scan.

Summarized Context Aware News

There are two kinds of news consumers.

wake up sid?

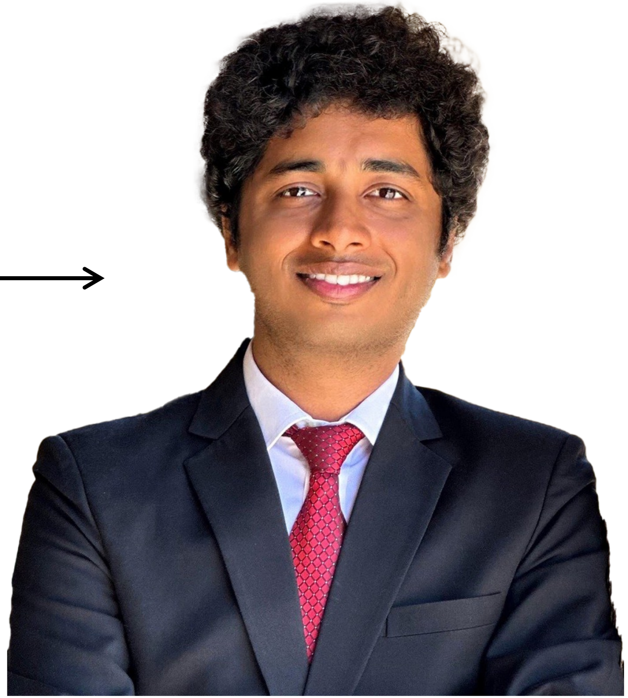
wake up sid?

Woke Sid.

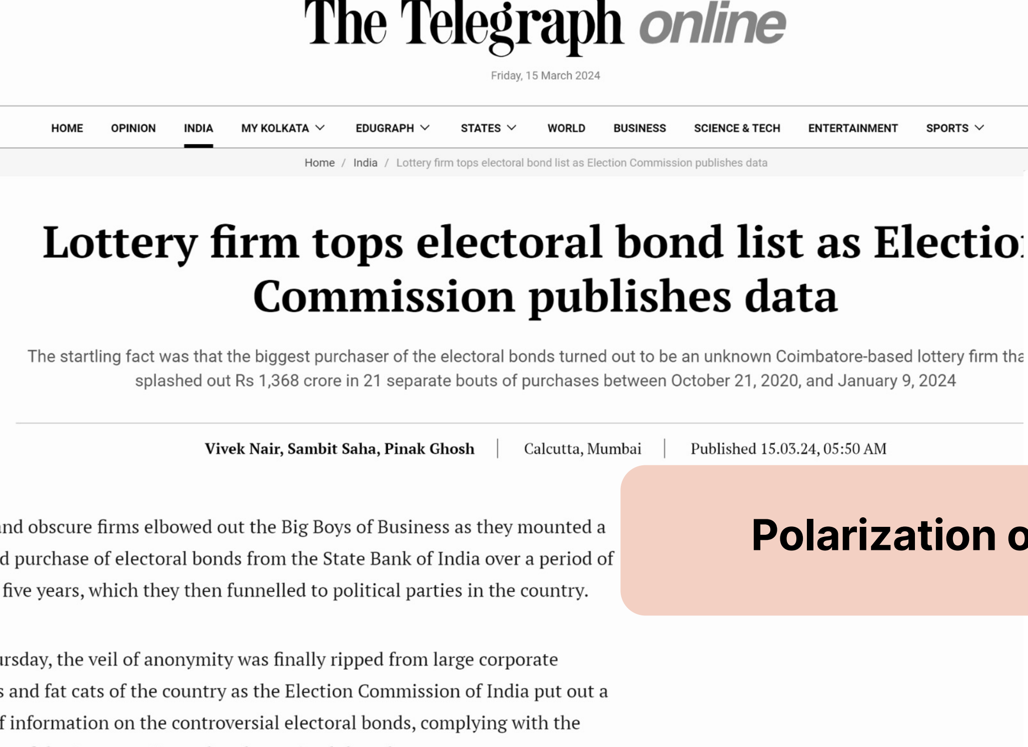
wake up sid?

?

Woke Sid.



Polarization of news



The Telegraph online

Friday, 15 March 2024

HOME
OPINION
INDIA
MY KOLKATA
EDUGRAPH
STATES
WORLD
BUSINESS
SCIENCE & TECH
ENTERTAINMENT
SPORTS
Home / India / Lottery firm tops electoral bond list as Election Commission publishes data

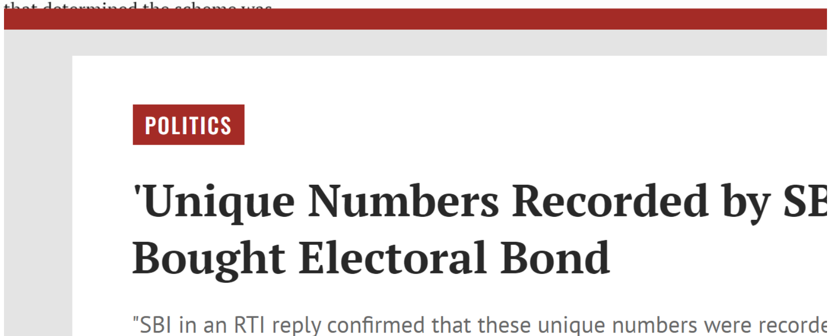
Lottery firm tops electoral bond list as Electio Commission publishes data

The startling fact was that the biggest purchaser of the electoral bonds turned out to be an unknown Coimbatore-based lottery firm tha splashed out Rs 1,368 crore in 21 separate bouts of purchases between October 21, 2020, and January 9, 2024

Vivek Nair, Sambit Saha, Pinak Ghosh | Calcutta, Mumbai | Published 15.03.24, 05:50 AM

nd obscure firms elbowed out the Big Boys of Business as they mounted a d purchase of electoral bonds from the State Bank of India over a period of five years, which they then funnelled to political parties in the country.

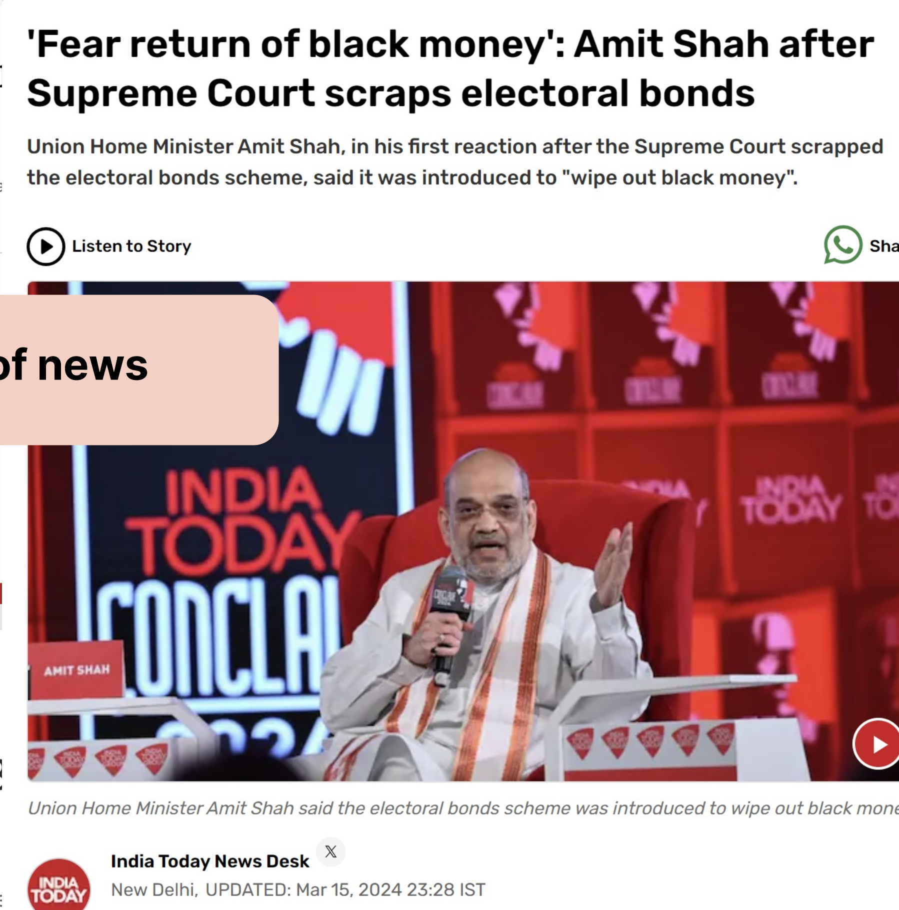
ursday, the veil of anonymity was finally ripped from large corporate s and fat cats of the country as the Election Commission of India put out a f information on the controversial electoral bonds, complying with the



POLITICS

'Unique Numbers Recorded by SE Bought Electoral Bond

"SBI in an RTI reply confirmed that these unique numbers were recorde



'Fear return of black money': Amit Shah after Supreme Court scraps electoral bonds

Union Home Minister Amit Shah, in his first reaction after the Supreme Court scrapped the electoral bonds scheme, said it was introduced to "wipe out black money".

