
Media Sentiment impact on oil prices
-: BY AAYUSH MULYE

Problem Statement

- **Objective** -: Perform sentiment analysis (using BERT) of oil price related tweets/News headlines and then quantify its impact to study the relationship between media sentiment and oil price movements.
- **Testing dataset** – Tweets pulled from twitter, News headlines from Newsapi.org, Datanews.io.
- **Training dataset** includes Financial Phrasebank, FiQA dataset and Stanford's sentiment140 dataset.

Literature Review

- Factors affecting oil prices -:
  - Supply and Demand
  - Geopolitical Situations
  - Interest rates and the US dollar
  - Speculation
- BERT for sentiment analysis -: Dogu Tan Araci in his paper (FinBERT: Financial Sentiment Analysis with Pre-trained Language Models) trained BERT on TRC2 financial, financial phrasebank and FiQA datasets. Our model implementation is based on further training BERT on financial phrasebank, FiQA datasets and Sentiment140 dataset.

Data Sources (training)

- Training datasets –

  - **Financial phrase bank** - The dataset contains 4,840 sentences selected from financial news. The dataset is manually labeled by 16 researchers with adequate background knowledge on financial markets. The sentiment label is either positive, neutral or negative.

  - **FiQA dataset** - FiQA is a dataset that was created for WWW '18's (france) financial opinion mining and question answering challenge (https://sites.google.com/view/fiqa/home). It includes 1,174 financial news headlines and tweets with their corresponding sentiment score. The targets for this datasets are continuous ranging between [−1, 1].

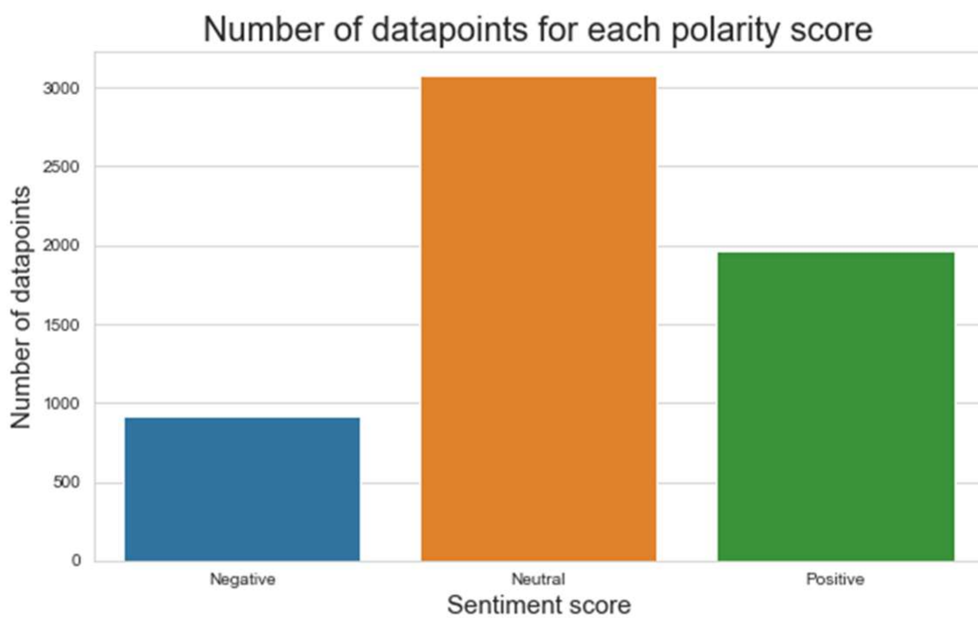
https://arxiv.org/pdf/1908.10063.pdf

Data Sources (training)

- Training datasets –

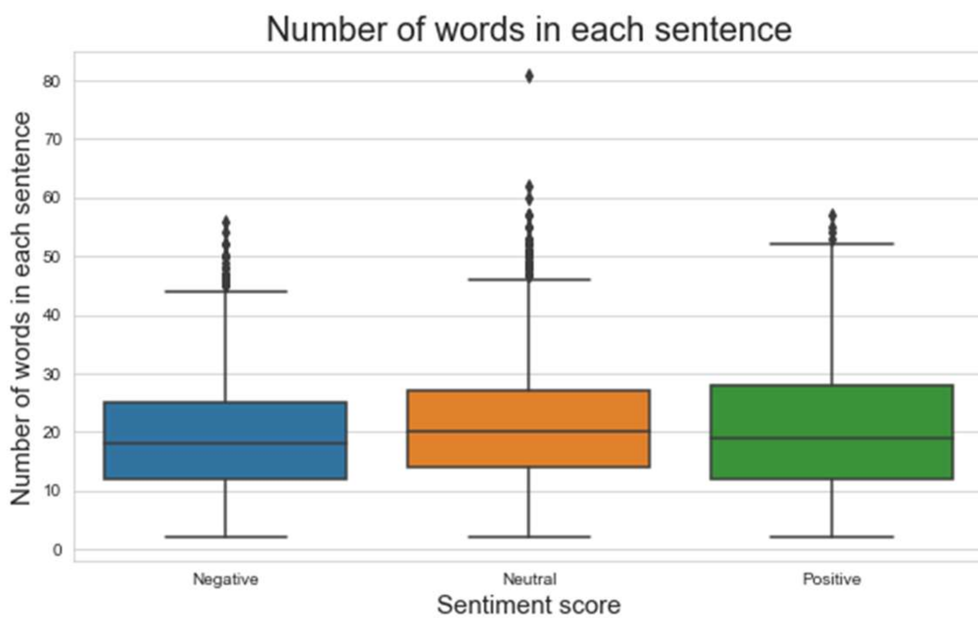
  - **Stanford's Sentiment140 –** This dataset contains 1.6 million labeled tweets extracted using twitter api (by using keyword search). The classification criteria for the labels is based on emoticons.
  - Out of these 1.6 million tweets, relevant tweets for the project have been extracted based on list of financial keywords extracted from university of Baltimore's website.
  - Additionally other keywords include stock tickers of companies with market cap > $2 bn.
  - Final dataset size ~23k tweets.

- Current model to be trained on combined Financial phrase bank, FiQA dataset (~6k sent.). Second model to be trained on dataset inclusive of filtered Sentiment140.



