Predictive Value of Sentiment Analysis from Headlines for Crude Oil Prices

Understanding and Exploiting Deep Learning-based Sentiment Analysis from News Headlines for Predicting Price Movements of WTI Crude Oil

Master thesis

Author:	Himmet Kaplan Burgstrasse 70 9000 St. Gallen Switzerland himmet.kaplan@stud.fhgr.ch
Under the supervision of	Prof Dr Peter-Ralf Mundani, FHGR ralf-peter.mundani@fhgr.ch
Co-Referent:	Prof Dr Heiko Rölke, FHGR heiko.roelke@fhgr.ch@fhgr.ch
	University of Applied Sciences of the Grisons (FHGR) Competence Centre for Data Analysis, Visualisation and Simulation (DAViS) Pulvermühlestrasse 57 7000 Chur Switzerland
In cooperation with:	Ilja Rasin, IBM Global Business Services ilja.rasin@ch.ibm.com
Editing period:	26th February 2021 – 13th August 2021

This thesis was written in partial fulfilment of the requirements for the degree of Master of Science FHGR in Business Administration with a Major in Information and Data Management.

Acknowledgements

This dissertation has been an enriching and educational experience for me, and I would not have been able to complete it without the support and advice of numerous individuals.

First and foremost, I'd like to express my sincerest gratitude to my supervisor Prof Dr Peter-Ralf Mundani. Whose expertise and encouragements were vital for developing the research question during our many discussions. Thanks to my co-referent, Prof Dr Heiko Rölke, for his support.

My thanks also go out to the support I received from the collaborative work with IBM Global Business Services. I am especially grateful to Ilja Rasin for believing in my research.

An exceptional thank you to Dr Karin Kneissl for her precious time and vital intelligence concerning the domain adaptation.

Also, special thanks to Dr Martin Frey for his helpful inputs regarding the evaluation part.

I am also very grateful to all my colleagues at the Institute for Mechatronic Systems (IMS, ZHAW), especially the institute director, Prof Dr Hans Wernher van de Venn and my manager, Mr Michal Jerzy Malinowski, for providing me with support, trust, and a flexible working environment while pursuing my research. Additional thanks to my colleagues at the Autonomous Systems Labs (ASL, ETHZ) for their encouragement and inspirational work ethic.

I am also profoundly grateful to the Hirschmann Foundation for granting me a scholarship, which was helpful to acquire the necessary hardware for this work.

Lastly, I would like to thank my family for always supporting, encouraging and believing in me.

Abstract

Deep learning has become a popular approach for sentiment analysis, a prevalent text classification task in natural language processing. It has been demonstrated to outperform traditional classification methods bringing enormous potential in various applications. This thesis aims to research and develop models that can perform sentiment analysis on the news related to crude oil and provide valuable insight. First, a literature review is provided on the potentials of news affecting crude oil prices and the current state-of-the-art deep learning-based sentiment analysis methods utilizing transformer architectures. Additionally, news data sources, as well as appropriate frameworks for conducting sentiment analysis, are examined. Next, based on the literature review, the models are identified, implemented and iteratively fine-tuned through domain adaptation. Finally, recommendations for implementing deep learning-based sentiment analysis methods for predicting crude oil prices are made.

Keywords: BERT, Deep Learning, Domain Adaptation, News Analytics, Natural Language Processing, Sentiment Analysis, Transformers

Contents

List of Figures[1]

[1] Figures with copyright were acknowledged for use, figures without source are own representations.

List of Tables

List of Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag-of-Words
CBOW	Continuous Bag-of-Words
DL	Deep Learning
ELMo	Embeddings from Language Models
EMH	Efficient Market hypothesis
GEPU	Global Economic Policy Uncertainty
GLUE	General Language Understanding Evaluation
GPT	General Purpose Transformer
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neuronal Network
SST	Stanford Sentiment Treebank
ULMFit	Universal Language Model Fine-tuning for Text Classification
WTI	West Texas Intermediate

1 Introduction

Crude oil is currently the primary source of energy production and one of the most necessary production materials and is thus of crucial importance for the global economy. Since it is a limited natural resource, it can serve as a booster and hazard for economies. Thus, predicting the price of such an important commodity could provide great value. Unfortunately, its price is determined mainly by demand and supply, which are notoriously difficult to forecast. According to the literature, the main factor why crude oil is one the most volatile markets in the world is that the demand is primarily affected by exogenous events such as armed conflicts and natural disasters (Buyuksahin & Harris, 2011, p. 2).

Traditionally, analysts statistically analysed the historical price movements to uncover potential insights for the future (also known as technical analysis). However, such historical data rarely contains perfect patterns for generating absolute probabilities. Thus, it can only identify probable outcomes (McCarthy et al., 2019, p. 197). In order to achieve more reliable forecasts, the data could be enriched with other relevant data, such as news which have a significant influence on almost all publicly listed, marketable assets, including crude oil. Therefore, many researchers were interested in incorporating news to improve predictions (Baboshkin & Uandykova, 2021, p. 1). For instance, one way to assess and incorporate news would be to quantify its content on how positive or negative it is through sentiment analysis. With the advancement in computer hardware, current natural language processing algorithms are capable of evaluating text for strategic forecasting. Current state-of-the-art sentiment analysis can achieve a remarkable accuracy of 97.5 percent (Jiang et al., 2020, p. 1).

Considering these facts, the focus of this thesis deals with the task of research and development of state-of-the-art sentiment analysis methods, which can potentially provide helpful quantification of news that can be used to assess the future price movements of crude oil.

The main contributions of this thesis for this particular task are (i) identifying suitable sources for obtaining news data, (ii) deep learning-based sentiment analysis models and (iii) iterative improvements of the selected model through the application of domain adaptation.

2 Literature review

The following chapter consists of a literature review regarding the efficient market hypothesis, news media properties, crude oil properties, the definition of sentiment itself, and a brief history concerning the evolution of applied sentiment analysis. The latter will provide a comparatively in-depth explanation regarding the transformer architecture leading to the development of BERT, which is also the main focus of this study.

2.1 Efficient Market Hypothesis

The efficient market hypothesis (EMH) is a theory from the field of capital market research that generally states that an asset's prices are mainly the result of all the available information. The EMH was divided by Eugene Fama (1970) into a strong, semi-strong and weak form. The strong form of the EMH states that the current prices of any asset reflect historical prices and all publicly available and confidential information. If markets are efficient in this sense, applying fundamental analysis based on all available information cannot lead to abnormal economic returns. The semi-strong form of the EMH behaves similarly, except that confidential information is not included in the price. In this case, insider knowledge could be an added value for a forecast, as it has not yet influenced the current price. (Malkiel, 1989, p. 127)

In contrast to the strong and semi-strong forms, the weak form of the EMH claims that the price results only from its historical price history. Hence investors cannot develop a profitable investment strategy using technical analysis, which is essentially based on past price movements. (Malkiel, 1989, p. 128)

However, several researchers have disputed the theory regarding the strong and semi-strong form of the EMH, arguing that price changes can be predicted with over 50% accuracy through fundamental analysis such as news analytics (Qian & Rasheed, 2007).

News analytics aims to provide instantaneous analysis and quantification of news content. For this purpose, unstructured news data are collected using text mining algorithms and analysed for their attributes such as novelty, relevance, sentiment and volume. In the case of crude oil, there are several ways in which news can influence its price. Certain events, such as announcements of embargoes, riots, natural disasters, oil discoveries or pipeline disruptions, influence future oil demand and supply expectations. The benefits of news analytics have been extensively studied in financial markets since they are strongly influenced by public news. (Feuerriegel & Neumann, 2013, p. 3–4)

2.2 Properties of News Media

News media comes in two different forms mass media and social media. Mass media can be further broken down into text-heavy print media (journals, newspapers, and news magazines) and broadcast media (radio, television, and movies). Social media is made out of many digital platforms (media sharing networks, discussion forums, and blogging networks) on which a large and increasing number of participants can share content in various digital formats (audio, video, photo, and text). (Candia & Mazzitello, 2008, p. 2–3)In theory, any form of news media could deliver predictive insight for a given subject. According to Gan et al. (2020), four primary information sources were considered helpful for researchers in finance: corporate filings, professional news presses, internet message boards, and social media platforms such as Twitter (Gan et al., 2020, p. 5). However, it becomes increasingly challenging to process all available news, particularly non-scheduled news such as social media. With non-scheduled news, separating information from noise and determining its' significance is not a trivial task.

As a result, empirical studies concentrate on specific and recognisable events, such as macroeconomic and geopolitical news from esteemed newspapers. It should also be noted that newspapers are not as easily accessible as other news media since many newspaper publishers demand premiums for their content. One possible way to bypass this limitation is to solely rely on the news headlines since they serve as an appetiser for the reader. Therefore, they are more likely to be accessible (Bai et al., 2021, p. 10). Another reason for focusing simply on the headlines is to further limit the noise from newspapers. The headlines are mostly summaries of the complete content. In addition, the entire articles contain more irrelevant words and repetition than headlines (Khadjeh Nassirtoussi et al., 2015, p. 306). Li et al. (2019), which analysed the impacts of news on crude oil, also focused on headlines. Additional reasons were the convenience of retrieving the information and requiring less storage and computational power to process it. (Li et al., 2019, p. 1550)

2.3 Properties of Crude Oil

Crude oil is vital to the world's economy from an industrial aspect since it is a crucial manufacturing element. However, according to the literature, its price is determined mainly by demand and supply, which are notoriously difficult to forecast. Hamilton (2008, p. 3) even claims that oil price gives the impression to be determined by a random walk with drift. According to Buyuksahin and Harris (2011), the critical factor why crude oil is one the most volatile market in the world is that the demand is affected mainly by exogenous events such as armed conflicts and natural disasters and the existence of speculators such as noise traders. They discovered a significant correlation between crude oil price movements and the behaviour of politically economically unstable countries, which are the leading contributors to such exogenous events (Buyuksahin & Harris, 2011, p. 2).