Union Home Minister Amit Shah said the electoral bonds scheme was introduced to wipe out black mone

India Today News Desk

New Delhi, UPDATED: Mar 15, 2024 23:28 IST

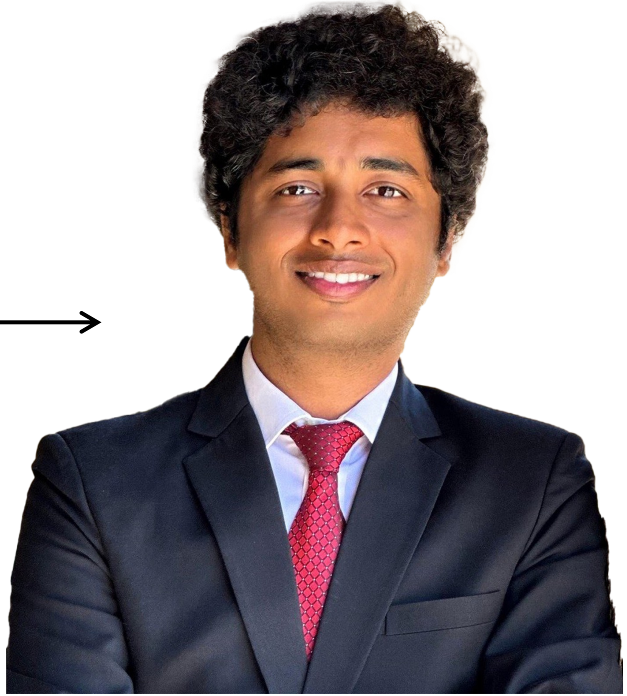
Lack of context

How can we combat **bias** and **misrepresentation** of news coupled with lack of **awareness**, to make information accessible?

wake up sid?

scan.

Woke Sid.



Potential Applications

- Can be used in collaboration with platforms like Google News & InShorts
- Implemented as a Browser extension for ease of use
- Media Bias and Objectivity Detection

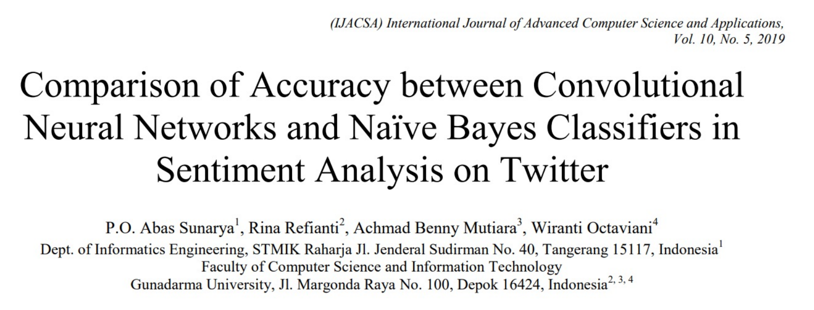
Potential Impact

- Regulation of media and news consumption
- Informed decision-making
- Increase in number of news readers

Literature Review

Sentiment Analysis

- Comparison of the classifier models CNN and NBC for sentiment analysis on English-language tweets concerning the Turkey Crisis of 2018.
- CNN outperformed NBC on various performance metrics.
- CNN: A Deep Learning algorithm.
- NBC: A simple probabilistic classifier based on the Bayes' theorem.



(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 10, No. 5, 2019

Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter

P.O. Abas Sunarya[1], Rina Refianti[2], Achmad Benny Mutiara[3], Wiranti Octaviani[4]
Dept. of Informatics Engineering, STMIK Raharja Jl. Jenderal Sudirman No. 40, Tangerang 15117, Indonesia[1]
Faculty of Computer Science and Information Technology
Gunadarma University, Jl. Margonda Raya No. 100, Depok 16424, Indonesia[2, 3, 4]

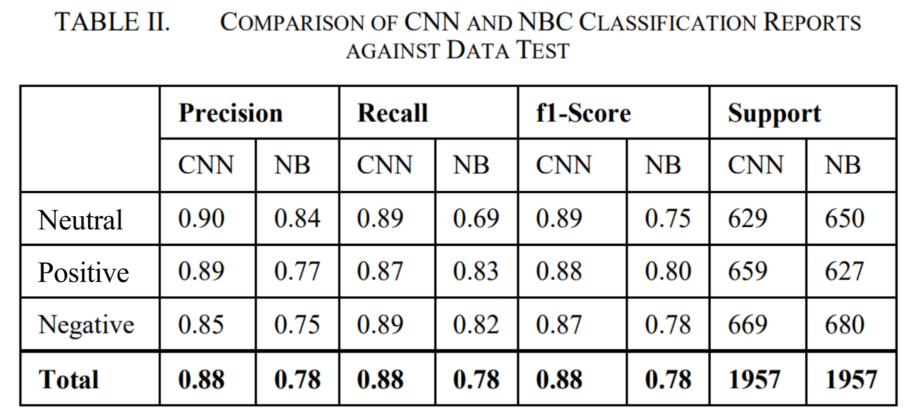
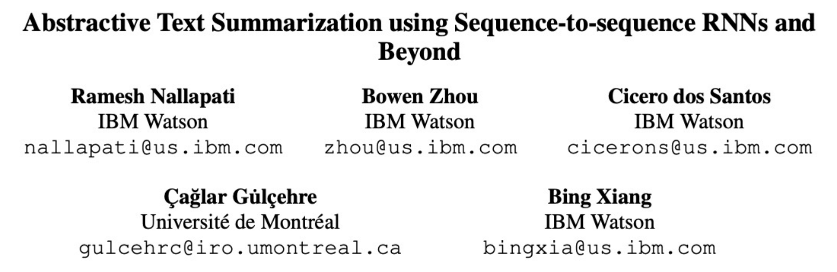


TABLE II. COMPARISON OF CNN AND NBC CLASSIFICATION REPORTS AGAINST DATA TEST

	Precision		**Recall**		**f1-Score**		**Support**	
	CNN	NB	CNN	NB	CNN	NB	CNN	NB
Neutral	0.90	0.84	0.89	0.69	0.89	0.75	629	650
Positive	0.89	0.77	0.87	0.83	0.88	0.80	659	627
Negative	0.85	0.75	0.89	0.82	0.87	0.78	669	680
Total	**0.88**	**0.78**	**0.88**	**0.78**	**0.88**	**0.78**	**1957**	**1957**

Abstractive Summarization

- Feature-rich encoder captures important keywords better using linguistic features
- Switching generator-pointer handles rare/unseen words effectively
- Hierarchical attention captures document structure for long inputs
- Authors primarily used data from CNN and Daily Mail



Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

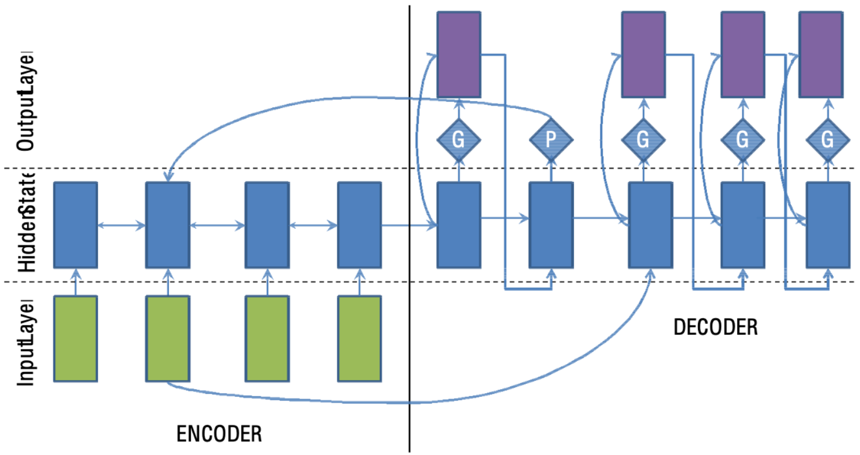
Ramesh Nallapati
IBM Watson
nallapati@us.ibm.com