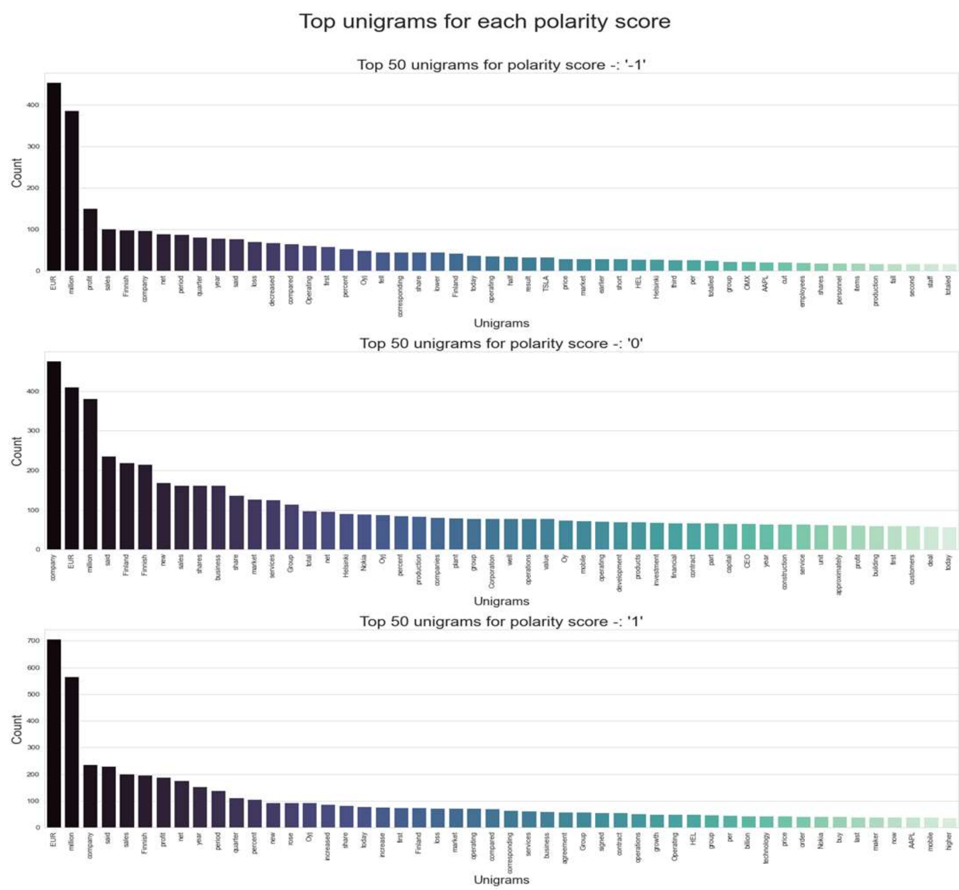
EDA of training dataset (excl. Sentiment140)

- Length of the dataset – 5957
- No. of datapoints for each polarity score



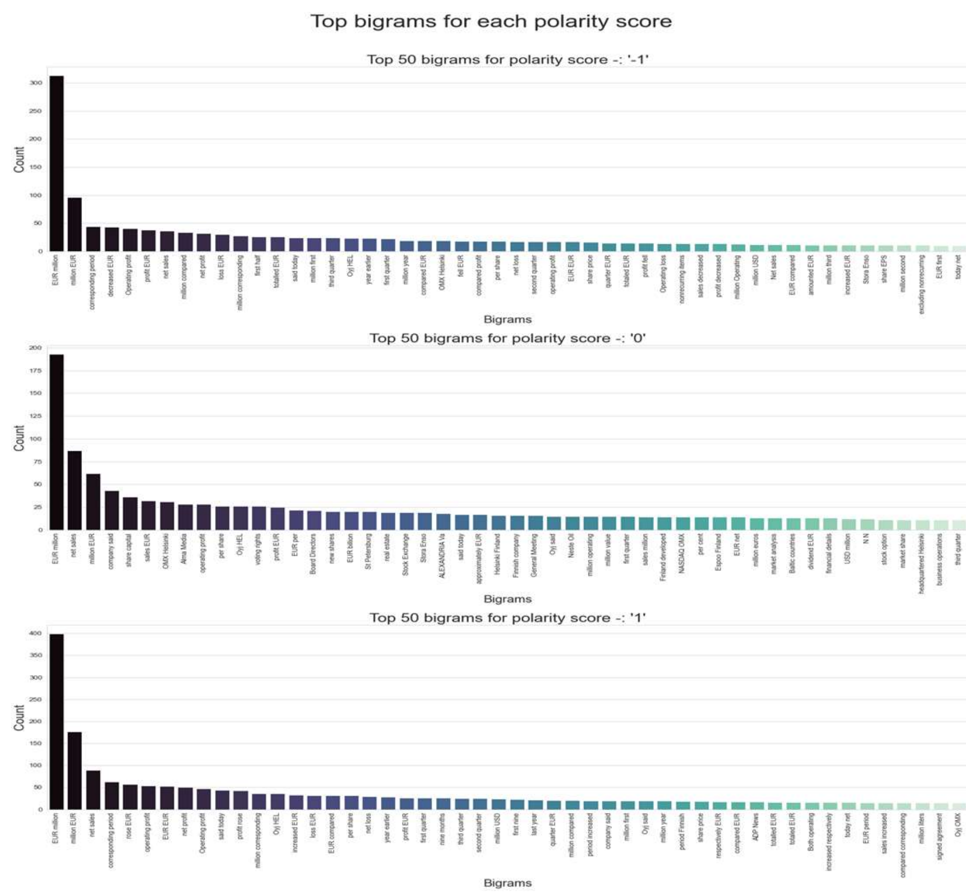
EDA of training dataset (excl. Sentiment140)

- Number of words in each sentence/tweet (distribution)



