Report on the outcomes of a Short-Term Scientific Mission

Action number: CA19102

Grantee name: Sarang Shaikh

Details of the STSM

Title: **Differentiating the human, machine-generated text generated by LLMs using the state-of-the-art NLP techniques**

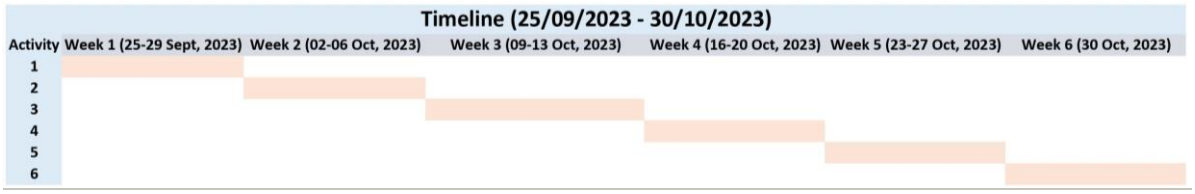
Start and end date: **25/09/2023** to **30/10/2023**

Description of the work carried out during the STSM

The purpose of this STSM is to use the black box testing for the review generated using by ChatGPT to differentiate them from human written review using SOTA NLP techniques such as using different linguistic patterns of the text, etc.

This collaboration includes meetings, presentations, research activities and professional network development in academia.

During my visit, I firstly met with the host Prof. Dr. Sher Muhammad Daudpotta. He is a senior faculry member at Dept. of Computer Science, Sukkur IBA University, Pakistan. As, he was already aware of what we were going to do in the proposed STSM, so we directly started with planning all the activities of the proposed STSM. We selected “ChatGPT” as the target LLM to consider for the scope of this STSM. The proposed weekly-based plan is shown below which was followed during the STSM activities.



The work will be carried out as per the objectives set in the goals of this STSM research stay.

The week 1 started with performing state-of-the-art of existing studies to understand how other people have been targeting the same problem scope as per this STSM. In order to start with existing studies, we search for relevant literature on “**Scopus**” database with the very specific query given below:

("machine generated text") AND ("large language models" OR "generative ai" OR "deep learning" OR "transformers" OR "machine learning" OR "natural language processing" OR "LLM")

Next, we will discuss some statistics about the research papers identified from the papers extracted based on above query. Figure 1 shows the year-wise frequency of the published research articles relevant to the scope of this

STSM. As we can see from the figure, there is a growing interest of research community in last 05 years for publishing papers into similar topic as per this STSM.

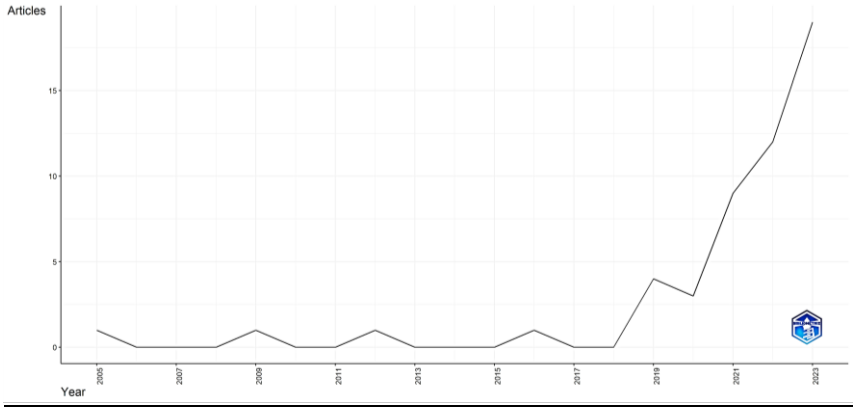


Figure 1 Year-wise research publications count

In order to see the impact of the topic we proposed in this STSM, we also looked into average citations of the articles gathered using the above query from the Scopus database. As we can see from the figure in the year 2021, there are maximum number of citations to the articles which again shows the potential of the proposed topic in this STSM.

