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 datasets –

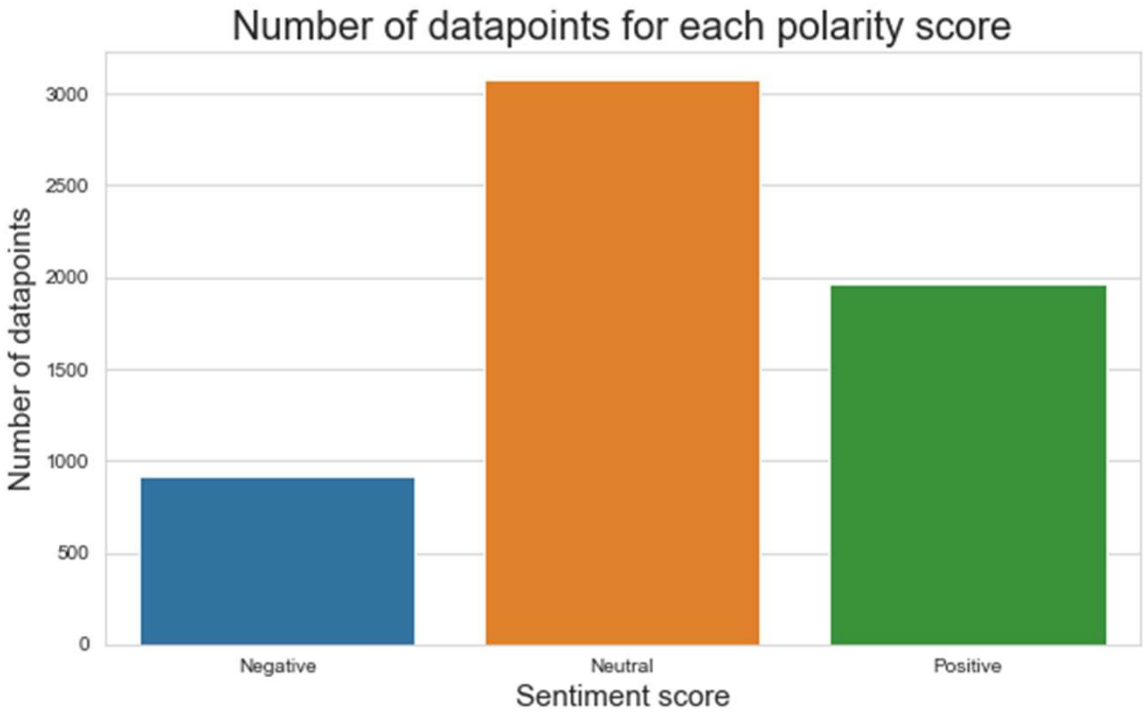
  - **Financial phrase bank -** The dataset contains 4,840 sentences selected from financial news. The dataset is manually labeled by 16 researchers with adequate background knowledge on financial markets. The sentiment label is either positive, neutral or negative.