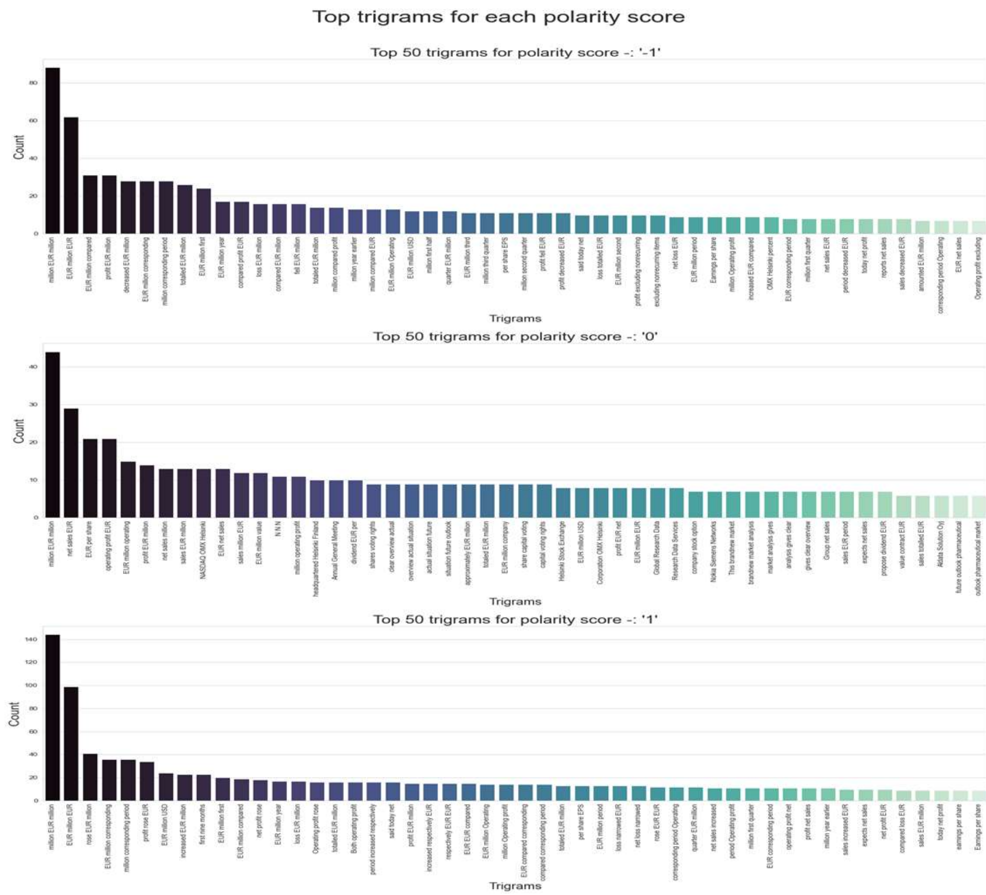
EDA of training dataset (excl. Sentiment140)

- Top unigrams for each polarity score



EDA of training dataset (excl. Sentiment140)

- Top Bigrams for each polarity score



EDA of training dataset (excl. Sentiment140)

- Top Trigrams for each polarity score

EDA of training dataset (excl. Sentiment140)

- Wordcloud for each polarity score

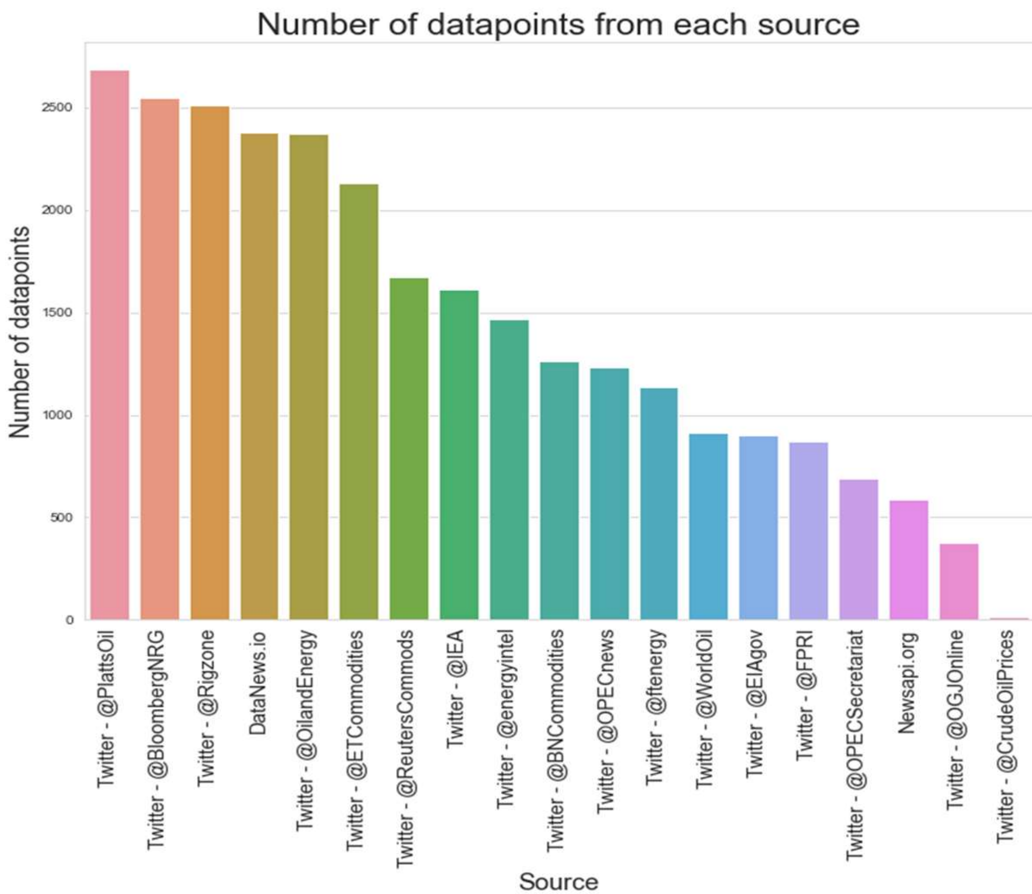
Testing dataset

- **Twitter** - Tweets pulled from the following categories of twitter handles using **twitter API** for the period **1st Sept 2019 to 1st Nov 2020 (16 Months)**. No. of datapoints ~24.5k after filtering on relevant keywords (Oil, COVID, companies name, Saudi, Russia etc.) – :

Oil and Gas news	OPEC News	Commodities News	USFP/US Gov./Think tanks
@OilandEnergy	@OPECnews	@BloombergNRG	@FPRI
@CrudeOilPrices	@OPECSecretariat	@ReutersCommods	@IEA
@PlattsOil		@ETCommodities	@energyintel
@OGJOnline		@BNCommodities	@EIAgov
@Rigzone		@ftenergy	