In addition, Brandt and Gao (2019) pointed out the difficulties of quantifying such news. For example, events like a financial crisis are almost impossible to quantify, despite their significant impact on oil prices. Moreover, adjusting a time series record to such an event is challenging since the explanatory power of such events varies over time (Brandt & Gao, 2019, p. 1). The following figure from the works of Wei et al. (2017) emphasises the volatility of West Texas Intermediate (WTI) crude oil concerning the Global Economic Policy Uncertainty (GEPU) index and global oil demand graphically:

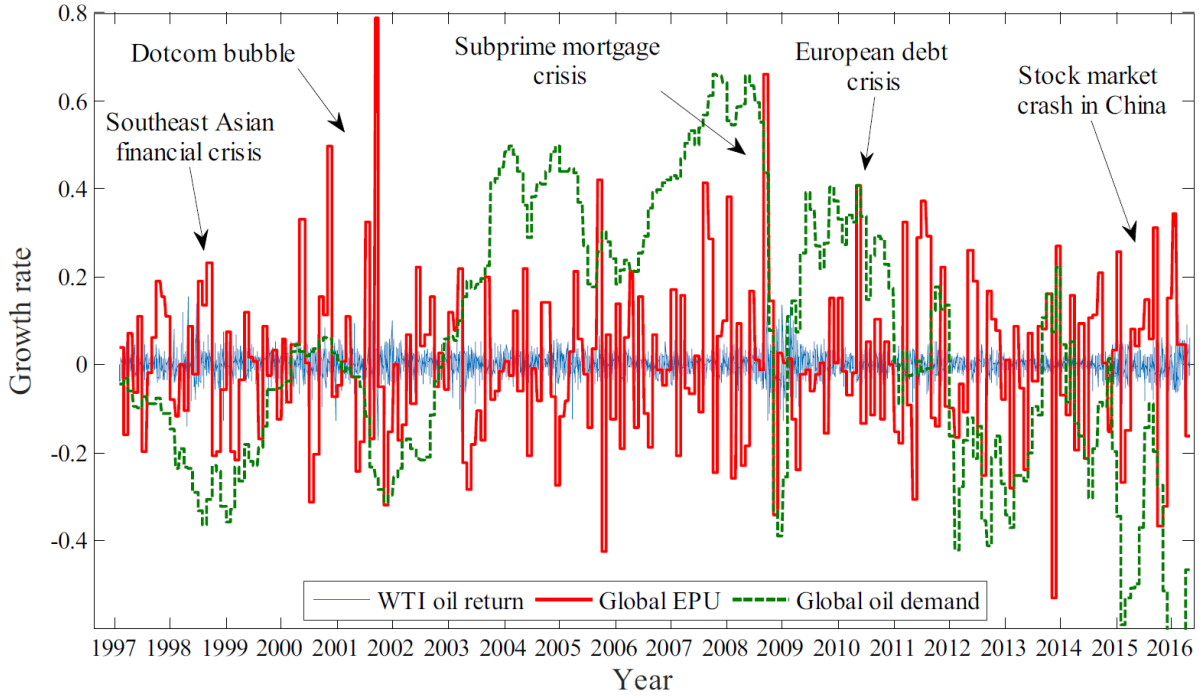

Figure 1: Time evolution of WTI oil returns and growth rates of GEPU and global demand (Wei et al., 2017, p. 5)

The GEPU-index is a GDP-weighted average of national EPU indices for 20 countries: Australia, Brazil, Canada, Chile, China, France, Germany, Greece, India, Ireland, Italy, Japan, Mexico, the Netherlands, Russia, South Korea, Spain, Sweden, the United Kingdom, and the United States of America (Baker et al., 2016, p. 1).

According to Wex et al. (2013), predictions based on sentiment scores of global news from such categories are statistically significant (Wex et al., 2013, p. 1). This conclusion is further supported by the more recent researches of Brandt and Gao (2019). They identified that macroeconomic and geopolitical news is especially influential on crude oil with different effects. For instance, macroeconomic news influences the near-term price changes and predicts the oil price in the long term. On the other hand, the influence of geopolitical news is usually robust and instantaneous, resulting in a greater trading volume. However, it produces no decisive insights in terms of forecasting. (Brandt & Gao, 2019, p.1)

2.4 Sentiment Analysis

Sentiment analysis is a prevalent classification task in natural language processing (NLP). It helps classify the sentimental content and polarities (positive, negative, or neutral) of entire documents, individual paragraphs, and individual sentences. It is becoming increasingly popular due to its enormous potential in various applications in the economy, finance, marketing, political science, psychology, human-computer interaction, and more (Mohammad, 2020). Depending on its application, it differs not only from the domain but also the purpose. Financial sentiment analysis would be the closest to the domain of crude oil since it is a limited publicly traded commodity. The purpose, on the other hand, is dependent on its user and his position. For instance, news about a new oil discovery may come across as “good” news. However, it also conveys a signal of an increase in oil supply, thus causing a negative influence on the oil price. Generally speaking, financial sentiment analysis attempts to guess how the market will react to the presented information. (Xiaodong Li et al., 2014, p. 14)

In practice, texts' contents are not always straightforward due to ambiguity, confusion, exaggeration, irony, slang, and grammatical errors. In other words, it takes a proper "language understanding" to correctly interpret such texts.

Nevertheless, recent improvements in NLP led to the publication of BERT (Devlin et al., 2019), which gave way to remarkable scores on the General Language Understanding Evaluation (GLUE) benchmark[2]. It is one of the main reasons why NLP gained new traction (figure 2).

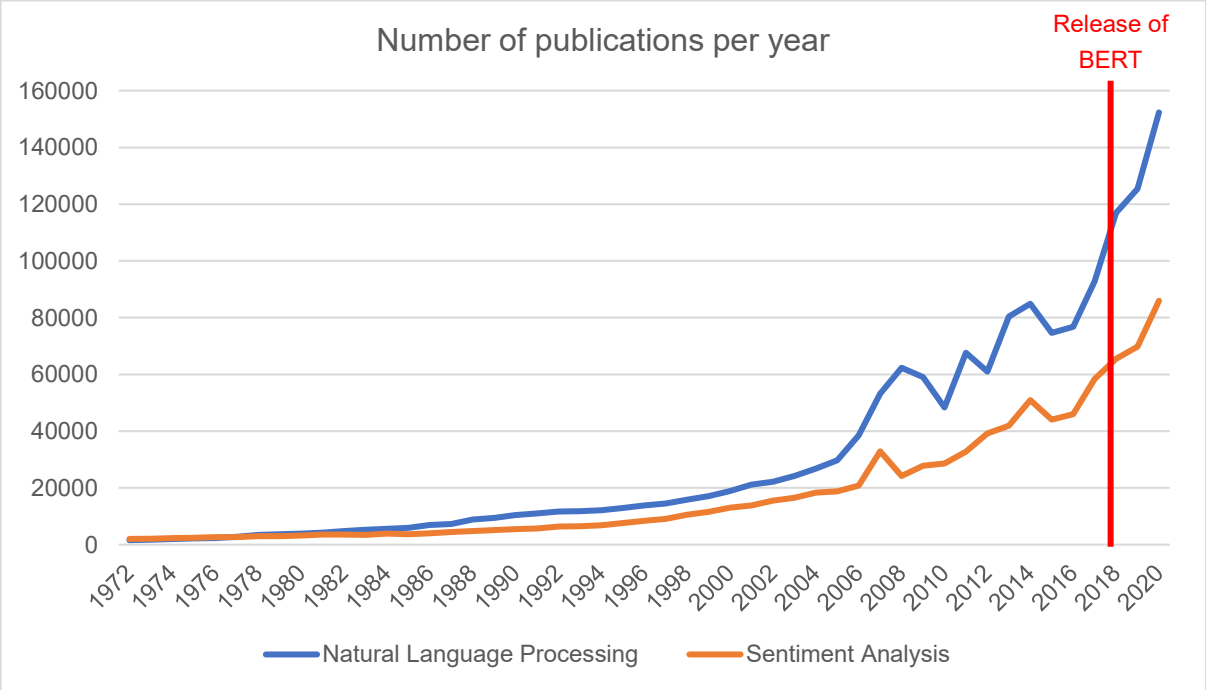

Figure 2: Number of publications published in each year (Source: Dimensions.ai).

[2] https://gluebenchmark.com/leaderboard

2.4.1 Definition of Sentiment

Sentiment in the context of sentiment analysis (also known as opinion mining) refers to a fixed practical value, typically, something good (positive) or bad (negative). Still, there is little discussion about what sentiment in the context of NLP represents (Hovy, 2015, p. 1–2). Generally, researchers assume that authors always express some sentiment while writing since emotions, opinions, and expressions in language are fundamental human traits (Taboada, 2016, p. 2). Most literature tends to break sentiment down to attitudes by means of positive, negative, and neutral polarities, which can be further divided into emotions. Analysing the latter is also known as emotion detection. (Liu, 2012, p. 8)

2.4.2 Definition of Attitude

Attitudes are considered the most basic affective states in natural language, composed mainly by a binary condition in either positive or negative polarity. In many cases, the neutral condition is also introduced and considered as a basic affective state. The classification of attitude is often universal. However, the interpretation of it depends on the domain in which it is used. (W. L. Hamilton et al., 2016, p. 8)

2.4.3 Definition of Emotion

At first glance, the term emotion might sound self-explanatory, yet it describes a complex and multifaceted concept on which scholars have not yet agreed on a single definition (Brandstätter et al., 2013, p. 130). Initial approaches towards the definition of emotions can be traced back to Charles Darwin. He as early as 1872 named numerous emotions such as "anger", "joy", "rage", "fear" in the chapter headings of his work "The Expression of Emotion in Man and Animals". Darwin compared the expression of emotions in humans and non-human primates and discussed the function of emotions in the evolutionary process but did not introduce a taxonomy. Robert Plutchik later took up and further refined this concept (1980). With his psycho-evolutionary thesis that emotions activate behaviours that assure the individual or species' survival, he established the groundwork for linking emotions, behavioural adaptations, and evolution. Plutchik postulated eight primary emotions, naming "fear," "anger," "joy," "sadness," "disgust," "surprise," "acceptance," and "anticipation". The last two were later renamed "trust" and "expectation." These eight emotions are organised bipolarly in the so-called "Plutchik wheel of emotions" (figure 3) with a graded intensity. (Lehmann et al., 2017, p. 7–8)

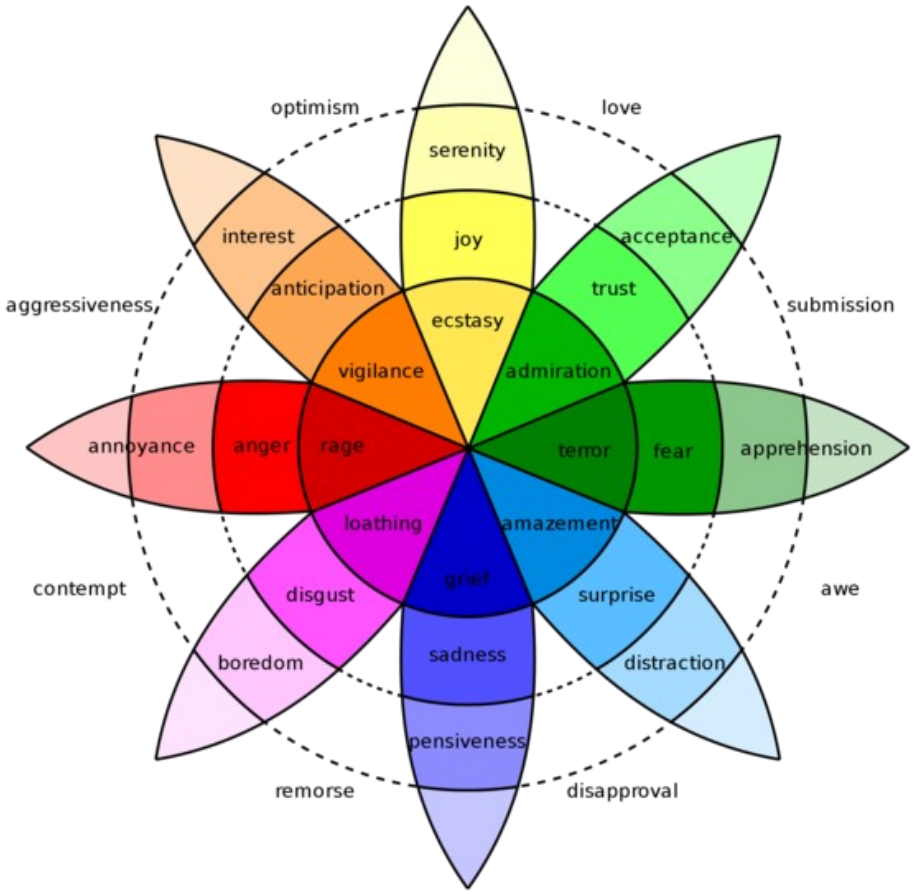

Figure 3: Plutchik wheel of emotions (Chafale & Pimpalkar, 2014)

This concept of basic emotions has been embraced by various researchers (table 1).