  - **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

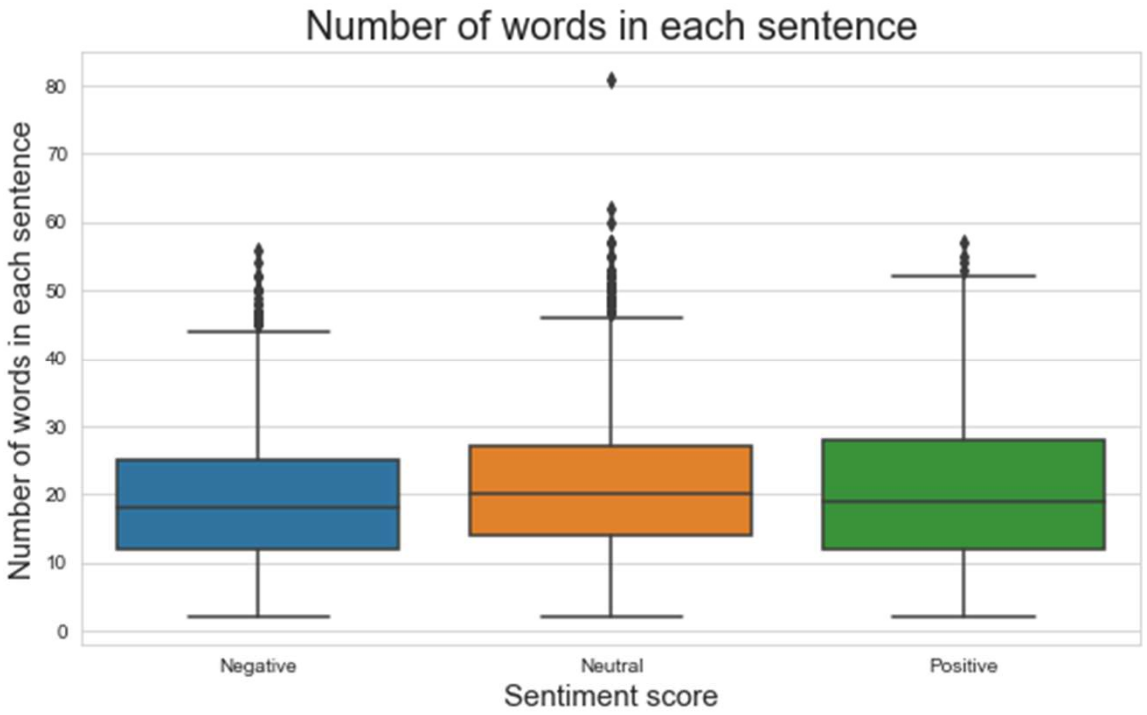
Data Sources

- Training 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.



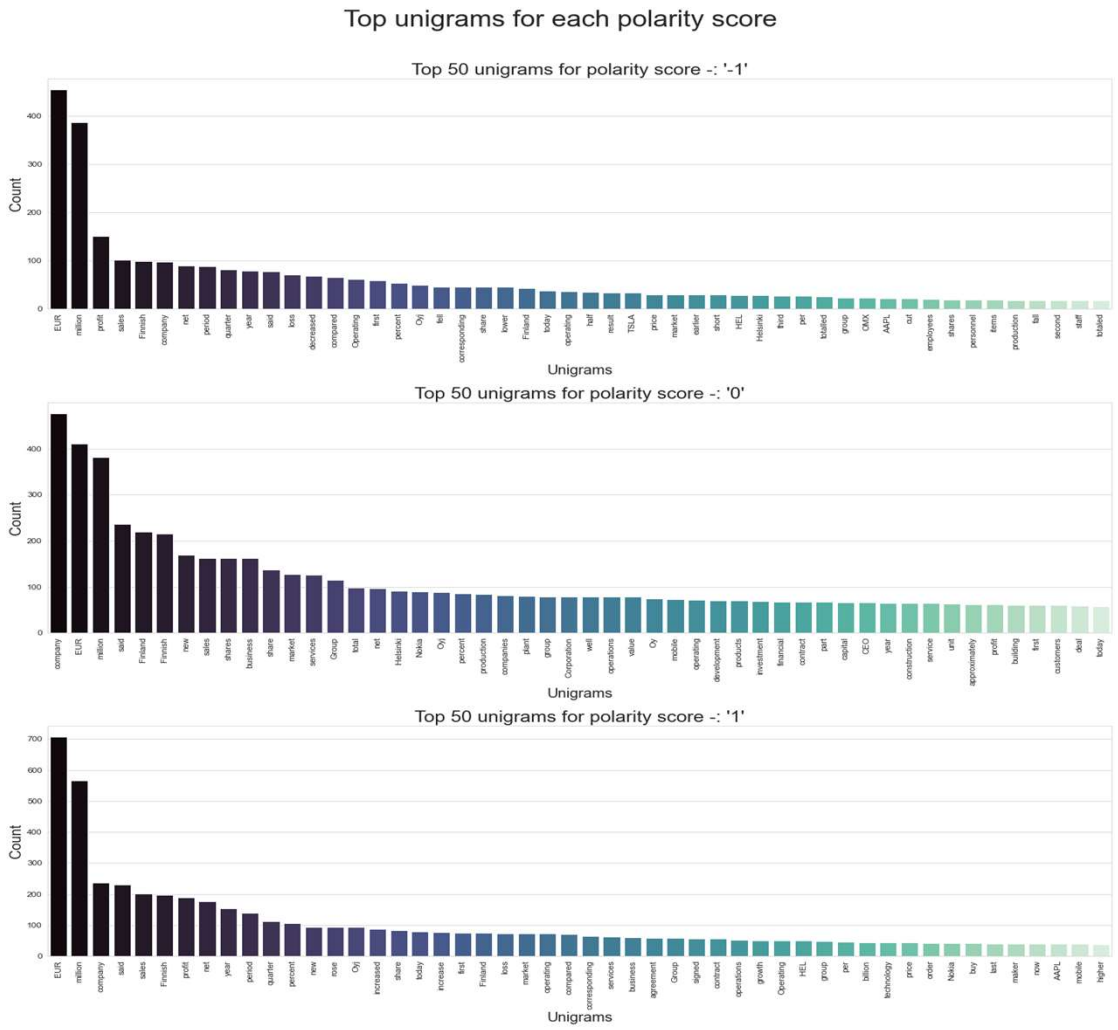
EDA of training dataset (excl. Sentiment140)