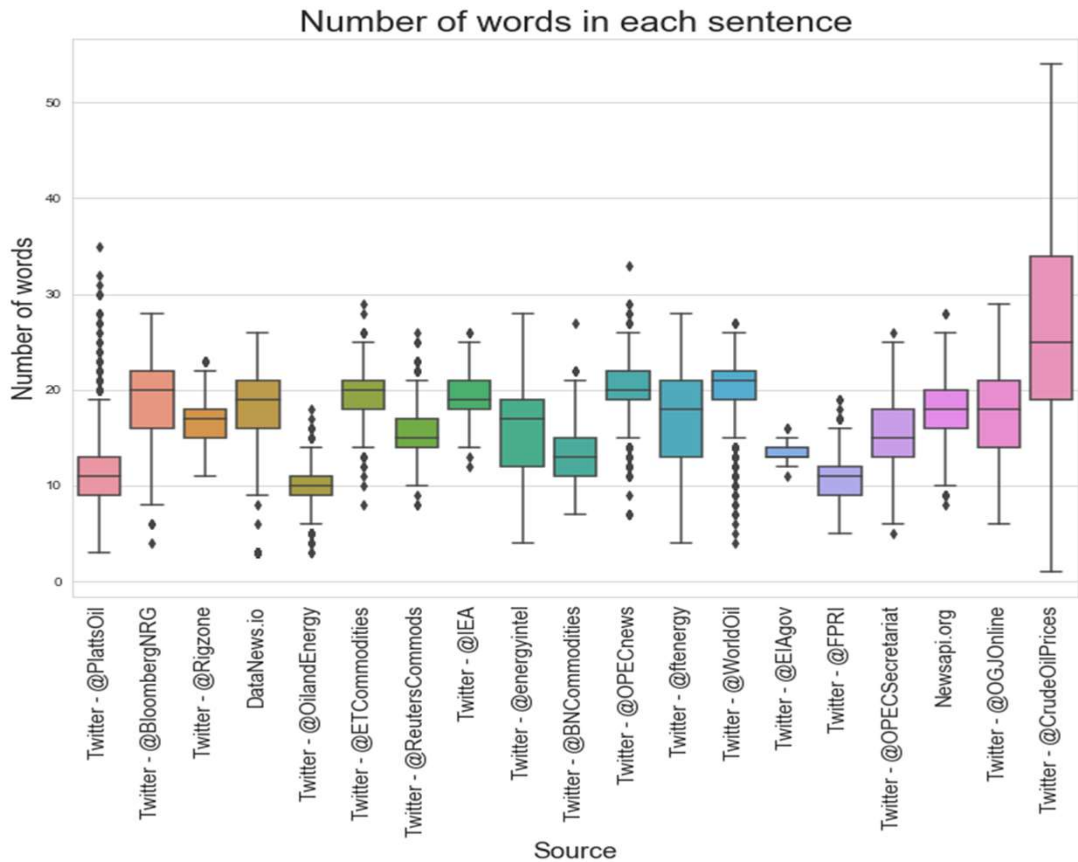
Testing dataset

- **Newsapi.org -** (limitation – 100 requests per day, back in time – 1 month). Extracted the news headlines for period - **1st Oct 2020 to 1st Nov 2020**. Includes articles from 'bbc-news', 'Bloomberg', 'business-insider', 'financial-post', 'google-news', 'the-huffington-post', 'the-wall-street-journal' etc. No. of datapoints = 620.
- **Datanews.io** – Limited number of requests. Extracted the news headlines for period - **1st Oct 2019 to 1st Nov 2020**. Sources include – 'forextv.com', 'uk.reuters.com', 'oilandgas360.com', 'washingtonpost.com', 'businessinsider.com', 'dailyfx.com', 'wsj.com', 'moneycontrol.com' etc. No. of datapoints = 2,381.



EDA of testing dataset – tweets from source

- Length of the testing dataset – 27,358
- No. of datapoints for each source.

