|| 26- 28th March, 2021 ||

INTER IIT TECH MEET'21

IIT Guwahati

HIGH PREP EVENT

(400 Points)

BRIDGEi2i'S AUTOMATED HEADLINE AND SENTIMENT GENERATOR

Digital content is expanding at a very rapid pace. Many activities that experts undertake today involve the ability to process digital content 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 STATEMENT:

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) We would need a one-sentence headline of max of 20 words for articles that follow the mobile technology theme. A headline for tweets 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/paragraph 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 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 approaches 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

3) Automated Headlines evaluation: (Note: The generated Headlines need to be in English irrespective of the language in the article)

a. 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 will be shared

b. Rough and BLEU score (https://en.wikipedia.org/wiki/ROUGE_(metric) (https://en.wikipedia.org/wiki/BLEU). Code for the same will be shared

4) Scoring Speed of all three algorithms (Ensure that the runtime is stored in a variable which can be called out later)

5) Innovative approach

6) Scalability of a solution to another 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.

Please note that we will provide final testing data ~ 1 day before the final submission date. There will be interim connects required to see and evaluate intermediate output.

A maximum of 10 participants (per team) shall be awarded participation /merit certificate.