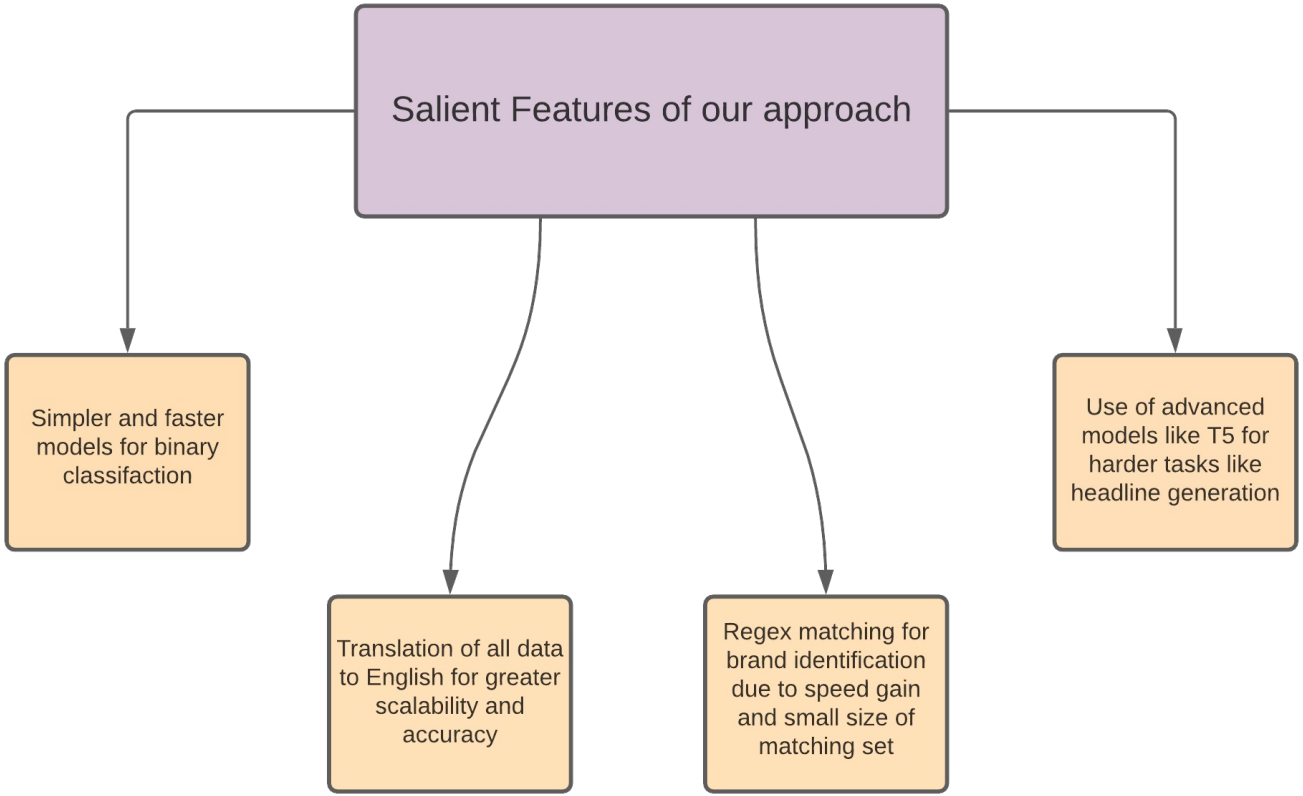
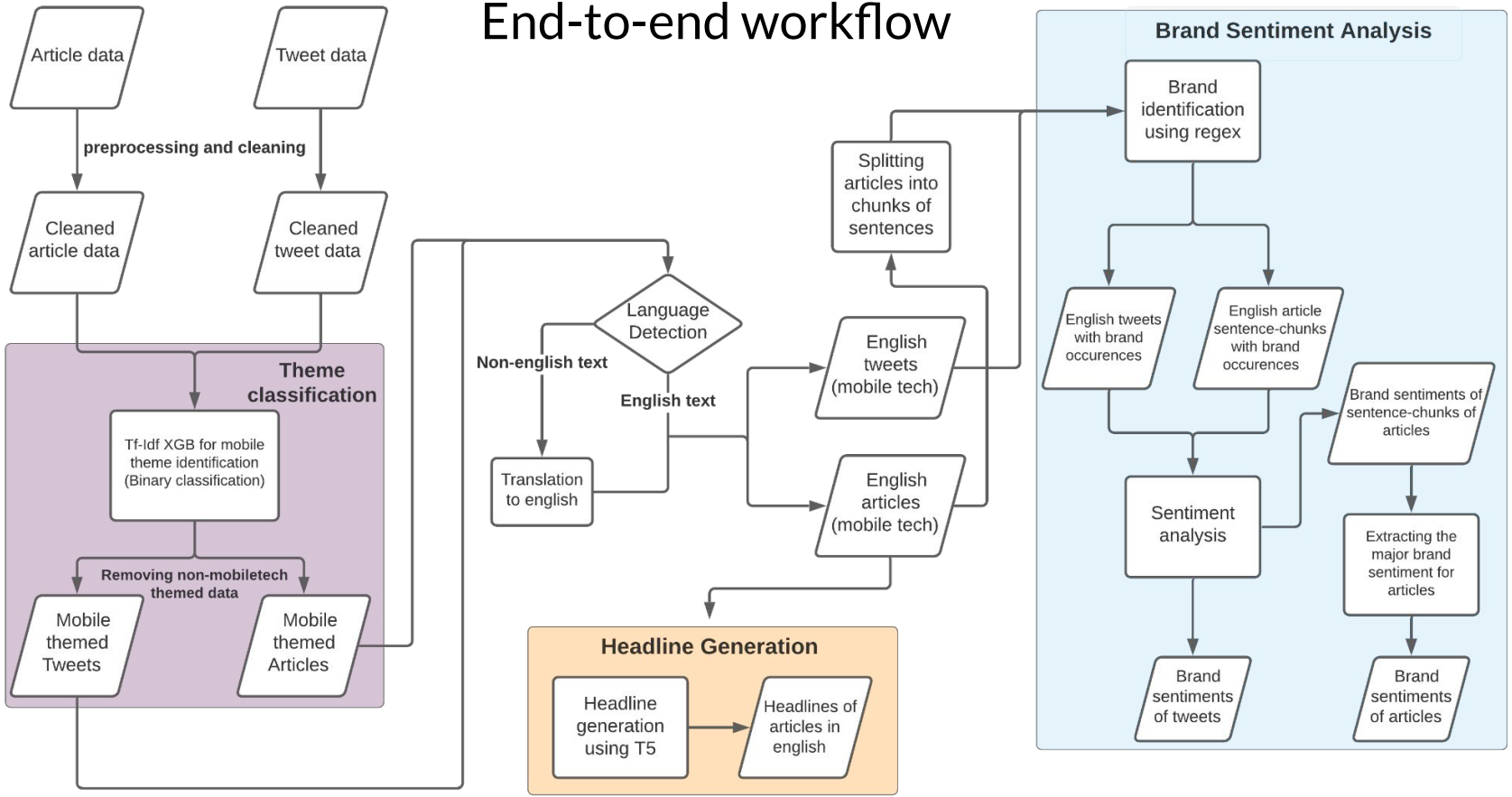
Bridgei2i's Automatic Sentiment and Headline Generator

