# Bridgei2i: Preliminary Report Submission

## Objective

Process digital content like emails, articles, reports, videos, tweets etc. The task is further broken down into:

- Theme Identification of tweets and articles.
- Headline generation for articles which follow the mobile technology theme.
- If the identified theme is mobile tech, assign a sentiment against the brand described.

## Data Description and Preprocessing

The data provided to us contains 4000 tweets and 4000 articles of both mobile tech and non-mobile tech themes. We apply the following preprocessing strategies to clean the text:

**Tweets:**

- Removed hyperlinks, numbers, symbols and emojis.
- Removed username.
- Removed RT (Re-Tweets).
- 90-10 stratified split for train and validation samples.

**Articles:**

- Removed symbols, hyperlinks.
- Translated target Hindi headlines to English (predicted headlines are to be in English). This was done using fine-tuned mBART (initialized from checkpoint) on 0.5 million parallel samples from IIT-B Hindi-English dataset.
- Removed the samples with article length < 32 and headline length < 1.
- 85-15 stratified split for train and validation samples.

The following table describes counts for each class theme.

| Data Type | Before pre-processing | | After Preprocessing | |
|---|---|---|---|---|
| | Mobile Tech | Non-Mobile Tech | Mobile Tech | Non-Mobile Tech |
| **Tweets** | 1000 | 3000 | 407 | 1660 |
| **Articles** | 1000 | 3000 | 858 | 2865 |

The following challenges were identified in both the datasets across all the tasks.

## Challenges

- Tweets
    - Architecture should be scalable across multiple languages.
    - The available data is very less and similar (e.g: the abundance of Samsung related tweets in the dataset), making it difficult to generalize.
    - Presence of multilingual imbalance due to less Hindi and Hinglish tweets as compared to English tweets
- Articles
    - Scalability across multiple languages.
    - Articles contain a large amount of text and building long-range context is difficult.
    - Most of the sentences are neutral and do not convey a polarizing sentiment.
    - Presence of multilingual imbalance due to fewer Hindi and Hinglish articles.

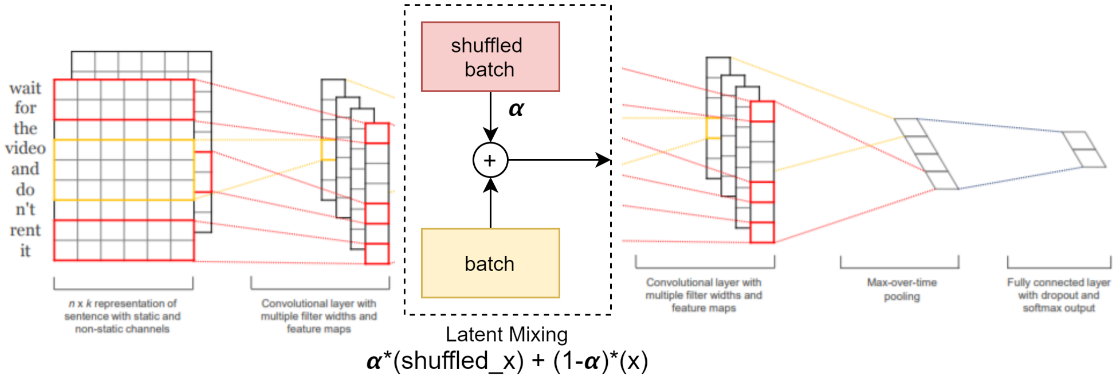
We shall now describe each of the modules in detail in the following sections.

## Theme identification

To identify the theme of the tweet or article, we propose a text classifier which can categorize any given text (in Hindi/English/Hinglish) into binary classes - mobile tech or non-mobile tech.

### Proposed Method:

1) Since we are working with multilingual text, it is important to identify a common representation scheme across all languages. We propose to use **phonemes** - the smallest unit of speech distinguishing one word. Hence, text in any language can be converted to its phoneme representation using International Phonetic Alphabet (IPA). This further **reduces the vocabulary size**, making it easier to train deep learning models.
2) Even though Transformer models are the current SOTA in language understanding, the task of identifying the theme as mobile tech does not require significant linguistic understanding and it suffices to identify certain keywords (e.g: “mobile”, “Samsung” etc). To justify this hypothesis, we finetune a pre-trained XLM-R model on the preprocessed dataset and we notice that the model classifies any tech-related tweet or article as “mobile tech” due to its capacity to capture linguistic information over specific keywords. Therefore we propose to use a variant of **Kim-CNN** to take a phoneme input and predict the theme. Further, this method is **computationally efficient** and showcases **faster inference speeds.**
3) It is difficult to ensure sufficient generalization capacity with the given small dataset. Hence taking inspiration from Cut-Mix regularization used in images, we propose a linear interpolation between latent spaces (**Latent Mix**) at a specific intermediate layer while simultaneously mixing the target classes as well with the same interpolation value. This forces the model to attend to less discriminative parts from the sentence for classification. Further, it increases the diversity of the training samples and induces the model to capture interdependencies between languages.


wait
for
the
video
and
do
n't
rent
it
*n* x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
shuffled batch
α
+
batch
Latent Mixing
α*(shuffled_x) + (1-α)*(x)
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

## Results

| Data Type | Train Accuracy | Validation Accuracy |
|---|---|---|
| **Articles** | 99.97% | 92.2% |
| **Tweets** | 96.12% | 93.72% |

**# of Params:** 0.94 M
**Inference Speed* (in secs):**
0.0019 (phoneme conversion) + 0.0015 (model inference) for a seq len of 550 phonemes.