Bowen Zhou
IBM Watson
zhou@us.ibm.com

Cicero dos Santos
IBM Watson
cicerons@us.ibm.com

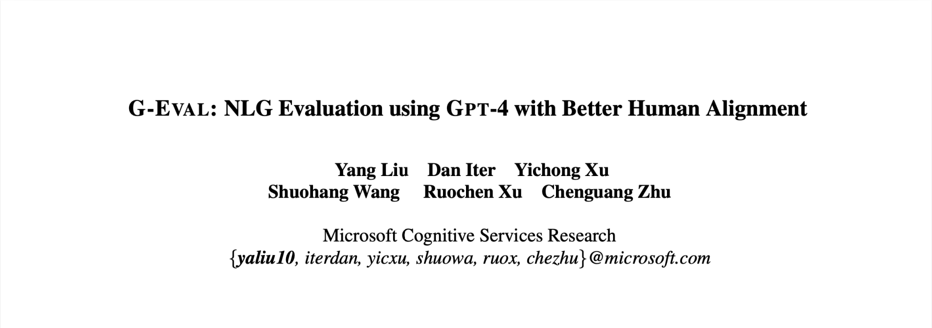
Çağlar Gülçehre
Université de Montréal
gulcehrc@iro.umontreal.ca

Bing Xiang
IBM Watson
bingxia@us.ibm.com



Evaluation Metrics

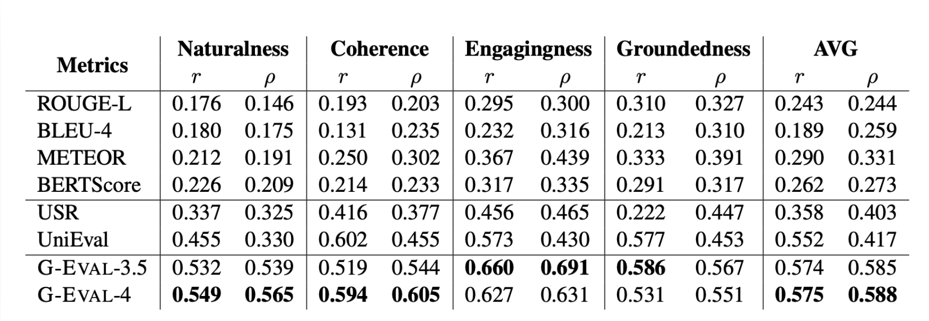
- ROUGE Score – Traditional Scoring method
- Alternative – G-Eval by OpenAI
- Uses LLMs and scoresheet-based prompt – Slightly Biased towards AI generated content
- Survey people at Plaksha to get summary scores and opinions
- Draw comparisons with Human written summaries for better evaluation



G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment

Yang Liu   Dan Iter   Yichong Xu
Shuohang Wang   Ruochen Xu   Chenguang Zhu

Microsoft Cognitive Services Research
{***yaliu10***, *iterdan, yicxu, shuowa, ruox, chezhu*}*@microsoft.com*



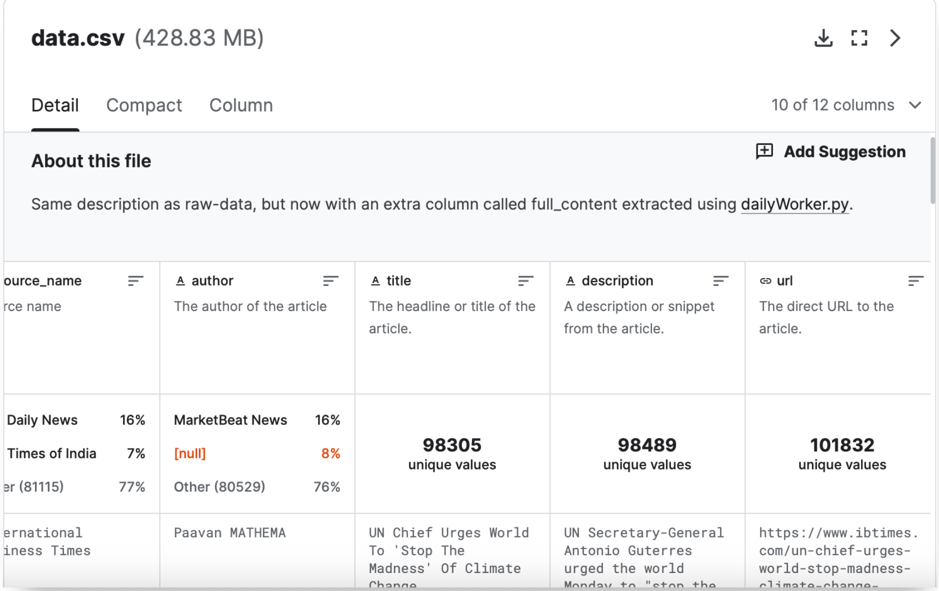
Metrics	**Naturalness**		**Coherence**		**Engagingness**		**Groundedness**		**AVG**	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
ROUGE-L	0.176	0.146	0.193	0.203	0.295	0.300	0.310	0.327	0.243	0.244
BLEU-4	0.180	0.175	0.131	0.235	0.232	0.316	0.213	0.310	0.189	0.259
METEOR	0.212	0.191	0.250	0.302	0.367	0.439	0.333	0.391	0.290	0.331
BERTScore	0.226	0.209	0.214	0.233	0.317	0.335	0.291	0.317	0.262	0.273
USR	0.337	0.325	0.416	0.377	0.456	0.465	0.222	0.447	0.358	0.403
UniEval	0.455	0.330	0.602	0.455	0.573	0.430	0.577	0.453	0.552	0.417
G-EVAL-3.5	0.532	0.539	0.519	0.544	**0.660**	**0.691**	**0.586**	0.567	0.574	0.585
G-EVAL-4	**0.549**	**0.565**	**0.594**	**0.605**	0.627	0.631	0.531	0.551	**0.575**	**0.588**

G-Eval Correlation Score Comparison

Features Preprocessing

Global News Dataset

- Open Source - Kaggle
- 98,305 Articles – Scraped by authors
- Sentiment Labels – Positive, Negative and Neutral
- Relevant Columns – Title, Description, Article, Title Sentiment
- No Summaries provided



data.csv (428.83 MB)

About this file

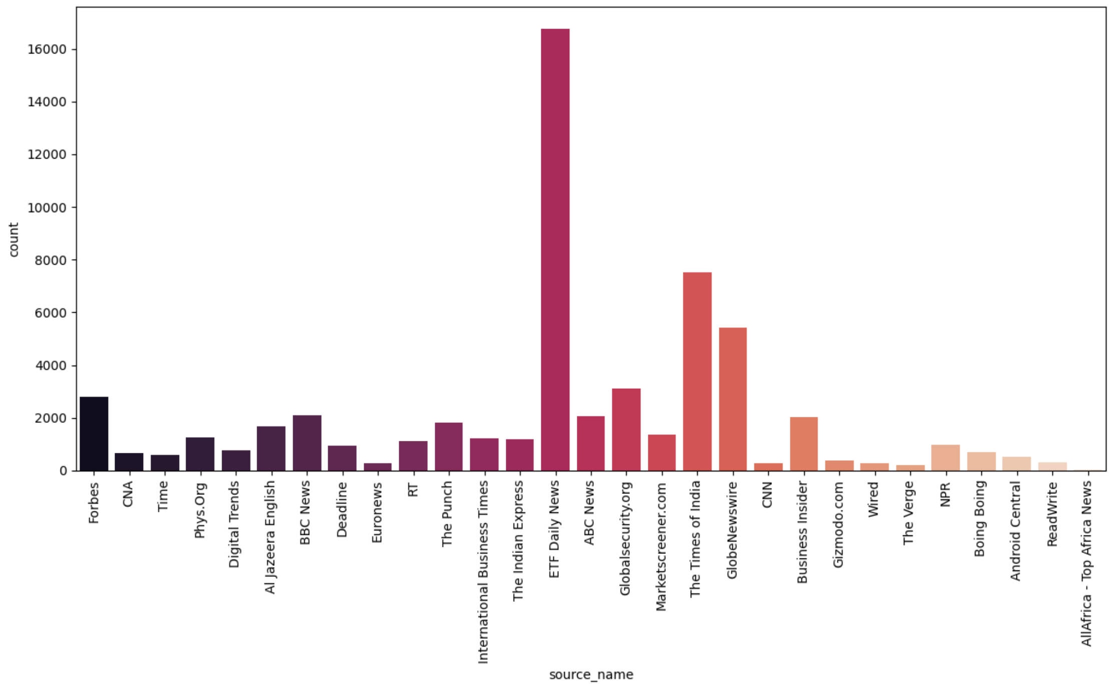
Same description as raw-data, but now with an extra column called full_content extracted using dailyWorker.py.

