both textual and numerical features.

Feature Extraction:

- **Textual Features:** Combined features from ad descriptions and transcriptions using NLP techniques such as tokenization provided by the `DistilBertTokenizer`.
- **Visual Features:** Extracted using the `extract_video_features` function, which utilized a pre-trained ResNet3D model for video frame analysis.
- **Audio and On-Screen Text:** Extracted speech using the Whisper model and on-screen text using EasyOCR.

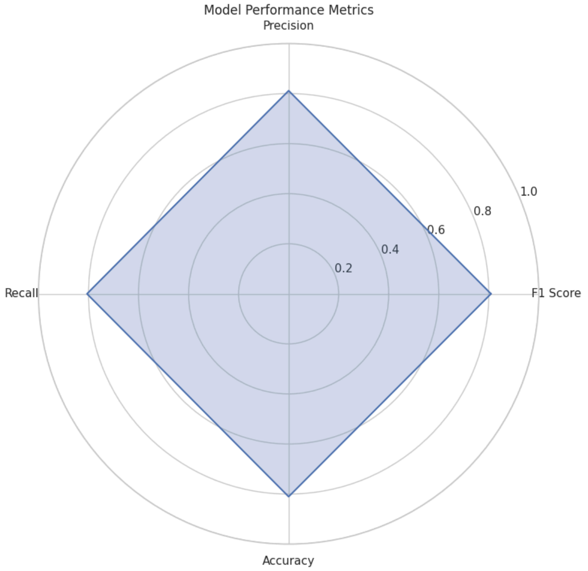
Model Architecture and Training: The classifier architecture, defined in the `VideoAdClassifier` class, combined textual features processed by a pre-trained DistilBert model and numerical features. The model included a fully connected layer to integrate these features and make predictions. The `train_model` function handled the training process, which included:

- **Optimizer:** AdamW
- **Loss Function:** Binary Cross-Entropy with Logits Loss
- **Evaluation Metrics:** F1 Score, Precision, and Recall

The training process involved splitting the data into training and validation sets, normalizing numerical features using `StandardScaler`, and preparing data loaders using the `prepare_dataloader` function. Training was conducted over multiple epochs, with the best model saved based on validation F1 score.

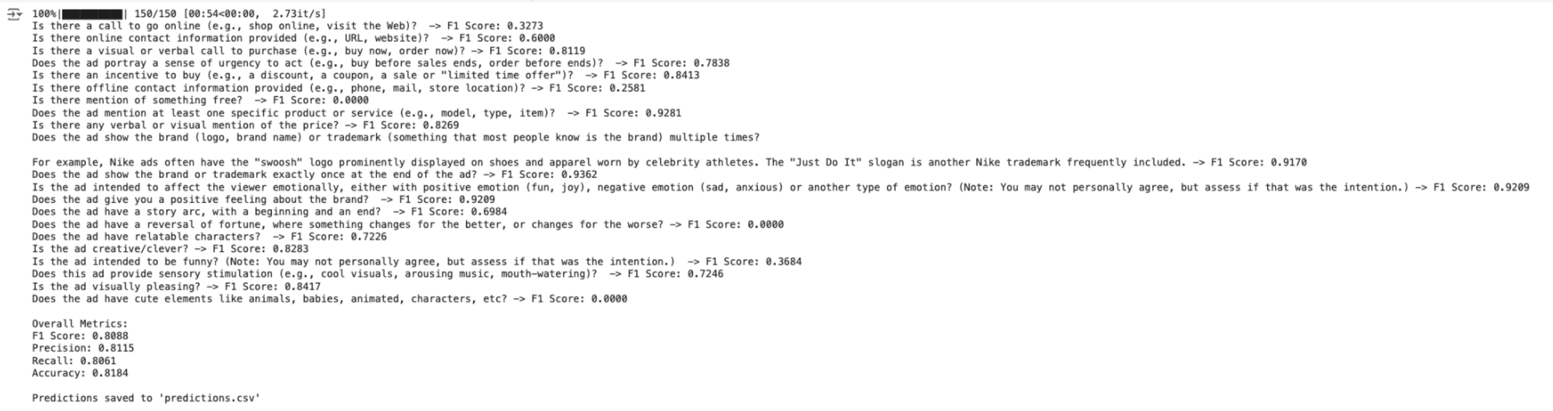
Evaluation and Results: Upon evaluating the trained classifier, the model achieved the following metrics on the test data:

- **F1 Score:** 0.8088
- **Precision:** 0.8115
- **Recall:** 0.8061
- **Accuracy:** 0.8184



These results indicated robust performance, especially in questions related to explicit features such as product mentions and calls to action. However, the model struggled with more subjective and nuanced questions, such as those related to emotional impact and creativity.