Author	#Emotions	Basic Emotions
Ekman	6	anger, disgust, fear, joy, sadness, surprise
Parrot	6	anger, fear, joy, love, sadness, surprise
Frijda	6	desire, happiness, interest, surprise, wonder, sorrow
Plutchik	8	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins	9	desire, happiness, interest, surprise, wonder, sorrow
Matsumoto	22	joy, anticipation, anger, disgust, sadness, surprise, fear, acceptance, shyness, pride, appreciate, calmness, admire, contempt, love, happiness, excitement, regret, ease, discomfort, respect, like

Table 1: List of some existing definitions of basic emotions (Cambria et al., 2012, p. 146)

A more recent approach to machine-readable categorisation of emotions comes from Eric Cambria et al. (2012). Inspired by Plutchik's theories, he introduced a new model, namely the hourglass of emotions (Cambria et al., 2012, p. 148–149). He arranged the primary emotions in so-called “Sentic Levels” in a hierarchical manner within the four independent dimensions "Pleasantness", "Attention", "Sensitivity" and "Aptitude" (table 2).

Sentic Level	Pleasantness	Attention	Sensitivity	Aptitude
+3	ecstasy	vigilance	rage	admiration
+2	joy	anticipation	anger	trust
+1	serenity	interest	annoyance	acceptance
0	-	-	-	-
-1	pensiveness	distraction	apprehension	boredom
-2	sadness	surprise	fear	disgust
-3	grief	amazement	terror	loathing

Table 2: The sentic levels of the hourglass model (Cambria et al., 2012).

Unlike the previous models, the hourglass of emotions could map *all* emotional states using the sentic level, thus providing a machine-readable framework for detecting emotions. Nonetheless, Cambria and his team recently revisited the hourglass of emotion (Susanto et al., 2020) by rearranging certain emotions and modifying the four dimensions, leading to the release of SenticNet 6 (Cambria et al., 2020). The differences can be seen in figure 4.

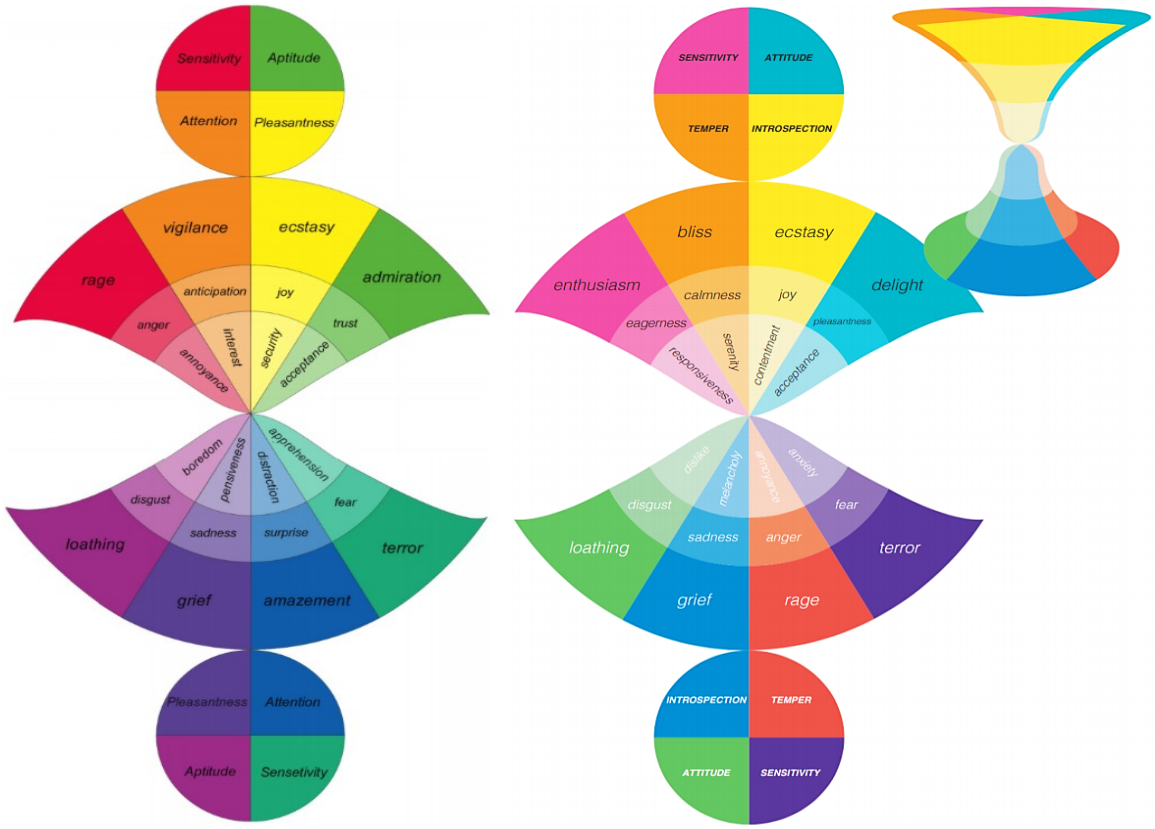

Figure 4: The old (left) and revisited (right) version of the Hourglass of Emotions Model (Susanto et al., 2020).

In general, the taxonomy of emotions in terms of their numbers and polarity is hard to generalise primarily because the interpretation of emotions is based on categories from conventional language. There is no stringently verifiable counterpart in the non-linguistic reality. (Lehmann et al., 2017, p. 9)

2.5 Evolution of applied Sentiment Analysis

One of the earliest works on NLP dates back to 1954 when researchers from IBM collaborated with Georgetown University and demonstrated machine translation from Russian to English. This demonstration led to a great deal of public interest in the field of NLP (Frederking, 2004). From such humble beginnings, NLP developed into a fruitful field of research that expanded to include several natural language understanding tasks. One prevalent NLP task is arguably sentiment analysis, the classification of sentimental and other affect states in natural language using computer algorithms. The main bottleneck in the performance of sentiment analysis, as in most NLP tasks, is the challenge of representing natural language in a machine-readable yet meaningful format. Over the years, many efforts have been proposed to address this issue.

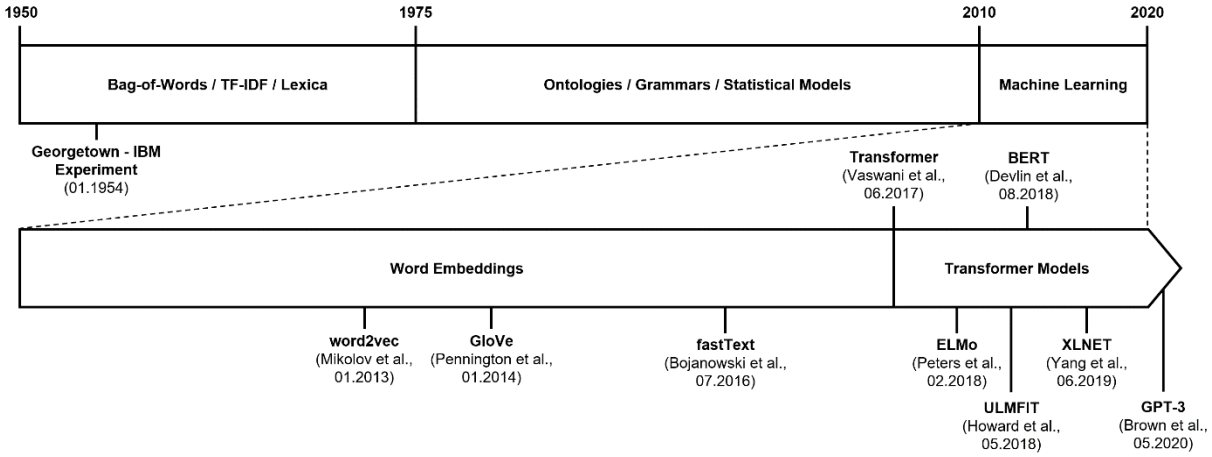

Figure 5: Evolution of applied sentiment analysis

In the Bag-of-Words (BOW) methods, the classification relied on annotated dictionaries. In comparison, machine learning methods rely on semantically meaningful word embeddings. With advances in NLP algorithms, one promising method for this matter was the use of contextualised word embeddings generated through transformers, which was first proposed by Vaswani et al. (2017). Building partly on the architecture of the transformer, Devlin et al. (2019) introduced the Bidirectional Encoder Representations from Transformers (BERT), a famous and influential NLP-tool

2.5.1 Bag-of-Words Approach

One of the first approaches to recognising sentiment in texts was the Bag-of-Words (BOW) methodology, often known as the Lexicon-based technique (Liew, 2016, p. 47). Because a text consists of several words (also known as tokens), every natural language processing model must first determine how to represent a single token. In the case of Bag-of-Words (BOW), it simply "counts" the positive and negative words that occur in the text and thus arrives at meaningful results for simple texts. As the name suggests, these methods utilise a lexicon consisting of words and their sentimental value, predetermined by multiple human annotators.

However, the capability to extracting all of the keywords or word combinations from a sentence correctly and have them appear in a lexicon is a challenge. One way to address this challenge is a semantic analysis of the syntactic composition of the sentences. Another possible solution is to use a brute-force approach. Possible combinations and permutations of n-grams in sentences are searched for that correspond to a Multi-Word-Concept (Biagioni, 2016, p. 40). In terms of sentiment analysis in finance, a popular and recent lexicon is the Loughran-Macdonald-dictionary containing words assigned to values such as "positive" or "uncertain. However, according to various efforts, such as the ones from Malo et al. (2014), the overall accuracy with around 60 % of such techniques is unsatisfactory in finance. (Malo et al., 2014, p. 1)

2.5.2 Machine Learning Approach

With the advancements in technology, artificial intelligence (AI) allowed machines to understand the semantics and, more recently, the context of words in natural language. These advancements enabled it to overcome the limitations of the BOW approach. A well-known subfield of AI is Machine Learning (ML), which is further divided into Deep Learning (DL), an increasingly popular field of research (Evans et al., 2017, p. 2).

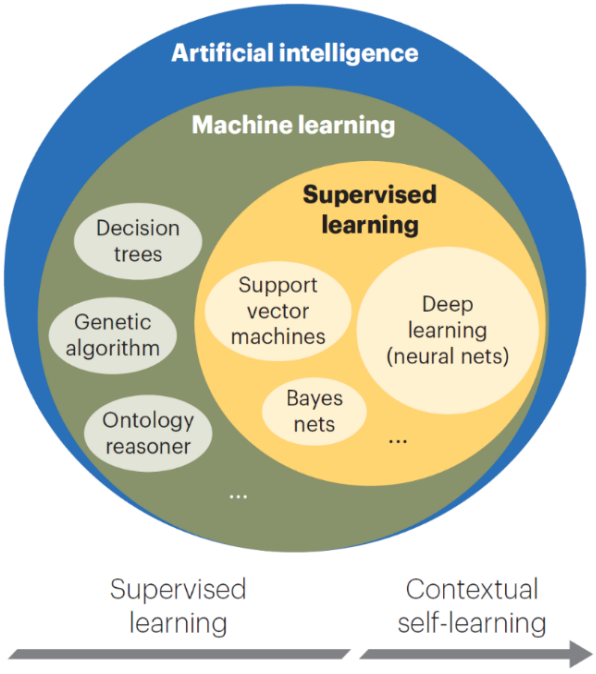

Figure 6: AI overview taken from (Evans et al., 2017, p. 2)

Initial ML approaches were mainly based on *supervised learning*, in which the learning progressed by training on annotated datasets, consisting of pairs of inputs and the corresponding solutions. As a result, the system could derive an applicable rule or pattern from the data by correcting and optimising itself based on the solution. The types of tasks relying on supervised learning can be narrowed down to solving classification and regression problems. (Chollet, 2018, p. 129 - 130)

In the case of sentiment analysis, supervised learning mainly relied on Recurrent Neuronal Networks (RNNs) and Long Short-Term Memory (LSTM) since these are suitable for sequential data (Tang et al., 2016). However, there were also promising approaches, such as the works from Zhang (2016) involving *Convolutional Neural Networks* (CNN), which formerly delivered remarkable outcomes in computer vision tasks. One downside of supervised learning, specifically for classification problems such as sentiment analysis, is the training dataset itself. A larger training dataset generally produces better performance, but the usage of large training datasets is often expensive (computationally and monetary) and not always obtainable.