ource_name	author	title	description	url
rce name	The author of the article	The headline or title of the article.	A description or snippet from the article.	The direct URL to the article.
Daily News 16% Times of India 7% er (81115) 77%	MarketBeat News 16% [null] 8% Other (80529) 76%	98305 unique values	98489 unique values	101832 unique values
ernational iness Times	Paavan MATHEMA	UN Chief Urges World To 'Stop The Madness' Of Climate Change	UN Secretary-General Antonio Guterres urged the world Monday to "stop the	https://www.ibtimes.com/un-chief-urges-world-stop-madness-climate-change-

News Sources

Insights

- ETF Daily News - largest news source
- Most articles are financial news
- Indian news outlets – Times of India, Wired, Indian Express etc.

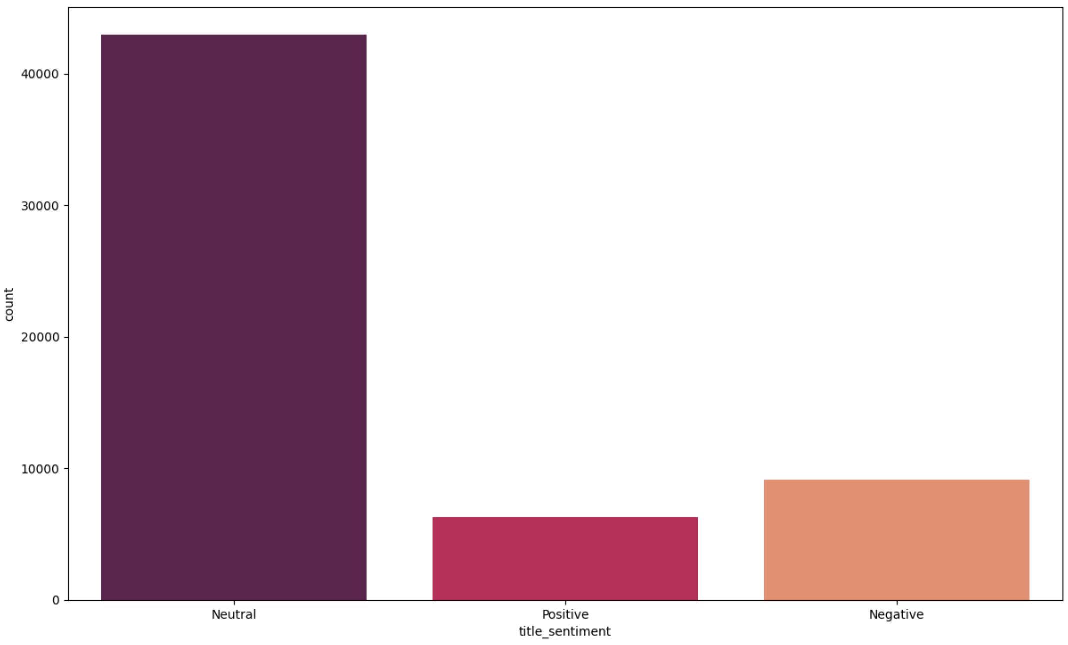


Title Sentiment

Insights

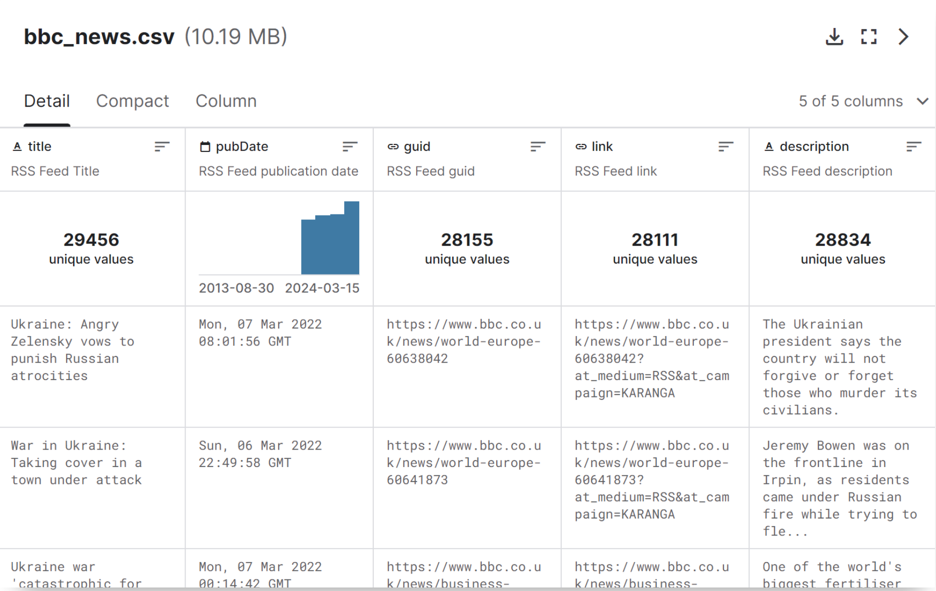
- Dataset is heavily biased towards neutral
- ETF Daily News – 90% Neutral

  Digital Trends – 51% Positive



BBC News Dataset

- Open Source - Kaggle
- ~30,000 Articles – Updated Daily
- 5 Columns in total
- Relevant Columns – Title, link, description
- No Summaries or Sentiment



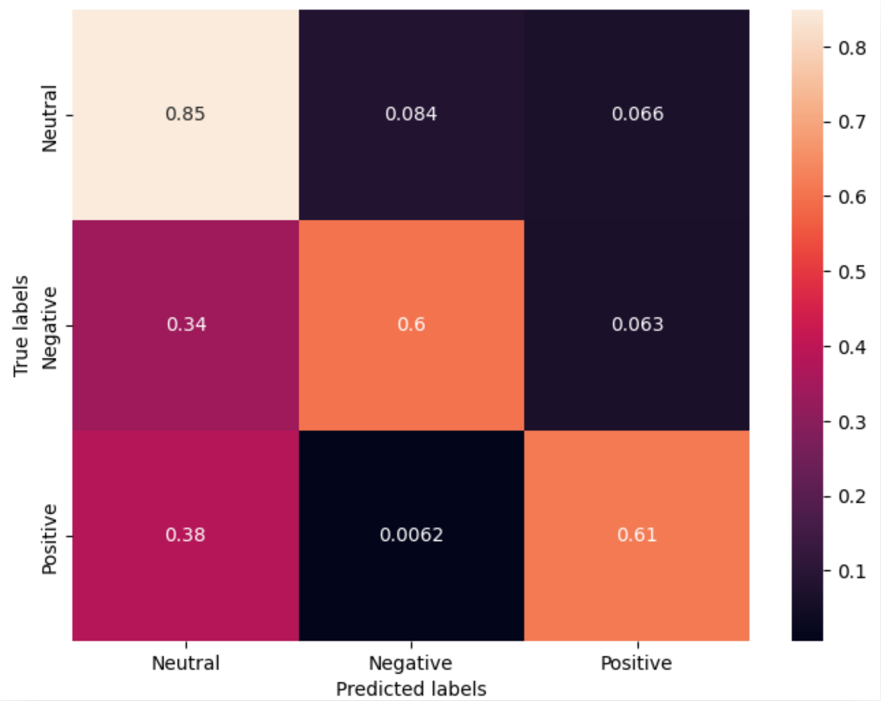
bbc_news.csv (10.19 MB)