From Notebook:



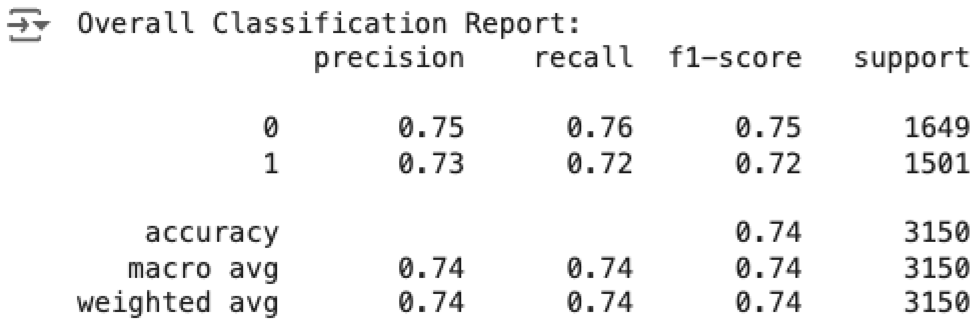
```
100%|██████████| 150/150 [00:54<00:00,  2.73it/s]
Is there a call to go online (e.g., shop online, visit the Web)?  -> F1 Score: 0.3273
Is there online contact information provided (e.g., URL, website)?  -> F1 Score: 0.6000
Is there a visual or verbal call to purchase (e.g., buy now, order now)? -> F1 Score: 0.8119
Does the ad portray a sense of urgency to act (e.g., buy before sales ends, order before ends)?  -> F1 Score: 0.7838
Is there an incentive to buy (e.g., a discount, a coupon, a sale or "limited time offer")?  -> F1 Score: 0.8413
Is there offline contact information provided (e.g., phone, mail, store location)? -> F1 Score: 0.2581
Is there mention of something free?  -> F1 Score: 0.0000
Does the ad mention at least one specific product or service (e.g., model, type, item)?  -> F1 Score: 0.9281
Is there any verbal or visual mention of the price? -> F1 Score: 0.8269
Does the ad show the brand (logo, brand name) or trademark (something that most people know is the brand) multiple times?

For example, Nike ads often have the "swoosh" logo prominently displayed on shoes and apparel worn by celebrity athletes. The "Just Do It" slogan is another Nike trademark frequently included. -> F1 Score: 0.9170
Does the ad show the brand or trademark exactly once at the end of the ad? -> F1 Score: 0.9362
Is the ad intended to affect the viewer emotionally, either with positive emotion (fun, joy), negative emotion (sad, anxious) or another type of emotion? (Note: You may not personally agree, but assess if that was the intention.) -> F1 Score: 0.9209
Does the ad give you a positive feeling about the brand?  -> F1 Score: 0.9209
Does the ad have a story arc, with a beginning and an end?  -> F1 Score: 0.6984
Does the ad have a reversal of fortune, where something changes for the better, or changes for the worse? -> F1 Score: 0.0000
Does the ad have relatable characters?  -> F1 Score: 0.7226
Is the ad creative/clever? -> F1 Score: 0.8283
Is the ad intended to be funny? (Note: You may not personally agree, but assess if that was the intention.)  -> F1 Score: 0.3684
Does this ad provide sensory stimulation (e.g., cool visuals, arousing music, mouth-watering)?  -> F1 Score: 0.7246
Is the ad visually pleasing? -> F1 Score: 0.8417
Does the ad have cute elements like animals, babies, animated, characters, etc? -> F1 Score: 0.0000

Overall Metrics:
F1 Score: 0.8088
Precision: 0.8115
Recall: 0.8061
Accuracy: 0.8184

Predictions saved to 'predictions.csv'
```

Post-Training Application: A function was developed to automate the process of extracting features from new videos. This function, `process_video`, integrates:

1. **Video Features Extraction:** Using ResNet3D.
2. **Speech Extraction:** Using Whisper model.
3. **Text Extraction:** Using EasyOCR.
4. **Summarization:** Using DistilBART model.