2.5.3 Word Embeddings

One potential solution to lower the requirements of large training datasets would be the usage of word embeddings. Word embeddings (also known as word vectors) attempt to represent words meaningful and machine-readable. One of the most well-known word vector models is word2vec from Mikolov et al. (2013). Here, the words w are encoded in a vocabulary V in form of one-hot-vectors $\epsilon\{0,1\}^{|V|}$. This is achieved by unsupervised training of a feed forward neural network. In the case of word2vec, there are two training models, the Continuous Bag-of-Words (CBOW) and Skip-Gram, both with a single embedding layer based on the co-occurrences of the words concerning the surrounding words. Consequently, the embedded representation $x_w \in \boldsymbol{R}^{d+1}$ where the words with semantically similar meanings are arranged closer to each other within the embedding matrix W. Like the BOW approach, this embedding matrix can then be used as a lookup table. (Mikolov et al., 2013 p. 1 - 6)

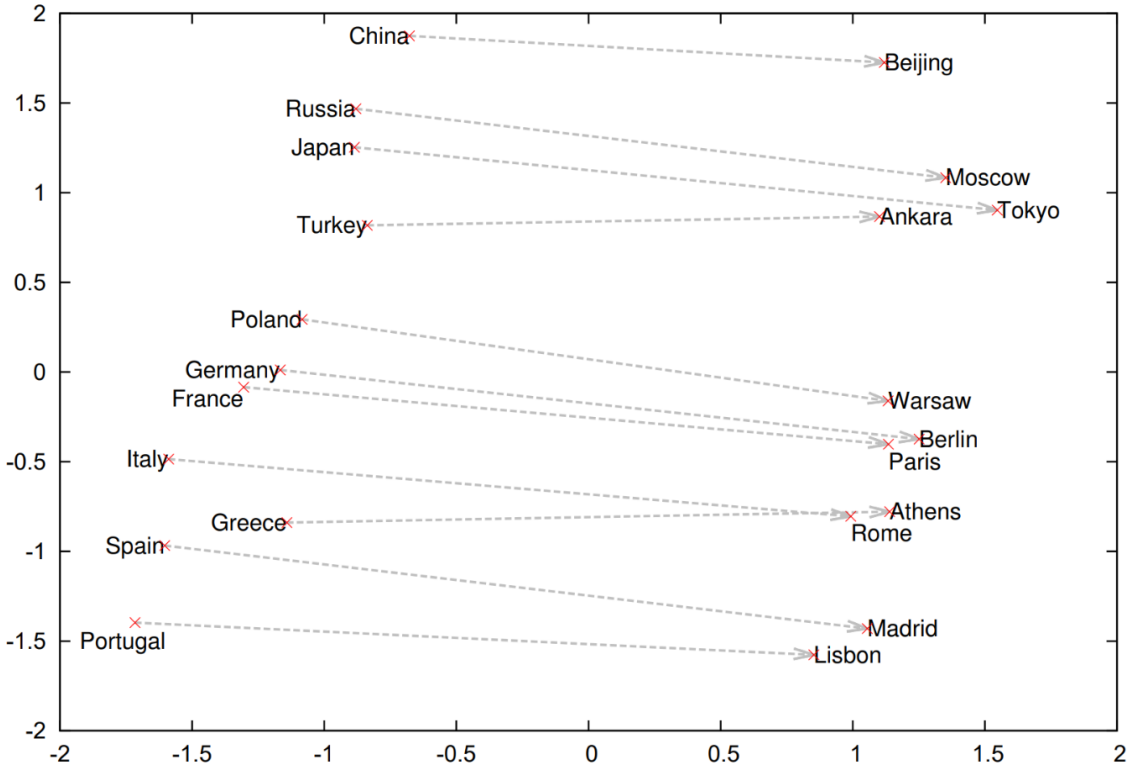

Figure 7: Two-dimensional PCA projection of vectors of countries and their capital cities (Mikolov et al., 2013 p. 5)

As a result, it can automatically organise words (figure 8) without the usage of annotated training dataset and enable arithmetic operations such as:

$$Tokyo - Japan + Turkey \approx Ankara \tag{1.1}$$

At this stage, it should be pointed out that word embeddings provide only semantic representations of words that are not sufficient for classification tasks. Therefore, they still require a training dataset. However, the advantage of word embeddings is that the model is already equipped with "prior knowledge", leading to a steeper learning curve and reaching their potential faster. Therefore, it can achieve good accuracy by training on a relatively more minor training dataset. According to various publications, the achieved accuracies with similar word embeddings such as GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2016) are around 85%.

2.5.4 Contextual Word Embeddings

It should also be noted that the earlier mentioned word embeddings in chapter 2.5.3 might capture the semantics of words in a corpus. However, once captured, they remain static. Thus, a model relying on word embeddings such as word2vec is not aware of the context of the words, which is primarily determined by the surrounding words. The following two sentences illustrate the limitations of such static word embeddings:

Jamie would like to visit Turkey next summer.
Jane collects the eggs of her favourite turkey.

In a static word embedding, the vector of the word "turkey" would have the same vector in both cases. In this case, it would be placed either near countries or near birds within the embedding.

Depending on the domain on which the word embedding was generated, the word apple is arranged differently. In this case, it is either arranged close to other fruits or close to other corporations within the embedding. For example, with a static word embedding the word vector, “apple” would yield the same values.

Contextual word embeddings, on the other hand, aim to differentiate words by not only relying on the target word w as an input but also consider the subset of the context c which are surrounding the target word mapping to a context-specific x_w vector:

$$\left(c_{previous}, w, c_{subsequent}\right) \mapsto (x_w) \in X^{d+1} \tag{1.2}$$

These could allow the model to distinguish the word apple by the words surrounding it since it contains the target two times with different vectors x_1 and x_2 in the embedding. (Yenicelik, 2020, p. 16)

There are various models for producing such contextual embeddings. For instance, Peters et al. (2018) proposed a relatively simple model to produce such embeddings, known as Embeddings from Language Models (ELMo).

Here, the language model refers to predicting the next word in a given piece of text. It achieves this by using various statistical and probabilistic techniques based on the sequence of words occurring in a sentence. These models are generally trained on very large corpora to capture many varieties of word sequences. In the case of ELMo, the model learns how to weigh distinct representations from different LSTM layers such that one contextualised vector per token can be calculated. Word embeddings may be calculated for any piece of text using ELMo's pre-trained weights. Once the contextualised representations have been retrieved, they may be utilised for most other NLP tasks, which surprisingly performed remarkably well compared to static word embeddings such as word2vec. (Peters et al., 2018, p. 2227 - 2228)

For instance, it achieved state-of-the-art performance on the Stanford Sentiment Treebank (SST) benchmark for fine-grained classification tasks (McCann et al., 2018, p. 1). As a matter of fact, this discovery that a model trained for language modelling can be effectively fine-tuned for most NLP applications, including classification, with modest alterations is one of the most critical findings in natural language processing. Although ELMo uses pre-trained language models to contextualise representations, the information collected using a language model is only present in any model's first layer, making it difficult to fine-tune task-specific NLP models.

Howard and Ruder (2018) introduced the transfer learning model Universal Language Model Fine-tuning for Text Classification (ULMFit) to solve this limitation. Novel training strategies that can pre-train the language model on domain-specific corpora allowed it to outfit it with task-specific layers, which, unlike ELMo, allows fine-tuning the model. For this, they involved domain-specific pre-training of the language model, reasoning that data originates from a different distribution for a specific task than the general corpus on which the initial model was built. (Howard & Ruder, 2018, p. 1)

However, both ELMo and ULMFit are RNN-based approaches. RNNs suffer from vanishing and exploding gradients, making them unsuitable for long texts. Besides, RNNs require input data from the previous state, which must be passed sequentially to allow operations for the current state. Since such a sequential flow is incompatible with parallel processing, RNNs are slower to train than other neural networks.

2.5.5 Transformer Model

The efforts behind pre-training through language modelling resulted in substantial improvements in a variety of NLP tasks. However, the activities behind machine translation, which delightfully also happens to be a foundational block in NLP dating back to the Georgetown – IBM experiment in 1954, laid the groundwork for today's state-of-the-art performance of NLP models. Machine translation was formerly achieved through BOW, and at later stages, it utilised RNN encoder-decoder, which Cho et al. (2014) proposed. However, as mentioned in chapter 2.5.4, RNNs have limitations. Bahdanau et al. (2016) addressed these partly by introducing the attention mechanism. In the context of neural networks, the attention mechanism is a strategy that simulates the cognitive process of selectively focusing on a distinct component of information while disregarding other perceptible information. It emphasises the relevant bits of the input data while fading out the rest, forcing the neural network to focus more computer resources on minor but relevant portions of the data (Bahdanau et al., 2015, p. 4).

Despite its novelty, there are numerous academic works of literature regarding the attention mechanism. The works of Weng (2018) provides a summary of several attention mechanisms:

Name	Function	Source
Content-based	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \text{cosine}[\boldsymbol{s}_t, \boldsymbol{h}_i]$	(Graves et al., 2014)
Additive (Concat)	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[\boldsymbol{s}_t; \boldsymbol{h}_i])$	(Bahdanau et al., 2015)
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a s_t)$	(Luong et al., 2015)
General	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \mathbf{W}_a \boldsymbol{h}_i$	(Luong et al., 2015)
Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \boldsymbol{s}_t^\top \boldsymbol{h}_i$	(Luong et al., 2015)
Scaled Dot-Product	$\text{score}(\boldsymbol{s}_t, \boldsymbol{h}_i) = \frac{\boldsymbol{s}_t^\top \boldsymbol{h}_i}{\sqrt{n}}$	(Vaswani et al., 2017)

Table 3: Summary of several popular attention mechanisms (Weng, 2018):

In addition to academic works, certain organizations provide a visually appealing and accessible explanation of the attention mechanisms. For example, the following explanation is partly derived from the research blog posts and educational videos from RASA[3] and PELTARION[4].

[3] https://blog.rasa.com/tag/research/
[4] https://peltarion.com/blog/data-science/

The following sentence will be used as an example to explain the simple attention mechanism:

Collect the turkey eggs.