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



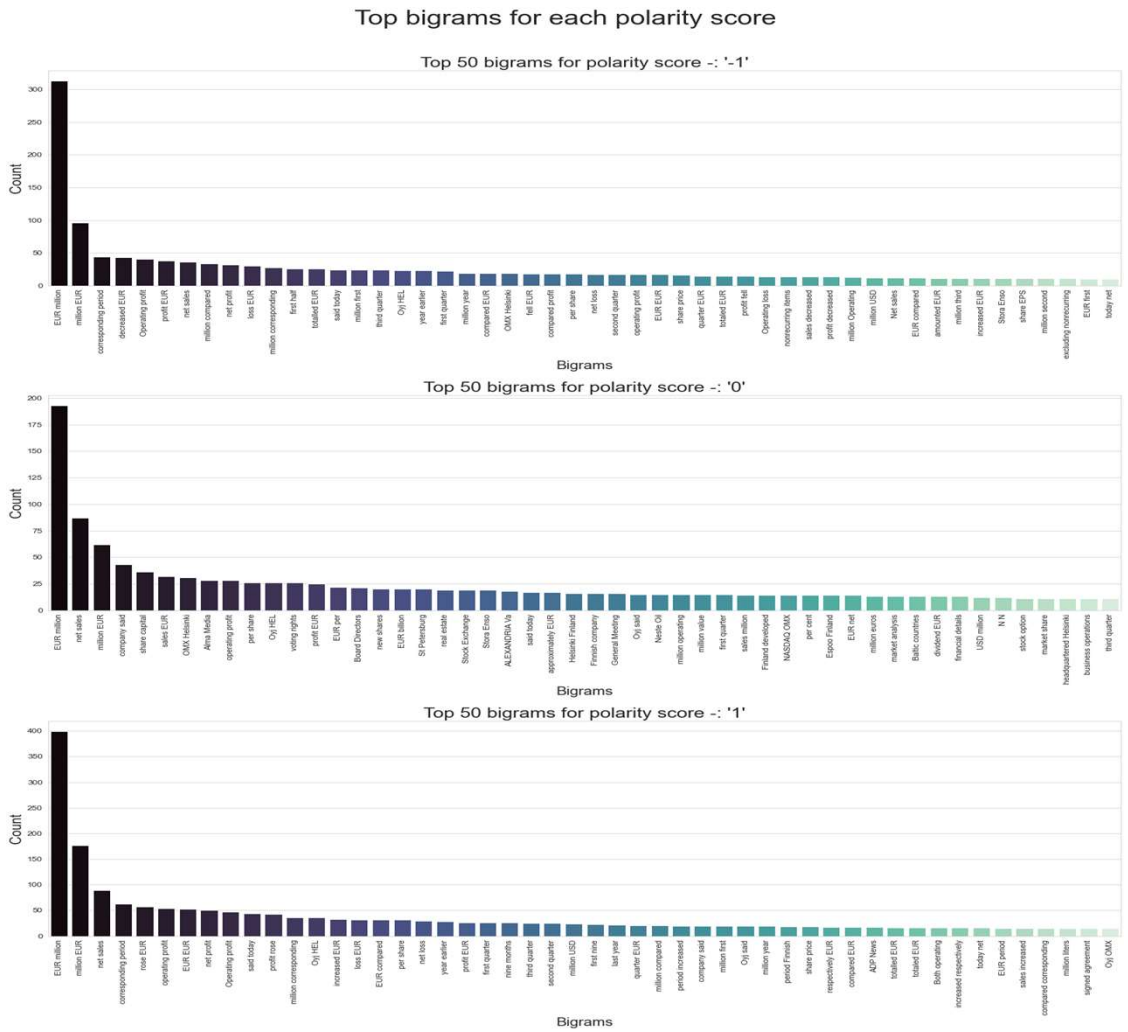
EDA of training dataset (excl. Sentiment140)

