Clickbait Intensity & Style Analysis

IRE Major Project - Monsoon 2022

- Gopichand Kanumolu   (2021701039)
- Lokesh Madasu   (2021701042)

This presentation contains following sections

- Problem statement & Why it is important?
- Dataset statistics & Plots
- Existing approaches
- Model Results
- Proposed Approaches
- Comparison and Analysis of results

Problem Statement

- Clickbait is a headline or a piece of text that creates enthusiasm in the reader and forces the reader to click on the link in order to satisfy their information need.
- We encounter clickbaits in almost all social media platforms like
  - Online News Websites
  - Youtube Thumbnails
  - Not trusted websites

- Clickbaits are used to increase the revenue of the social media platforms by increasing their page views to show advertisements

Why it is important to control clickbait intensity?

Clickbaits wastes users time and loses user trust by failing to satisfy the information gap that is created by the headline or the piece of text.


It is important for the news agencies to control the Intensity of the clickbait by maintaining the right balance between the curiosity gap they create and the information they provide

Dataset Statistics

Field	Train	Test
Number of sentences in the dataset	17506	4341
Number of clickbait sentences in the dataset	4281	1104
Minimum Sentence Length	1	1
Maximum Sentence Length	267	476
Average Sentence Length	12	11
Minimum Intensity	0	0
Maximum Intensity	1	1
Average Intensity	0.327	0.331

Sentence Length Distribution
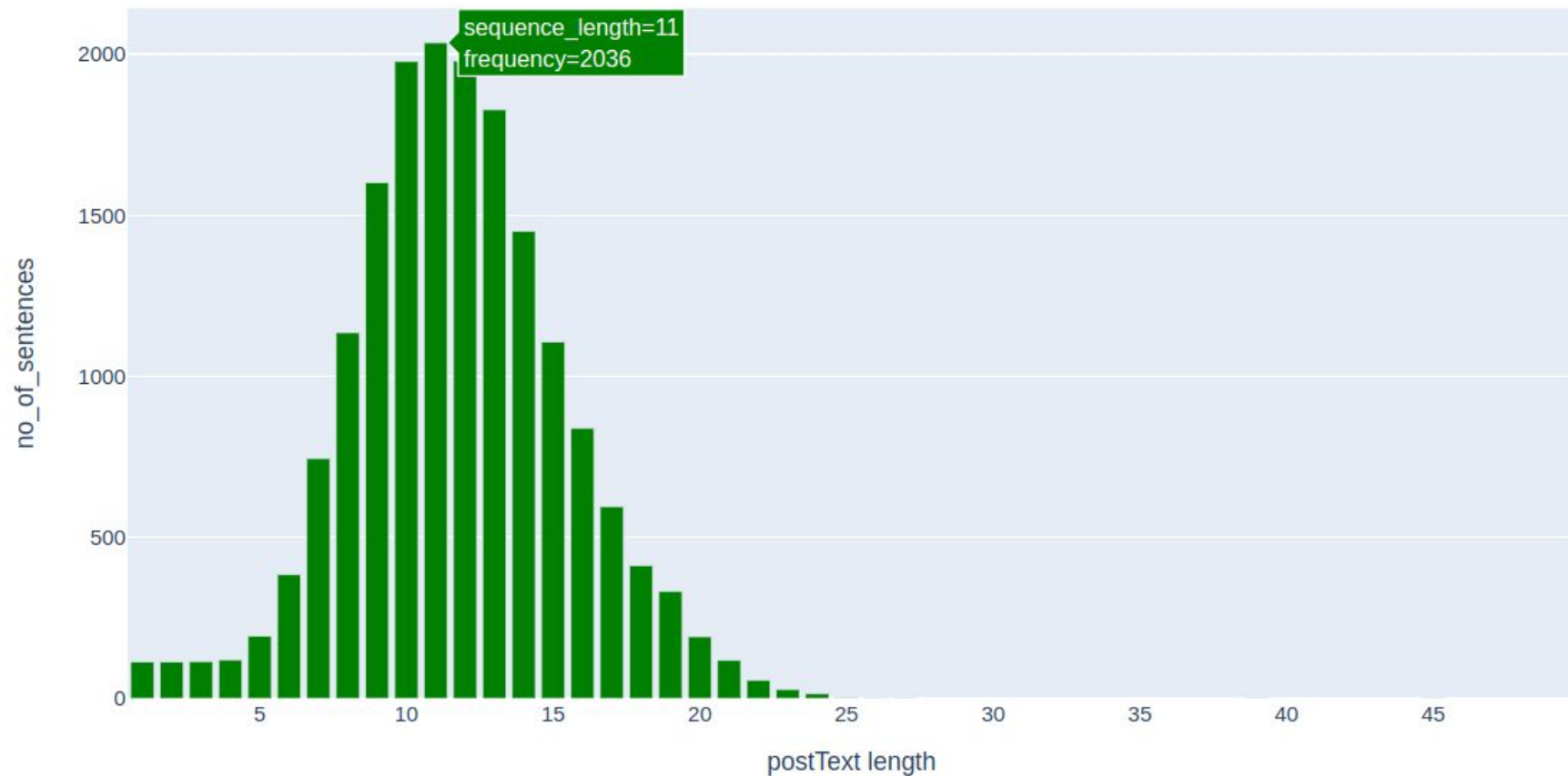
Existing Approaches

