# Advances in Natural Language Processing

## Dr Diptesh Kanojia, Prof Constantin Orăsan

Surrey Institute for People-centred AI, Centre for Translation Studies
d.kanojia@surrey.ac.uk, c.orasan@surrey.ac.uk

**People-Centred AI** — UNIVERSITY OF SURREY

**CENTRE FOR TRANSLATION STUDIES** — UNIVERSITY OF SURREY

## 💡 Ongoing Projects 💡

With collaborations from researchers in academia and industry alike:
- Evaluating **Machine Translation Quality**
- Automatic Post-editing of Machine Translation
- Identify **Aggressive / Offensive** Language
- **Acronym Expansion** in Scientific Domain
- **NLP** applied to the **Legal** Domain
- **Sarcasm** and **Humour Understanding**
- Cognitively-aided Language Understanding

## Language Understanding

→ Natural Language Processing (NLP) is moving at a fast pace. **Language Understanding is a core sub-area of NLP**, which has seen tremendous advances since the advent of pre-trained 'foundation models'.

→ Surrey is leading research efforts on many fronts in this direction. **Our interdisciplinary approach addresses NLU problems in multiple domains** such as social, legal and scientific [1].

→ Using **state-of-the-art NLU**, we are naming entities [2] and expanding acronyms [1] while estimating the quality of machine translation [3]. That didn't make sense? **Read our papers!**

Generate Human Language

Analyse Human Language

Understand Human Language

## Collaborations 🤝

CFILT

EAMT — EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

Unbabel

ai4bharat

ebay

Samsung Research

IBM Research

Aston University

JUST: Access

Our diverse NLP interests have led to exciting collaborations with researchers worldwide!

## Surrey Natural Language Processing

https://github.com/surrey-nlp

https://huggingface.co/surrey-nlp

## Research Advancements 🏅

→ We released the **most extensive acronym identification and expansion dataset** in the scientific domain [1]. Using pre-trained language models, we achieve SoTA performance on the task! We also released the **most comprehensive Hindi Named Entity Recognition dataset** with the CFILT [2] Lab.

→ Ongoing investigations report **SoTA QE scores** using our **novel multi-task learning based approach**.

→ We are curating **QE data for six new language pairs** investigating the MT performance from English to Indian Languages. Last year, we co-organized the QE and APE shared tasks while **contributing two datasets**. Additionally, we are creating **APE data for two**, and **MQM evaluation data for one language pair**.

## NLP Café ☕

The **most popular café in Guildford!** We roast our beans once a week for an hour with many researchers from both NLP and non-NLP spaces. We started as a reading group but evolved into something more. In one year, we undertook a deep learning module, discussed interesting programming libraries, and looked at NLP advancements through a critical lens; all in the café.

Are you interested in joining us? **Scan the QR code to know more!**

## Multimodality 👁

Rohit: Well, no offense, but I didn't go to art school.
Rohit: I have a PhD in astrophysics

Melvin: Oh, great, it'll be easier for you to look up at the stars without a roof over your head.

**Sarcastic**

Bianca: Speaking of George,
Bianca: how's single life treating you?

**Non Sarcastic**

Alice: Fine, I guess. I've been focusing on me.

⇒ **Cognitive Data as a New Modality**

Eye-tracking (with human participants)

Gaze-based Features

Textual Utterance

Transformer-based Text Features

Low-level Audio Features

Preprocessing

(Raw Audio)

ResNet-152 Video Features

(Raw Video)