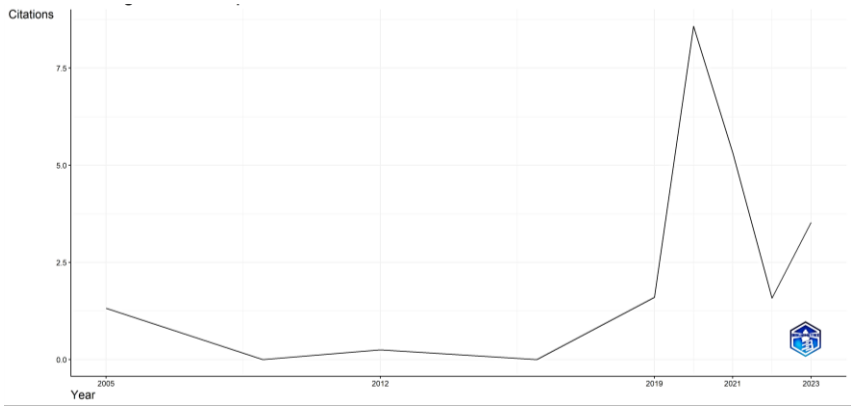


Figure 2 Year-wise average citations of research publications

In addition to this, to understand the existing approaches/methods/techniques used for differentiating human-written-text from machine-generated-text we created the co-occurrence network, word cloud and thematic map of all the different computational approaches used to address the similar problem as per this STSM. Figure 3 shows the word cloud of different techniques used such as: deep learning, text processing, computational linguistics, classification, etc. Furthermore, Figure 4 and Figure 5 shows the co-occurrence networks and thematic map of different techniques used in the existing studies for differentiating human-written-text from the machine-generated text.

Figure 3 Wordcloud of different methods used in the research publications

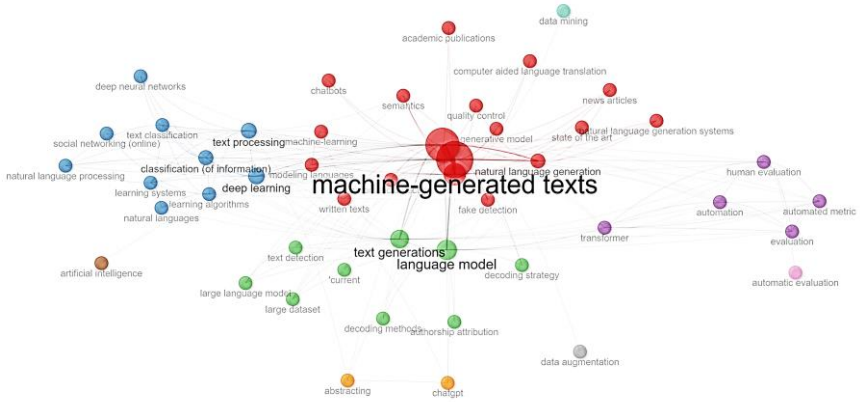


Figure 4 Co-occurrence network of different methods used in the research publications

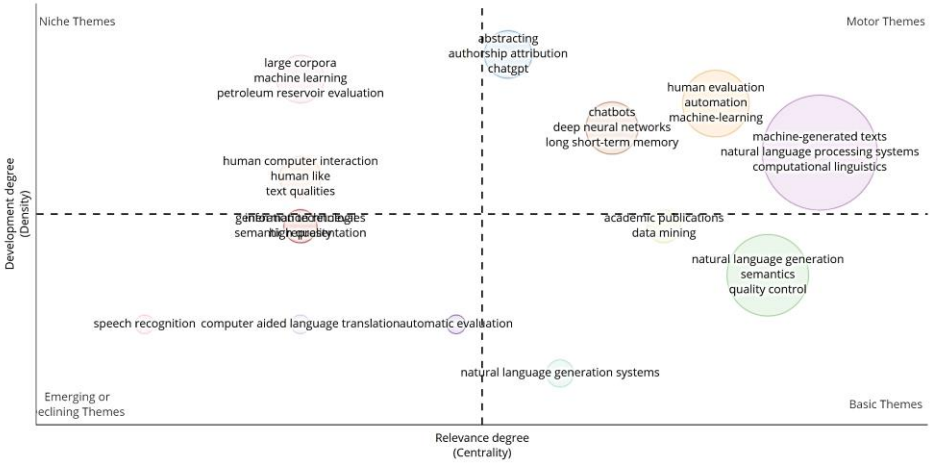


Figure 5 Thematic map of different methods used in the research publications

Figure 6 shows the world map of different countries by highlighting the regions from where the most of the publications are published as per similar domains to this proposed STSM.