When applied to new data, the model achieved an overall **F1 score of 0.74**.

From notebook:



```
Overall Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.76      0.75      1649
           1       0.73      0.72      0.72      1501

    accuracy                           0.74      3150
   macro avg       0.74      0.74      0.74      3150
weighted avg       0.74      0.74      0.74      3150
```

Key Functions and their purposes:

a) `load_and_preprocess_data()`:

- Loads various datasets (sample data, ground truth, video text, video features, video descriptions)
- Preprocesses and merges the data
- Normalizes numerical features

b) VideoAdClassifier (neural network model):

- Combines BERT-based text features with numerical features
- Uses a fully connected layer for final classification

c) `prepare_dataloader()`:

- Tokenizes text data
- Combines text and numerical inputs
- Creates PyTorch DataLoader for efficient batching

d) `train_model()`:

- Trains the model using AdamW optimizer and BCE loss
- Implements early stopping based on validation F1 score

e) `extract_video_features()`:

- Uses a pre-trained 3D ResNet model to extract video features

f) `extract_text_from_video()`:

- Uses EasyOCR to extract text from video frames

g) `generate_video_description()`:

- Transcribes speech using Whisper
- Performs object detection on video frames
- Summarizes the transcription using a pre-trained summarization model

h) `process_video()`:

- Combines video feature extraction, text extraction, and description generation

i) `predict_and_evaluate()`:

- Makes predictions on new data
- Calculates evaluation metrics (F1 score, precision, recall, accuracy)

Multi-Modal Approach:

The multi-modal approach is used because video advertisements contain information in multiple formats: visual, audio, and textual. By combining these different modalities, the model can capture a more comprehensive understanding of the ad content. Here's why and how this approach is beneficial:

Why:

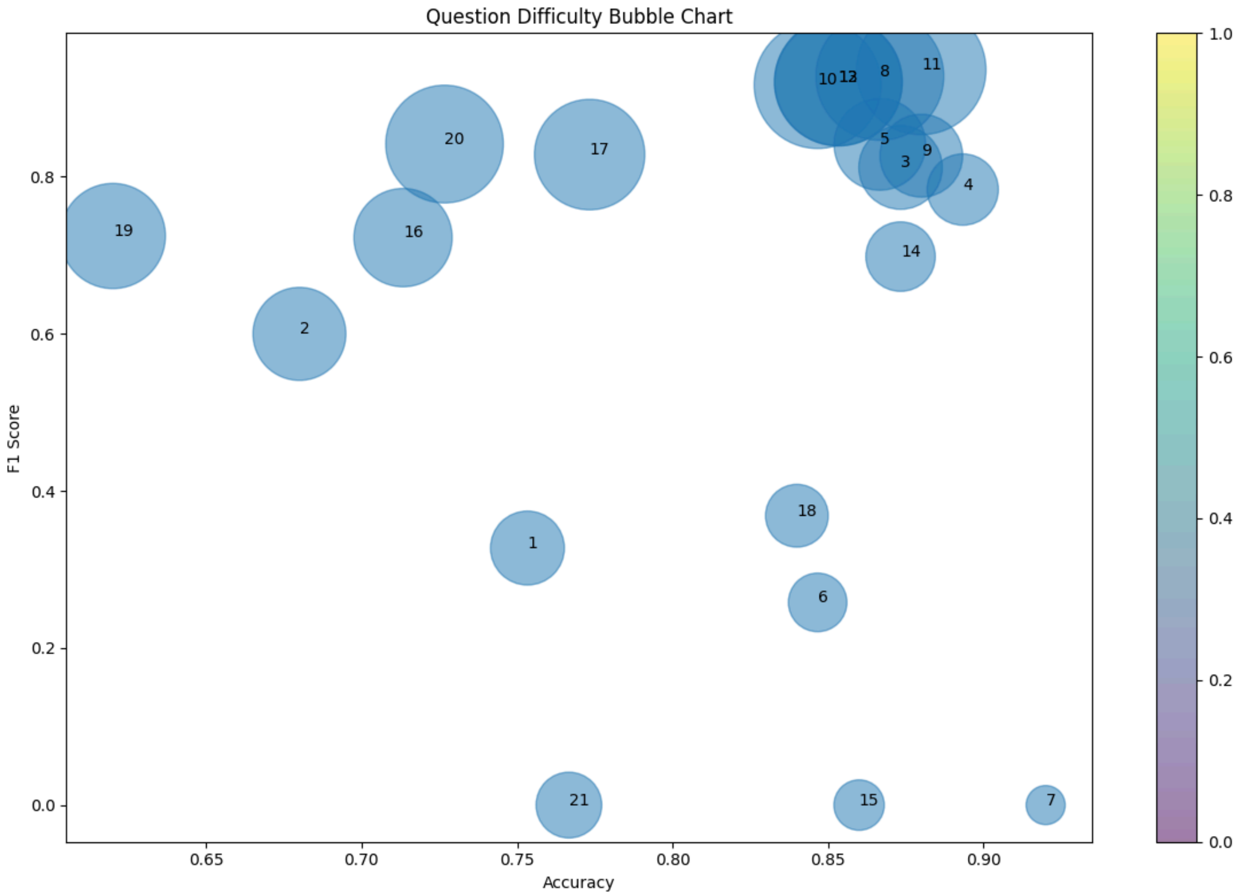
- Visual features capture the imagery, scenes, and visual style of the ad
- Audio features (through speech transcription) capture spoken content and potentially tone
- Textual features from descriptions and extracted on-screen text provide additional context
- Numerical features like duration and spend estimates offer metadata about the ad campaign

