GSOC 2023 Project Proposal at RedHenLab

Archit Mangrulkar

April 4, 2023

Summary of the Proposal

In this project, I propose to develop a **stance detection method for television news using multimodal data**. I will focus on the topics of gun control, abortion, immigration, and covid/vaccination at the stance level, i.e., promoting, neutralizing, or debunking. I will use a combination of segmentation, annotation, feature extraction, and supervised and unsupervised models to perform stance detection. I will evaluate the performance of our models using standard evaluation metrics and analyze the impact of different features on the performance of the models. I will also address ethical considerations such as bias and fairness in our approach. The proposed method will provide valuable insights into the prevalence of biases in cable news networks.

Background

- Literature Review
  In the past, there has been significant work in the field of multimodal stance detection-
  - In a study by Hassan et al. (2017) [5], they used SVMs to classify the stance of video clips towards various topics such as religion, terrorism, and politics. They used a combination of audio, visual, and text features to train the SVM model. The audio features included prosodic features, such as pitch, energy, and duration, while the visual features included facial expressions and body gestures. The text features included unigrams and bigrams extracted from the speech transcript.
  - In a study by Habib et al. (2018) [4], they used a CNN-based model that combined audio and visual features to detect the stance of video segments towards various political issues. The audio features included prosodic features, such as pitch, energy, and intensity, while the visual features included facial expressions and head movements. The CNN model was trained to learn the joint representation of the audio and visual features and classify the stance of the video segments.
  - Another deep learning method that has been used for multimodal stance detection is the Recurrent Neural Network (RNN). In a study by Khattab et al. (2018) [6], they used an RNN-based model that combined audio and text features to detect the stance of Arabic news segments towards various topics such as religion, politics, and social issues. The audio features included prosodic features, such as pitch, energy, and duration, while the text features included unigrams and bigrams extracted from the speech transcript. The RNN model was trained to learn the joint representation of the audio and text features and classify the stance of the news segments.
  - Another approach to multimodal stance detection is to use ensemble models that combine the predictions of multiple classifiers. In a study by Sharghi et al. (2019) [14], they used an ensemble model that combined the predictions of SVM, NB, and Random Forest classifiers for multimodal stance detection in videos. They used a combination of audio, visual, and text features to train the classifiers. The audio features included prosodic features, such as pitch, energy, and duration, while the visual features included facial expressions and head movements. The text features included unigrams and bigrams extracted from the speech transcript.
- Inspired by this work, I will work towards coming up with models to optimally combine different features to generate condensed representations that can be used for stance detection effectively in weakly supervised & unsupervised environments

Goal and Objectives

The goal of this research is to develop a stance detection method for television news using multimodal data

Goal 1 will be extracting speech features from the videoes using openSMILE & librosa package in Python

Goal 2 will be to achieve a representation of the visual features using pre-trained CNNs & facial features using the OpenFace toolkit

Goal 3 will be to combine the audio, video & facial features using deep variational autoencoders

Goal 4 will be deploying supervised & unsupervised models that use these condensed features for stance detection

Goal 5 will be evaluation of our model according to use of different features and analysing for presence of bias

Methods

The overall approach to this project can be broken down into several steps:

1. Data Collection:
   The first step is to collect a large corpus of television news programs from the three major cable news networks. We can use online databases, such as the **Internet Archive and TV News Archive**, to collect the news programs.

2. Preprocessing and Segmentation:
   The news programs need to be preprocessed and segmented into coherent stories. This can be achieved using mixture topic models, which can identify consecutive sentences relating to the same topic. This segmentation will enable stance detection to be performed on a topic-by-topic basis. We can use topic modeling techniques, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), to segment news programs into coherent stories.

3. Annotation:
   A small set of videos will be hand-coded to identify whether they promote, neutralize, or debunk the four topics under study. As manual annotation is time-consuming, hence only a small sample of videos will be selected for annotation using a stratified sampling approach to ensure that we have a balanced representation of each topic and network. We will annotate the videos using the Argument Annotation Scheme proposed by Palau and Moens (2009) [13] to classify arguments into claim, evidence, and stance. A claim is a proposition that the speaker is advocating, opposing, or evaluating. Evidence is the information that the speaker presents to support or undermine the claim. Stance is the speaker's position with respect to the claim.