- The authors (Vijayasaradhi, Vasudeva Varma et al) of the research paper "**Predicting Click-bait Strength in Online Social Media**" pose the problem as a regression which predicts the clickbait intensity of the give text.
- Uses various regression algorithms like linear regression, ridge regression, random forest regression, gradient/adaboost regression
- With various contextual word embedding techniques BERT, RoBERTa, Universal Sentence Embeddings etc


Clickbait or Not?

How intensive the clickbait is?

Example: 0.7

Results

Embedding	Model	MSE	MedAe	F1-Score	Accuracy
BERT_Large	LR	0.02935	0.1096	0.64638	0.84459
BERT_Large	RR	0.0292	0.10893	0.64451	0.84459
BERT_Large	GBR	0.03157	0.1196	0.61249	0.83837
BERT_Large	RFR	0.04623	0.15415	0.32181	0.77842
BERT_Large	ABR	0.04325	0.16596	0.5614	0.82707
RoBERTa	LR	**0.02765**	0.10627	**0.67324**	**0.85474**
RoBERTa	RR	**0.02741**	0.10668	**0.67535**	**0.85635**
RoBERTa	GBR	0.02983	0.115	0.63494	0.84436
RoBERTa	RFR	0.04243	0.14348	0.41534	0.78902
RoBERTa	ABR	0.03934	0.15569	0.63169	0.84298
RoBERTa_Large	LR	0.02723	0.10461	0.68863	0.8605
RoBERTa_Large	RR	0.02695	0.1037	0.68705	0.86073
RoBERTa_Large	GBR	0.02987	0.11596	0.62916	0.84344
RoBERTa_Large	RFR	0.04127	0.13801	0.5018	0.80862
RoBERTa_Large	ABR	0.04043	0.16543	0.59579	0.83606
UniversalSentenceEmbeddings	LR	0.03108	0.11351	0.63563	0.84298
UniversalSentenceEmbeddings	RR	0.03074	0.11423	0.63276	0.84183
UniversalSentenceEmbeddings	GBR	0.03147	0.11891	0.62179	0.83675
UniversalSentenceEmbeddings	RFR	0.04483	0.15534	0.41237	0.78972
UniversalSentenceEmbeddings	ABR	0.04397	0.1671	0.56019	0.82984
BERT	LR	0.02881	0.1119	0.6649	0.85382
BERT	RR	0.02876	0.11202	0.66526	0.85405
BERT	GBR	0.03168	0.11822	0.60346	0.83606
BERT	RFR	0.0448	0.14852	0.32143	0.78095
BERT	ABR	0.04256	0.16794	0.56838	0.83191

This is what we have so far…..

IT'S GOOD
BUT NOT ENOUGH

Input Text ⟹ | Model | ⟹ Intensity
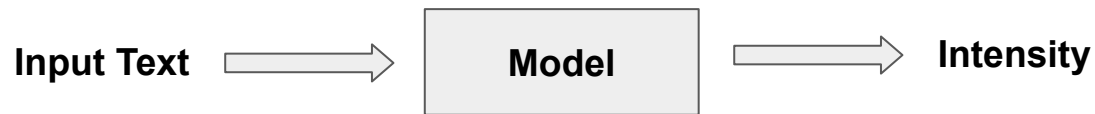
- Hillary Clinton's gut-wrenching day ****** 0.44

- How President Obama spent his final day in office ****** 0.43

- 11 reasons why women's football is better than men's ****** 0.56

- **Can we reduce the intensity of the text ?**

- **Can we change the style of the text to achieve that ?**

Naive approaches

- Randomly choose word(s) and mask them and check whether the intensity is changing or not
  - **Actual Text**: Hillary Clinton's gut-wrenching day
  - **New Text:** Hillary Clinton's day
  - **Problem**: **We miss the important information** (gut-wrenching)

- Apply Named Entity Recognition (NER) and mask the non-entity word(s) and predict the intensity
  - **Actual Text:** How President Obama spent his final day in office
  - **New Text :** President Obama spent final day office
  - **Problem: The sentence is not grammatically correct / not fluent**

Key Challenge: We need a system which generates a new text that retains the overall meaning of the actual text, which in turn reduces the intensity

A more better approach using Paraphrasing

- Paraphrasing a sentence means, you create a new sentence that expresses the same meaning using a different choice of words.