How:

- Video features are extracted using a 3D ResNet model, capturing temporal and spatial information
- Text is extracted from video frames using OCR and from audio using speech recognition
- These features are combined with pre-existing textual descriptions and numerical metadata
- The BERT model processes the textual data, while the numerical features are directly input to the classifier
- The neural network learns to weight these different sources of information for optimal classification

Overview on Video Classification Metrics:

Questions	Agreement Score	F1 Score	Precision	Recall
Is there a call to go online (e.g., shop online, visit the Web)?	0.7133	0.0444	0.5	0.0233
Is there online contact information provided (e.g., URL, website)?	0.5467	0.6222	0.5	0.8235
Is there a visual or verbal call to purchase (e.g., buy now, order now)?	0.7267	0.6095	0.64	0.5818
Does the ad portray a sense of urgency to act (e.g., buy before sales ends, order before ends)?	0.7267	0.4058	0.4828	0.35
Is there an incentive to buy (e.g., a discount, a coupon, a sale or 'limited time offer')?	0.74	0.7417	0.6588	0.8485
Is there offline contact information provided (e.g., phone, mail, store location)?	0.7867	0.2	0.3077	0.1481
Is there mention of something free?	0.9133	0.0	0.0	0.0
Does the ad mention at least one specific product or service (e.g., model, type, item)?	0.86	0.9247	0.86	1.0
Is there any verbal or visual mention of the price?	0.7267	0.6917	0.5823	0.8519
Does the ad show the brand (logo, brand name) or trademark (something that most people know is the brand) multiple times?	0.8267	0.9044	0.8483	0.9685
Does the ad show the brand or trademark exactly once at the end of the ad?	0.8467	0.917	0.8759	0.9621
Is the ad intended to affect the viewer emotionally, either with positive emotion (fun, joy), negative emotion (sad, anxious) or another type of emotion?	0.8533	0.9209	0.8533	1.0
Does the ad give you a positive feeling about the brand?	0.84	0.913	0.8571	0.9767
Does the ad have a story arc, with a beginning and an end?	0.7	0.0426	0.1111	0.0263
Does the ad have a reversal of fortune, where something changes for the better, or changes for the worse?	0.8667	0.0	0.0	0.0
Does the ad have relatable characters?	0.4933	0.3091	0.5	0.2237
Is the ad creative/clever?	0.5267	0.5419	0.7119	0.4375
Is the ad intended to be funny? (Note: You may not personally agree, but assess if that was the intention.)	0.7867	0.0588	0.3333	0.0323
Does this ad provide sensory stimulation (e.g., cool visuals, arousing music, mouth-watering)?	0.52	0.64	0.5664	0.7356
Is the ad visually pleasing?	0.7467	0.8516	0.7415	1.0
Does the ad have cute elements like animals, babies, animated, characters, etc?	0.7667	0.0	0.0	0.0