Individual Feature Vectors

Collaborative Gating Module

Attention based Multi-modal Fusion

Fused Feature Vectors

FFNN-based Prediction Module

Output Label (Sarcastic / Non-Sarcastic)
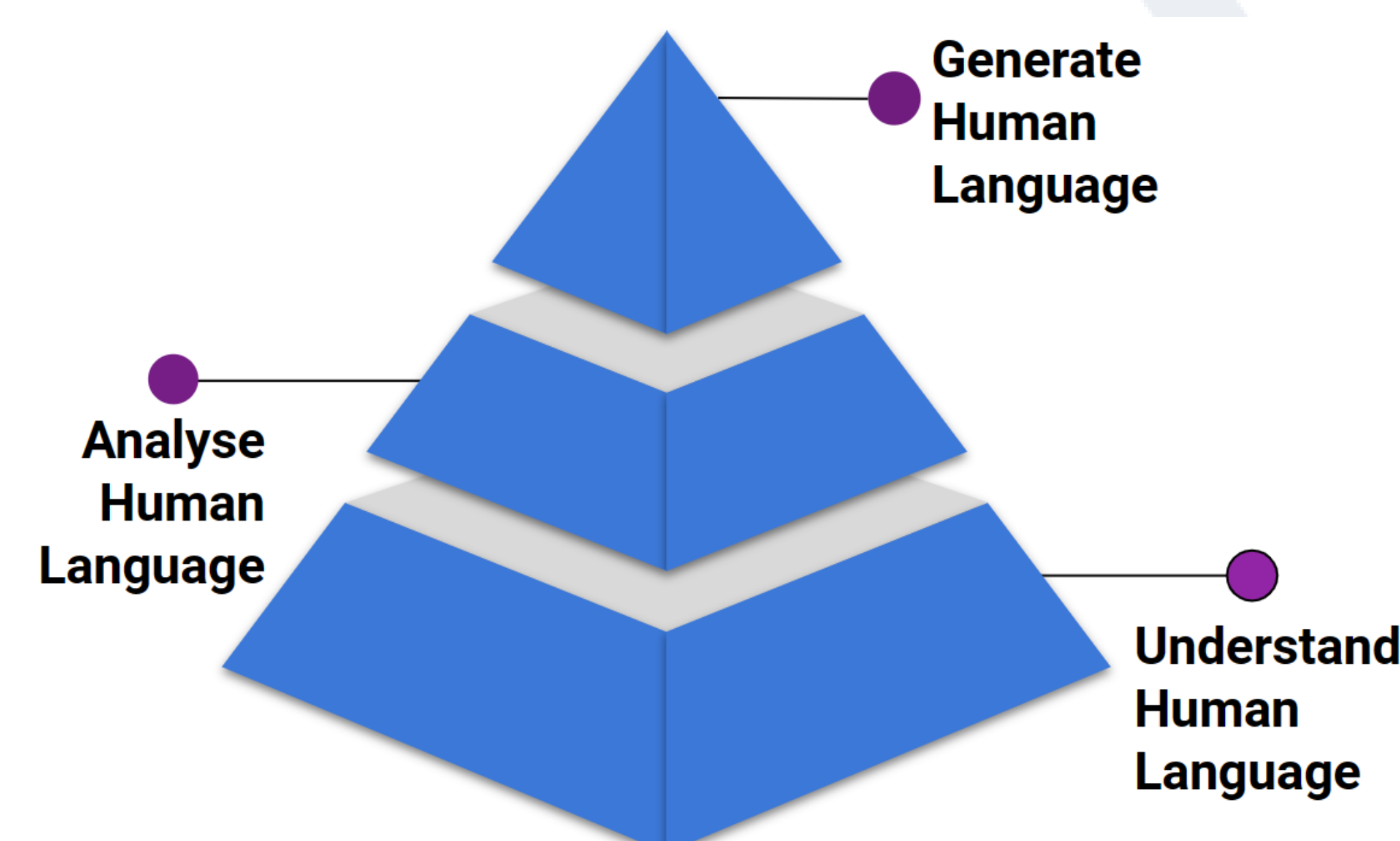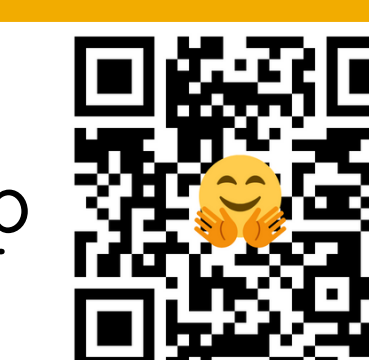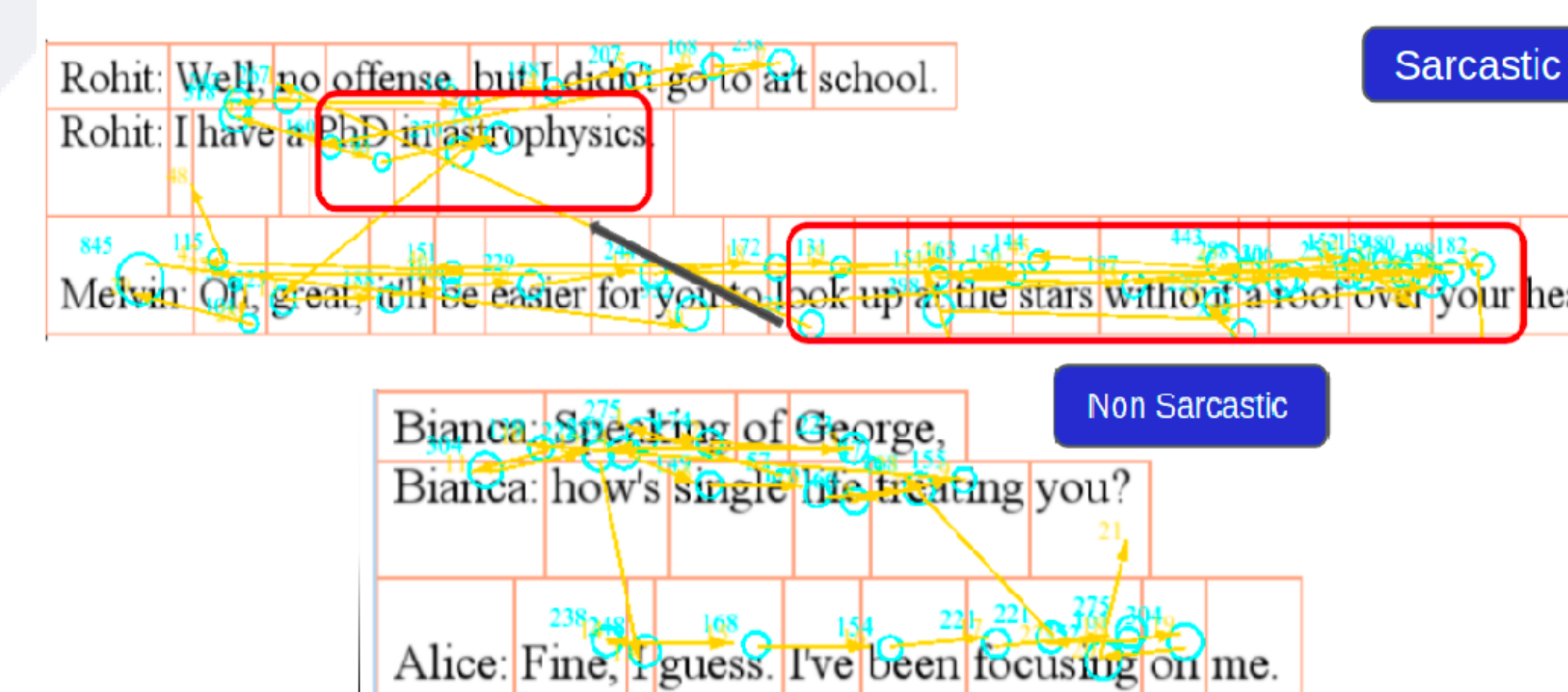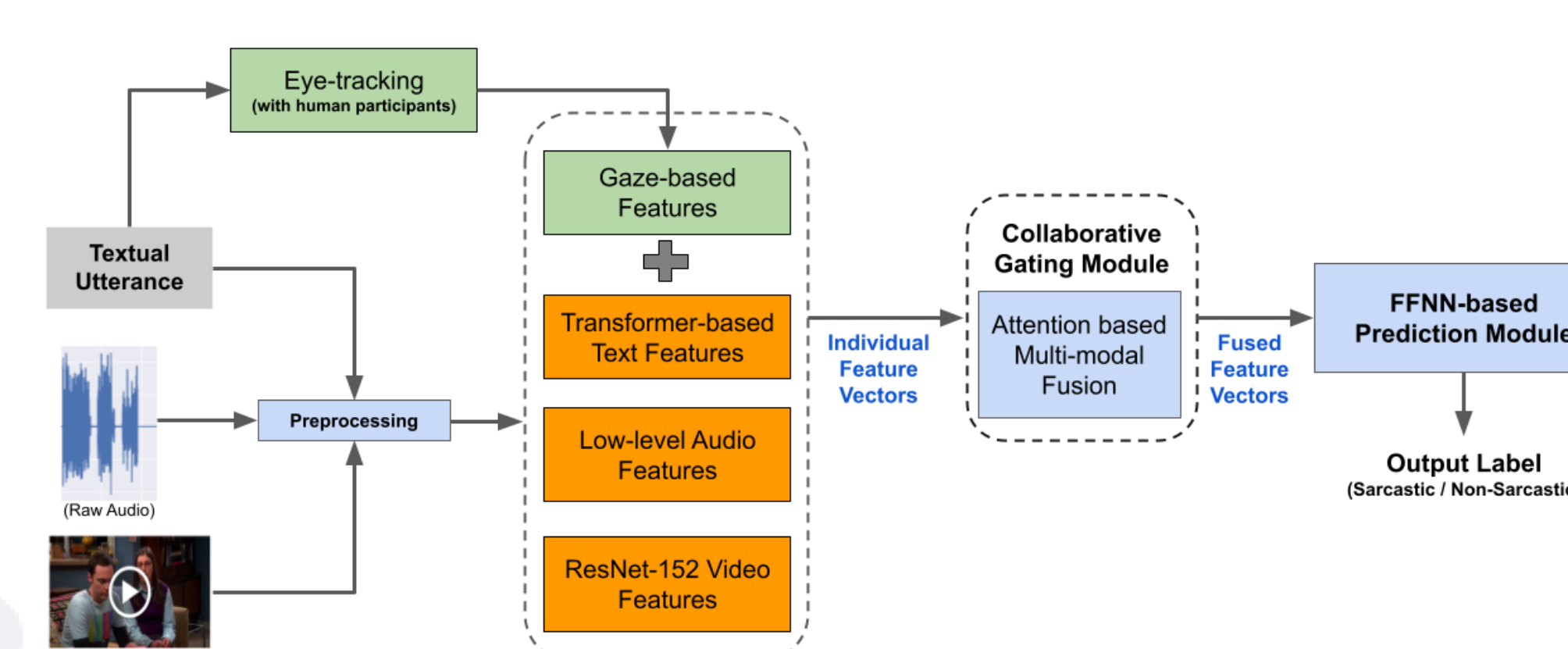
⇒ **Novel Approach / Multimodal Sarcasm**