title	pubDate	guid	link	description
RSS Feed Title	RSS Feed publication date	RSS Feed guid	RSS Feed link	RSS Feed description
29456 unique values	2013-08-30 2024-03-15	28155 unique values	28111 unique values	28834 unique values
Ukraine: Angry Zelensky vows to punish Russian atrocities	Mon, 07 Mar 2022 08:01:56 GMT	https://www.bbc.co.uk/news/world-europe-60638042	https://www.bbc.co.uk/news/world-europe-60638042?at_medium=RSS&at_campaign=KARANGA	The Ukrainian president says the country will not forgive or forget those who murder its civilians.
War in Ukraine: Taking cover in a town under attack	Sun, 06 Mar 2022 22:49:58 GMT	https://www.bbc.co.uk/news/world-europe-60641873	https://www.bbc.co.uk/news/world-europe-60641873?at_medium=RSS&at_campaign=KARANGA	Jeremy Bowen was on the frontline in Irpin, as residents came under Russian fire while trying to fle...
Ukraine war 'catastrophic for	Mon, 07 Mar 2022 00:14:42 GMT	https://www.bbc.co.uk/news/business-	https://www.bbc.co.uk/news/business-	One of the world's biggest fertiliser

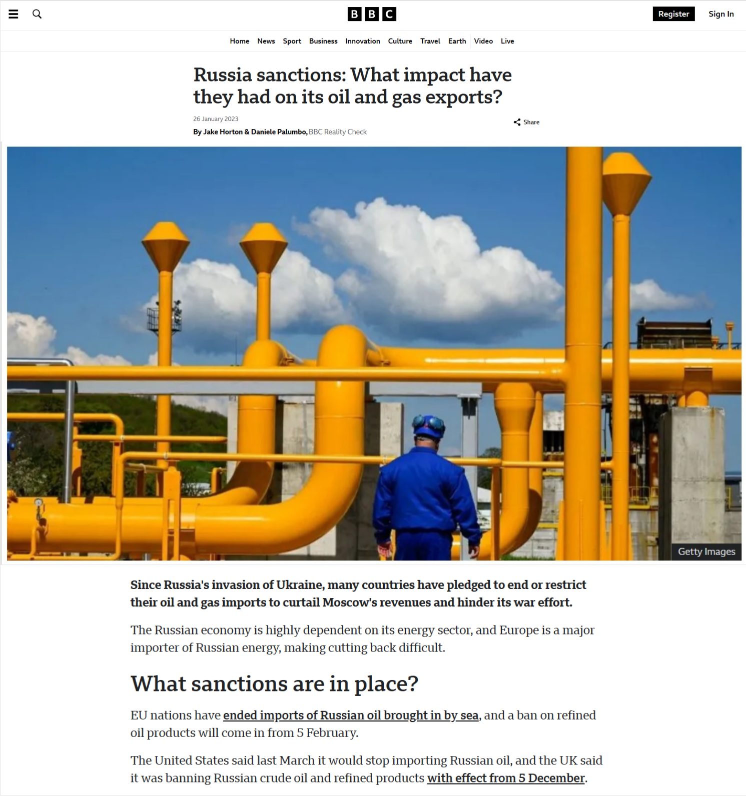
Labelling unlabeled data

- Self Training Classifier with

  Random Forest as Base Model
- 58K labeled samples → 80-20 split
- 78K unlabeled samples
- Accuracy Score → **~73%**
- Not using for further training.



Web Scraping


BBC
Register
Sign In
Home News Sport Business Innovation Culture Travel Earth Video Live
Share
Getty Images

Russia sanctions: What impact have they had on its oil and gas exports?

26 January 2023

By Jake Horton & Daniele Palumbo, BBC Reality Check

Since Russia's invasion of Ukraine, many countries have pledged to end or restrict their oil and gas imports to curtail Moscow's revenues and hinder its war effort.

The Russian economy is highly dependent on its energy sector, and Europe is a major importer of Russian energy, making cutting back difficult.

What sanctions are in place?

EU nations have ended imports of Russian oil brought in by sea, and a ban on refined oil products will come in from 5 February.

The United States said last March it would stop importing Russian oil, and the UK said it was banning Russian crude oil and refined products with effect from 5 December.

What our scraper does:

- Fetches article content using NewsAPI endpoint with the provided URL.
- Scrapes article content from the webpage using BeautifulSoup with specified HTML tags.
- Reads CSV file with news article URLs into a pandas data frame.
- Iterates through the data frame rows, extracts article content with BeautifulSoup, and updates the data frame.

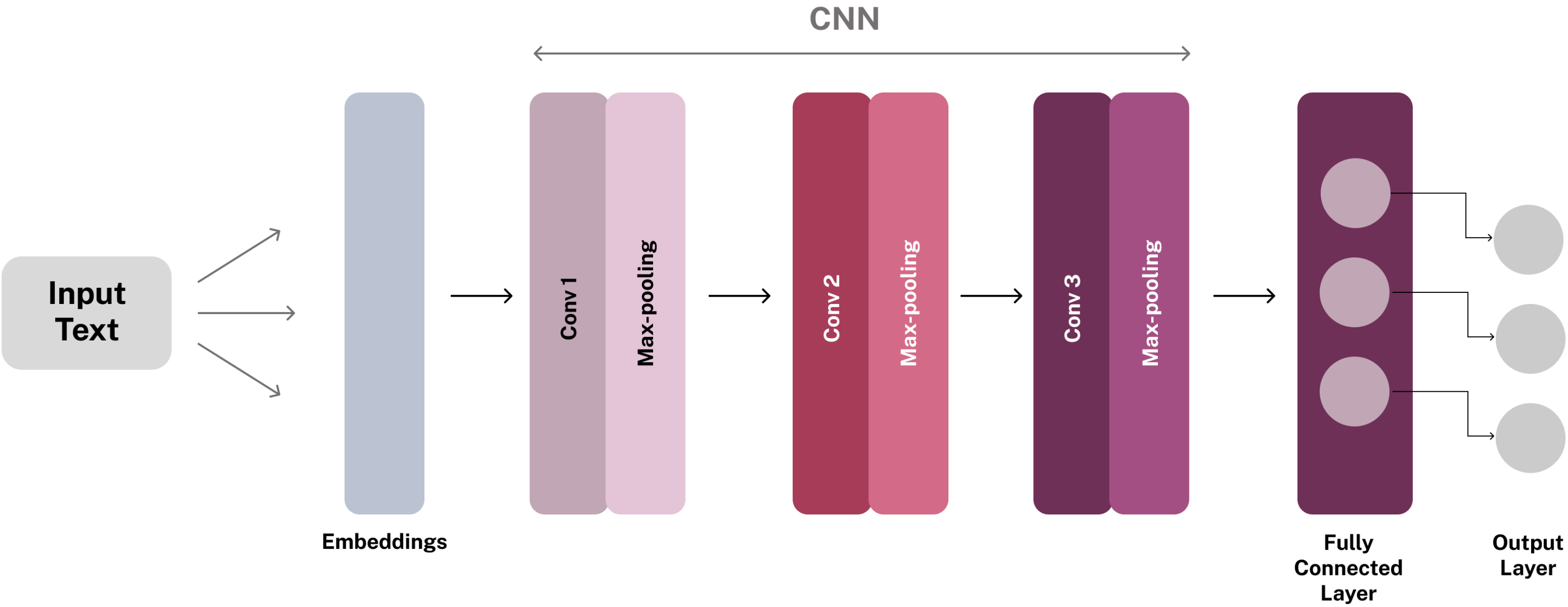
Gemini Summarization

Advantages

- Human-like linguistic capabilities for summarization
- Processing rate - 12 articles per minute
- Temperature – 0
- Top_k – 1
- Top_p – 1
- Max Output – 400 tokens