Bonus Questions:

1. Analyze why certain videos might not work well with the classifier or some questions yield inconsistent answers:

Advance Marketing Language Poses Challenges for Classifier:

Our model struggles to interpret complex marketing language, as evidenced by the low F1 score (0.3273) for detecting calls to go online. Phrases like "Not all zeros are created equal" likely confuse the classifier, which may be searching for more explicit online-related terminology.

Visual Context: A Significant Blind Spot:

The classifier's inability to process visual elements effectively is glaringly apparent. Notably, three questions yield F1 scores of 0.0000, indicating complete failure in classification:

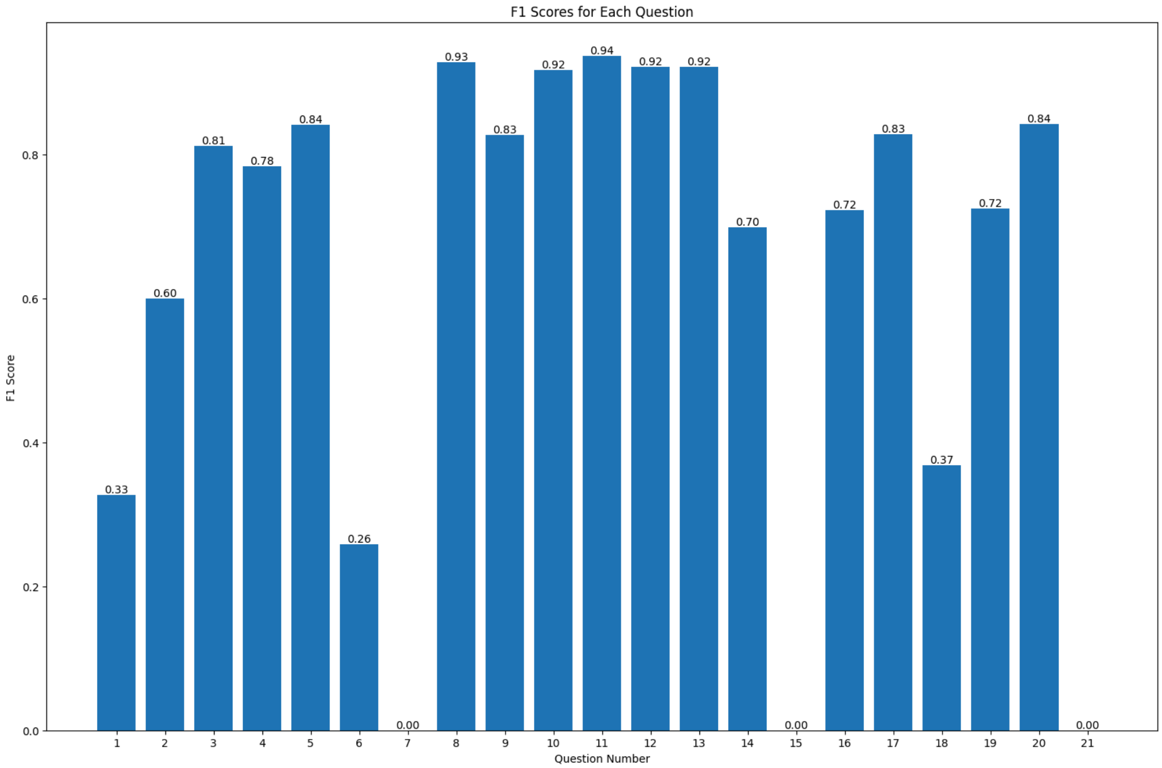
1. "Is there mention of something free?"
2. "Does the ad have a reversal of fortune, where something changes for the better, or changes for the worse?"
3. "Does the ad have cute elements like animals, babies, animated, characters, etc?"

These zero scores strongly suggest that the model is missing crucial visual cues and context, severely impacting its performance on visually-dependent questions.