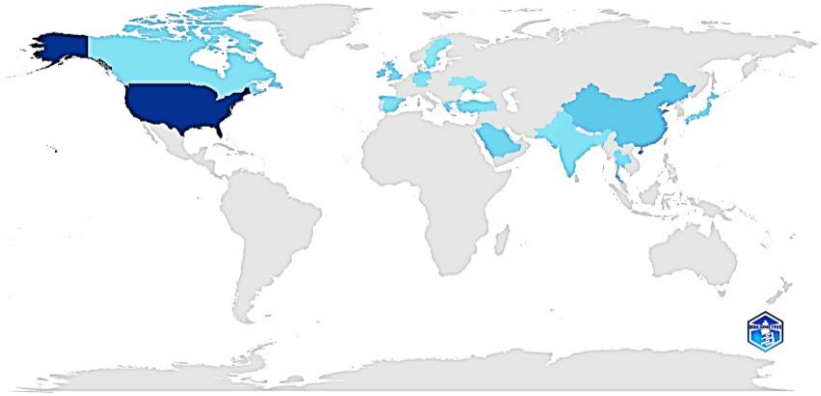


Figure 6 Countries map from where the research publications are produced

All the extracted research papers used for performing state-of-the-art studies in existing research are available at the link[1].

Figure 7 shows the overall proposed approach which we designed, developed and executed to achieve the proposed objective of this STSM which is to differentiate human-written-text from machine-generated text.

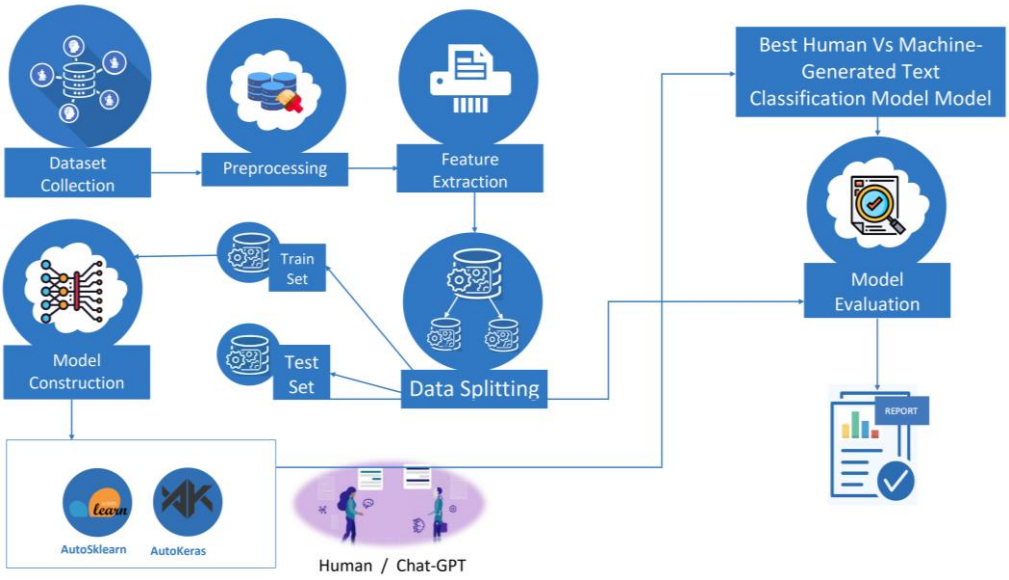


Figure 7 Approach followed in proposed STSM

The week 2 **(i.e., Data Collection and Preprocessing steps in the proposed approach in Figure 7)** started with gathering a large dataset of text from various sources, including human-written-text and machine-generated-text by ChatGPT. To ensure that the dataset covers a diverse range of topics and domains we started first look for existing datasets if available any. Finally, after doing the search for couple of days we identified an existing dataset suitable for our purpose. The name of the dataset is **"small-GPT-wiki-intro-features"** is available open-source at the link[2]. The dataset contains 100000 randomly selected texts (50000 from Wikipedia and 50000 generated by ChatGPT) comprising of overall 50000 wikipedia topics. The authors of the dataset combined the "title" of the original wikiepdia page with the prompt used for generating text by the ChatGPT. Below is the example of the used prompt.