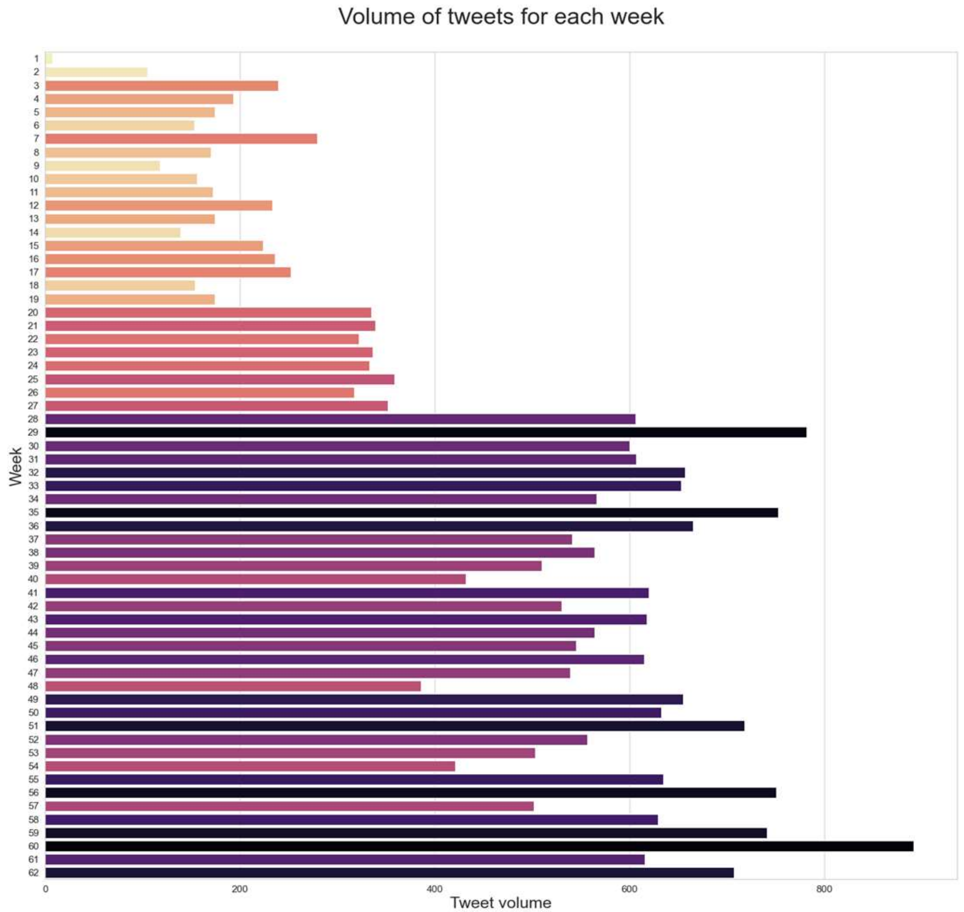


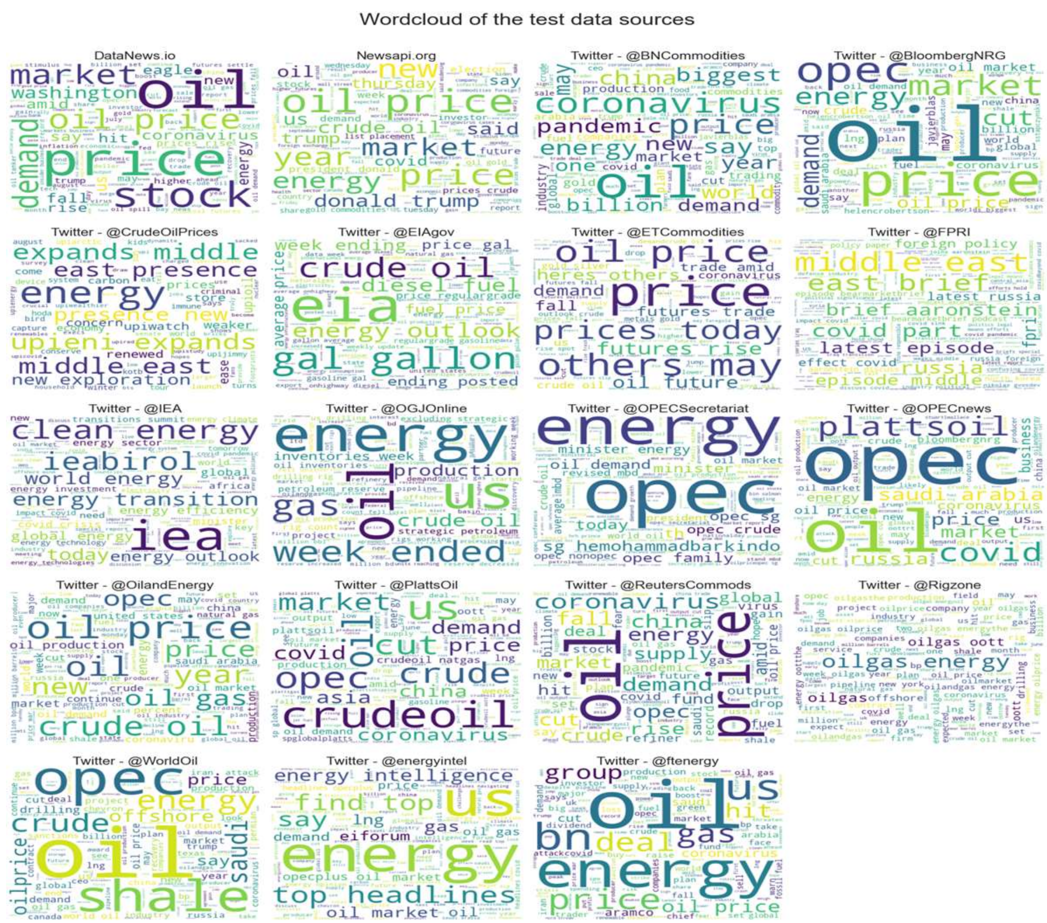
EDA of testing dataset – words per tweet

- Number of words in each sentence/tweet (distribution) w.r.t source).

EDA of testing dataset – tweets volume

- Volume of tweets for each week under consideration.





EDA of testing dataset - Wordcloud

- Wordcloud of all the test data sources

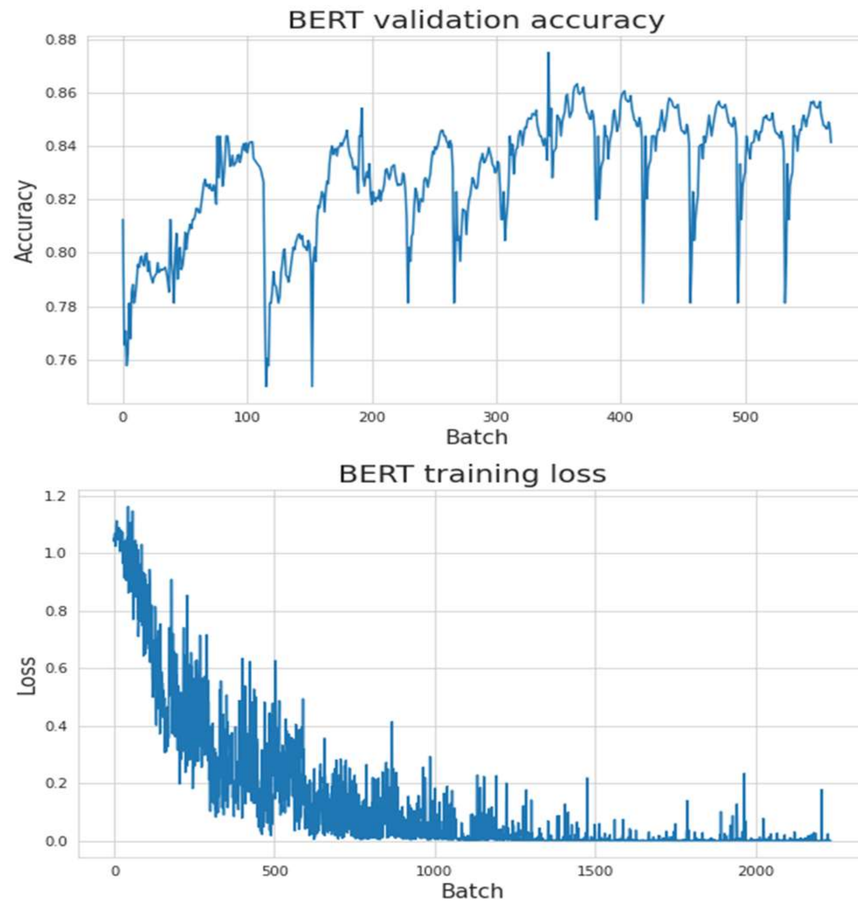
Model Implementation (BERT)

- Implemented BERT (Transformer-based machine learning technique for NLP developed by Google) There are two models – BERT Base, and BERT Large, both have been trained on English Wikipedia with about 2500Mn words.
- Transformers module from Huggingface provides the general architecture for BERT, GPT2, DistilBert… etc.
- And so, the pretrained BERT was retrieved from Huggingface.
- Current status of the Validation accuracy ~85%.

BERT training

After retrieving the pretrained model from Huggingface, the model was again trained on the combined Financial phrasebank + FiQA dataset after pre-processing the textual data.

Val Accuracy ~85%



Model Implementation (other)