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



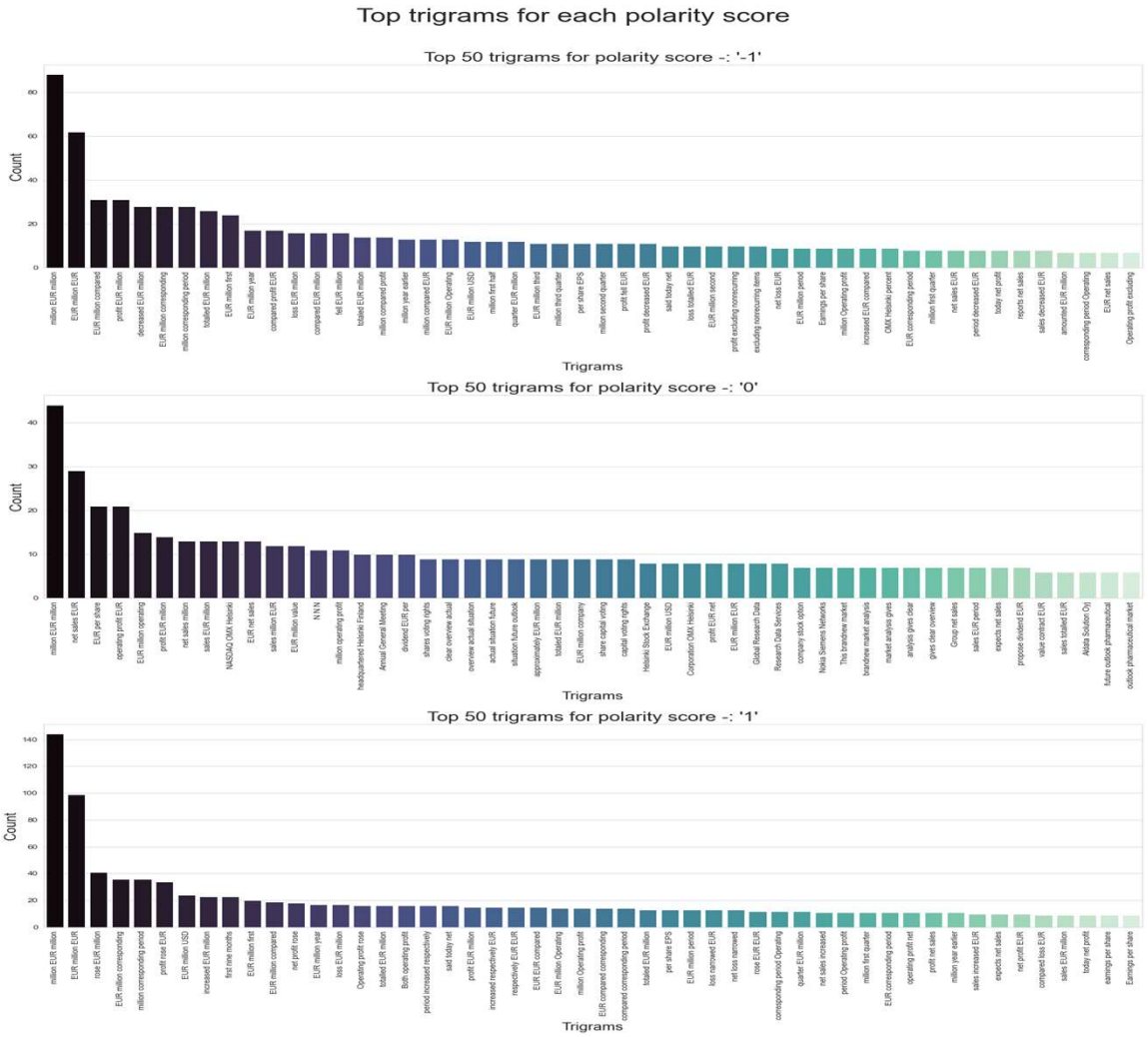
EDA of training dataset (excl. Sentiment140)