In a well-trained word embedding, the word “turkey” should be present two times, one close to countries and the other close to birds. In order to determine the correct version of “turkey”, it could use the attention mechanism by utilising the dot product of the surrounding vectors. The following illustration shows the simplified process of self-attention for the first word:

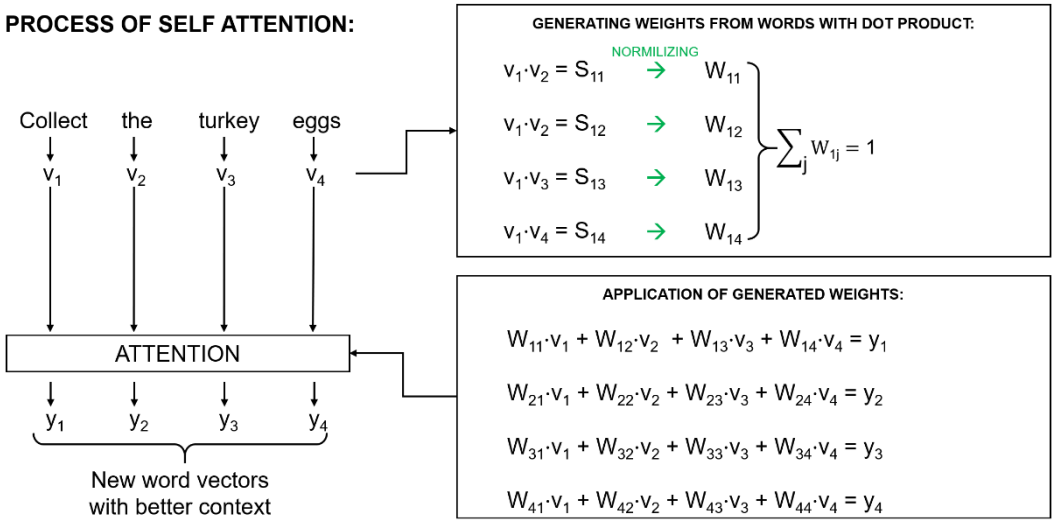

Figure 8: Simplified process of self-attention mechanism

Here, steps involving the generation of weights are repeated for each word in the sentence. The dot product of each word should result in 16 weights, which could be used to calculate the output vectors with a possibly higher context. By calculating the dot product of the two vectors of “turkey” and “eggs”, some of their vector values should be in accordance with each other. For instance, they could be linked with subjects like nature, animals, and other similarities. The two vectors v_3 and v_4, which in theory, should emphasise the linked vector values with a relevance score S_{34}. This score can then be used as a weight W_{34} allowing the neural net to pay attention to such contingencies and assign context to related words. (Rasa, 2020)

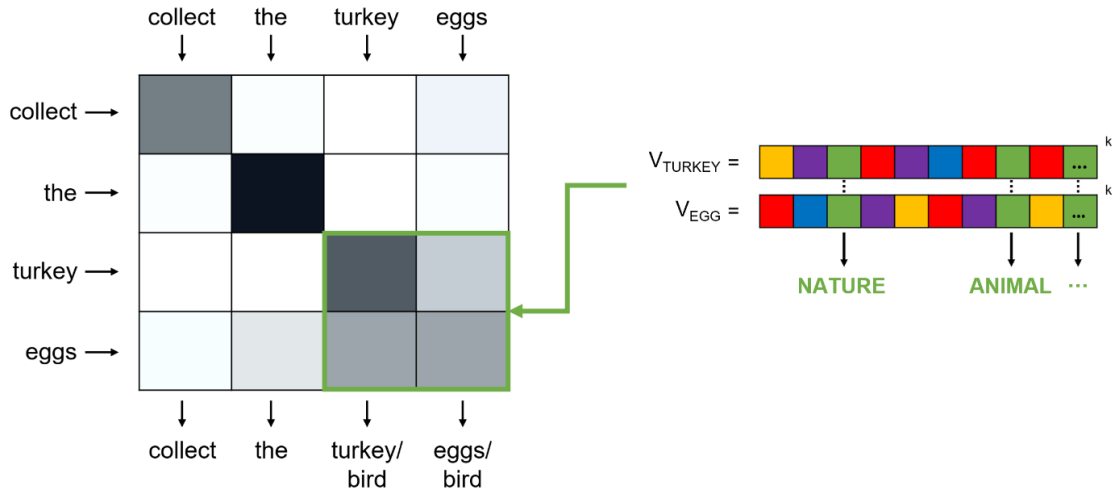

Figure 9: Visualization of the comparatively high relevance between the word vectors of “turkey” and “eggs”

Ideally, the final output made by this mechanism should produce word vectors with a higher context in which the word “turkey” is further linked to birds in this case. The main advantage here is that the matrix multiplications are not required to be processed sequentially, as in RNNs, meaning it can be done in a parallelised manner with a faster training speed. (Futrzynski, 2020)

However, in this attention mechanism, there is no absolute control about the selection of the weights themselves. This lack of control could result in paying attention to less relevant aspects of the sentence. In the works of Vaswani et al. (2017), the self-attention function was enhanced by the mapping of additional abstract weight vectors, namely values V and a set of pairs consisting queries Q and keys K. They allow to extract different contextual components of any given input word. With these extractions, certain words can give each other even more context. The outcome is a weighted sum of the values, with the weight assigned to each value calculated by the query's compatibility function with the relevant key. (Vaswani et al., 2017, p. 4)

However, similar to the BOW approach, this attention mechanism has no awareness of the position of each word. In order to address this issue, Vaswani et al. introduced another vital component in their transformer architecture, namely the Positional Encoding. The positional encoding returns information of the position of a word in a sentence by generating a position embedding derived from the cyclic nature $\sin(x)$ and $\cos(x)$ of each word x. This position embedding is added to the word embedding (figure 10).

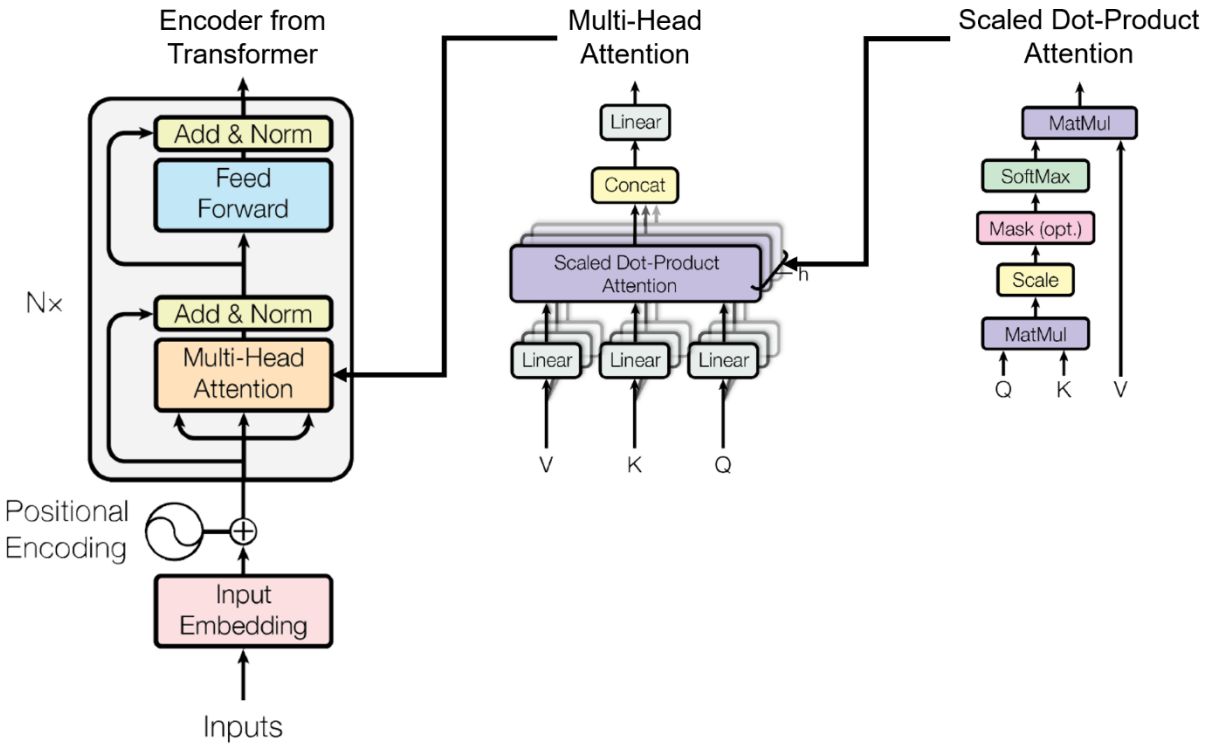

Figure 10: Components of the Multi-Head Attention. (Vaswani et al., 2017)

Later, this whole attention process, or layer, can be done multiple times, also known as Multi-headed Attention (figure 11). It is a simple yet clever and important innovation in NLP. This architecture allows the same network to learn different Q, K, and V matrices for different semantic meanings of attention (grammar, vocabulary, conjugation, etc.). (Futrzynski, 2020)

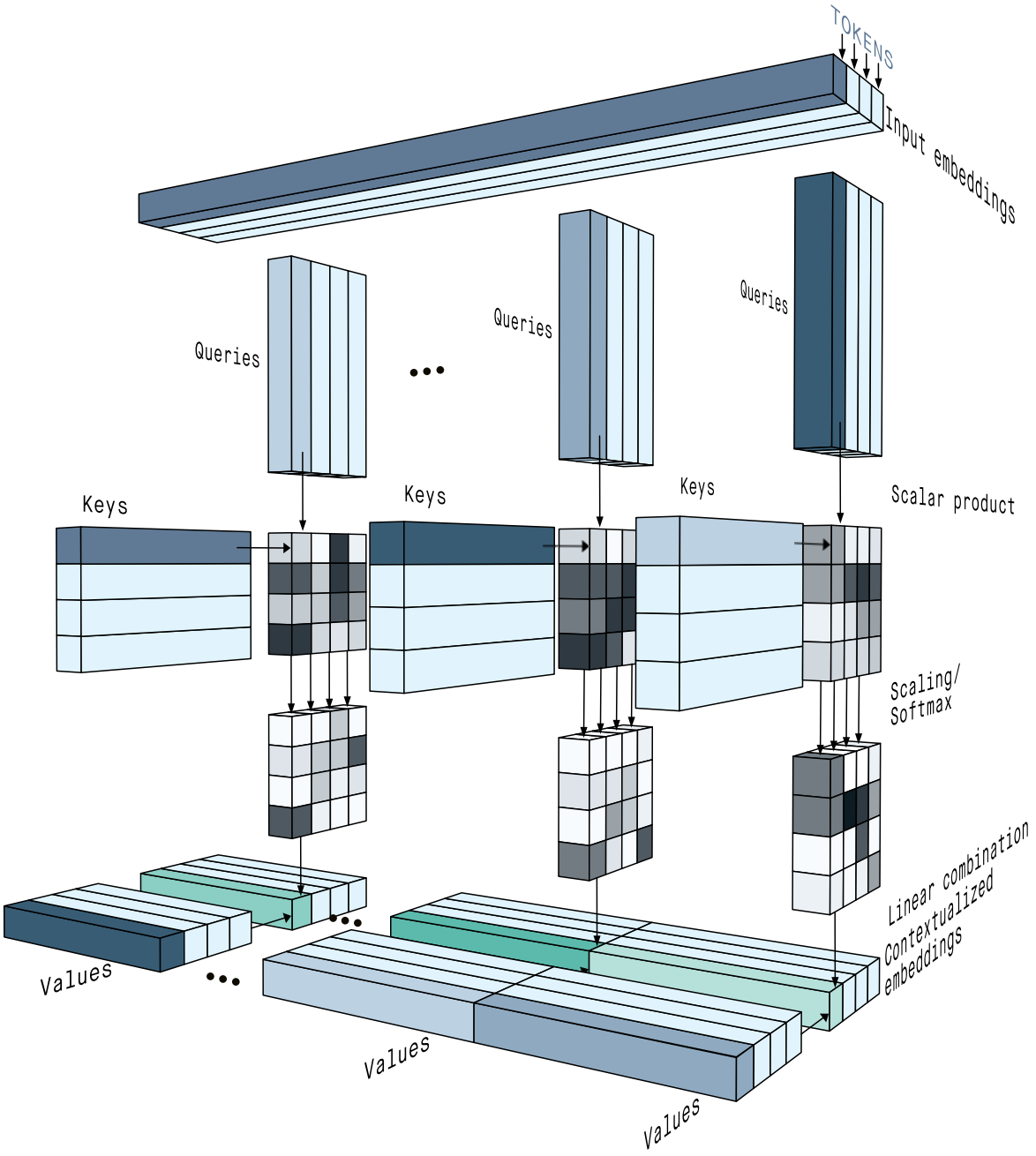

Figure 11: Visualization Multi-headed self-attention (Futrzynski, 2020)

