NEWS TONE ANALYSIS

NAME	: BHAVYA PANDYA
ID	: a1785085
Course	: Master Data Science

SEMESTER 1 - 2021

UNIVERSITY OF ADELAIDE

NEWS TONE ANALYSIS

Sentiment Analysis using Text Classification Techniques

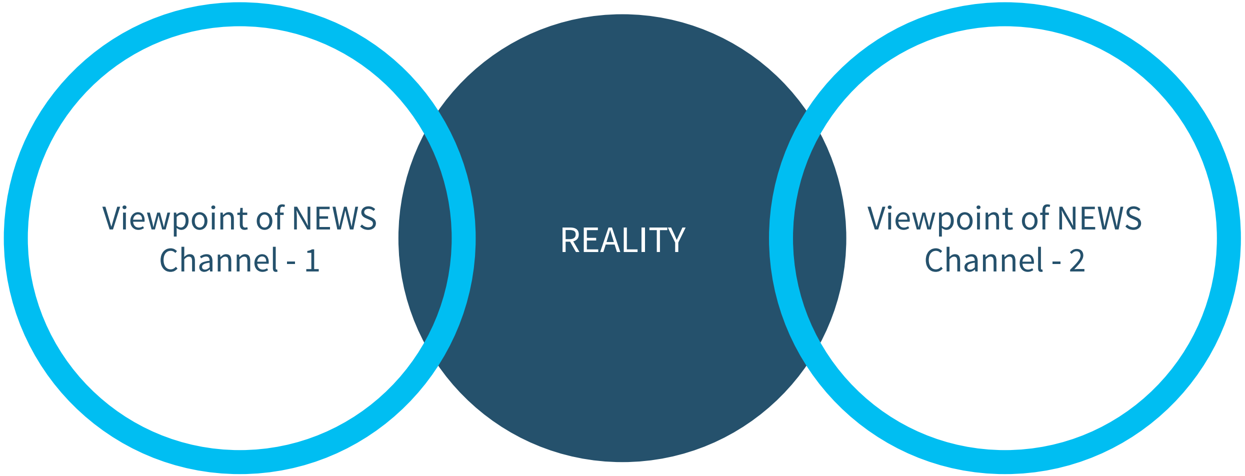
RESEARCH PROBLEM OVERVIEW

Text based sentiment analysis of NEWS data, by comparing different models to review the sentiment scores and determine if some aspects of the NEWS are more polarised than others.

RESEARCH OUTLINE

- Are headlines more polarised than the article?
- Are NEWS articles in certain categories more polarised?
- Does polarity of NEWS data depend on events?
- Compare sentiment score for different ML and deep learning models?