- 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)

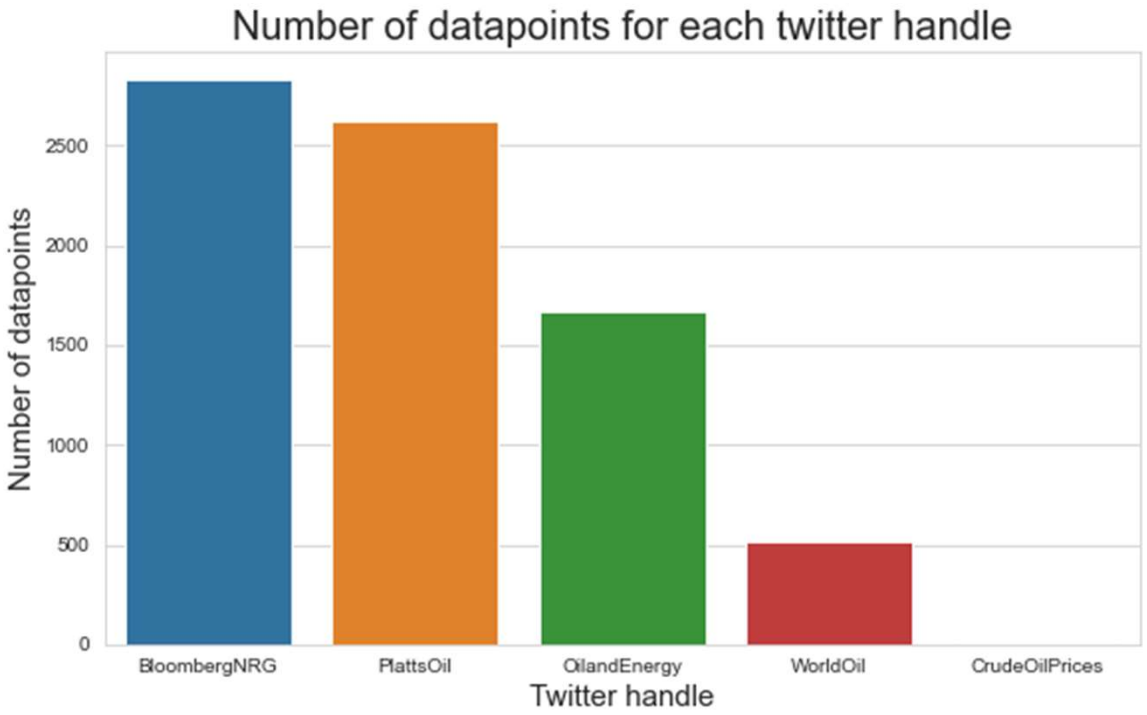
- Wordcloud for each polarity score

Model Implementation

- 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 ~82% (without tuning).