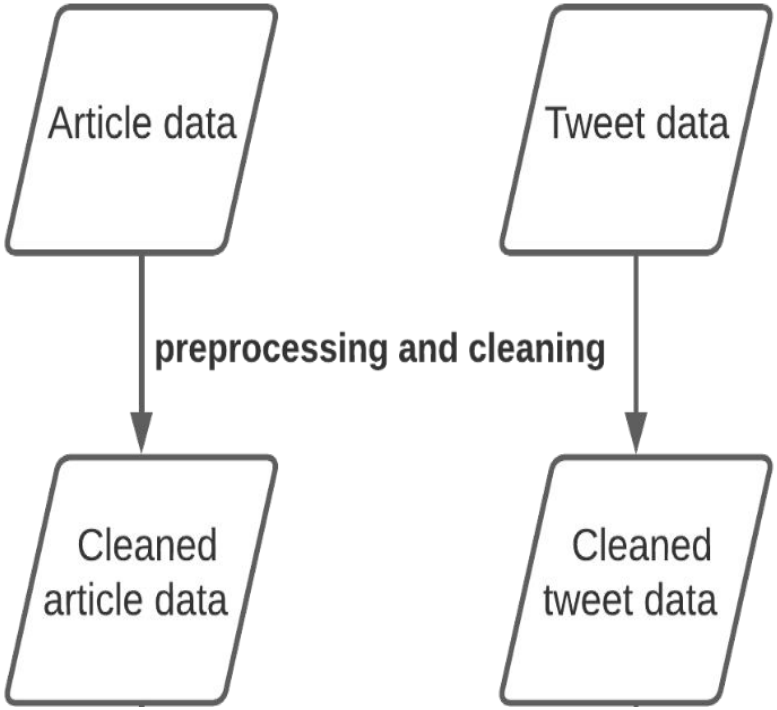
H2_B2I_4

Introduction



End-to-end workflow



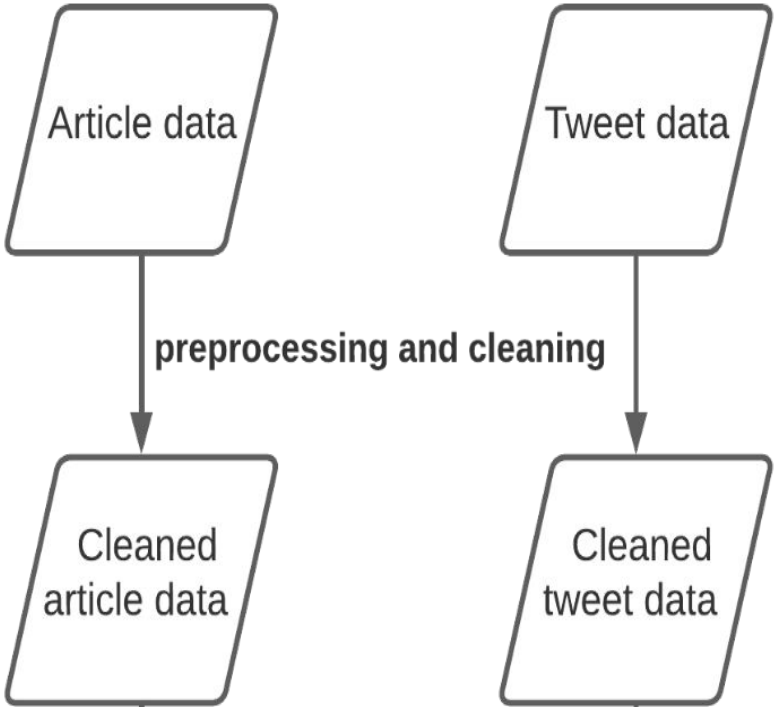


Tweets:

- Remove tweet specific terms like RT and QT
- Remove URLs and extra whitespace
- Remove mentions if they don't contain any brand names
- Replace Emojis with their text descriptors

"@airtelindia SIM card ke sath 5G phone bhi de dena.. kuch mahine pahle hi to smartphone liya tha.. 😒"

"@airtelindia SIM card ke sath 5G phone bhi de dena.. kuch mahine pahle hi to smartphone liya tha.. **:unamused face:**"