MOTIVATION

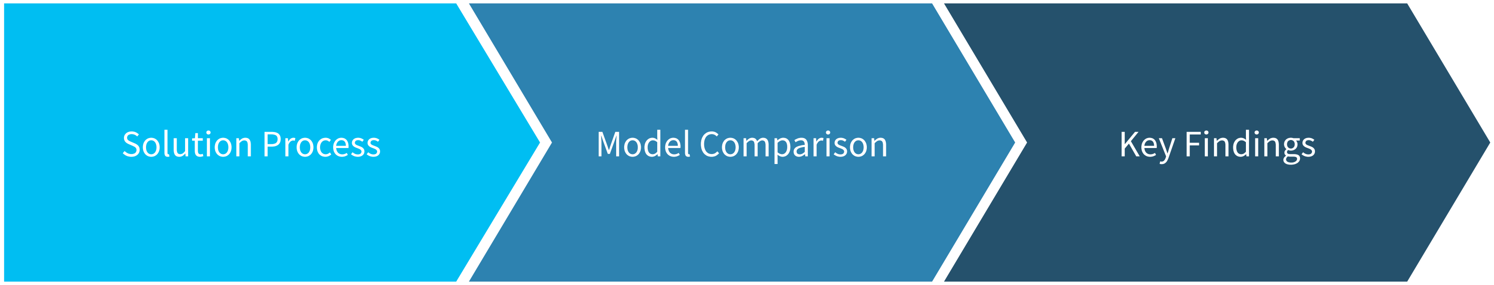


Viewpoint of NEWS Channel - 1

REALITY

Viewpoint of NEWS Channel - 2

OUR PROCESS IS EASY

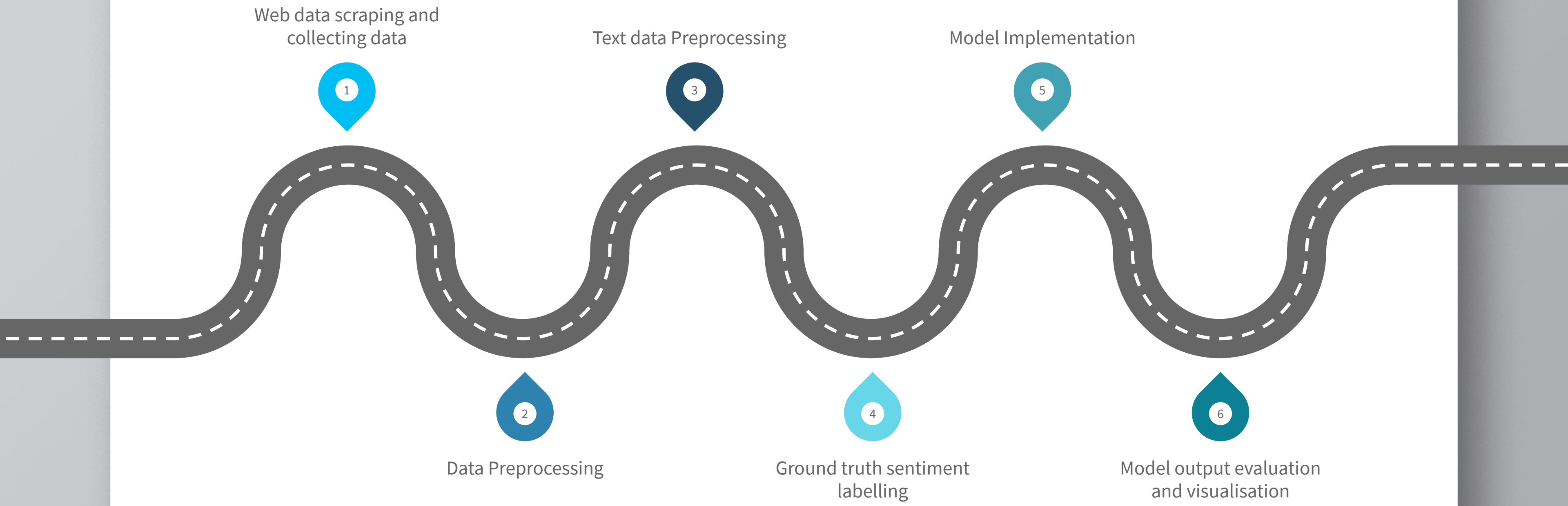


Solution Process

Model Comparison

Key Findings

SOLUTION PROCESS



1 Web data scraping and collecting data

2 Data Preprocessing

3 Text data Preprocessing

4 Ground truth sentiment labelling

5 Model Implementation

6 Model output evaluation and visualisation

WEB SCRAPING AND DATA COLLECTION

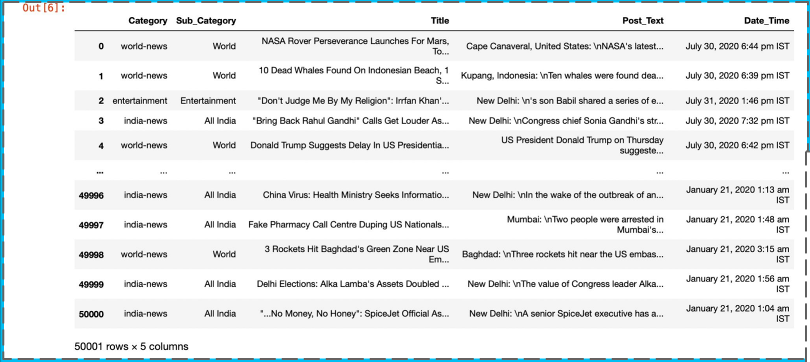
Language used : Python

Module used : Scrapy

Websites :

- NEWS Publication 1 - (50k)
- NEWS Publication 2 - (70k)

DATASET - NEWS PUBLICATIONS

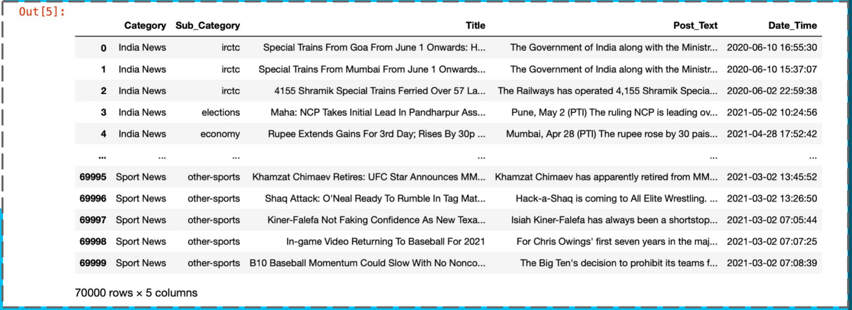


Out[6]:

	Category	Sub_Category	Title	Post_Text	Date_Time
0	world-news	World	NASA Rover Perseverance Launches For Mars, To...	Cape Canaveral, United States: \nNASA's latest...	July 30, 2020 6:44 pm IST
1	world-news	World	10 Dead Whales Found On Indonesian Beach, 1 S...	Kupang, Indonesia: \nTen whales were found dea...	July 30, 2020 6:39 pm IST
2	entertainment	Entertainment	"Don't Judge Me By My Religion": Irrfan Khan'...	New Delhi: \n's son Babil shared a series of e...	July 31, 2020 1:46 pm IST
3	india-news	All India	"Bring Back Rahul Gandhi" Calls Get Louder As...	New Delhi: \nCongress chief Sonia Gandhi's str...	July 30, 2020 7:32 pm IST
4	world-news	World	Donald Trump Suggests Delay In US Presidentia...	US President Donald Trump on Thursday suggeste...	July 30, 2020 6:42 pm IST
...	...	...	...	...	...
49996	india-news	All India	China Virus: Health Ministry Seeks Informatio...	New Delhi: \nIn the wake of the outbreak of an...	January 21, 2020 1:13 am IST
49997	india-news	All India	Fake Pharmacy Call Centre Duping US Nationals...	Mumbai: \nTwo people were arrested in Mumbai's...	January 21, 2020 1:48 am IST
49998	world-news	World	3 Rockets Hit Baghdad's Green Zone Near US Em...	Baghdad: \nThree rockets hit near the US embas...	January 21, 2020 3:15 am IST
49999	india-news	All India	Delhi Elections: Alka Lamba's Assets Doubled ...	New Delhi: \nThe value of Congress leader Alka...	January 21, 2020 1:56 am IST
50000	india-news	All India	"...No Money, No Honey": SpiceJet Official As...	New Delhi: \nA senior SpiceJet executive has a...	January 21, 2020 1:04 am IST

50001 rows × 5 columns

PUBLICATION - 1



Out[5]:

	Category	Sub_Category	Title	Post_Text	Date_Time
0	India News	irctc	Special Trains From Goa From June 1 Onwards: H...	The Government of India along with the Ministr...	2020-06-10 16:55:30
1	India News	irctc	Special Trains From Mumbai From June 1 Onwards...	The Government of India along with the Ministr...	2020-06-10 15:37:07
2	India News	irctc	4155 Shramik Special Trains Ferried Over 57 La...	The Railways has operated 4,155 Shramik Specia...	2020-06-02 22:59:38
3	India News	elections	Maha: NCP Takes Initial Lead In Pandharpur Ass...	Pune, May 2 (PTI) The ruling NCP is leading ov...	2021-05-02 10:24:56
4	India News	economy	Rupee Extends Gains For 3rd Day; Rises By 30p ...	Mumbai, Apr 28 (PTI) The rupee rose by 30 pais...	2021-04-28 17:52:42
...	...	...	...	...	...
69995	Sport News	other-sports	Khamzat Chimaev Retires: UFC Star Announces MM...	Khamzat Chimaev has apparently retired from MM...	2021-03-02 13:45:52
69996	Sport News	other-sports	Shaq Attack: O'Neal Ready To Rumble In Tag Mat...	Hack-a-Shaq is coming to All Elite Wrestling. ...	2021-03-02 13:26:50
69997	Sport News	other-sports	Kiner-Falefa Not Faking Confidence As New Texa...	Isiah Kiner-Falefa has always been a shortstop...	2021-03-02 07:05:44
69998	Sport News	other-sports	In-game Video Returning To Baseball For 2021	For Chris Owings' first seven years in the maj...	2021-03-02 07:07:25
69999	Sport News	other-sports	B10 Baseball Momentum Could Slow With No Nonco...	The Big Ten's decision to prohibit its teams f...	2021-03-02 07:08:39

70000 rows × 5 columns

PUBLICATION - 2

DATASET FEATURES

Features Columns :

- Category
- Sub - Category
- **Article Headline**
- **Article Text**
- Publishing Date

TEXT DATA PREPROCESSING

SYMBOLS and PUNCTUATION REMOVAL

@,#,url links, punctuations using RegEX

Convert numbers to words

For standardised text data throughout

Convert to lower case

Convert all alphabets of words to lowercase

StopWord Removal

Remove most regularly occurring words like a, the, an etc.

Stemming and Lemmatization

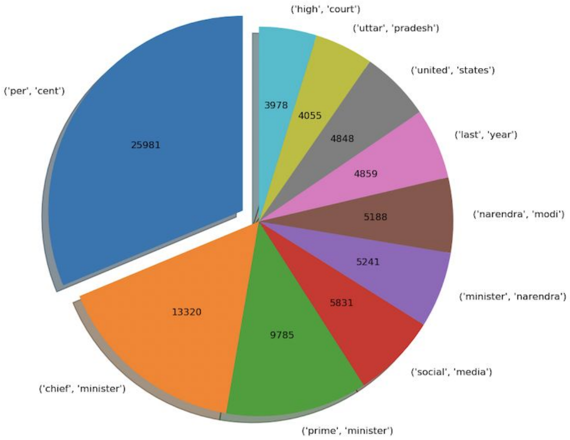
Stemming chops up end of words while lemmatization is to convert text into base form of word

N-GRAM Model

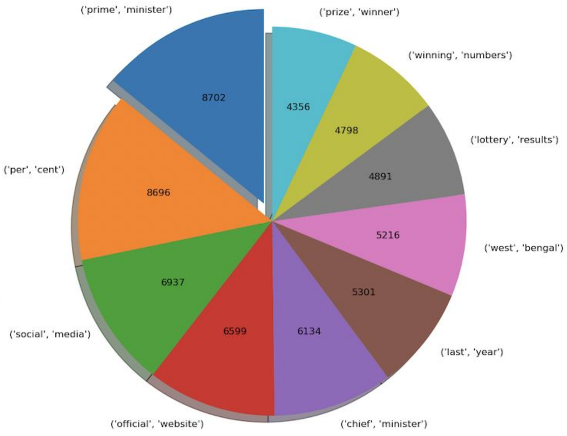
Convert the text data in token engram pair of 2 tokens (bi-grams) and 3 tokens(tri-grams)