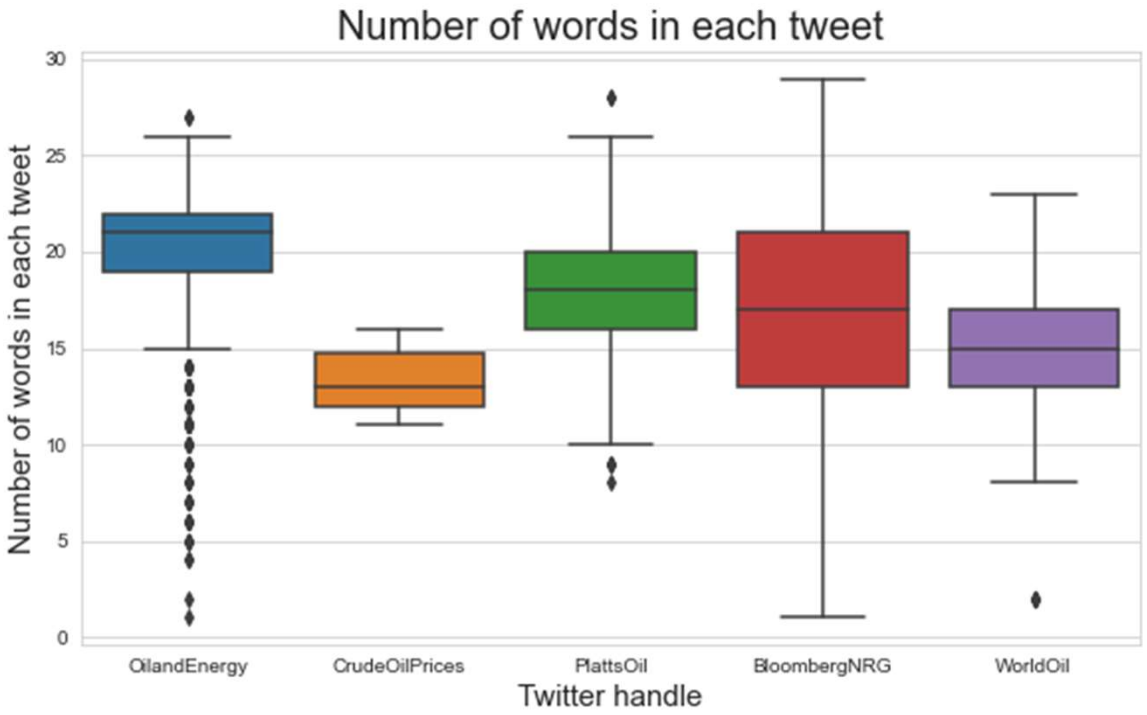
Testing dataset

- **Twitter -** Testing dataset currently is comprised of tweets pulled from the following twitter handles - '@OilandEnergy', '@CrudeOilPrices', '@PlattsOil', '@BloombergNRG', '@WorldOil'. (technicalities – start date, end date, twitter handle).
- Additional possible sources
  - **Newsapi.org**. (limitation – 100 requests per day, back in time – 1 month). Includes articles from 'bbc-news', 'bloomberg', 'business-insider', 'financial-post', 'google-news', 'the-huffington-post', 'the-wall-street-journal' etc.
  - **Datanews.io** – limit of 3k requests. Similar sources as above.