4. Feature Extraction:
   Audio and visual information can be used to extract features that capture different aspects of the news program, such as the tone of the speaker's voice, facial expressions, and body language. The features extracted can be used as input to the machine learning models that will be trained to classify the stance of the news program.

   - Text Features:
     Textual features will include bag-of-words, word embeddings such as GloVe and FastText, and syntactic features.

   - Acoustic Features:
     Audio features can be extracted using tools such as openSMILE (Eyben et al., 2013) [3], which is an open-source toolkit for extracting features from speech signals. We can extract features such as pitch, loudness, spectral features, and prosody features, which capture aspects such as the speaker's intonation and rhythm. Further, the librosa package in Python can be used to extract audio features like Mel-frequency cepstral coefficients (MFCC) which are commonly used in speech recognition tasks (McFee et al., 2015) [9], in speech recognition and audio analysis tasks (Logan, 2000) [8]. MFCCs are a set of features that represent the spectral envelope of a signal. The resulting MFCCs can then be used to represent the audio signal in a low-dimensional space.

- Visual Features:
  Visual features can be extracted using pre-trained convolutional neural networks (CNNs). We will use the pre-trained CNN as a feature extractor, and pass the frames of the video through the CNN to obtain a set of features for each frame. To capture the facial features, we can use the OpenFace toolkit, which provides a deep neural network-based facial recognition system that extracts facial landmarks and expressions. We can also use Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) can also be used to extract visual features (Dalal and Triggs, 2005 [2]; Ojala et al., 2002 [12]).

- Multimodal Features:
  Multimodal features can be extracted by concatenating the audio and visual features or by training a multimodal deep learning model that can learn to combine the audio and visual features in a way that is optimal for the task at hand (Ngiam et al., 2011) [11]. We can use a deep variational autoencoder (VAE) (Kingma & Welling, 2013 [7]) that will minimize the reconstruction error for images and audio jointly. The encoder will take input video frames (images) and audio frames (spectograms) and map it to a low-dimensional latent space which will be a compact representation of our features & can be used as inupt to the weak-supervised and unsupervised models.

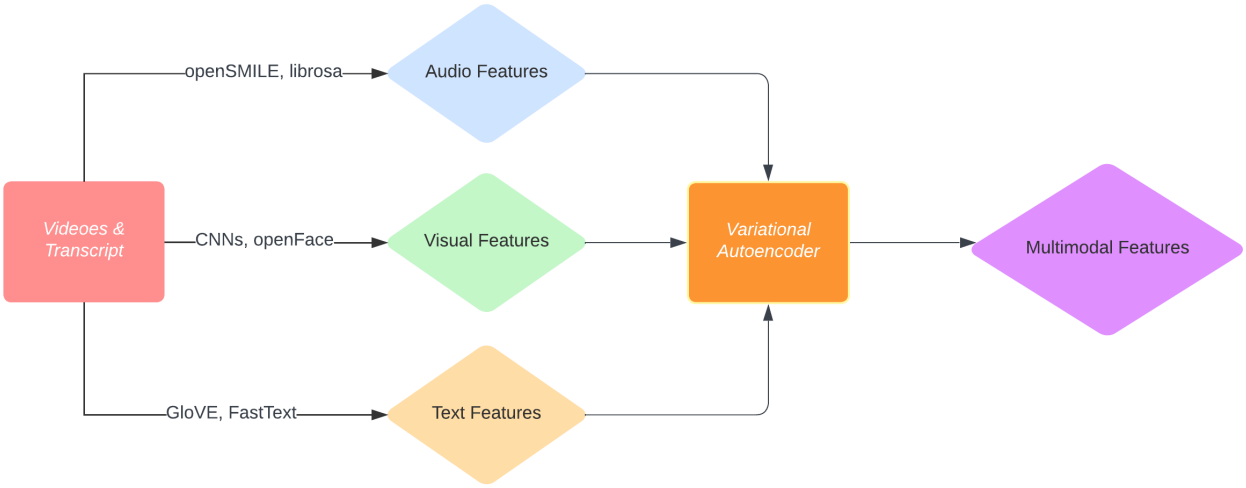


Figure 1: The Feature Extraction Pipeline

5. Stance Detection: In this module, we will make use of weak-supervised and unsupervised models to infer the stance of a story.