In summary, the transformer is a positionally aware, attention-based architecture that allows parallelised matrix multiplications making it much faster than RNNs. Initially, transformer architecture was proposed for neural machine translation. Therefore, it contains an encoder and decoder. The encoder is a fully connected feed-forward network made out of multiple identical multi-headed attention layers. The multi-headed attention layer allows the sequence to be evaluated from many varying "perspectives". (Araci, 2019, p. 3)

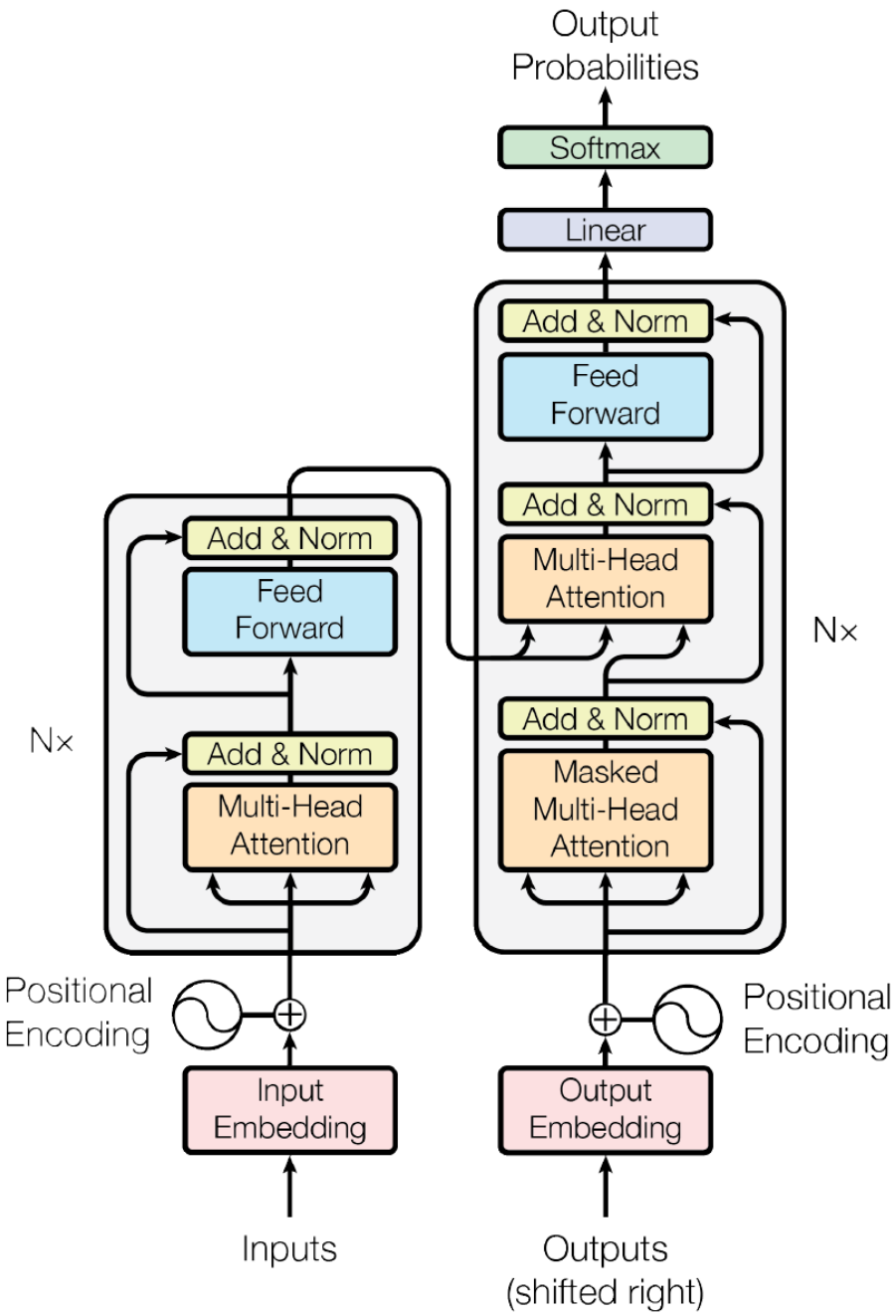

Figure 12: The Transformer Architecture with the encoder (left) and decoder (right) (Vaswani et al., 2017, p.3)

The original paper Attention is all you need, by Vaswani et al. (2017), the works of Weng (2018) and the publications from Rasa (2020) and Futrzynski (2020) provides a more detailed treatment of the above-shown processes.

2.5.6 BERT Model

Shortly after the release of the transformer architecture, Devlin et al. (2019) discovered that the encoder, when stacked, can also serve as a powerful representation learning model. Hence the name *Bidirectional Encoder Representations from Transformers* (BERT). There are two versions of BERT, namely BERT-base and BERT-large. BERT-base with 12 encoder layers, hidden size of 768, 12 multi-head attention heads and 110M parameters. BERT-large with 24 encoder layers, hidden size of 1024, 16 multi-head attention heads, and 340M parameters. (Yenicelik, 2020, p. 24)

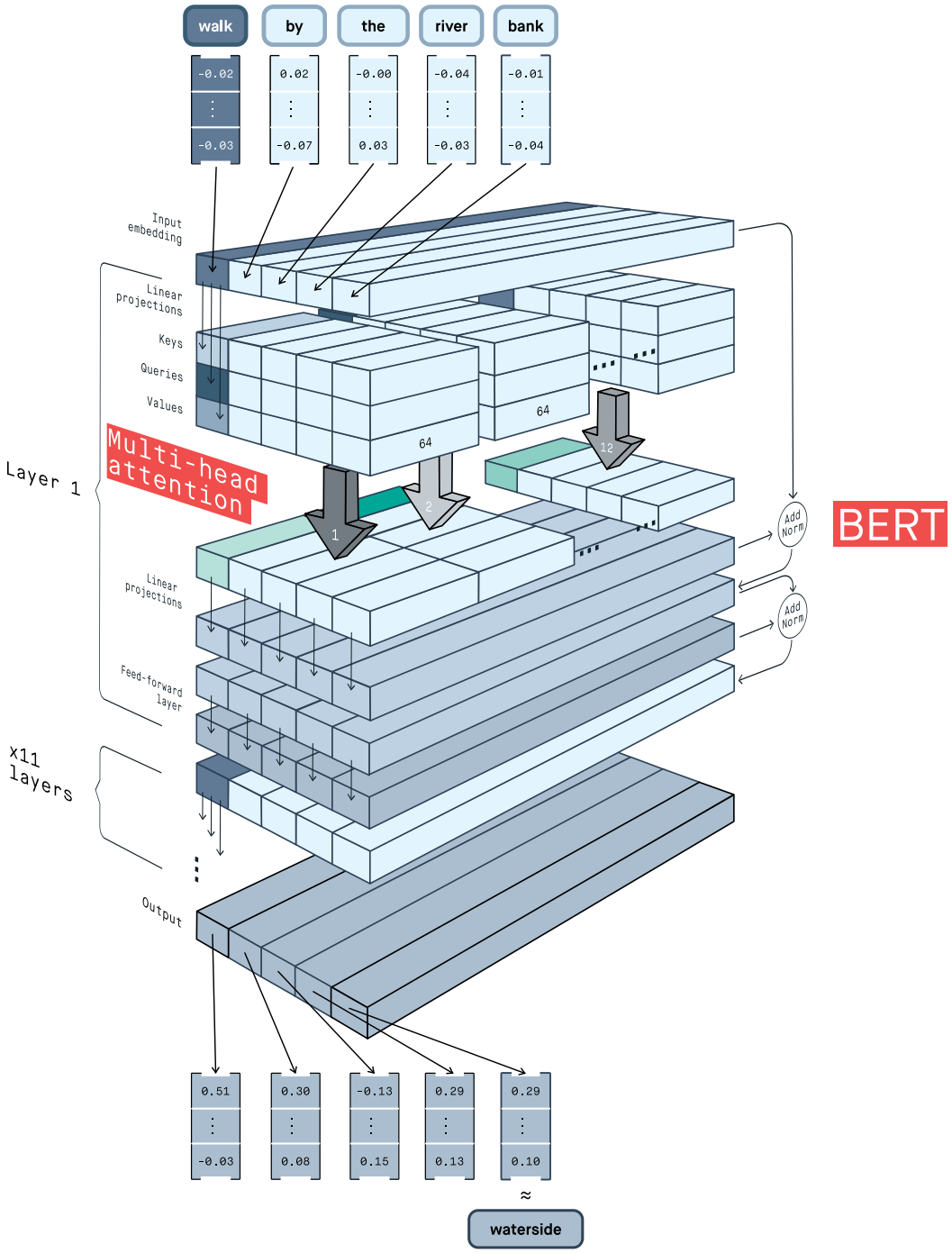

Figure 13: Visualization of the Token Encoder of the BERT-base Version (Futrzynski, 2020)

Besides BERT, there are also other prominent language models which are utilizing transformers, such as OpenAI's General Purpose Transformer (GPT) (Radford et al., 2018) and the more recent GPT-2 (Radford et al., 2019). In contrast to BERT, which is built on stacked encoders, GPT uses a stack of decoders. Aside from that, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach (Devlin et al., 2019, p. 13).