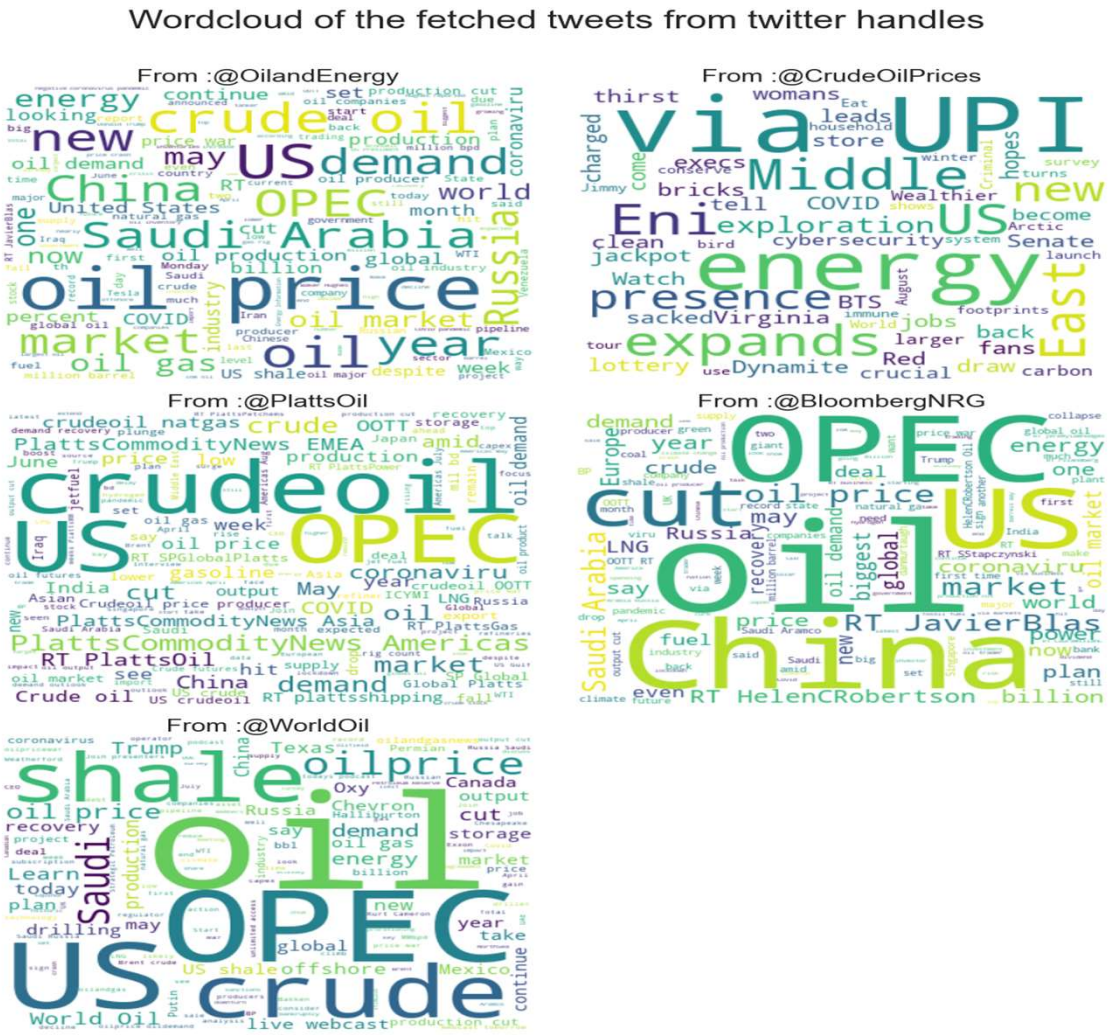
Testing dataset analysis

- Length of the testing dataset - 7639
- No. of datapoints for each twitter handle.



Testing dataset analysis

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



Testing dataset analysis

- Wordcloud of tweets for each twitter handles

Challenges/Next steps

- Tune the parameters on the training dataset (currently excl. filtered Sentiment140) to improve the accuracy.
- Finalise the testing dataset – Combining data from Twitter, Newapi.org and Datanews.io. Additional transformations (Preprocessing, data filtration).
- Run the models on testing dataset to generate sentiments.
- Figure out better way to filter the sentiment140 dataset.
- Compare the model performance based on the two training datasets.

Thank You