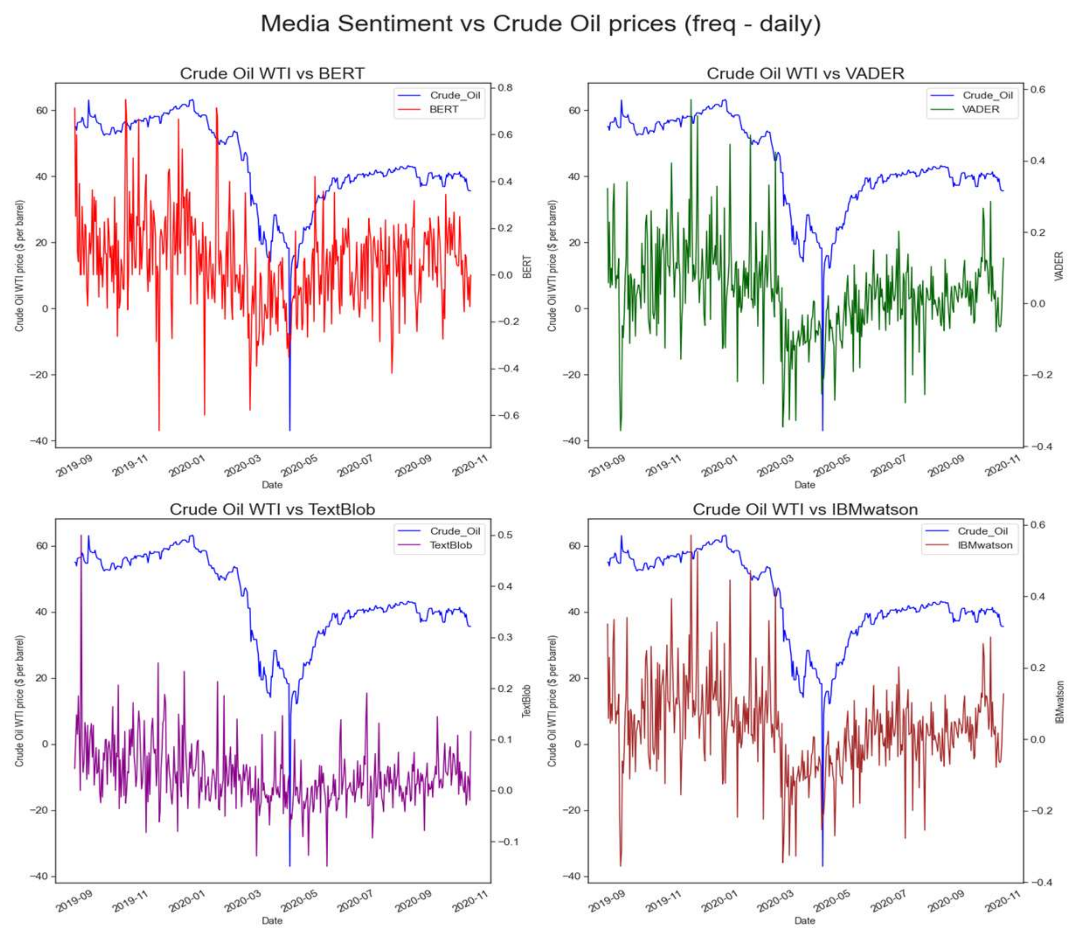
- Apart from BERT, also evaluated sentiment score based on -:
  - VADER sentiment analyser.
  - Textblob sentiment analyser.
  - IBM Watson sentiment analyser – Limitation – limited number of API calls.

Results

- BERT Sentiment scores were generated by feeding the testing data into the BERT model.
- Similarly, other sentiment scores were generated from the respective sentiment analysers.
- Since we have multiple datapoints for each day, sentiment score has been taken as the average across all the datapoints.
- Crude oil -: Crude oil spot prices retrieved from FRED -> *Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma (DCOILWTICO)*