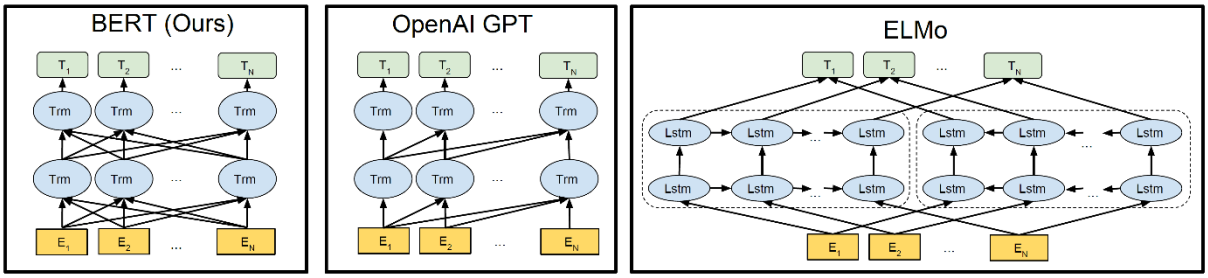

Figure 14: Differences in pre-training model architectures (Devlin et al., 2019, p. 13)

Unlike the works of Vaswani et al., where pre-training was achieved through language translation, BERT was pretrained through masked token prediction task, where 15 % of words were masked. The second pre-training was, similar to ELMo, based on language modelling. However, instead of predicting the next word, its task was to predict whether two given sentences follow each other. (Devlin et al., 2019)

2.5.7 FinBERT Model

One key feature of BERT is the possibility of customizing it for a large variety of NLP tasks. The initial pre-training of BERT is done on English Wikipedia and the BookCorpus (Zhu et al., 2015) to provide the model with a general understanding of the natural language. This model can be further adapted to the targeted domain, for instance, sentiment analysis of news, such as FinBERT (Araci, 2019). The primary purpose of FinBERT is to analyse the sentiment of the financial text. To achieve this, Araci (2019) used a subset of Thomson Reuters Text Research Collection (TRC2)[5] to adapt the model on the domain of financial news, a domain in which slang and spelling errors are rare. The training dataset Financial PhraseBank from Malo et al. (2014) was used for fine-tuning purposes. (Araci, 2019, p. 2)

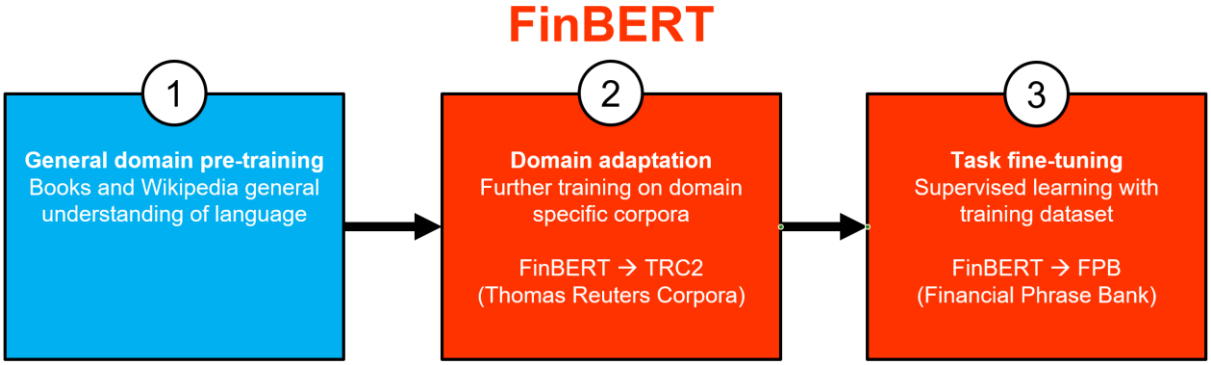

Figure 15: Process of generating FinBERT

[5] https://trec.nist.gov/data/reuters/reuters.html/

3 Research Objective

Investigating the influence of news sentiment could help forecast crude oil prices, thus providing potential value to economic, finance, and political science researchers. The key findings of the literature review are as follows:

- There is strong opposition against the strong form of the EMH. Thus, fundamental analysis such as news analytics can be advantageous for predicting crude oil price changes (chapter 2.1).
- News headlines could serve as an accessible, beneficial and feasible resource for carrying out such fundamental analysis (chapter 2.2).
- There is evidence that sentiment analysis from news delivers a statistically significant impact for predicting the price of crude oil (chapter 2.3).
- Financial sentiment analysis could be a promising domain for crude oil. Thus, a binary (positive, negative, and neutral) sentiment classification should be sufficient (chapter 2.4).
- With the advancements in NLP, pre-trained language models delivered high accuracies of around 95% in the field of finance (chapter 2.5). Therefore, deep learning-based sentiment analysis could be considered reliable for binary classification of news.

To the best of our knowledge, utilizing deep learning-based sentiment analysis on news headlines for predicting crude oil price movements has not been proposed in the literature (though the individual steps are not novel). When considering the findings above, the following research question arises:

Can deep learning-based sentiment analysis of news headlines help to predict price movements of crude oil?

The research question is then further divided into the following questions, including the chosen method of research:

Questions	Research method
Which platforms can provide archived headlines dating back decades, with high frequency and relevance to crude oil?	Literature Review Experiments
Which pretrained text classification model is most suitable for analysing headlines related to crude oil, and is there any relationship?	Literature Review Experiments
Are news headlines alone rich enough in content to provide insights into the price development of crude oil?	Literature Review Experiments

Table 4: Questions with the appropriate research methods

4 Methodology

The following chapter covers the available hardware and software resources, the process of collecting, preprocessing and merging news headlines and WTI crude oil prices, and finally, a treatment of suitable deep learning frameworks for the application of sentiment analysis.

4.1 Available Resources

For the implementation and evaluation of the experiments, a personal computer from the author with the following configuration is available:

- **CPU**: AMD Ryzen 9 5950X 16-Core Processor
- **GPU**: Nvidia GeForce RTX 3070 with 8 GB VRAM
- **RAM**: 32 GB
- **Storage**: 1 TB NVMe SSD

As the operating system, Windows 10 Pro was being used. The codes were written in Python's programming language and executed through the integrated development environment (IDE) Jupyter Notebook[6]. The bulk of the data analysis and data visualisation was conducted through the Python library Pandas[7]. However, a minor part of the was carried out in MS Excel for convenience reasons. The relevant Jupyter Notebook files are available on GitHub[8].

4.2 Sources for News Headlines

As mentioned in chapter 2.2, news headlines could serve as an insightful and accessible resource. For this matter, a variety of platforms were examined. The priorities for determining the platform were providing a source of frequent, possibly daily, unique headlines with high relevance to crude oil, reaching back a few decades to the past.

4.2.1 The New York Times Developer Network

The New York Times provide access to their archive for free with user-friendly API's. In this case, the archive API9 delivers monthly arrays of New York Times stories from 1851 (NYT Developers, n. d.) However, since this approach limits the source to only one news publisher, it is doubtful whether it can provide daily news for crude oil. Therefore, a platform representing multiple publishers should be favoured. Nevertheless, despite this limitation, it is still a noteworthy platform when considering the far-reaching period.

[6] https://jupyter.org/
[7] https://pandas.pydata.org/
[8] https://github.com/Captain-1337/Master-Thesis/
[9] https://developer.nytimes.com/docs/archive-product/1/overview/

4.2.2 Wharton Research Data Services

Wharton Research Data Services is a web-based data research service utilized by academic, government, and non-profit organizations worldwide. It aims to provide its users with various research databases through a common interface on a single location. For example, one of its research databases, Capital IQ Key Developments[10], provides structured summaries of news and events for more than 800'000 companies worldwide going as far back as 1964. (Wharton Research Data Services, n. d.)

However, a query search in which headlines should contain at least one of the following terms, "crude oil", "diesel", "gasoline", "kerosene", "petrol" and "petroleum" (figure 16) resulted in around 3'000 hits, including many duplicates. Such a low number of unique hits for such a long period cannot provide sufficiently frequent coverage needed to track changes in the sentiment in the news. It should also be noted that the usage of these services is for academic and non-commercial research purposes only.

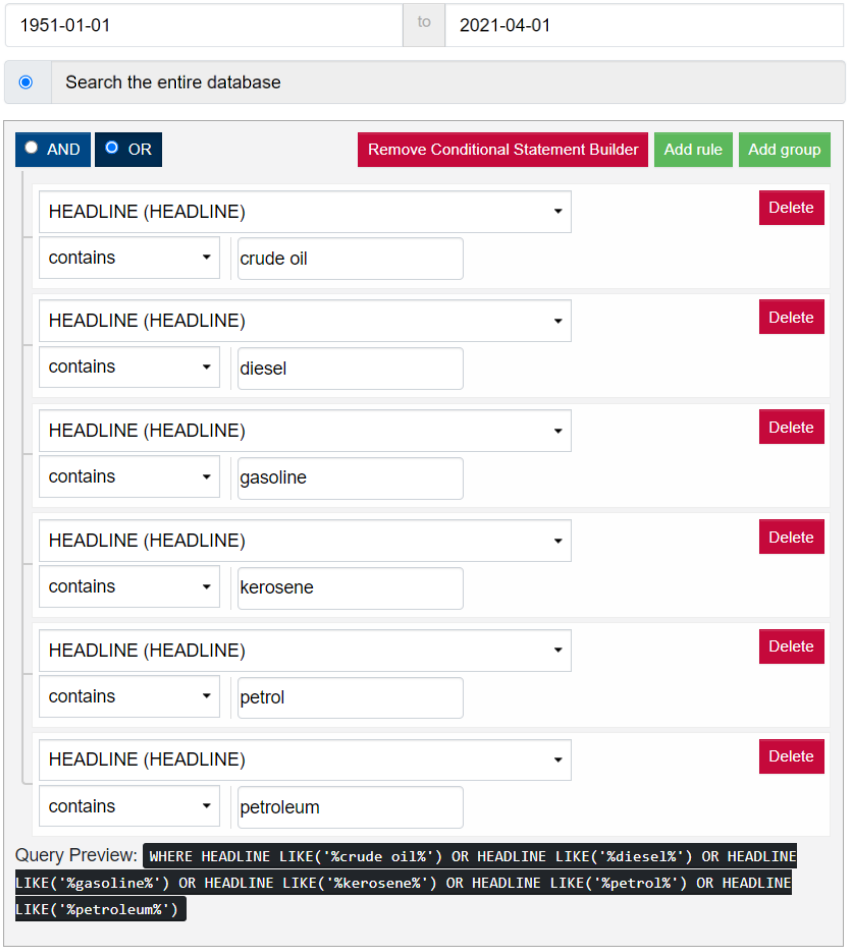

Figure 16: Screenshot of WRDS Capital IQ Key Developments displaying the selected filters

[10] https://wrds-www.wharton.upenn.edu/pages/get-data/compustat-capital-iq-standard-poors/capital-iq/

4.2.3 Nexis Uni Academic Search Engine

LexisNexis is a company that offers data mining platforms, computer-assisted legal research, and information on large groups of customers around the globe through online portals. The academic variant of LexisNexis called Nexis Uni[11], which the author's institute has access to, offers more than 17,000 news, business, and legal sources dating back to 1790. In addition, it allows to search across all content types in a single search and deliver unlimited search results. (Nexis Uni | Designed for Collaborative Research, n. d.)

In order to find news related to crude oil, the data types were limited to only news, and the query search included the terms “crude oil”, “diesel”, “gasoline”, “kerosene”, “petrol”, and “petroleum”. This search yielded several hundred thousand hits (figure 17). Despite the filtering option of Nexis Uni, there were many duplicates present, which would require an additional preprocessing step. However, the main bottleneck was the cumbersome retrieval of the data since there is a limit of 100 articles per download.