ML Methodology

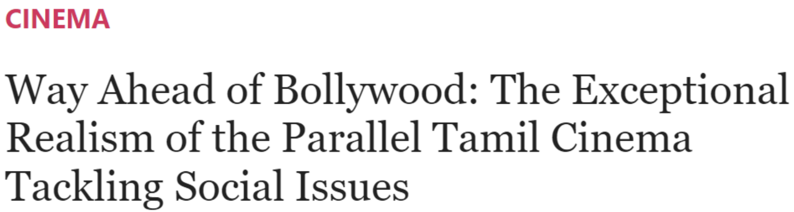
Sentiment Analysis Model Architecture



Model Configuration

- 3 Convolutional Layers
- Kernel Sizes - 3, 4, and 5 (for each conv layer)
- Vocabulary Size - Number of unique tokens in the vocabulary
- Embedding Dimension – 100
- Dropout - 0.5
- Number of Classes - 3 (Positive, Negative, Neutral)
- Number of Channels - 100 (output channels in each conv layer)

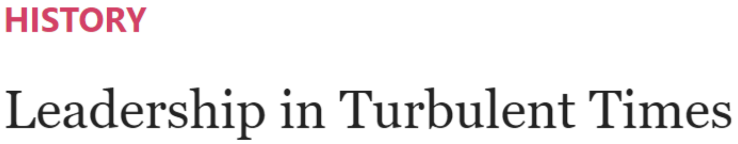
Model Outputs



CINEMA

Way Ahead of Bollywood: The Exceptional Realism of the Parallel Tamil Cinema Tackling Social Issues

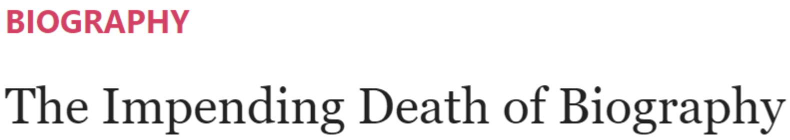
Predicted: Positive, we believe: Positive



HISTORY

Leadership in Turbulent Times

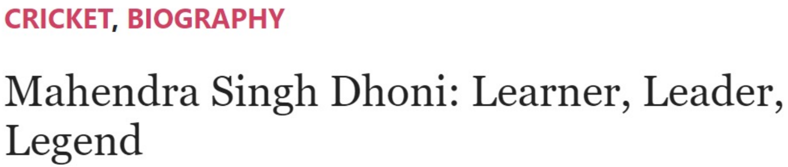
Predicted: Neutral, we believe: Neutral



BIOGRAPHY

The Impending Death of Biography

Predicted: Negative, we believe: Negative

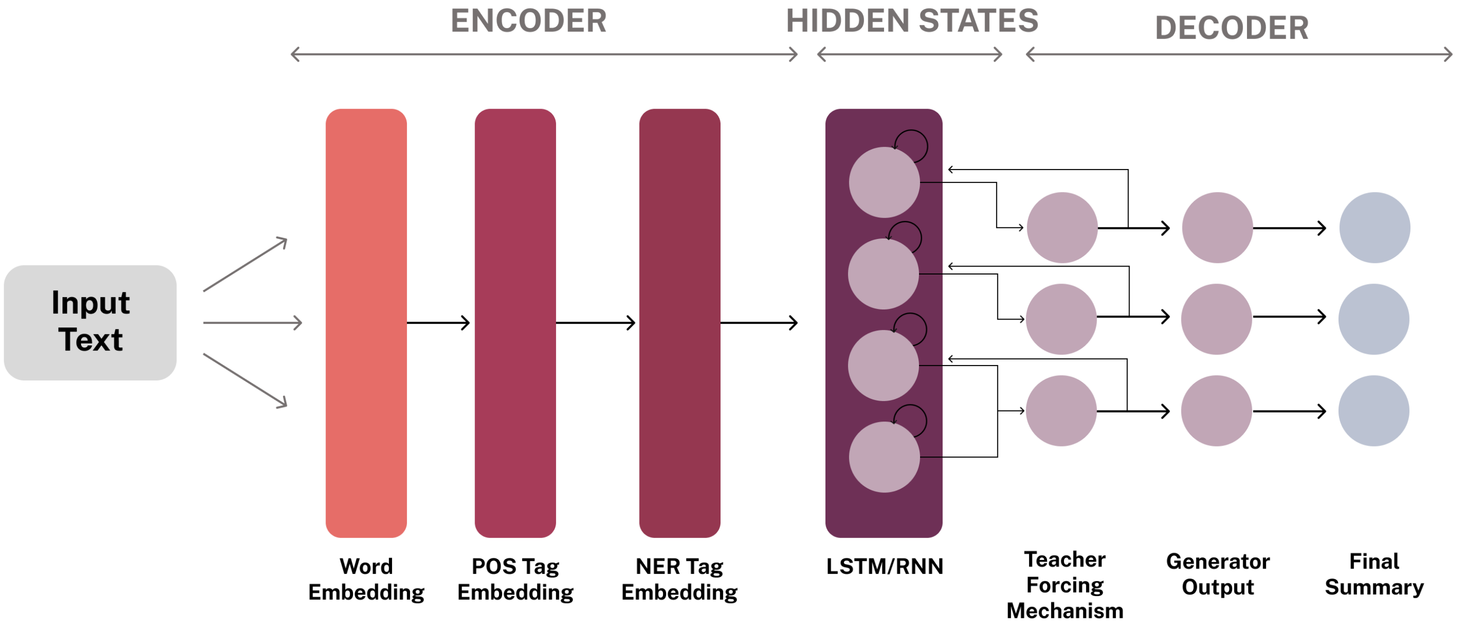


CRICKET, BIOGRAPHY

Mahendra Singh Dhoni: Learner, Leader, Legend

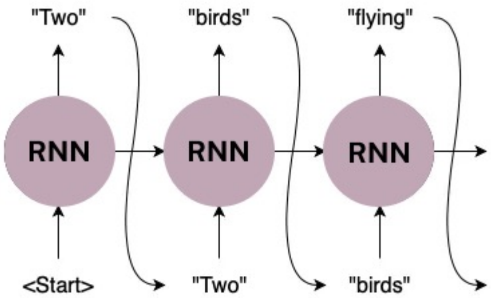
Predicted: Neutral, we believe: Positive

Summarization Model Architecture

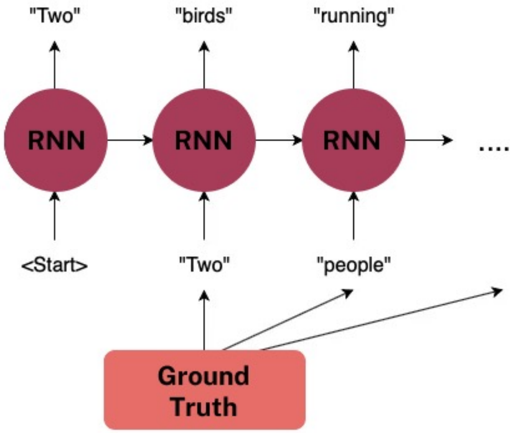


Model Configuration

- Hidden Layer Dimension – 256
- Embedding Dimension – 300
- Vocab Size – 75,000 words
- Max Summary length – 100 words
- Teacher Forcing Ratio – 0.5
- Dropout – 0.5



Without Teacher Forcing



With Teacher Forcing

CINEMA

Way Ahead of Bollywood: The Exceptional Realism of the Parallel Tamil Cinema Tackling Social Issues