Input Text ⟹ **Paraphrase Model** ⟹ **New Text** ⟹ Clickbait Intensity Model ⟹ **Intensity**

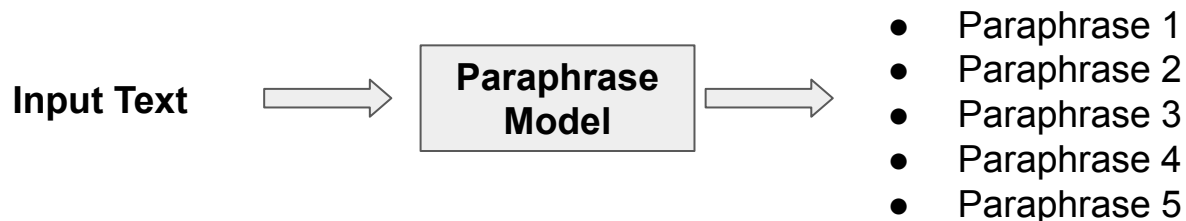| Input Text	Actual Intensity	Paraphrased Text	Predicted Intensity
Hillary Clinton's gut-wrenching day	0.733	Hillary Clinton had a difficult day.	0.362
How President Obama spent his final day in office	0.666	President Obama spent his last day in office.	0.301
Why you 100% need to start watching British crime drama 'Broadchurch'	1	British crime drama 'Broadchurch' is a must watch.	0.332

Issues with Paraphrasing

- The quality of the paraphrase
  - Not all the paraphrases retain the important information of the actual text
    - **Actual Text:** Donald Trump senior adviser Kellyanne Conway: I'm getting death threats
    - **Paraphrased Text:** I'm getting death threats.
    - **Result :** Increased Intensity (from 0.1 to 0.66)
  - Of the 4339 sentences in the test set
    - For 1493 sentences the intensity has reduced by paraphrasing the sentence
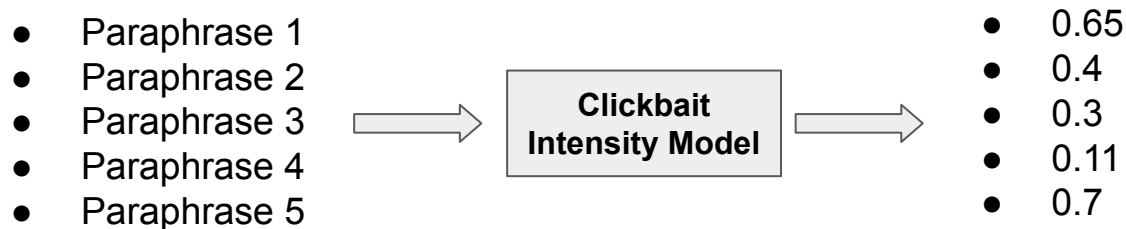   - For 2846 sentences the intensity has increased

Key Challenge: generate a text (paraphrase) such that it always reduces the intensity

Possible Solutions

Step-1: Generate different paraphrases for the actual text (say 5)

Input Text ⟹ **Paraphrase Model** ⟹
- Paraphrase 1
- Paraphrase 2
- Paraphrase 3
- Paraphrase 4
- Paraphrase 5

Step-2: Predict the Intensity of each paraphrased text

- Paraphrase 1
- Paraphrase 2
- Paraphrase 3
- Paraphrase 4
- Paraphrase 5

⟹ **Clickbait Intensity Model** ⟹

- 0.65
- 0.4
- 0.3
- 0.11
- 0.7

Step-3: Select the paraphrased sentences whose predicted intensity is less than the actual intensity

Say actual intensity is 0.6, then select the sentences
- Paraphrase 2
- Paraphrase 3
- Paraphrase 4

Step-4: Form a training set for each actual sentence

Actual Sentence (Source)	Paraphrased Sentence (Target)
actual sentence	Paraphrase 2
actual sentence	Paraphrase 3
actual sentence	Paraphrase 4

Step-5 : Train a T5 (Text-To-Text Transfer Transformer) model. After training, i.e during inference the model should generate a less clickbaity sentence when given a clickbaity sentence.

Thank You