BI - GRAMS COMPARISON

PUBLICATION - 1



PUBLICATION - 2



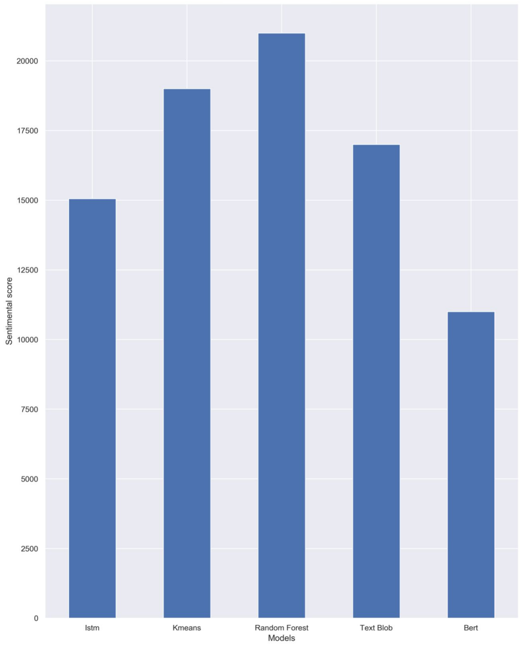
SENTIMENT ANALYSIS

Ground Truth Sentiment label using **Textblob**

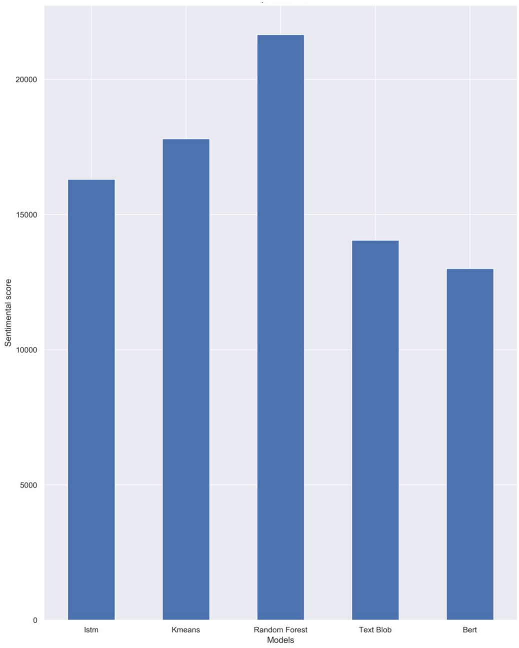
Models used for comparison :