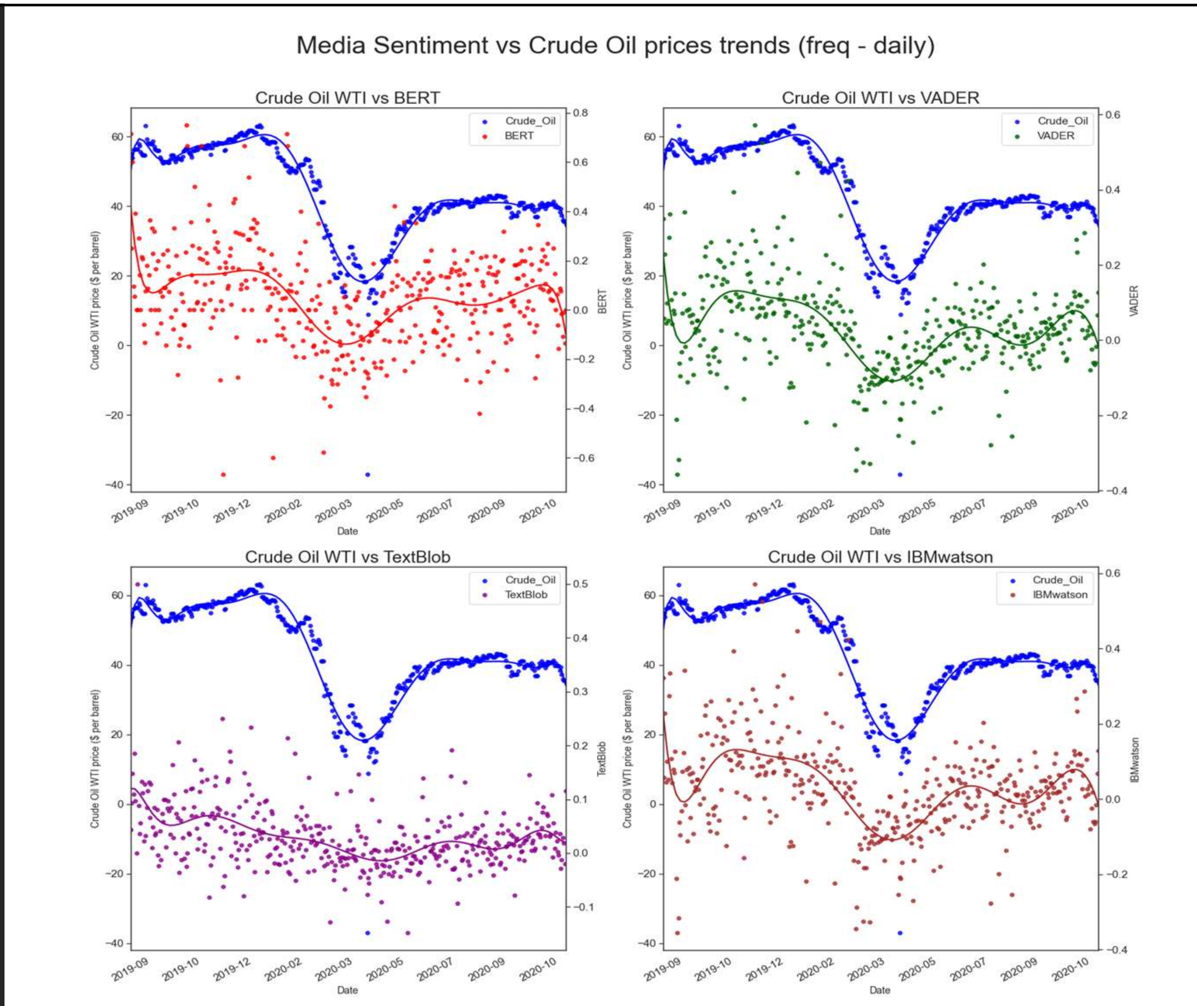
Avg. daily sentiment w.r.t Crude Oil

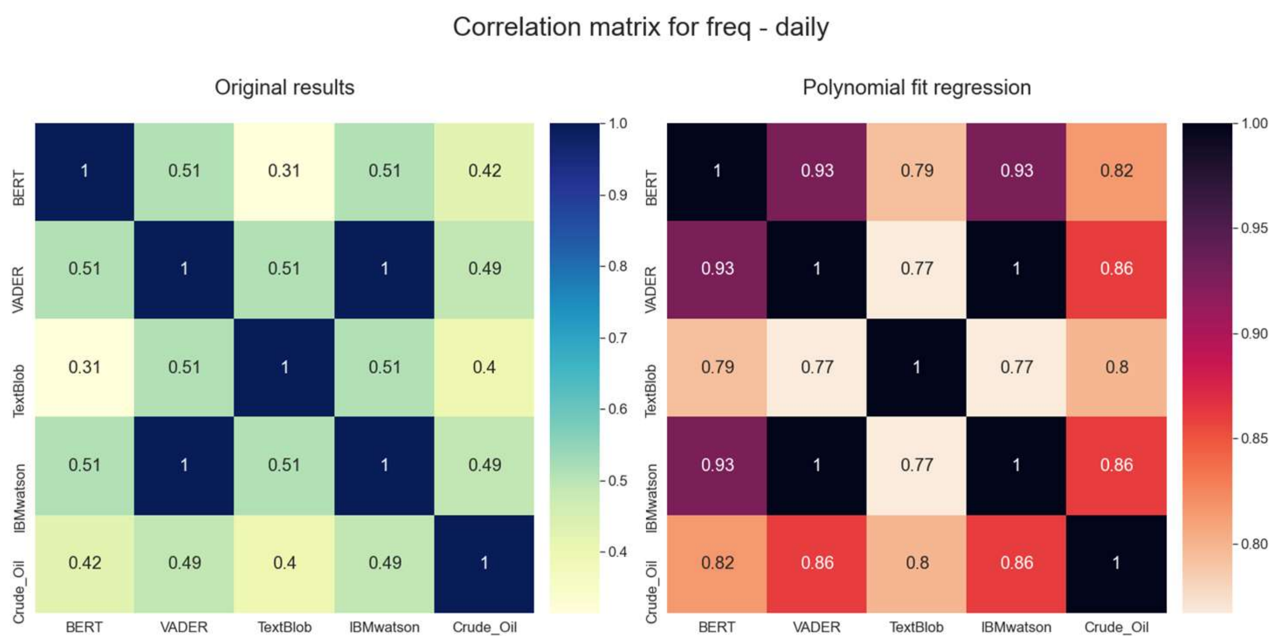
- The huge observed during the months March – June is was the result of Russia-Saudi Arabia oil price war accompanied by the decrease in demand due to COVID-19.



Avg. daily sentiment w.r.t Crude Oil

- From the plots, it can be observed that sentiment score and the daily crude oil spot prices follow similar trends.

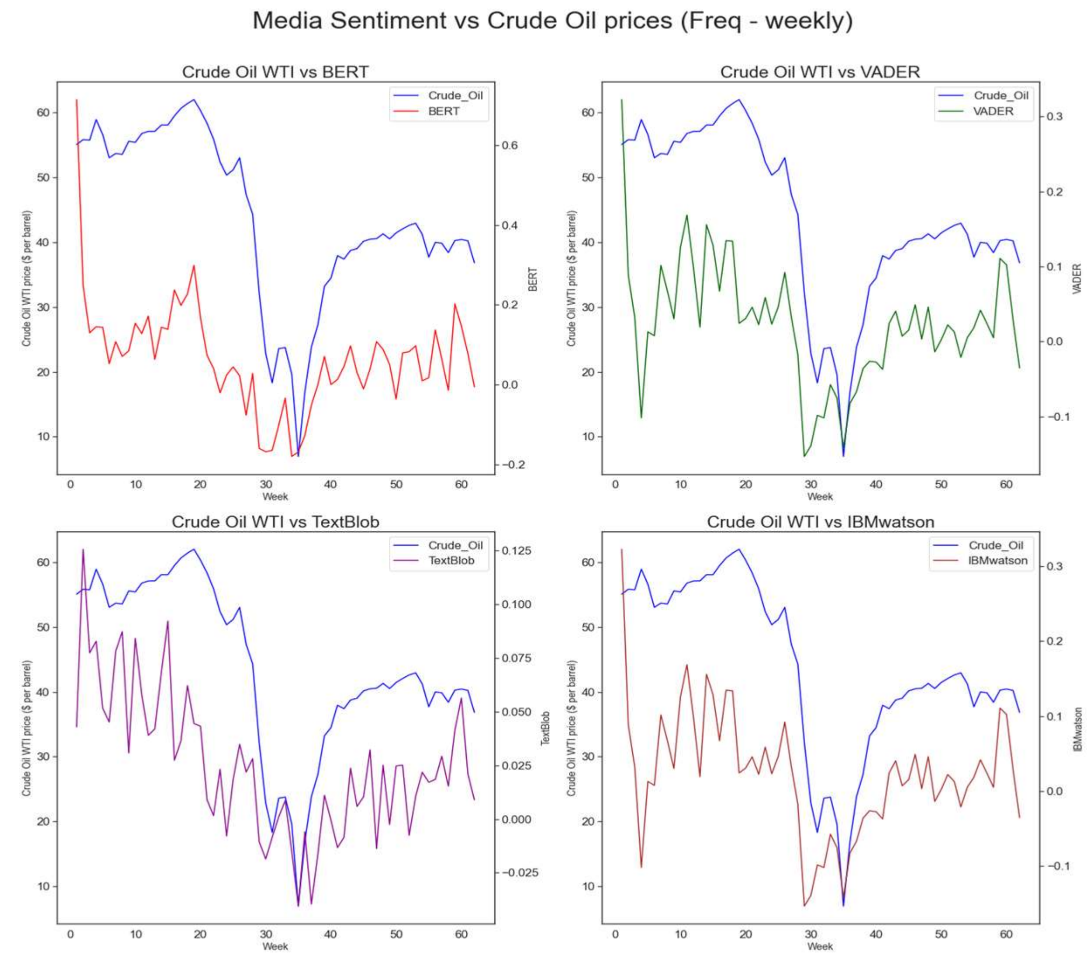




Correlation matrix (Freq - daily)