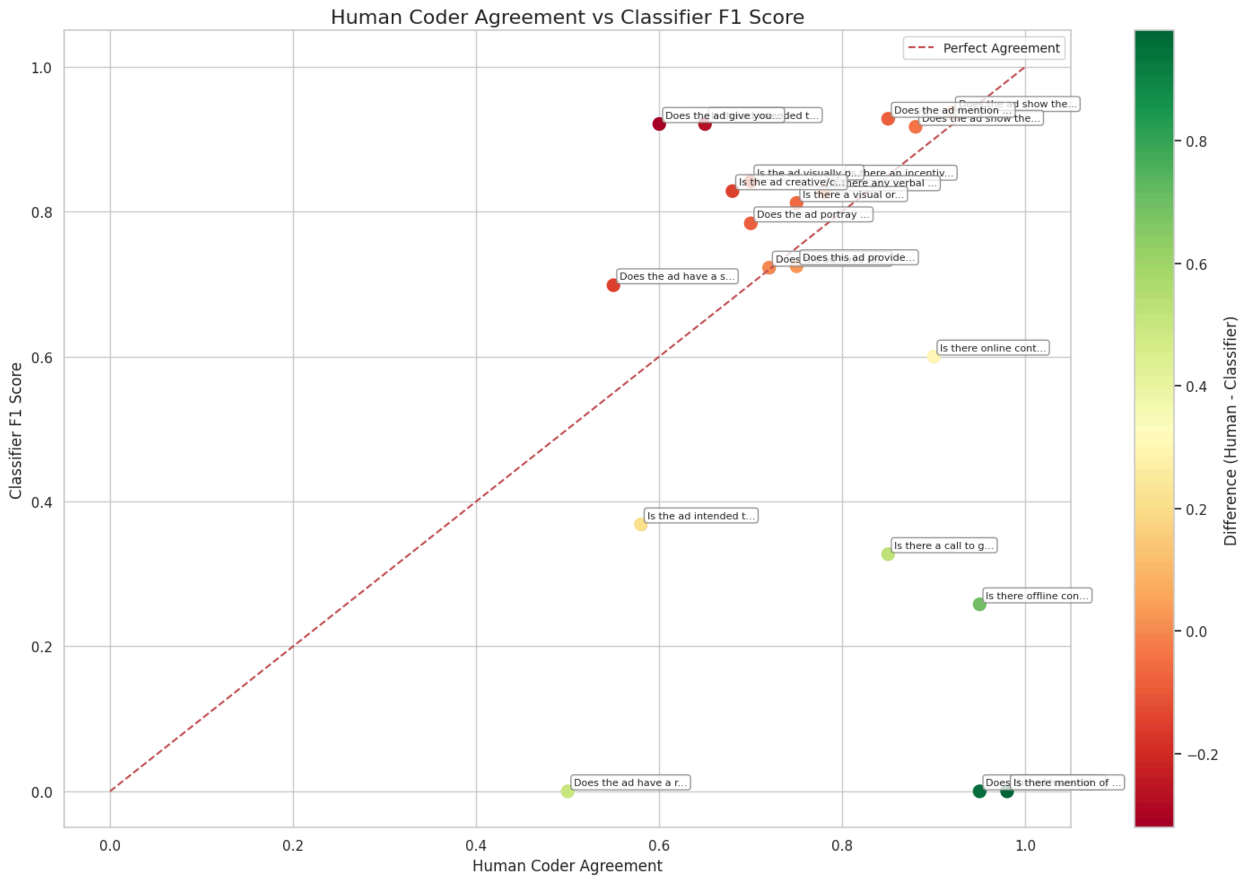
In conclusion, the Multi-Model classifier faces significant hurdles with videos featuring:

1. Metaphorical or complex marketing language
2. Heavy reliance on visual storytelling
3. Subtle or indirect calls to action
4. Content requiring subjective interpretation

The model's strengths lie in identifying straightforward, objective elements, such as specific product mentions (F1 score: 0.9281). However, its struggles with nuanced content and complete failure in certain visual aspects highlight areas for substantial improvement.



2. Provide an in-depth analysis of the human coders' responses and the performance of your classifier :



Overall Performance Comparison:

- Most questions cluster around the diagonal line, indicating general alignment between human coder agreement and classifier performance.
- There are notable outliers in both directions, suggesting areas where the classifier either excels or struggles compared to human coders.

