# **Automated identification, summarization and entity based sentiment analysis of mobile technology articles and tweets**

**Context**: Digital content is expanding at a very rapid pace. Many activities that experts (like editors, auditors, judges, doctors, underwriters) undertake today involve the ability to process digital content (emails, articles, reports, videos, tweets etc.) and synthesize them to make decisions. This is a complex activity that experts have mastered over decades of experience and expertise.

This problem explores a foundational aspect of enabling machines to assist experts in taking decisions by helping them synthesize digital content effectively.

**Problem**:

1) Develop an intelligent system that could first identify the theme of tweets and articles.

2) If the theme is mobile technology then it should identify the sentiments against a brand (at a tweet/paragraph level).

3) Participants to generate a one sentence headline of max 20 words for articles which follow the mobile technology theme. Headline for tweets and non-mobile tech articles is not required.

The articles and tweets would be in multiple languages (will focus on English, Hindi, Hinglish, as a start).

**Input Dataset**:

1) A mix of 4000 newspaper articles in English, Hindi & Hinglish along with their headlines will be provided

2) A mix of 4000 tweets in English, Hindi & Hinglish will be provided

Data would be shared in a csv format with the students. List of themes/topics for classification with corresponding #tag. Students have to scrap the tweets corresponding to the #tag.

**Infrastructure**: Google Colab notebook with Python/R code, Readme.txt file, Requirement.txt file is preferable. No paid API or Services should be used.

**Skills Required**: Advanced NLP & Deep Learning, Web-scraping, Knowledge of any coding platform (Python, R, etc)

**Output:**

1. Binary Classification of the article & tweets to a *‘mobile\_tech’* or *‘other’* theme
2. For all articles and tweets where the classified theme is ‘mobile\_tech’ you would need to identify the brand name and its corresponding sentiment at a tweet/Article level. For e.g.  Tweet -> *‘Apple phones have a better battery life compared to Samsung phones #APPLEROCKS #SAMSUMGSUCKS’* should recognise **Apple** & **Samsung** as two brands along with a **positive** & **negative** sentiment for them respectively
3. Automatically generated headline on *‘mobile\_tech’* themed articles in English,  Hindi & Hinglish
4. Approach note summarizing the algorithmic approach used for developing the solution, other solutions evaluated and considerations behind the choice of this specific approach

**Evaluation**: A dataset of 100-500 articles & tweets in English, Hindi & Hinglish will be used to validate all the algorithms and approach adopted. Following would be the key criteria:

1. Theme classification evaluation: Precision, Recall and F1 Score
2. Entity based sentiment evaluation:  Accuracy of Brand identification and Precision, Recall and F1 Score Sentiment(Note: The Brand names need to be in English irrespective of the language in the article)
3. Automated Headlines evaluation: (Note: The generated Headlines need to be in English irrespective of the language in the article)
   1. Average similarity scores of AI generated headlines compared with actual headlines would be used as a metric for evaluation. Embedding based similarity score generating code is available in the Python Notebook
   2. Rough and BLEU score (<https://en.wikipedia.org/wiki/ROUGE_(metric)> (<https://en.wikipedia.org/wiki/BLEU> ). Code is available in the Python Notebook
4. Scoring Speed of all three algorithms (Ensure that the runtime is stored in a variable which can called out later)
5. Innovative approach
6. Scalability of solution to other language

A sample output dataset with 10-20 articles and tweets will be shared for your understanding. Make sure the output is in the given format. No adherence to the required structure would lead to disqualification.