200 word wikipedia style introduction on '{title}'
{starter_text}
where title is the title for the wikipedia page, and starter_text is the first seven words of the wikipedia introduction. Here's an example of prompt used to generate the introduction paragraph for 'Secretory protein' -
'200 word wikipedia style introduction on Secretory protein
A secretory protein is any protein, whether'"

By doing this for 50000 topics, the authors developed a combined dataset of all the original Wikipedia texts with respective ChatGPT generated texts with overall count of 100000 rows. Additionally, we also performed the necessary preprocessing steps by removing irrelevant information, special characters, and formatting inconsistencies. We also performed standard text preprocessing techniques such as tokenization, lowercasing, and removing stop words.

The week 3 **(i.e., Feature extraction step in the proposed approach in Figure 7)** started with looking into existing feature extraction techniques available for understand text data. From the research studies discussed above, we identified that most of the studies used common text-understanding feature extraction techniques such as Bag-of-Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), Word Embeddings, N-grams, and LDA (Linear Discriminant Analysis). But, to make the work done in this STSM different from the existing studies we focused on using **"Linguistic Features"** of the text for differentiating human-written vs machine-generated text in the collected dataset. Hence, after looking into the existing research studies, we identified a study conducted by Lee et al. and is available at the link[3]. In this study, the authors developed and proposed a list of handcrafted linguistic features which could possibly be extracted from any kind of text to different it from the other text. The authors developed a complete list of 219 features which were possible to extract for any text using the open-

[1] https://s.ntnu.no/U5CtpGOd
[2] https://huggingface.co/datasets/julia-lukasiewicz-pater/small-GPT-wiki-intro-features
[3] https://aclanthology.org/2023.bea-1.1.pdf

source library provided in the Python programming language and is available at the link[4]. Some of the examples of the used features are total_number_of_words, total_number_of_stop_words, total_number_of_punctuations, total_number_of_syllables, total_number_of_words_more_than_two_syllable, total_number_of_words_more_than_three_syllables, total_number_of_unique_words, total_number_of_sentences, etc. We extracted all the 219 features for the full dataset mentioned in the week 2 activities and filtered out those features whose correlations were below zero with the "class" column in the dataset. The class column in the dataset represents either the text is human-written or machine-generated. Hence, we only considered the 38 positive correlated features from all 219 features extracted using LFTK for the dataset to use these for differentiating human-written vs machine-generated text. The correlation values for each of 219 features is available at the link[5]. Once, we decided upon the input features, we split the dataset into standard 80/20 split where 80% of the dataset was used to train the supervised classification model and 20% of the data was used to test the trained model in terms of how efficient the model is in differentiating human-written vs machine-generated text. Figure 8 shows the comparison of average scores of 38 positive correlated features for both human-written vs chatgpt-generated text from the dataset which shows difference in values for both type of texts.

The week 4 and 5 consisted of exploring and evaluating state-of-the-art supervised classification models suitable for the task of differentiating human and machine-generated text based on our dataset. We started to consider SOTA models in machine learning and deep learning for this proposed STSM. According to "No Lunch Free" theorem, there is no single classifier which works for all kind of data. Hence, in order to save the time, we applied the "AutoML" approach using the python programming language.

Automated Machine Learning, or **AutoML** for short, involves the automatic selection of data preparation, machine learning model, and model hyperparameters for a predictive modeling task. It refers to techniques that allow semi-sophisticated machine learning practitioners and non-experts to discover a good predictive model pipeline for their machine learning task quickly, with very little intervention other than providing a dataset. In specific to the AutoML approach, we selected "**AutoKeras**" and "**AutoSklearn**" for performing our experiments. **AutoKeras** is an open-source library for performing AutoML for deep learning models. The search is performed using so-called Keras models via the TensorFlow tf.keras API. **AutoSklearn** provides out-of-the-box supervised machine learning. Built around the scikit-learn machine learning library, auto-sklearn automatically searches for the right learning algorithm for a new machine learning dataset and optimizes its hyperparameters. Thus, it frees the machine learning practitioner from these tedious tasks and allows her to focus on the real problem.