   - Supervised Learning
     We will use a variety of supervised learning models such as Support Vector Machines (SVMs), Random Forests, and Neural Networks to classify the stance of the news program based on the extracted features. The topic of the clip will be used as an additional feature for the stance detection model. Since manual annotation of large amounts of data is expensive and time-consuming, we will utilize weak-supervised and unsupervised methods to reduce the annotation effort.

   - Weakly Supervised Learning
     We will use a combination of manual annotation and weak supervision techniques to label a small sample of the videos. The weak supervision techniques include distant supervision and self-supervision. Distant supervision (Mintz et al., 2009) [10] involves using an external knowledge base (in this case, a set of manually annotated videos) to label a large corpus of unlabeled data. Self-supervision involves using the video itself to learn the stance. We can perform self-training to iteratively improve model performance (Zhou & Li, 2005 [15]), where the model is first trained on the labeled data, and then used to classify the unlabeled data.

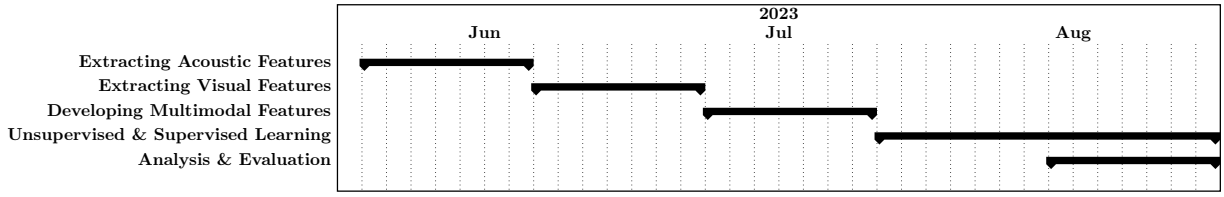
The highly confident predictions are then added to the labeled data for retraining the model. Another technique that can be used is co-training, where multiple classifiers are trained on different sets of features, and each classifier labels a subset of the unlabeled data. The labeled data from each classifier is then combined to train the final model (Blum & Mitchell, 1998 [1]).

- Unsupervised Learning
  To perform unsupervised stance detection, we will use a variety of clustering and topic modeling techniques. We will use the same textual, visual, and acoustic features as in supervised stance detection. We will cluster sentences based on their feature representation to identify groups of sentences with similar stances. We will also use topic modeling techniques to identify topics and their associated stances.

6. Evaluation: To evaluate the performance of our stance detection models, we will use standard evaluation metrics such as precision, recall, F1 score, and accuracy. We will also perform cross-validation and test the models on different datasets to evaluate their generalizability. We will compare the performance of our supervised and unsupervised models and analyze the impact of different features on the performance of the models.

7. Analysis: The analysis will also investigate whether there are any biases in the coverage and how they manifest in terms of promoting, neutralizing, or debunking arguments. Bias and fairness are important ethical considerations in this project. We will ensure that our annotated dataset is balanced and representative of different perspectives and viewpoints. We will also investigate the presence of bias in the data and the models and take steps to mitigate it.

Tentative Timeline



References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT '98)*, pages 92–100, 1998.

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[3] Florian Eyben, Björn Schuller, Martin Wöllmer, and Gerald Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, 2013.

[4] Muhammad Habib, Hussain Almerekhi, Syed Muhammad Ali Shah, Tareq Y Al-Naffouri, Abdullah Al-Dhelaan, and Yahya Al-Ohali. Multimodal classification of stance in political videos using lexical, acoustic and facial expressions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2349–2353. IEEE, 2018.