## Sample outputs

RT @aajtak: 20 हजार के अंदर 5G स्मार्टफोन्स आ चुके हैं, क्या इससे भी सस्ता 5G स्मार्टफोन लाएगा Nokia? पढ़ें ये रिपोर्ट.. #Smartphones #Nokia https://t.co/of7KIH3lKS
**Predicted probabilities:** [1.6838817e-25, 1.0000000e+00] **(predicted theme: Mobile Tech)**

February is American Heart Month! ❤️ Celebrate with Raleigh Parks by taking time to show your heart some love! We'll be sharing tips all month to improve your overall health and lessen your risk for heart disease! https://t.co/HwzyaGTIgy #HeartMonth #RaleighParks https://t.co/xkiBzP6DND
**Predicted probabilities:** [9.9999875e-01, 1.3529364e-06]**(predicted theme: Non-Mobile Tech)**

*Inference times are computed on an RTX 2060 GPU

# Headline Generation

To summarize articles, we propose a text-summarization model scalable to multiple languages.

### Proposed Method:

- **Augmenting mobile tech samples** by randomly replacing a part of the sentence with their synonyms. This doubled the size of the dataset to help the model generalize better while also introducing some noise in the training dataset.
- **mBART**, a transformer based model from facebook is used for headline generation. It's an encoder-decoder model which was pretrained with the objective of denoising multiple languages simultaneously. This model is finetuned on the given dataset. Pretrained checkpoints can be chosen based on the required languages.
- We plan on introducing a **routing layer** to automatically route word tokens (ref: paper). This allows the model to learn interdependencies between multiple languages and learn transformations unique for each language.

### Results

| Data Type | Rouge-L | BLUE Score |
|---|---|---|
| **Validation samples with Mobile_Tech_Flag = 1** | Precision = 0.46<br>Recall = 0.49<br>F1 Score = 0.52 | 23.46 |
| **All validation samples** | Precision = 0.39<br>Recall = 0.43<br>F1 Score = 0.42 | 20.53 |
| **Training samples** | Precision = 0.45<br>Recall = 0.49<br>F1 Score = 0.47 | 28.23 |

**# of Params:** 610.85 M
**Inference Time*(in secs):** 0.82

### Sample outputs

**ID:** article_0176
**Predicted headline:** Global Vertical Farming Market to Increase Demand for Organic Food in the Forecast Period 2021-2026
**ID:** article_0179
**Predicted headline:** ISL 7: Jamshedpur FC puts winless run to bed with spectacular Mobashir's goal
**ID:** article_1605
**Predicted headline:** Samsung Galaxy M02, Galaxy A32 and Galaxy A32 Pro 5G can be purchased online at Rs. 10,000.

*Inference times are computed on an colab T4 GPU

# Identification of mobile brands and their corresponding sentiment

## Proposed Method:

- The Hindi/Hinglish tweets are translated to English using the mBart model described in the previous section. In the case of articles, English headlines are generated using the previous module.
- Identified mobile brand entities using NER (Named Entity Recognition) for each mobile tech tweet/article.
- If the entity identified belongs to the exhaustive list of mobile companies, Aspect Based Sentiment Analysis (ABSA) is used to identify a sentiment towards each aspect in the text.
- Else, the sentiments for the recognized entities are computed by **weighing the sentiments** of corresponding adjectives, adverbs, amods, advmod, and verbs of the noun from the dependency graph. This helps capture the required sentiment for any given entity.

## Results:

There are no available ground truth labels for the training data and hence we cannot compute the performance metrics.

**Inference time* (in secs):** 0.82 (Translation) + 0.17 (NER + sentiment)

## Sample outputs:

Our observations on the predictions suggest that most of the given data had single entities. Out of the remaining multiple entities, there were very few entities with contrasting sentiments.

1) **Tweet**: '#Samsung is now making another addition into it's A series, and the word is out that it's upcoming variant Galaxy A72 4G has awesome battery and charger with brilliant specs . . . For more tech updates Follow 👉 @MobileNerve . . . #smartphone #mobile #mobilephone #mobilephoto https://t.co/b8UO2Z6mC4'
   **Predicted Entity**: ['samsung'] **Predicted Sentiment**: ['Positive']

2) **Tweet :** 'Vivo X60 5G: The snapdragon 888 is coming to India in March or April. Check its full specs, price, pros & cons. https://t.co/fRMzGq4Glt #Vivo #VivoX60ProPlus #vivoY51s #VivoY31 #VivoX60 #VivoX60Pro #X60 #X60Pro #X60ProPlus #vivov20 #vivov20se #VivoIndia #Smartphone #thewidgetguy'
   **Predicted Entity**: ['Vivo'] **Predicted Sentiment**: ['Neutral']

3) **Tweet**: 'RT @RethinkWireless: LG may exit smartphone market after 23 quarters of losses...https://t.co/PwCJnJlsY7 #Wireless #5G #wifi #wifi7 #cellular #operators #spectrum #RAN #rethinktechnology #cloud @LGUS https://t.co/ztKZhmRzzH'
   **Predicted Entity**: ['LG'] **Predicted Sentiment**: ['Negative']

*Inference times are computed on an colab T4 GPU