Meeting Notes: 21/01/2025

1. **Learn LIME and SHAP**:
   - Read about LIME and SHAP to understand their methodologies
   - Focus on using LIME to analyze linear combinations (LC) and SHAP to validate feature importance.
2. **Build the Neural Network Classifier**:
   - Input: BERT-based embeddings (TinyBERT, DistilBERT)
   - Output: Linguistic features like **gender**, **number**, and **PoS**.
3. **Apply LIME**:
   - Use LIME to explain the classifier's predictions.
   - Extract information for each result and evaluate LC values to identify which dimensions encode linguistic features.
   - Aggregate LC values to observe trends across multiple predictions
4. **Apply SHAP**:
   - Use SHAP to provide additional explanations and compare feature importance results with LIME.
5. **Compare with Ekaterina's Work**:
   - Check whether LIME and SHAP identify the same dimensions and features as Ekaterina's experiments.
6. **First Experiment**:
   - Start with **simple perceptron models** or embeddings from small BERT-based models.
   - Focus initially on **gender classification** for nouns at the first
   - Test multiple words and analyze whether dimensions align across predictions.
   - Reference Ekaterina's work for baseline comparisons.