Papers that are useful:

* [[1905.12681] What Makes Training Multi-Modal Classification Networks Hard?](https://arxiv.org/abs/1905.12681)
* [Survey on SOAT Multimodal](https://pdf.sciencedirectassets.com/271506/1-s2.0-S0957417423X00256/1-s2.0-S0957417423021942/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEML%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQDR3GXCx8bKQEtBeIZXDiaTJCx0rbpNVM%2BlAvywAT%2BdRAIgfXHJJS5cdwmZ70gS%2FcDLZOEujssTZKfEUOEuGTHc998qswUIShAFGgwwNTkwMDM1NDY4NjUiDKV54Gal0PCFggU7cSqQBXmUhqnJ9cahbL4Qs7b502BiaUwAS9VZYmhNycpibHD2iUzBQEsdrfyR%2FGJriBsgKaJ54vd5Q5P1umYwO%2FCGLlLdlg3MaqmCUEbgn%2BlNZbsKZRoENs09We800Vk4DbvUDjcTUI2FlxYgcKJ4LX2CGaphCe%2Buic2yl%2Fwln6txS5D6uhKKFIEtcx7CZ7Aw8QChK08lLU7lqQE7QcgU8PM0RGDktaFZTyog1oOKjEf9N923U5CcwVDDkbuy80VRHsq292cDx%2B5L4P%2FaxgeLiNFrrrWvRJZma0%2Bb3nFPP1w7v7YApKM6FoESlPOjx2kdHgGaAMjV3ILcsF8JL5NK7Lq29YIpf9k4L3%2BiOtBRal6giMyvhskDhEHw8QC%2FMK4vW0RAUkr2i5ebaGuunW96hBg%2Fq%2BVCxsf9fQcDavqAbh76IqvwWqoWg8WahQb4KnX79BewahAzojs97v6SQrhX4js5VI8CNIYNLwQhLr%2Fq9qrKrGoUVokKTJ6XizwgAUieTRmkfXALtxzpwKrn22xDnYhb%2BEKe3nlka2KbCEBhwzyRPOGMCQ4KfEpEcZo%2BsvO2MaQ7o3c7SaxNF5hE9WhsAWEtG%2FYXJDNr%2FlPwdHB8J9WjSfnfkCHLNqwquEWhmoAjOhorCkZBLn0U3uUBk8VJUjVcw8vL%2Fq2WU2b%2BwdbQeDY7oGu9ZUVDD1FW%2FnVpBDykVPi8ylz7yvMlhqZRMzII2A74lSbIA4ptIHz6b3cP2dMkStaIlA5VtWHgSb%2Bxjy73Y4NJut7VWtgEaugwuOgqhKDFQHmUkVu4TfLOgI7vOIDx7c1Fq43uErv6fSMGSBexzzY2UaWUo2KWRJDg5%2BEoHtvKQJwt71qg4e8RmlIaLcRxrwXAMODrs7kGOrEB6%2B%2B9uunRFImiVEf1hCL7n%2BD%2FewlsKe9uTG5zdAjvd%2BgS0SDRvlk6Vs8WbqKlHSG3bGCJE%2BgQM2yMflVKnW2hNWXSmXsNotx%2B2gWD5S0mhfQVGWCOWIW8nXXxBE08%2FiOWgeYoHWsSXkGwDyEnxUjiGPS2NDfNkMvMQ0jA2JYN3%2F6bWb7AZDFEBC3x1XGfrZ%2BQ6RkSHGMbdKPkWCcyL68Fyrxtuax8pIiZQ2kQIPxM0iwo&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20241107T180218Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYWGM6ANG5%2F20241107%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=197232c9f220421b7c734fa2223720cd98695c3f5449b085fb559dc151db0383&hash=694ba555223b5cb30e79733748eea0c02c498f0e8a8ddd6430b6ad0f38a24bb8&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S0957417423021942&tid=spdf-0d338d43-045a-4e8f-a3e7-fb574daf169a&sid=b141fd3684c3374edc791bb87d4cd60d06bdgxrqa&type=client&tsoh=d3d3LnNjaWVuY2VkaXJlY3QuY29t&ua=13125906060850050c53&rr=8def3bc7f86c2b8e&cc=us)
* [Multimodal Models and Fusion - A Complete Guide | Medium](https://medium.com/@raj.pulapakura/multimodal-models-and-fusion-a-complete-guide-225ca91f6861#b39c)
* [Emotion recognition in Greek speech using Wav2Vec2.ipynb - Colab](https://colab.research.google.com/github/m3hrdadfi/soxan/blob/main/notebooks/Emotion_recognition_in_Greek_speech_using_Wav2Vec2.ipynb#scrollTo=6UqlIV3uGxDA)

Resources:

1. Feature Extraction
   1. Text: RoBERTa / DistilBERT or Text2Emoji embedders
   2. Video:
   3. Audio: Wav2Vec2
2. Fusion
   1. Early Fusion: Combine features at input level (not sure how lol)
   2. Intermediate Fusion: Extract features using encoders/embeddings then concatenate together before adding to the next FCN layer or classifier (more popular)
   3. Late Fusion: Train unimodal models on each modal then combine output together (computationally expensive, similar to ensembling)
3. Classifier
   1. Train a classifier to classify from extracted features
4. Other notes:
   1. Modality Dropout: To help in the cases where some modals aren't present, then we could use modality dropout → dropout some modalities when training
      1. This creates more robust features

Initial Plan:

* Let's try intermediate fusion because it's the most common and see first
* I have split for train-val-test, but I need to create script to get all audio, text, and video
* Since videos are far away, object detection might be needed at training then make it closer, but try first without anything