**Scientific background**

Linguistic cues in text, such as variations in word choice and syntax and frequent repetition of certain words, are valuable for detecting deception. These cues often reflect changes in complexity, specificity, and emotional tone that may indicate dishonesty. Techniques like TF-IDF highlight these differences by weighing words based on their distribution across texts, and techniques like Word2Vec can build rich feature vectors that can help the models understand the semantics of deception. Classifiers like SVMs, Random forests and Artificial Neural Networks (ANNs) effectively analyze these weighted linguistic features, identifying patterns associated with deceptive behavior. This approach combines natural language processing with machine learning. (ANNs are used rather than RNNs because the data size is limited to 120 samples which is not sufficient for most RNN algorithms).

**Preprocessing**

The videos are transcribed using Google’s speech recognition to extract text from speech, then the text is passed through the following NLP pipeline to get cleaned:

1. Punctuation is removed.
2. Words are converted to lower case.
3. Tokenization is done.
4. Stop words are filtered out.
5. Words are lemmatized.

The cleaned text is then fed to a TF-IDF vectorizer to represent the sentences in a numerical format suitable for the ML and DL models.

**Model Training**

After preprocessing, the TF-IDF vectors are fed to various machine learning models like SVMs and XGB and SGD etc.. and deep learning models like a simple ANN. RNNs were not viable here due to the limited number of samples (only 120 samples) which is not sufficient to train an embedding layer.

**Important remarks and discoveries**

Other more complex approaches such as using Glove for vectorizing the text yield similar results to using TF-IDF so we chose to use TF-IDF because it is so much more efficient computation wise, the best model in this modality was the simple ANN yielding an accuracy of 79%, the machine learning algorithms’ were not so far behind achieving peek accuracies of around 75%.