Siddharth Follow

Oct 21, 2022 · 5 mins read

Share this

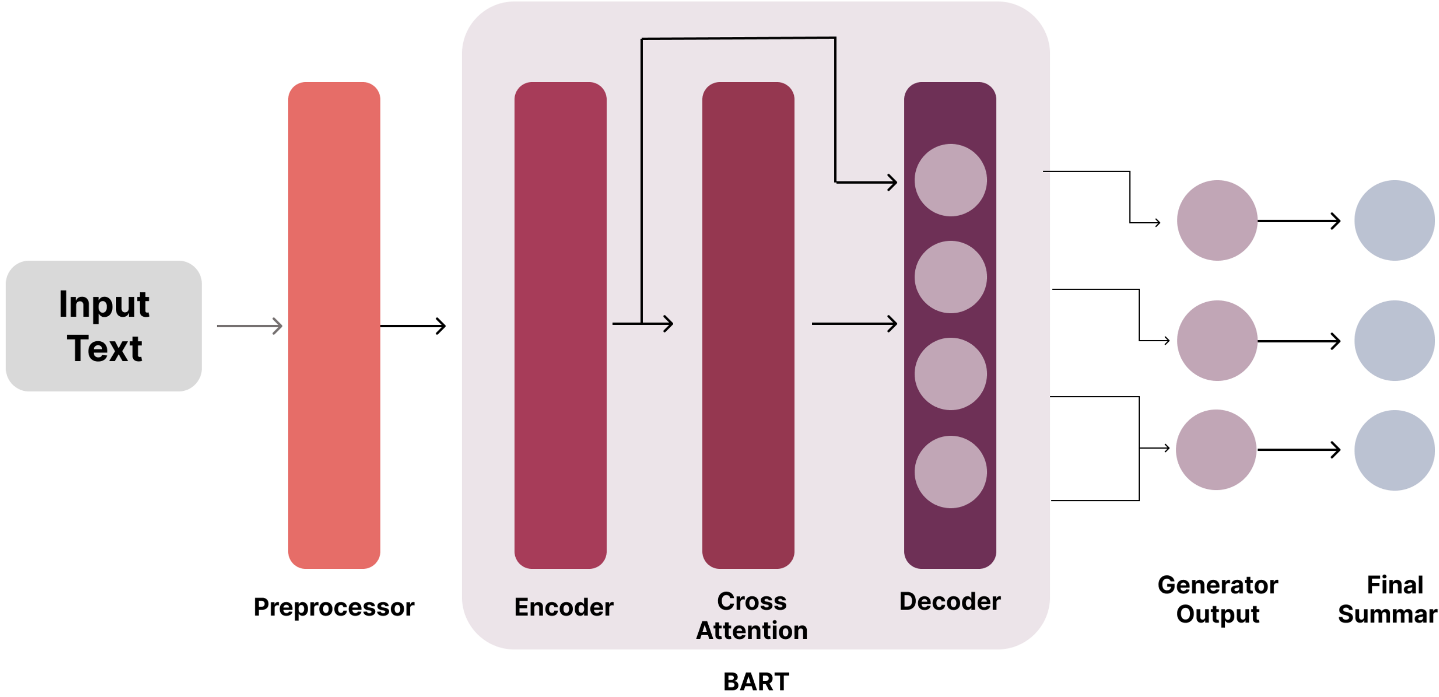
Since last December, four of the five highest-grossing Indian movies have been from South India namely *KGF: Chapter 2*, *RRR*, *Ponniyin Selvan: Part I*, and *Pushpa: The Rise - Part 1*. These hugely entertaining movies have enthralled audiences not just in the South but rather Indians across the world. The ripple effect originating from the success of these movies is not just limited to the profits they have made for their producers or the cultural shift that they have ushered in the visual depiction of Hindu epics

Example Output

```
<unk> the ag ceasefire on been to the kong on been to a for the sell of the south india film india the the the the the of world audience of the effect the
```

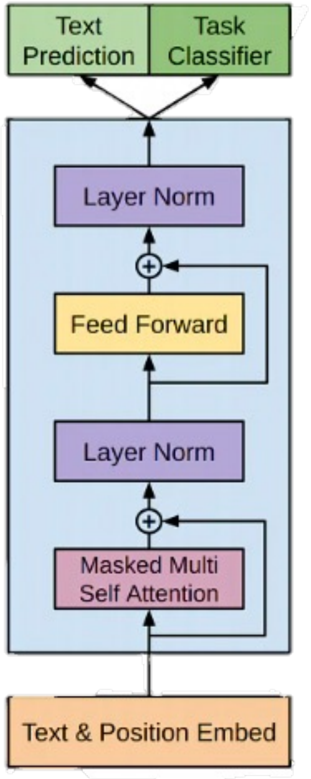
okay, toh ab?

Summarization Model Architecture



Model Configuration

- Frozen Encoder Weights
- Max Encoding Sequence Length - 512
- Max Decoder Sequence Length – 128
- Max Generation Length – 40
- Learning Rate – 5e-5
- Gradient Clipping Threshold – 1.0
- Excluded bias, gamma, and beta terms



Loshchilov & Hutter, 2017

Example Output

five of the top performing indian movies are based in the south including kGFs Chapter 2 and rRRs two films have been compared to other southbased bollywood films these movies have challenged the box office dominance and inspired a new appreciation of cinema in the indian context

Challenges Faced

- Models were extremely computationally expensive to train.
- Retaining context over long text sequences proved to be difficult with limited resources
- Tensor Dimensionality issues due to library compatibility.
- Coordinating and integrating the multiple parallel decoder components during training and inference is complex.

Performance Metrics

Evaluating Sentiment Classifier

- Trained for 10 Epochs
- Accuracy: 83%

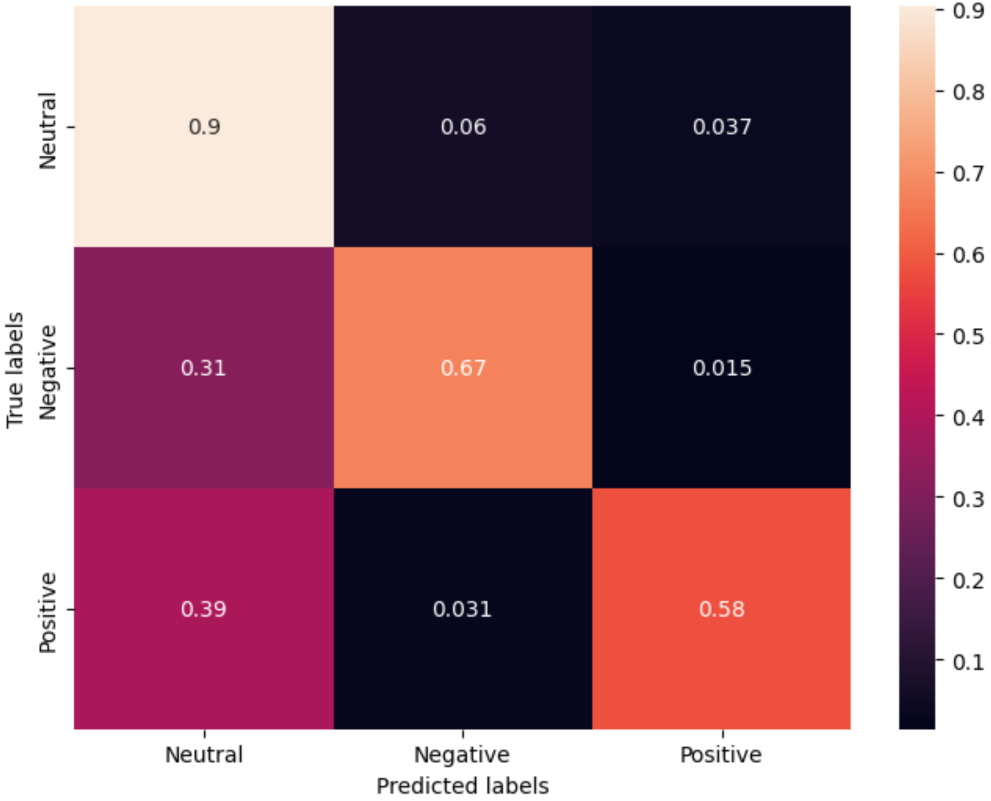
  Precision: 83%

  Recall: 83%

  F1 Score: 83%

^Our model is a Kapil Dev fan

Neutral: 8,586 | Negative: 1,827 | Positive: 1,259



ROUGE Scores

Recall-Oriented Understudy for Gisting Evaluation

ROUGE was designed to evaluate the quality of machine-generated summaries by comparing them to reference summaries provided by humans.

ROUGE-N

- Measures overlap of **n-grams** between the candidate and the target text.
- Used to evaluate the grammatical correctness and fluency of generated text.

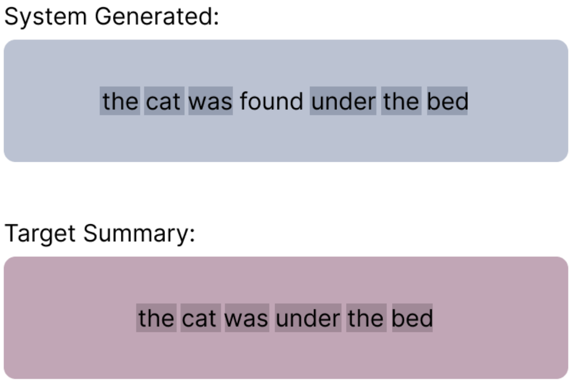
ROUGE-L

- Measures the **longest common subsequence** (LCS) between the candidate & reference text.
- Used to evaluate the semantic similarity and content coverage of generated text.