Feature	Human Text Scores (Avg. Scores)	ChatGPT Text Scores (Avg. Scores)
simple_punctuations_variation	0.19	0.25
simple_adpositions_variation	0.38	0.46
flesch_kincaid_reading_ease	67.48	77.79
average_number_of_auxiliaries_per_word	0.04	0.05
simple_determiners_variation	0.19	0.26
simple_coordinating_conjunctions_variation	0.27	0.33
average_number_of_named_entities_art_per_word	0.00	0.00
average_number_of_named_entities_date_per_word	0.02	0.03
average_number_of_numerals_per_word	0.03	0.04
simple_type_token_ratio	0.53	0.55
simple_type_token_ratio_no_lemma	0.53	0.55
average_number_of_stop_words_per_word	0.37	0.38
average_number_of_spaces_per_word	0.01	0.01
simple_auxiliaries_variation	0.27	0.30
average_number_of_coordinating_conjunctions_per_word	0.03	0.03
average_number_of_named_entities_quantity_per_word	0.00	0.00
total_number_of_named_entities_art	0.45	0.63
average_number_of_pronouns_per_word	0.04	0.04
average_number_of_named_entities_gpe_per_word	0.02	0.02
average_subtlex_us_zipf_of_words_per_word	4.07	4.11
average_number_of_named_entities_per_word	0.12	0.12
average_number_of_named_entities_norp_per_word	0.01	0.01
average_number_of_named_entities_art_per_sentence	0.06	0.07
average_number_of_determiners_per_word	0.09	0.09
average_number_of_named_entities_quantity_per_sentence	0.02	0.02
average_number_of_named_entities_loc_per_word	0.00	0.00
average_number_of_named_entities_event_per_word	0.00	0.00
average_number_of_named_entities_language_per_word	0.00	0.00
average_number_of_punctuations_per_word	0.13	0.13
simple_pronouns_variation	0.56	0.56
average_number_of_named_entities_product_per_word	0.00	0.00
corrected_determiners_variation	0.57	0.57
average_number_of_named_entities_money_per_word	0.00	0.00
average_number_of_named_entities_percent_per_word	0.00	0.00
root_determiners_variation	0.80	0.81
average_number_of_named_entities_law_per_word	0.00	0.00
average_number_of_interjections_per_word	0.00	0.00
average_number_of_named_entities_ordinal_per_word	0.00	0.00

Figure 8 Average feature scores (LFTK) of human-written vs chatgpt-generated text

[4] https://github.com/brucewlee/lftk
[5] https://s.ntnu.no/PTOUhsWO

Hence, for the “AutoKeras” approach, we selected 10 different deep learning models using AutoML approach by applying these models on train set of the dataset. Finally, we selected the best performed model by validating it on test set of the dataset. The best selected model achieved overall **82%** accuracy in differentiating human-written vs chatgpt-generated text as compared to the models trained via “AutoSklearn” approach which achieved overall **80%** accuracy. In addition to this, we also compared our best model with one state-of-the-art model available at the link[6] which is also used for similar kind of purpose. The model available at this link achieved overall **68%** accuracy for test set of our dataset. In order to further validate the performance of our best performing model, we selected few other metrics such as f1-score, kappa score and mcc score. Table 1 shows the different scores for all these metrics where we can see that our best performing model with AutoKeras has the maximum performance numbers for the test set of the dataset.

Table 1 Performance comparison of the models

	Accuracy	F1-Score	Kappa Score	Mathew Correlation Coefficient (MCC)
AutoKeras	**82%**	**0.82**	**0.63**	**0.63**
AutoSklearn	80%	0.80	0.59	0.59
ChatGPTDetector[6]	68%	0.65	0.35	0.41

In addition to this, we also visualized the performance of our model and other state-of-the-art model using confusion matrix which shows for each class label (i.e., human or chatgpt) how many number of instances are correctly or incorrecty predicted as compared to the actual lablels. Figure 9, 10, and 11 shows the confusion matrix for the models, where for our best performing model out of 9983 instances which are chatgpt-generated; the 7935 are correctly predicted.

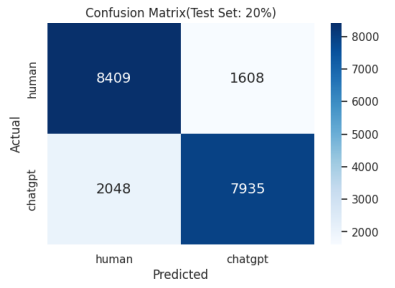


Figure 9 Confusion matrix (AutoKeras)

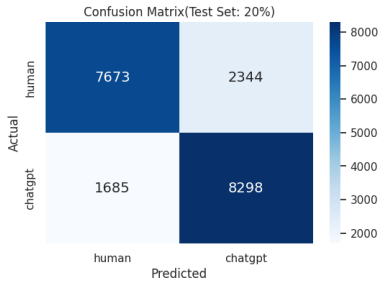


Figure 10 Confusion matrix (AutoSklearn)