[5] Najadat Hassan and Ftoon Abu Shaqra. Multimodal sentiment analysis of arabic videos. *Journal of Image and Graphics*, 6:39–43, 01 2018.

[6] Omar Khattab, Osama Zahran, Hamdy Mubarak, Rami Al-Sabbagh, and Khaled Shaalan. Multimodal stance detection in arabic broadcast news. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6149–6153. IEEE, 2018.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] Beth Logan. Mel frequency cepstral coefficients for music modeling. *ICASSP'00. Proceedings. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing.*, 2:II–1015, 2000.

[9] Brian McFee, Colin Raffel, Dawen Liang, and Daniel PW Ellis. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer, 2015.

[10] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 1003–1011, 2009.

[11] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[12] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 2002.

[13] Raul M Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 592–600. Association for Computational Linguistics, 2009.

[14] Ehsan Sharghi, Nattapong Pou-Prom, Paradee Luangvilai, Taweechai Phanchaipetch, and Nattapong Toochinda. Multimodal stance detection in videos using ensemble of classifiers. *IEEE Access*, 7:124912–124922, 2019.

[15] Ding Zhou and Chong Li. Improving text categorization by using unlabeled documents. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*, pages 1557–1564, 2005.

Archit Mangrulkar

Github: https://github.com/architmang
LinkedIn: https://www.linkedin.com/in/archit-mangrulkar-033327199/
Skype ID: live:.cid.1b4f4b4dd4dc8b69
Email: archit.mangrulkar@gmail.com

EDUCATION

11.2020 - Present **Bachelor of Technology** at Indian Institute of Technology Kharagpur
Department of Computer Science & Engineering
Expected graduation in April of 2024

5.2018 - 4.2020 **Higher Secondary** at New Greenfield Public Academy, Indore

CONFERENCES

10.2022 **29th International Conference on Computational Linguistics** | **Paper**

- Co-authored and presented paper on Stance & Premise Detection shared task at the Social Media Mining 4 Health workshop (SMM4H 2022)
- Achieved 0.636 F1 for stance classification using the CovidTwitterBERT model, 0.664 F1 for premise classification using BART-base model
- Experimented with Parts of Speech, dependency parsing, Tf-Idf & performed contrastive pretraining using a supervised contrastive loss

WORK EXPERIENCE

5.2022 - 7.2022 **Research Internship** at the Center for Advancing Electronics Design, **TU Dresden**

- Worked in the field of **Approximate Computing** in Neural Networks
- Designed C++ based approximate multipliers using tensorflowlite in different Network layers for **8-bit inference** of Multi-Layer Perceptrons
- Used 8-bit **lookup tables**, **EvoApproxLib's** 8 bit multipliers, polynomial **regression modelling** to approximate the multiplication operator

SKILLS AND QUALIFICATIONS

Languages Python, C++, C, Java
Libraries PyTorch, Tensorflow, OpenSMILE, OpenCV, PIL, NLTK, spaCy, HuggingFace

PROJECTS

12.2022 - 2.2023 **DevRev's Domain-Specific QA Challenge** PS | Report| Code

- Developed **Closed Domain QA system** on SQuAD-like Datasets
- Pioneered sentence-level improvisatoin over the DrQA retriever to improve retriever latency by 3 times
- Used **FedAvg algorithm** over semantically clustered themes, meta-learning, Incremental Replay Mechanism to handle Domain Adaptation
- Achieved **0.85 F1** using Electra-BERT REPTILE ensemble with **2.65x improved runtime** using caching, **ONNX** & 8-bit **quantization**

9.2022 - 11.2022 **Sigmoid's Emotion Detection Challenge** Code

- Developed multi-label emotion classifier for detecting emotions in tweets
- Leveraged feature engineering by concatenating **Parts of Speech** (PoS) **dependency parsing features** with BERT encodings
- Achieved an F1-score of 0.59 with a union ensemble of three Roberta models implementing **few-shot classification**