- K - means clustering
- Random forest
- LSTM
- Bert

SENTIMENT SCORE (vs) ALL MODELS

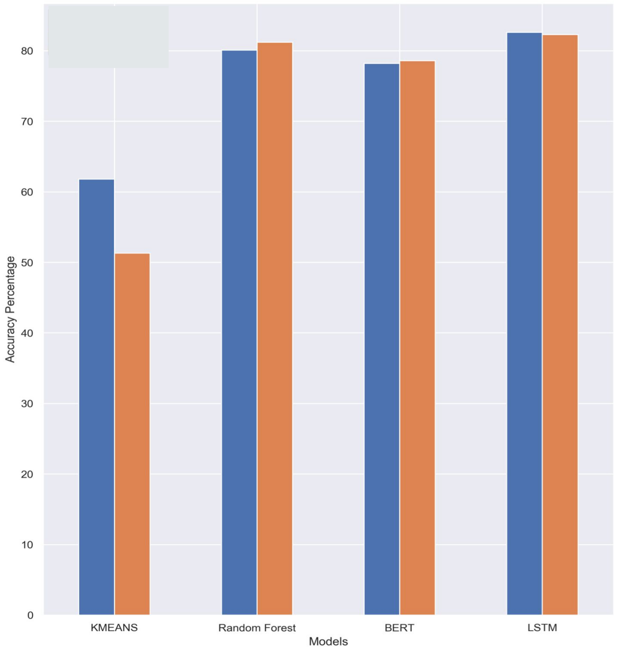


PUBLICATION - 1

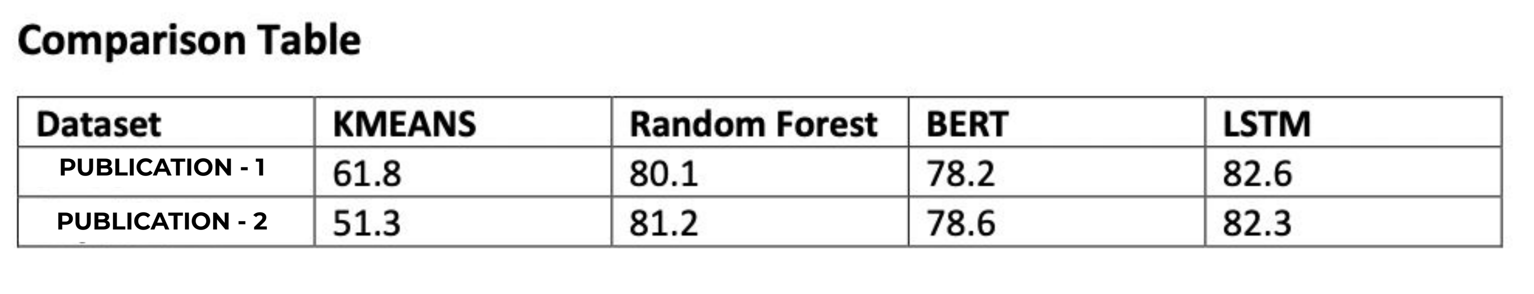


PUBLICATION - 2

ACCURACY (vs) ALL MODELS



ACCURACY (vs) ALL MODELS



Comparison Table

Dataset	KMEANS	Random Forest	BERT	LSTM
PUBLICATION - 1	61.8	80.1	78.2	82.6
PUBLICATION - 2	51.3	81.2	78.6	82.3

MODEL COMPARISON - Machine Learning Models

K-Means

- UNSUPERVISED clustering technique
- TF-IDF vectorization model is used to generate input to clustering process
- 3 - clusters [+ve,-ve and neutral]
- Batch size : 1000
- No. of epochs :

Random Forest

- SUPERVISED classification technique
- Extended decision tree models
- Bag of words vectorization model is used as input to classification process

MODEL COMPARISON - Neural Network Models

BERT - Model

- Neural network based transformer model fine tuned to be used for classification
- Use pre-trained BERT model and thus its own BERT tokenizer
- 3 category layers [+ve,-ve,neutral]
- Dense layers to train data for sentiment analysis

LSTM - Model

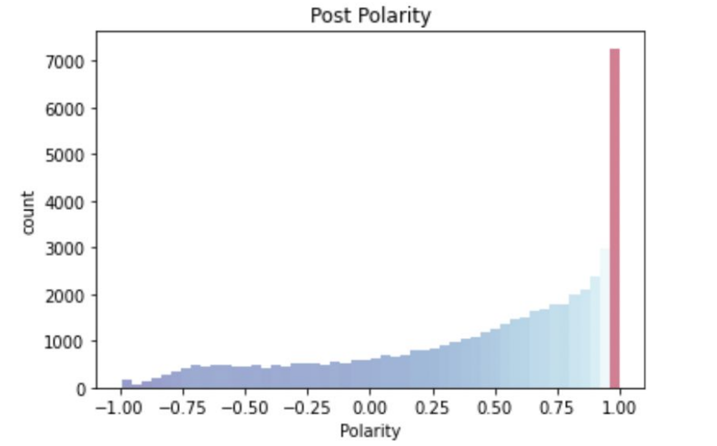
- Neural network based classification technique
- Model fined tuned from scratch to train model to work with sequence of text data
- Sentence embedding for vectorization using Glove vectors
- Dense layer on top for classifying text data which LSTM model has processed

1.

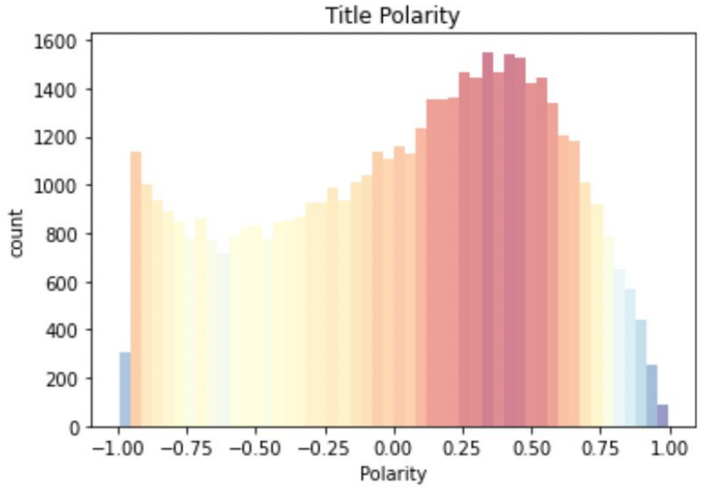
ARE HEADLINES MORE POLARISED THAN THE NEWS ARTICLES?

KEY FINDINGS

Is title More polarized the Post?