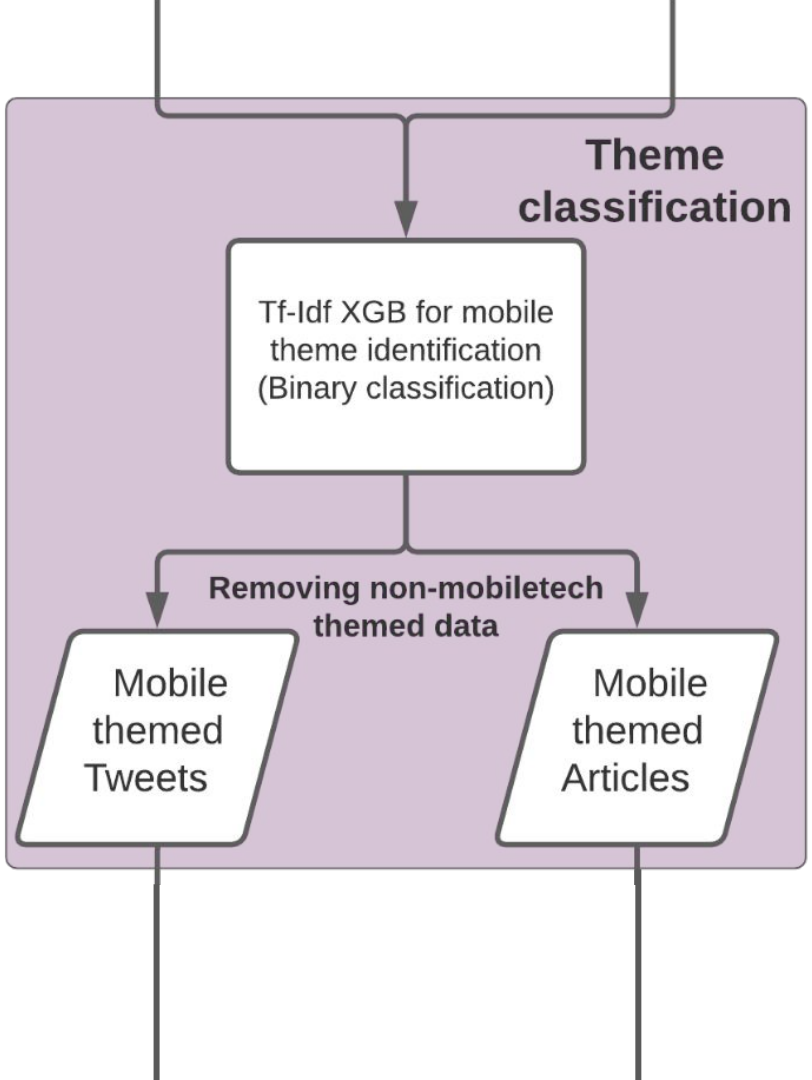


Articles:

- Remove formatting symbols like | and ^
- Remove URLs and extra whitespace
- Remove any space in usage of “ . ” to denote decimals to disambiguate them from full stops.

“The company, however, has announced the Mi 11 launch alongside MIUI 12 . 5 on February 8. Mi 11 specifications Xiaomi Mi 11 sports a 6. 81-inch AMOLED panel with a QHD+ resolution and a 120Hz refresh rate.”

“The company, however, has announced the Mi 11 launch alongside MIUI 12.5 on February 8. Mi 11 specifications Xiaomi Mi 11 sports a 6.81-inch AMOLED panel with a QHD+ resolution and a 120Hz refresh rate.”



Tf-Idf Vectorizer + *XGBoost* Classifier.

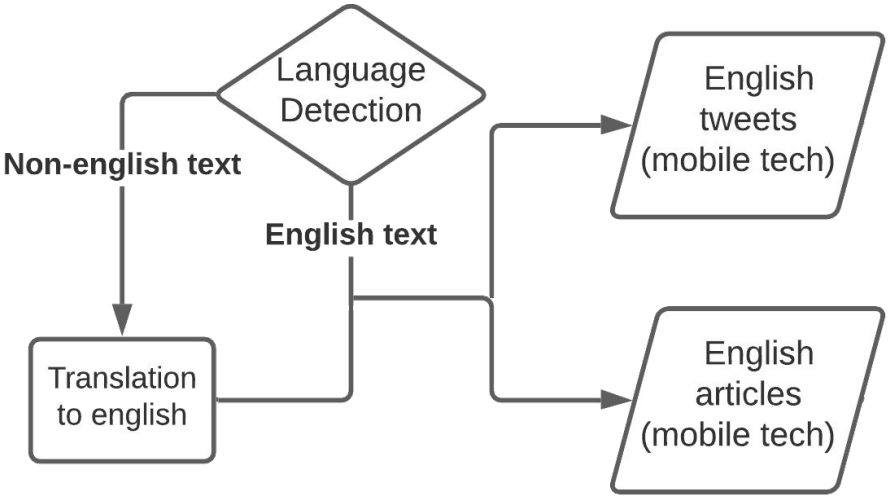
Tf-Idf - Term frequency-inverse document frequency. Better feature extraction as it is a statistical method.

XGBoost Classifier for binary classification.
Boosting to increase accuracy without significant size increase.

Advantages of using Tf-Idf XGBoost:

- Simple and easily customizable - The model is simple and therefore can be more easily customized to support the current dataset or data distribution.
- Consumes much less time and space - The comparisons to a BERT model are shown below:

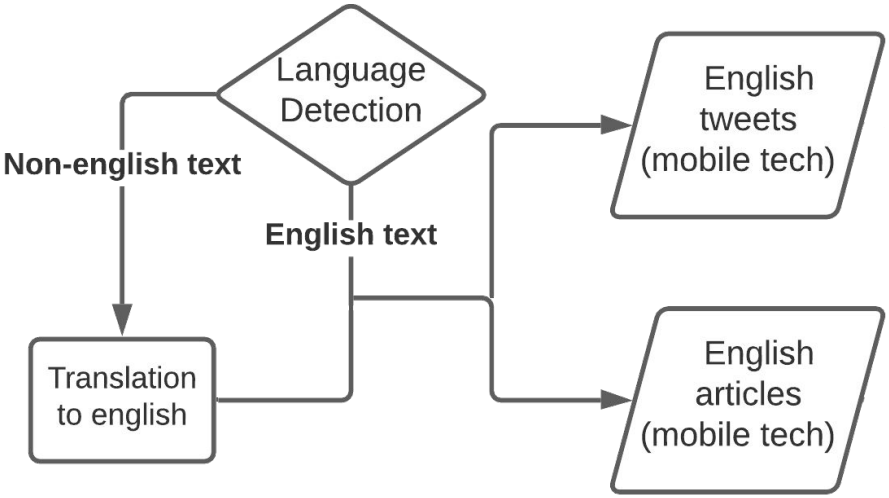
	Time	Space
BERT	Training: 10 minutes Inference : 2 minutes	Over 600 MB
Tf-IDF XGB	Training + Inference : 15 seconds	Less than 10 MB



Language Detection

Combined several open-source libraries with some custom rule-based logic:

- **langdetect** library (n-gram naive bayes) for english
- Detection of Devanagari characters for Hindi using **indic_transliteration**
- Low probability in both Hindi and English points to Hinglish

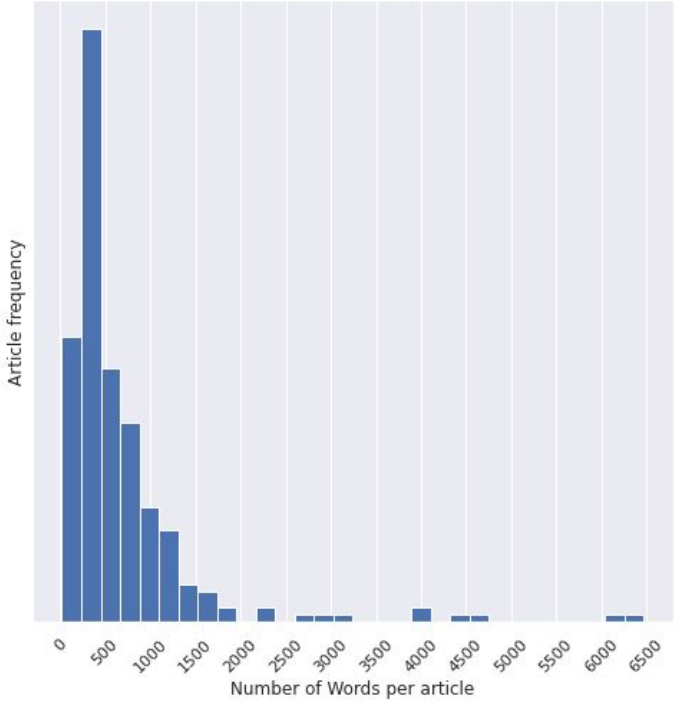


Language Translation

- Why?
  - Scalability
  - Accuracy
- Why Google API ?
- Speed: a bottleneck in the free API