## Open Access 📖

We are committed to open-access research in alliance with the UKRI open-access policy. Our **datasets**, like *PLOD, HiNER, S3D* are all **publicly available** along with the **code** and the **models** generated from our research. Check out our 🐙 and 🤗!

## Other Members 👨‍👩‍👧

→ Dr Félix do Carmo, Prof Nishanth Sastry, and Prof Helen Treharne

→ Dr Leonardo Zilio, and Dr Hadeel Saadany

→ Shenbin Qian, Archchana Sindhujan, Sourabh Deoghare, Vibhor Agarwal, Param Choudhary, Divyank Tiwari, Nazia Nafis, Akash Rawat, and Jordan Painter

## Significant Publications 📜

[1] L Zilio, H Saadany, P Sharma, D Kanojia, and C Orăsan. PLOD: An abbreviation detection dataset for scientific documents. In *Proceedings of the 13th LREC*, June 2022.

[2] R Murthy, P Bhattacharjee, R Sharnagat, J Khatri, D Kanojia, and P Bhattacharyya. HiNER: A large Hindi named entity recognition dataset. In *Proceedings of the 13th LREC*, June 2022.

[3] D Kanojia, M Fomicheva, T Ranasinghe, F Blain, C Orăsan, and L Specia. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the 6th WMT*, November 2021.

## Want to know more?

http://dipteshkanojia.github.io
https://dinel.org.uk/

UNIVERSITY OF SURREY