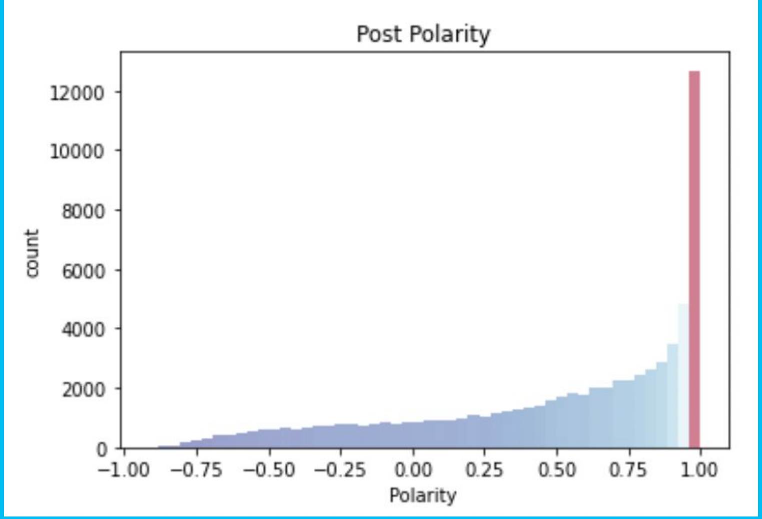


PUBLICATION - 1

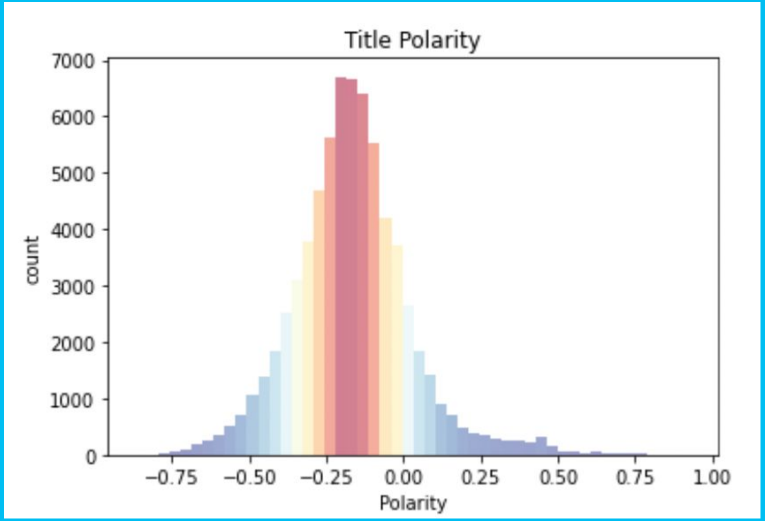


PUBLICATION - 2

KEY FINDINGS



PUBLICATION - 1

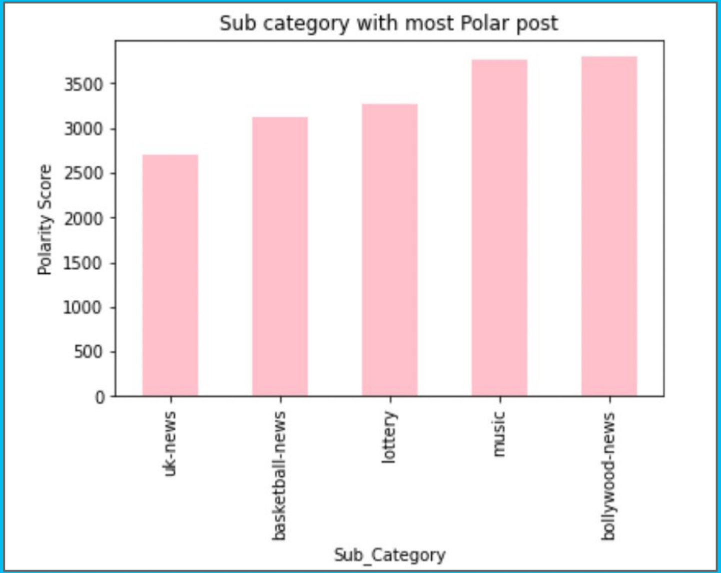


PUBLICATION - 2

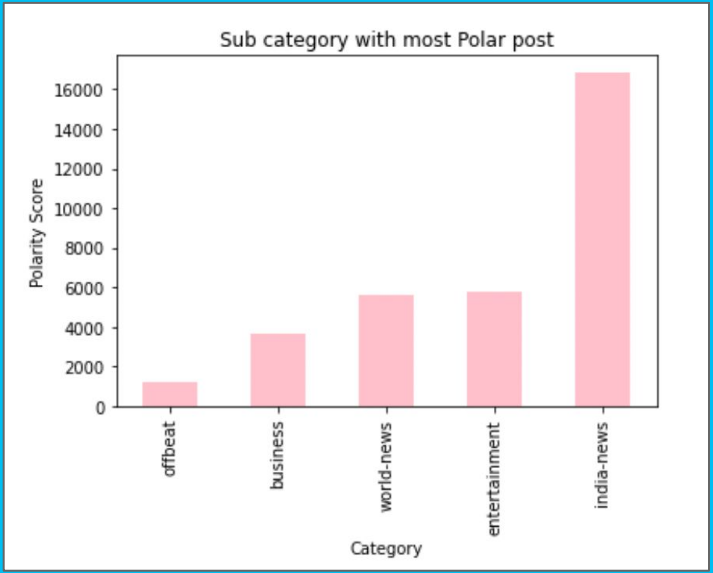
2.

DOES A PARTICULAR PUBLICATION TARGET SPECIFIC CATEGORIES WITH POLARISED NEWS?