Splitting articles into chunks of sentences

Splitting

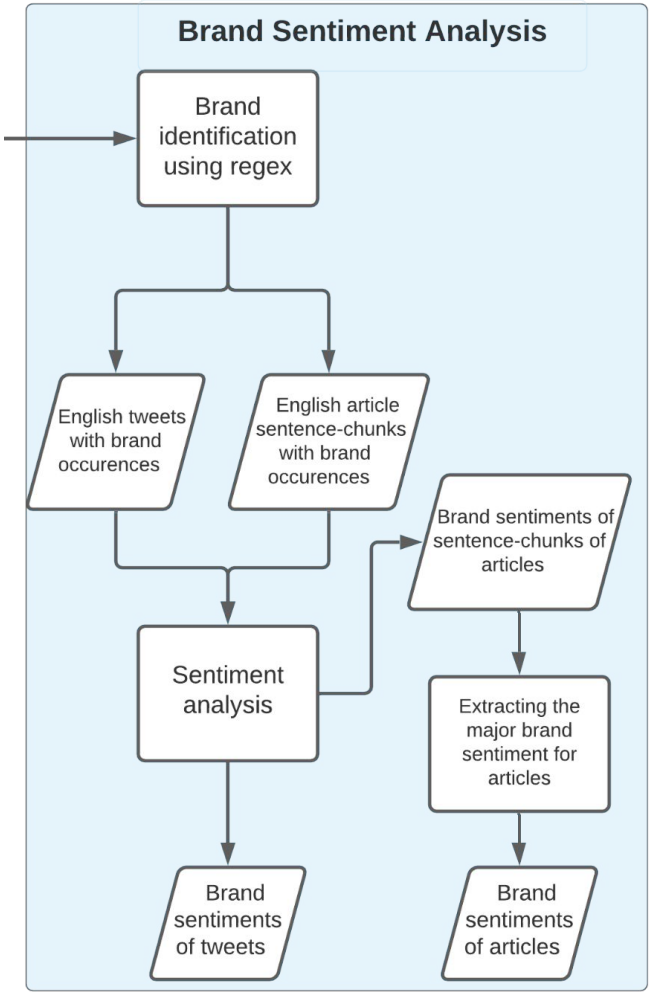

Article frequency
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
Number of Words per article

Splitting

Greedy approach:

The *span of the context* of a brand has been approximated by including the words starting from the first occurrence of the **brand** and ending with the occurrence of the **next brand**. This is carried out on a sentence level.

POCO is all set to launch a new X series smartphone with a Snapdragon 855 So C and four rear cameras. Read more to find about **POCO** X3 Pro Price in India , Specifications , and Features Price and Availability Several details of the **POCO** X3 Pro have surfaced online. However , there is no information about its price tag and launch date. Also read : **ITEL** A47 With HD Display , 32GB Storage Launched At Rs 5,499 Specifications and Features Four rear cameras 48MP primary sensor Qualcomm SM8150 (Snapdragon 855). The **POCO** X3 was launched in India in September last year with a starting price of Rs 16,999.



Brand identification

Why Regex?

- Huge speed gain
- Modestly-sized matching set which is publicly available
- Often more reliable