Lin, 2004

Recall & Precision

how much of the **reference summary** is the **system summary** capturing?



System Generated:

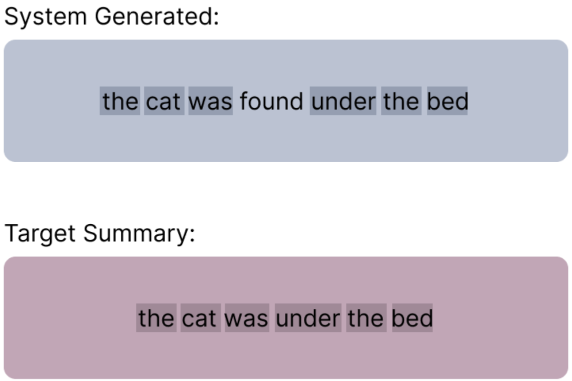
the cat was found under the bed

Target Summary:

the cat was under the bed

$$\text{recall} = \frac{\text{no. of overlapping words}}{\text{total words in } \textbf{target} \text{ summary}} = 1.0$$

how much of the **system summary** was in fact **relevant**?



System Generated:

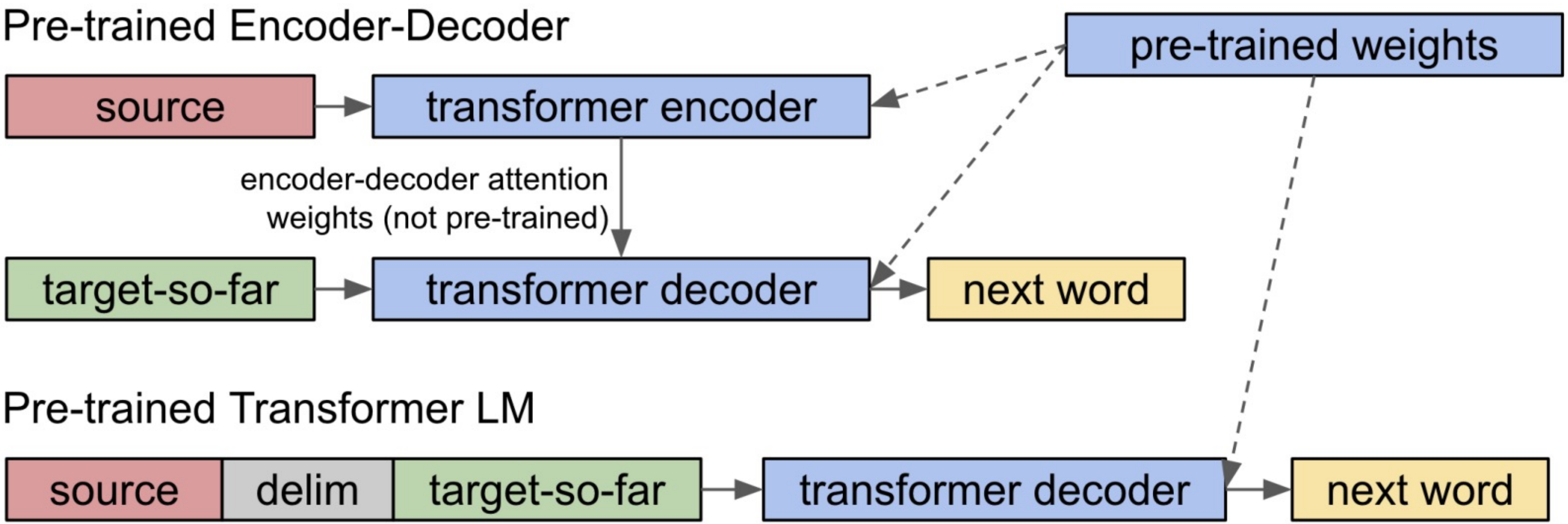
the cat was found under the bed

Target Summary:

the cat was under the bed

$$\text{precision} = \frac{\text{no. of overlapping words}}{\text{total words in } \textbf{output} \text{ summary}} = 0.85$$

BERT by
Khandelwal et al., 2019



Pre-trained Encoder-Decoder

pre-trained weights

source

transformer encoder

encoder-decoder attention weights (not pre-trained)

target-so-far

transformer decoder

next word

Pre-trained Transformer LM

source

delim

target-so-far

transformer decoder

next word

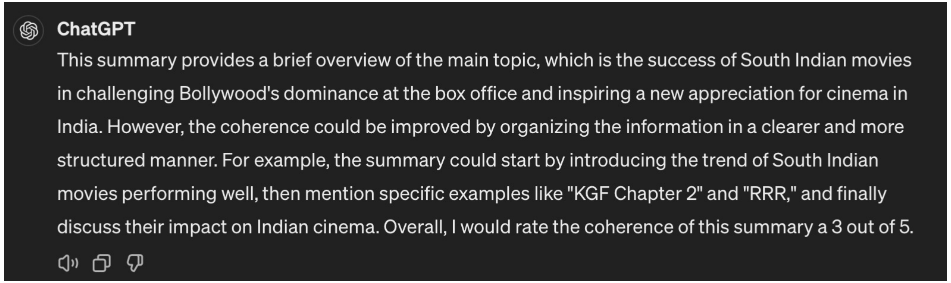
ROUGE Model Comparison

F1-Score	**LSTM - 1**	**IBM**	**SCAN**	**Stanford**
ROUGE-1	0.33	0.24	0.48	0.37
ROUGE-2	0.03	0.09	0.23	0.16
ROUGE-L	0.2	0.2	0.38	0.34

Khandelwal, U., Clark, K., Jurafsky, D., & Kaiser, L. (2019).

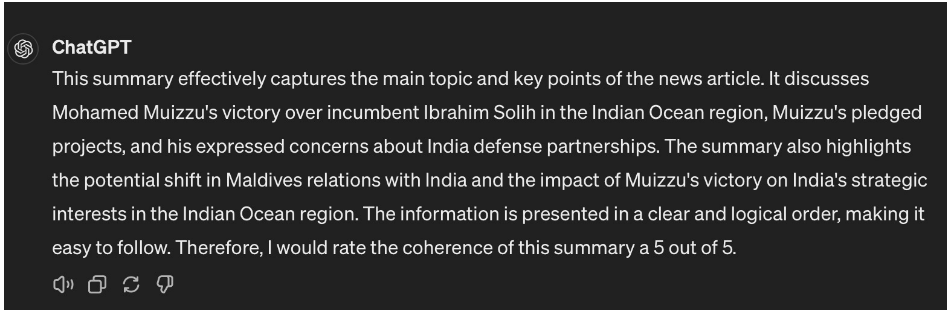
G-Eval Metric

- Coherence - 4/5
- Engagingness – 3/5
- Factual Consistency – 4/5
- Overall – 3.5/5



ChatGPT

This summary provides a brief overview of the main topic, which is the success of South Indian movies in challenging Bollywood's dominance at the box office and inspiring a new appreciation for cinema in India. However, the coherence could be improved by organizing the information in a clearer and more structured manner. For example, the summary could start by introducing the trend of South Indian movies performing well, then mention specific examples like "KGF Chapter 2" and "RRR," and finally discuss their impact on Indian cinema. Overall, I would rate the coherence of this summary a 3 out of 5.



ChatGPT

This summary effectively captures the main topic and key points of the news article. It discusses Mohamed Muizzu's victory over incumbent Ibrahim Solih in the Indian Ocean region, Muizzu's pledged projects, and his expressed concerns about India defense partnerships. The summary also highlights the potential shift in Maldives relations with India and the impact of Muizzu's victory on India's strategic interests in the Indian Ocean region. The information is presented in a clear and logical order, making it easy to follow. Therefore, I would rate the coherence of this summary a 5 out of 5.

Bibliography

- Lin, C. (2004, July 1). ROUGE: a package for automatic evaluation of summaries. ACL Anthology. https://aclanthology.org/W04-1013
- Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
- Khandelwal, U., Clark, K., Jurafsky, D., & Kaiser, L. (2019). Sample efficient text summarization using a single pre-trained transformer. arXiv preprint arXiv:1905.08836.
- https://www.freecodecamp.org/news/what-is-rouge-and-how-it-works-for-evaluation-of-summaries-e059fb8ac840/

thank you.