Classifier Outperformance:

- The classifier significantly outperforms human agreement on subjective questions like "Is the ad visually pleasing?" and "Is the ad creative/clever?"
- Most strikingly, for emotional impact questions ("Does the ad give you a positive feeling about the brand?" and "Is the ad intended to affect the viewer emotionally?"), the classifier's F1 score is much higher than human agreement. This suggests the model may have identified consistent patterns that humans find more subjective.

Human Coder Superiority:

- Humans significantly outperform the classifier on questions related to specific content like "Is there mention of something free?" and "Does the ad have cute elements?"
- The largest discrepancies are in these objective, content-specific questions, where human agreement is nearly perfect (0.95-0.98), but the classifier completely fails (F1 score of 0.0000).

Challenging Questions:

- "Does the ad have a reversal of fortune?" shows low performance for both humans and the classifier, indicating it's a difficult concept to consistently identify.
- "Is the ad intended to be funny?" also shows relatively low agreement and performance, highlighting the subjective nature of humor.

Potential Biases or Limitations:

- The classifier seems to struggle with detecting specific elements (free offers, cute elements) that humans easily identify. This could indicate a limitation in the feature extraction or model architecture for these specific tasks.
- Conversely, the classifier's strong performance on subjective questions might suggest it's picking up on subtle cues that even human coders find difficult to consistently agree upon.

Areas for Improvement:

- The largest gaps in performance are in content-specific questions where the classifier performs poorly. This suggests a need for improved feature extraction or additional training data for these categories.

- Questions with low overall performance (e.g., "Does the ad have a reversal of fortune?") might benefit from clearer definitions or guidelines for both human coders and the machine learning model.

3. Discuss any observed patterns or anomalies in the data and their potential causes:

Distribution of F1 Scores:

- Pattern: The distribution is bimodal, with peaks at very low (0-0.2) and high (0.8-1.0) F1 scores.
- Anomaly: There's a significant gap in the middle range (0.4-0.6).
- Potential causes: a) The classifier might be very effective for certain types of questions but struggle significantly with others. b) This could indicate a "all-or-nothing" performance, where the model either understands a concept very well or fails to grasp it entirely.

F1 Scores by Question Type:

- Pattern: Subjective questions show a higher median F1 score and smaller range compared to objective questions.
- Anomaly: Contrary to expectation, objective questions have more variability and lower overall performance.
- Potential causes: a) The model might be better at capturing nuanced, subjective aspects of ads than specific, objective elements. b) There could be more consistency in how subjective aspects are represented across different ads.

Top and Bottom Performing Questions:

- Pattern: Questions about brand presence, emotional impact, and specific product mentions perform very well.
- Anomaly: Questions about free offers, cute elements, and calls to go online perform poorly.
- Potential causes: a) The model might be particularly adept at recognizing brand imagery and overall ad tone. b) Specific elements like "free offers" or "cute elements" might be underrepresented in the training data or require more complex feature extraction. c) The

poor performance on "calls to go online" could indicate difficulty in interpreting implicit calls to action.

Question Length vs F1 Score:

- Pattern: There's no clear correlation between question length and F1 score.
- Anomaly: Some of the shortest questions have both very high and very low F1 scores.
- Potential causes: a) Question complexity doesn't seem to be related to its length. b) The nature of the question (what it's asking about) is more important than how it's phrased.

Zero F1 Score Questions:

- Anomaly: Three questions have an F1 score of 0, indicating complete failure.
- These questions are:
    - "Is there mention of something free?"
    - "Does the ad have a reversal of fortune, where something changes for the better, or changes for the worse?"
    - "Does the ad have cute elements like animals, babies, animated, characters, etc?"
- Potential causes: a) These concepts might be too complex or varied for the current model to capture. b) There could be severe class imbalance in the training data for these questions. c) The feature extraction process might not be capturing the relevant information for these specific elements.

High Performance on Branding Questions:

- Pattern: Questions about brand presence consistently perform well.
- Potential causes: a) Brand elements (logos, names) might be easier for the model to detect consistently. b) There could be a strong emphasis on brand-related features in the model's training.