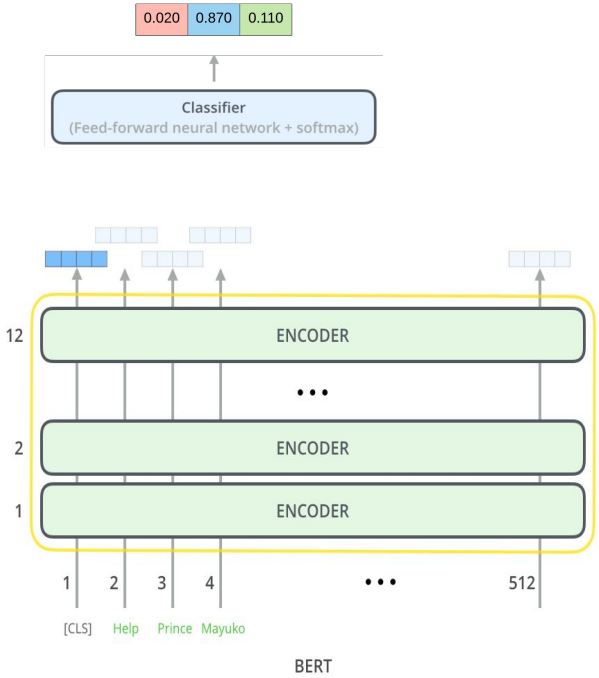
```
pattern = r'\b Oppo \b'
```

Matches word boundaries

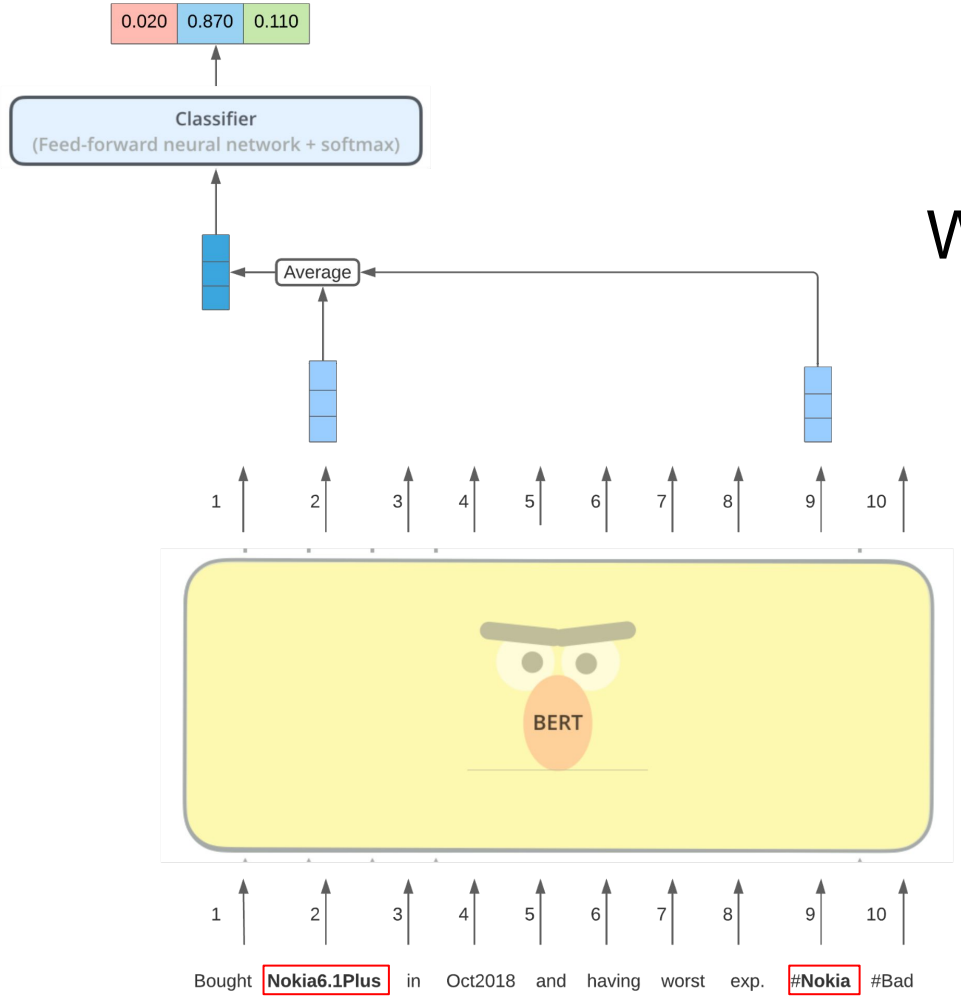
Oppo

Opportunity

Sentiment classification using Contextual Embeddings

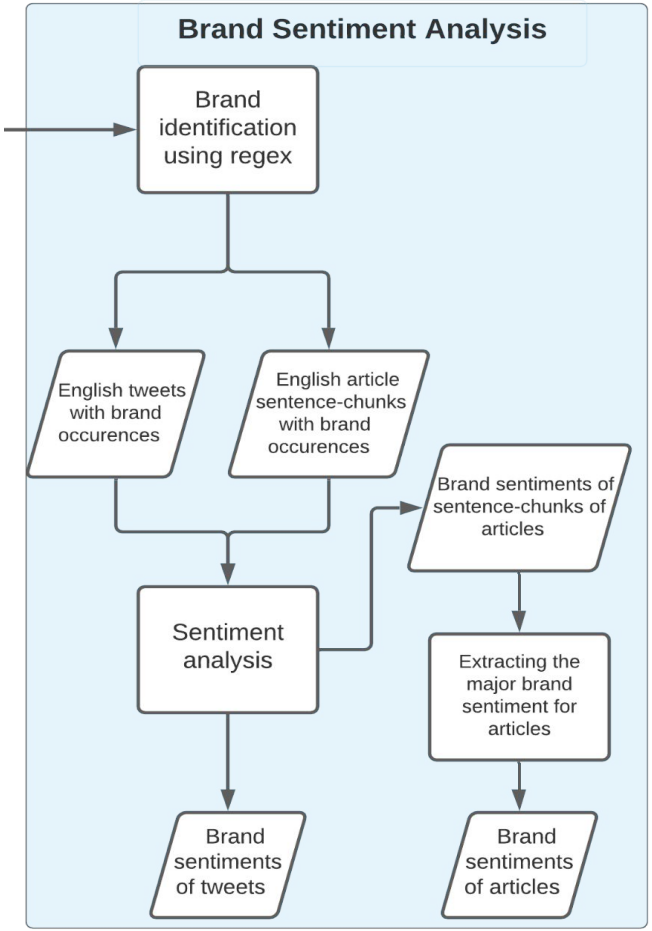


- Used context embeddings for identifying sentiment
- Final brand embeddings are calculated by averaging individual embeddings of all brand occurrences
- Final embeddings is maxpool of all tokens' embeddings from last layer



What embeddings to select?