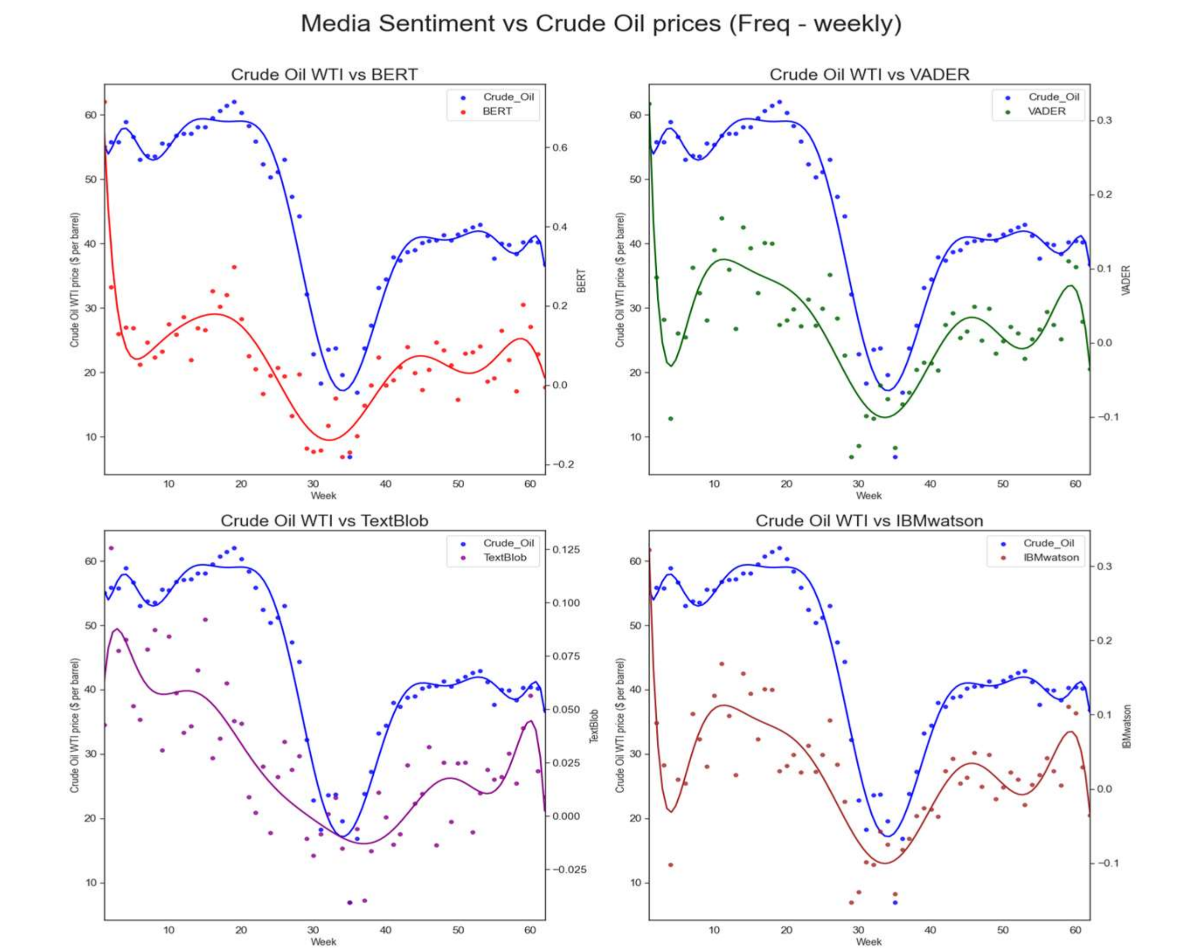
Avg. weekly sentiment w.r.t Crude Oil

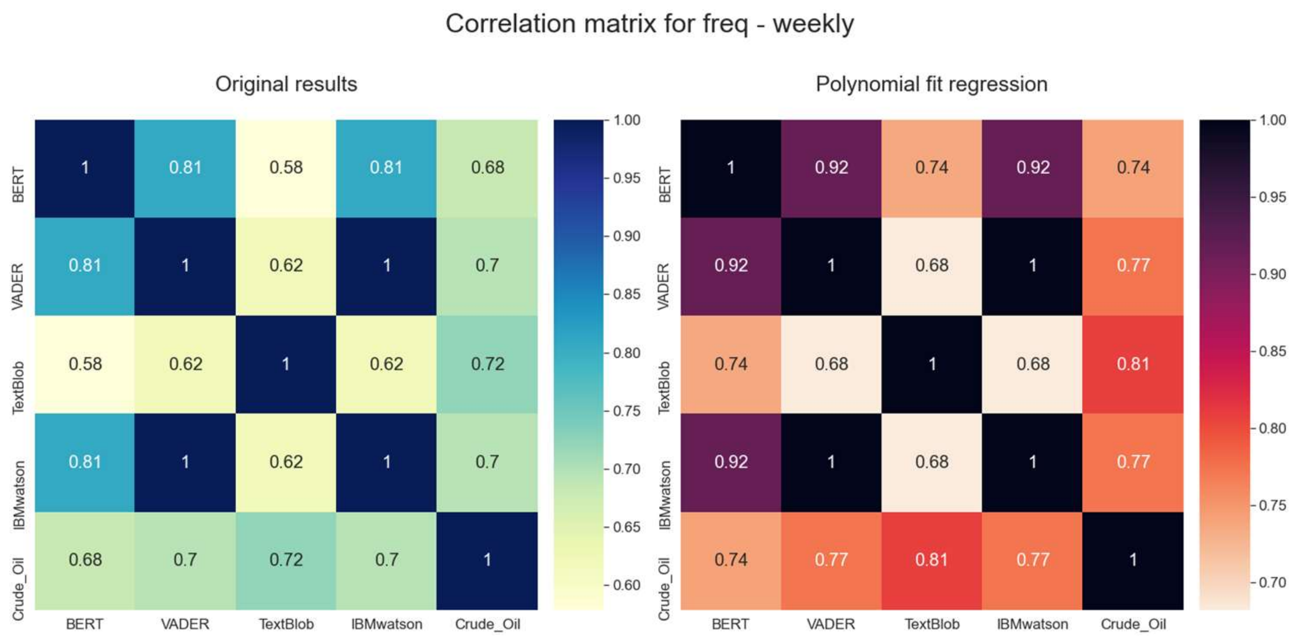
- ❖ Both the crude oil prices as well as the sentiment scores have been generated by averaging them over the week.
- ❖ Similar to the daily results, we see a drop for the months March – June as a result of Russia-Saudi Arabia oil price war and the decrease in demand due to COVID-19.



Avg. weekly sentiment w.r.t Crude Oil

- Similar trendlines are observed for the weekly plots as well.





Correlation matrix (Freq - daily)

	MI daily actual	MI daily regressed	MI weekly actual	MI weekly regressed
BERT_label	0.190096	1.192871	0.501348	0.990601
VADER_label	0.185157	1.262531	0.482781	1.058539
TextBlob_label	0.047880	1.083756	0.354188	0.879916
IBMwatson_label	0.185157	1.262531	0.482781	1.058539

MI score

Thank You