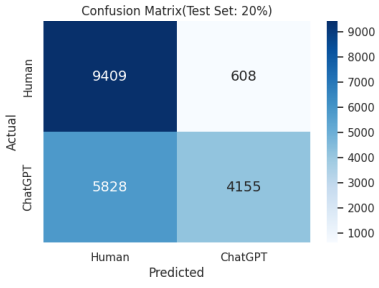


Figure 11 Confusion matrix (ChatGPT Detector)

The week 6 was consisted of deploying the best performing model in a real-world scenario to differentiate between human and chatgpt-generated text. We deployed a web application using **Streamlit** and **HuggingFace** library from python language where user can input the text and get the predicting result as human-written or chatgpt-generated text. The developed application is available at the link: **https://huggingface.co/spaces/sarangs/human_vs_chatgpt**. Figure 12 shows the screenshot of the developed application.

[6] https://huggingface.co/Hello-SimpleAI/chatgpt-detector-roberta

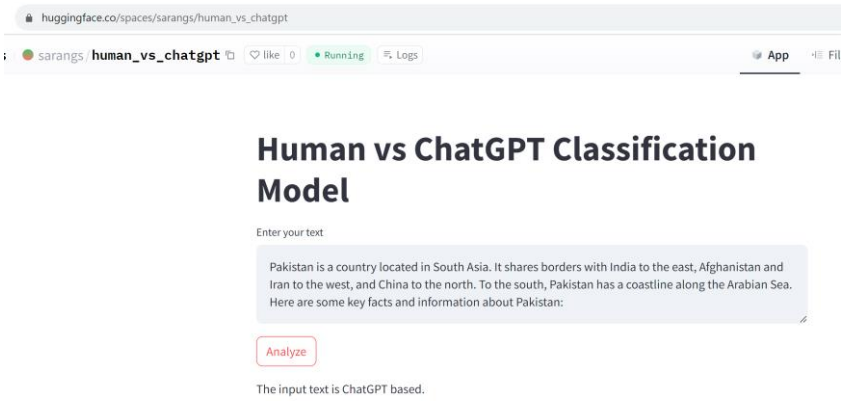


Figure 12 Screenshot of the deployed application

Finally, I presented the whole work to the host, and we started discussing the possible options to report all the work in form of publishing some peer-reviewed scientific papers. Information about this is given in the next section below.

Description of the STSM main achievements and planned follow-up activities

As discussed in the weekly based details in the above section, one can see that all the planned goals and expected outcomes of the proposed STSM are achieved.

Specifically, the planned goals and outcomes were:

1. Looking into the existing studies of differentiating human-written vs chatgpt-generated text – **this was achieved with the activities conducted in the week 1.** The data related to this is available at the link[1]
2. Collecting/gathering relevant dataset for the purpose/scope of this STSM - **this was achieved with the activities conducted in the week 2.**
3. Performing experiments using extracted linguistic features and building supervised learning classifier by incorporating AutoML - **this was achieved with the activities conducted in the weeks 3, 4, and 5.**
4. All the coding files related to experiments and results are available at[7] and[8]
5. The solution will be in form of a web-based application deployed online. - **this was achieved with the activities conducted in the week 4, 5 and 6.** The developed application is available at: **https://huggingface.co/spaces/sarangs/human_vs_chatgpt**
6. Discussion and future plan for minimum one paper publication from the overall activities of the proposed STSM - **this was achieved with the activities conducted in the week 6.**
    i) A scientific paper to report all the experiments performed on the gathered dataset along with the details on the deployed web application (Tentative Venue: IEEE Access) – Status: In Progress

The whole activity and expected results will contribute towards below specific Action objectives.

1. Research coordination objectives
    a) Develop: (i) methods, (ii) theory to study language in the human-machine era.
    b) Generate substantive guidelines for equitable development of emerging technologies.
    c) Advance understanding of emerging technologies likely to influence language.
2. Capacity building objectives
    a) Create a collaborative network with critical mass to drive scientific progress

[7] https://colab.research.google.com/drive/12U6sR8Hz1sSRdhgJqBmUGMotNPJ3x_vu#scrollTo=ot6KbgsbFRA-
[8] https://colab.research.google.com/drive/1v6GBtOmY41bpt2oW-04LhaXsm0KBXCm8#scrollTo=0190c085

b) Achieve breakthroughs by building much-needed bridges between computational linguists and a range of other linguists, alongside developers and stakeholders.