KEY FINDINGS

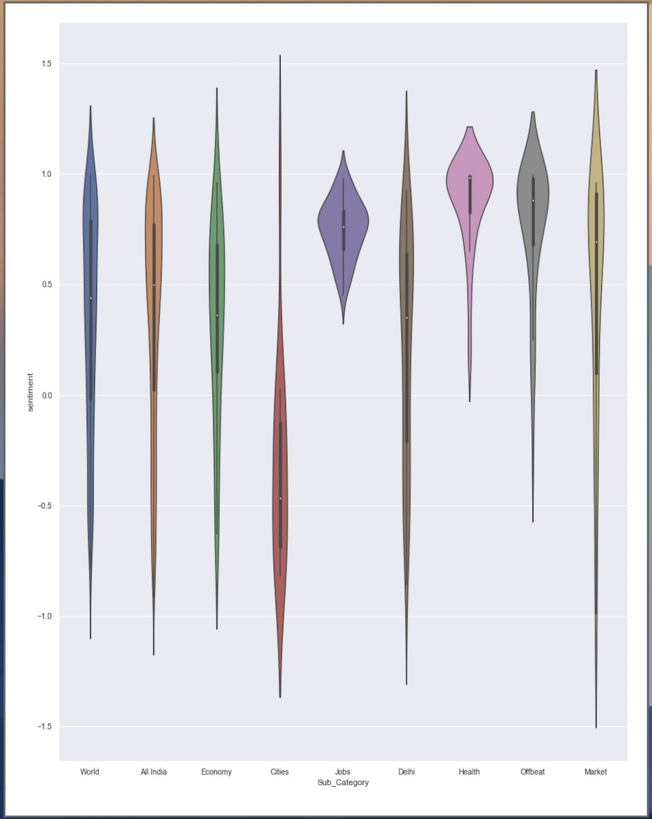


PUBLICATION - 1

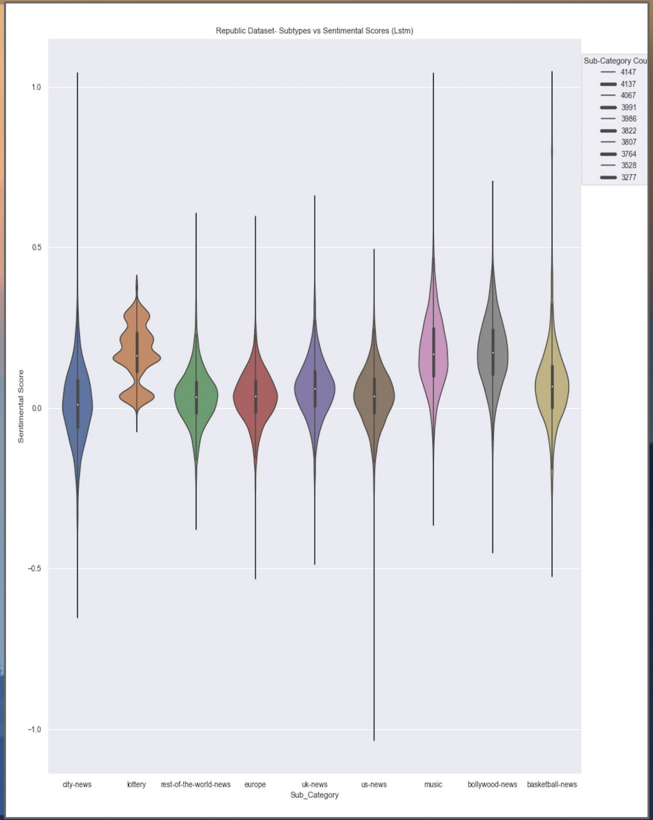


PUBLICATION - 2

SENTIMENT SCORE (vs) ALL MODELS



PUBLICATION - 1

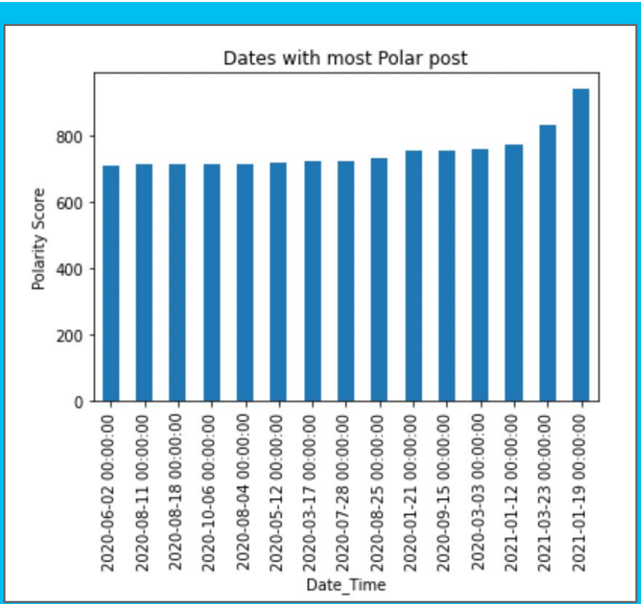


PUBLICATION - 2

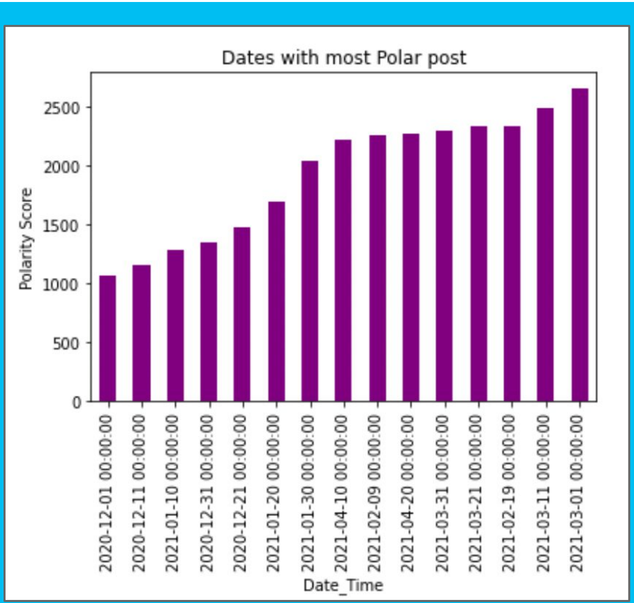
3.