- Final **brand embeddings** are calculated by averaging individual embeddings of all brand occurrences
- Final embeddings is maxpool of all tokens' embeddings from last layer



Fine-tuning & Validation

- **Validation Data:**
  - We hand labelled around 200 tweets for model selection and evaluation

- **External Dataset:**
- **Curated dataset using Amazon Mobile Reviews**
  - → Didn't perform well on validation data

- **Used BERT checkpoints trained on Hinglish and Indian domain text**
  - → Gave satisfactory performance on validation

Why Employ Abstractive Headline Generation ?

In the following example the abstractive headline is clearly more informative and requires lesser number of words:

Abstractive headline: "*motorola edge+ users report seeing purple patches oh the display*"

Extractive headline: *"motorola edge+ users on verizon have reportedly got a new firmware with display changes mentioned in the log*"

- Generates more catchy headlines.
- Unlike Extractive approaches requires lesser number of words to do so.
- Abstractive headlines can be more informative.

Consider the more creative headlines generated by our model:

"*risk of cancer in those who eat before 9 a.m.*"

"*playstation 5 review: sony's ps5 is an upgrade worth your money*"

Why Employ Abstractive Headline Generation ?

Here the model copies a sentence from middle of the article to generate the headline:

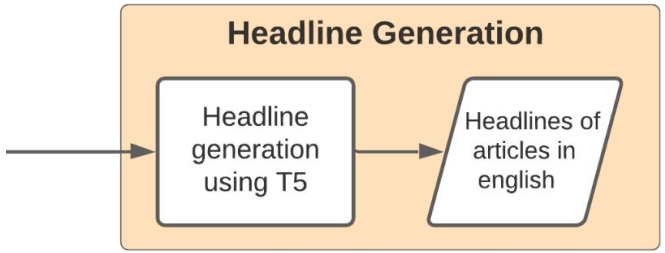
"*rs 64, 180 crore to boost healthcare infrastructure across the country aming ongoing covid*"

In the following example the model copies the first sentence from the article thus becomes an extractive model if required:

"*progressive care completes expansion, launches covid-19 rapid testing at new orlando location*"

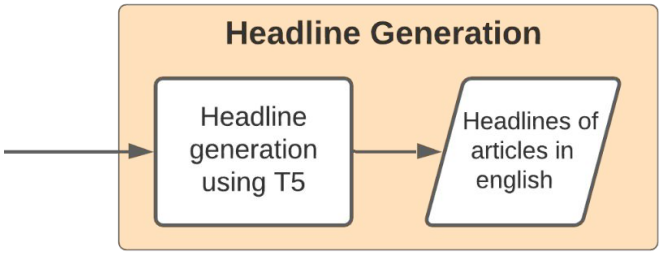
- In case only a few lines of the article capture its essence abstractive model restricts itself to these lines. These lines may be anywhere in the article.

Our Approach



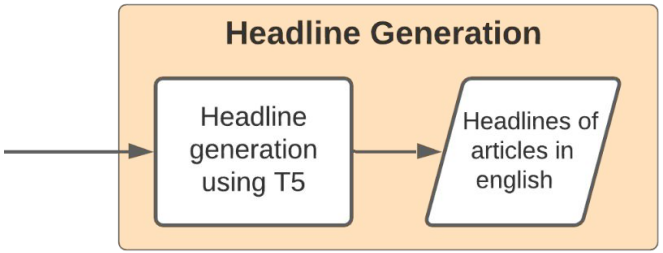
- Since we use pure English data, we need not worry about the article language.
- Hence we use T5-base model for headline generation. We fine tune it on the translated data.
- T5 being a model that can adapt to a variety of language generation and understanding tasks captures a wide range of domains.

Comparison of T5-base with other models



- We compare T5 with a few other models to get an idea of performance vs speed tradeoff.
- We compare it with PEGASUS and mT5(a multilingual version of T5).
- In Spite of being nearly three times bigger in size than T5-base, PEGASUS doesn't outperform it.
- Since the data is extremely noisy mT5 performs poorly on the untranslated data.

Comparison of T5-base with other models



- T5-base offers best performance, time and memory tradeoff.
- We fine tuned our model in 15 minutes on a Kaggle GPU. Inference took just over 2 minutes for 64 samples without batching.
- When compared to PEGASUS, T5's memory requirements are significantly lower.
- Thus our model is apt for instantaneous headline generation.

Thank You!