ARE THERE MORE POLARISED NEWS AROUND CERTAIN DATES/EVENTS?

KEY FINDINGS

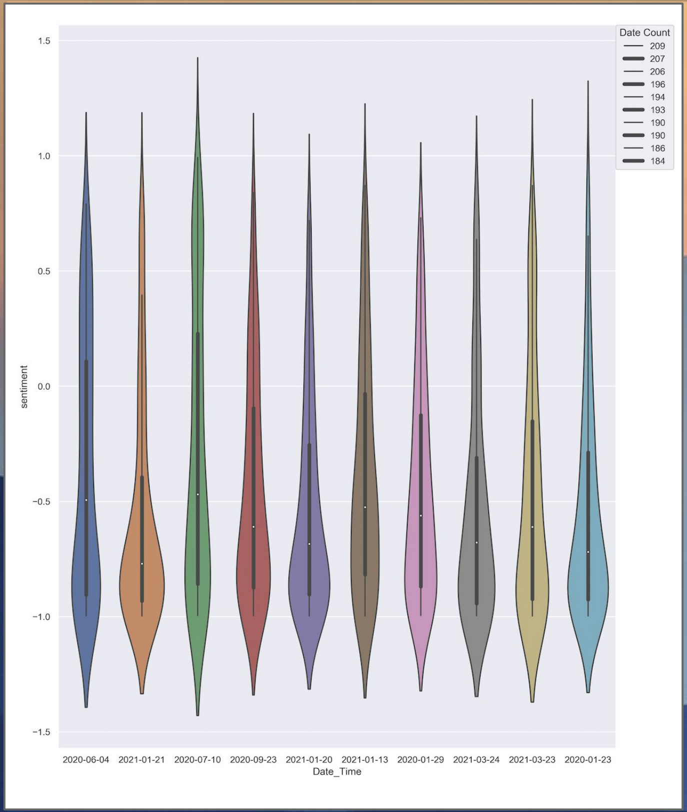


PUBLICATION - 1

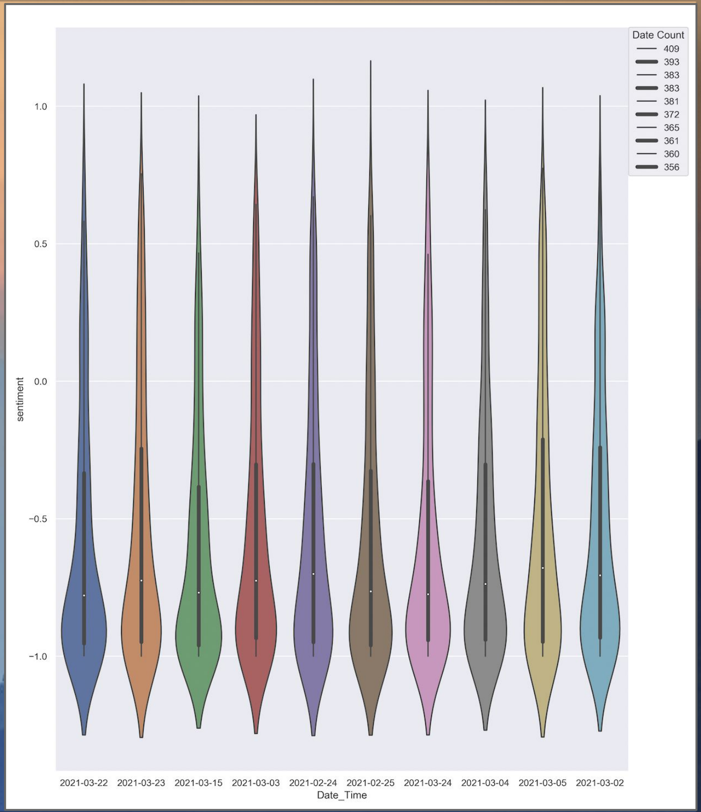


PUBLICATION - 2

SENTIMENT SCORE (vs) ALL MODELS



PUBLICATION - 1



PUBLICATION - 2

RESEARCH BENEFICIARIES

- News Readers
- News Publications
- News Aggregator platforms
- Social Media platforms
- ML/Neural Network developers

SPECIAL THANKS AND REGARDS

PROJECT SUPERVISOR

Prof. Nickolas Falkner

THANK YOU

THANKS!

Any questions?

You can find me at:

- » bhavya.pandya@student.adelaide.edu.au