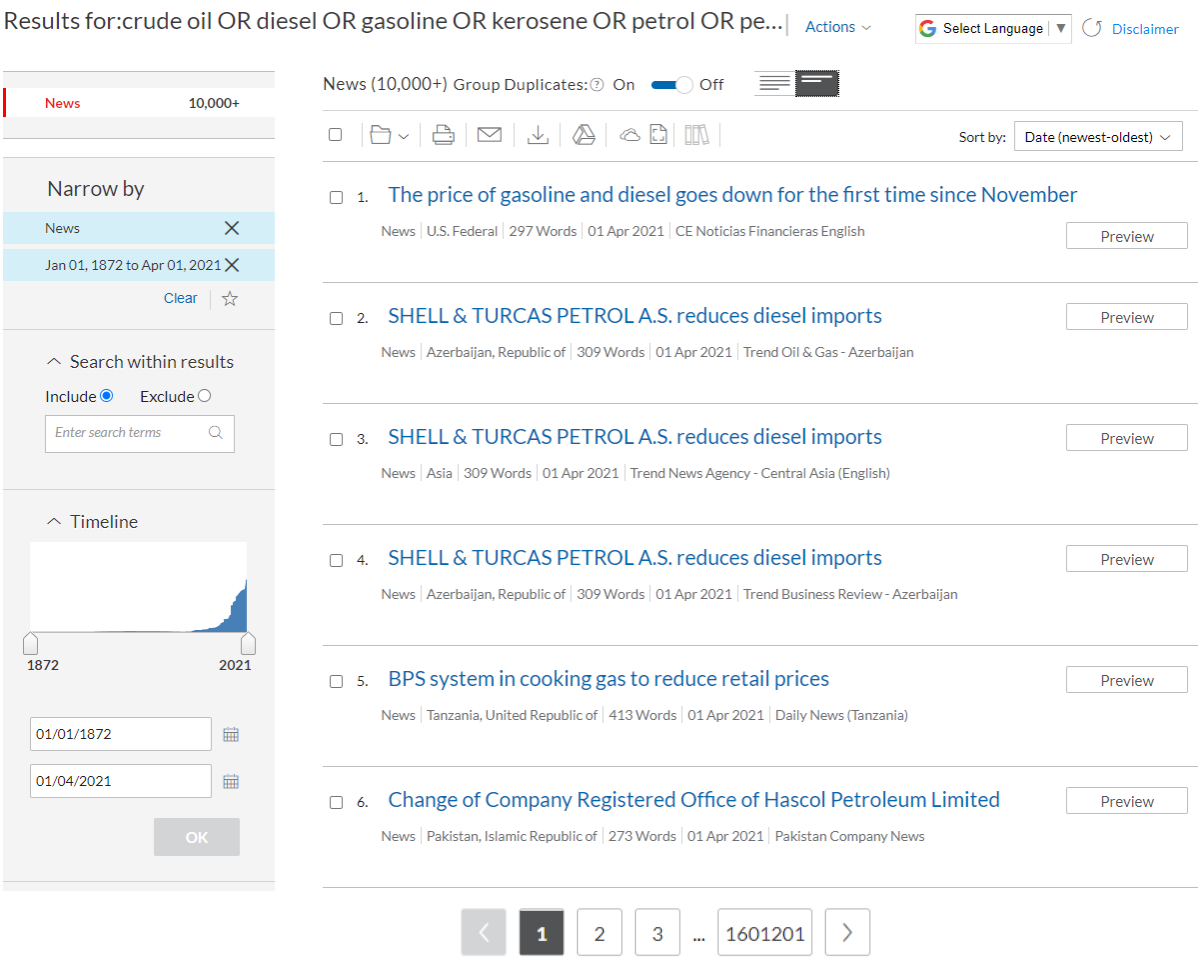

Figure 17: Screenshot of Nexis Uni displaying the selected filters

[11] https://www.lexisnexis.com/en-us/professional/academic/nexis-uni.page/

4.2.4 Investing.com Financial Markets Platform

Founded in 2007, Investing.com is an online global financial markets platform that offers real-time data, prices, charts, financial tools, breaking news, and analysis from 250 exchanges in 44 languages. It covers cryptocurrencies, world indices, world currencies, bonds, funds, interest rates, ETF's, futures, options, and also distinct commodities such as crude oil.
(About Investing.com, n. d.)
Unlike the earlier mentioned platforms, Investing.com contains a dedicated category for crude oil that conveniently provides news with high relevance, making the process of a query search obsolete. The news rubric[12] for crude oil (figure 18) contains around 35'000 unique news headlines, of which the earliest date back as far as 2009. However, there is no download function available. A possible way to acquire this data could be by utilizing a web scraping process.

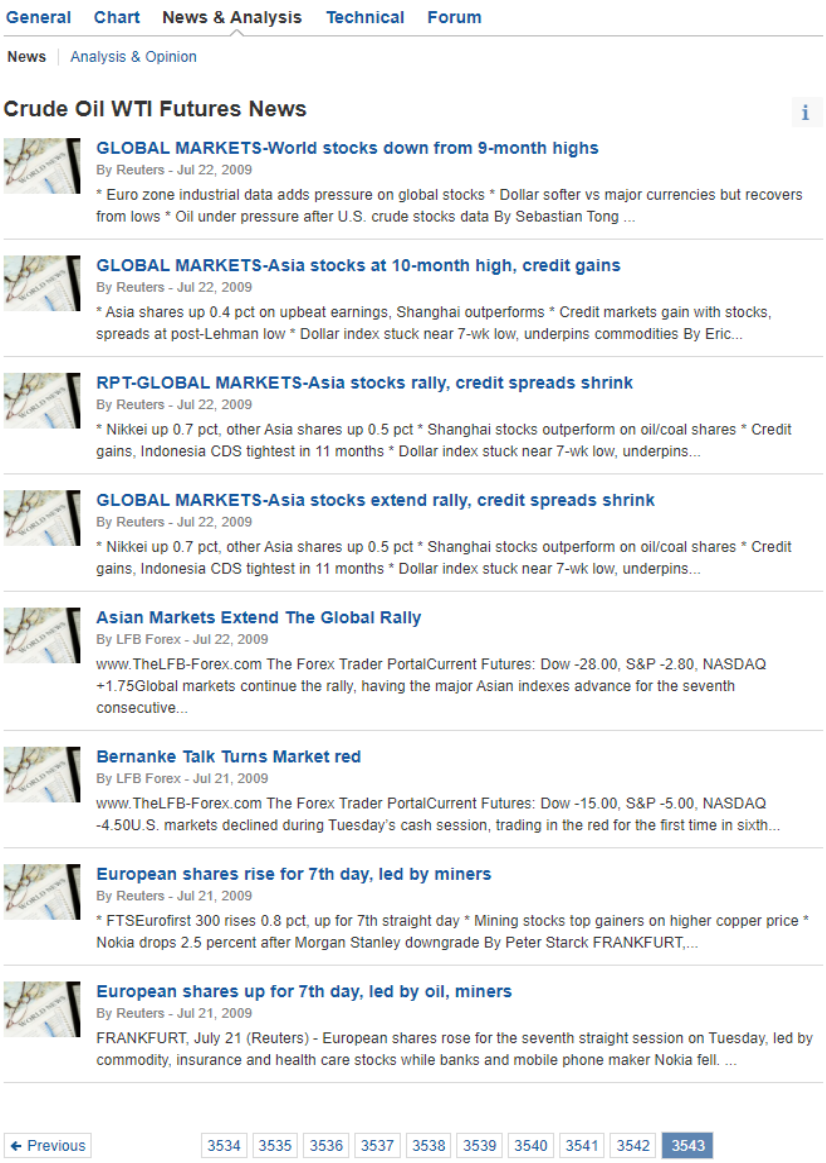

Figure 18: Screenshot of Investing.com displaying the news rubric of crude oil

[12] https://www.investing.com/commodities/crude-oil-news/

4.2.5 RavenPack Data Analytics Platform

RavenPack is a data analytics platform for financial professionals and academics. It offers a structured feed of news from credible content sources in real-time. Among the publishers are Dow Jones Newswires, the Wall Street Journal, and over 25,000 other conventional and social media sites with 20+ years of historical time-stamped data. (RavenPack, 2021)

Besides a structured feed of text, it also provides data analytics such as rule-based systematic identification of potentially market-moving events (out of 7'000) of the text. In addition, it can also recognize the entities (out of 300'000) such as companies, currencies, commodities, organizations, places, persons mentioned in a text. For instance, if there is news about a company suing another company, it can identify who is defending through named entity recognition and role detection and classify the phase of the event whether the lawsuit is still pending or settled. Besides event detection and entity recognition, it also provides many scores for relevance, novelty and sentiment, such as event sentiment score (ESS). (RavenPack, 2016)

The academic variant of RavenPack[13], which the author had access to, provided the ability to select crude oil as a distinct entity. Through, this selection all available data with high relevance to only crude oil was made available. In addition, the output was filtered to show the news headlines, their timestamp and their source. This search resulted in almost 46'000 headlines with high relevance to crude oil, reaching back to January 1st, 2000 (figure 19).

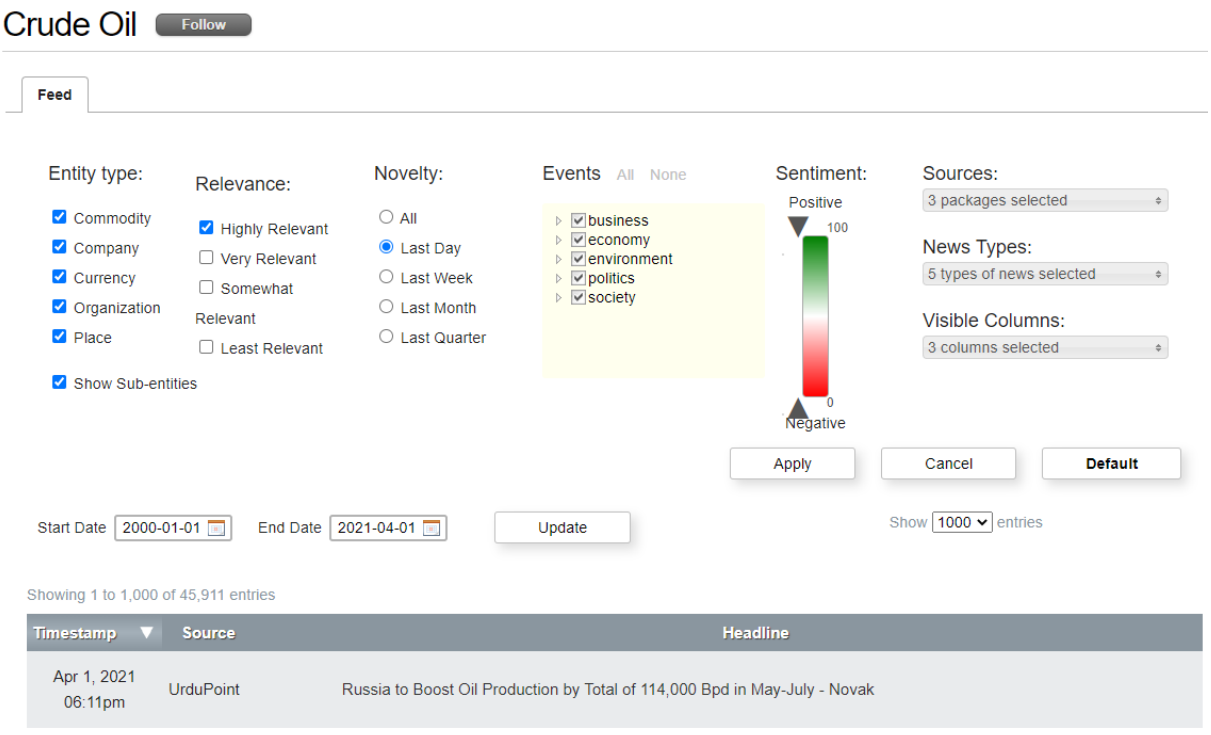

Figure 19: Screenshot of RavenPack Realtime News Discovery displaying the selected filters (RavenPack, n. d.)

[13] https://ravenpack.com/discovery/news_analytics_realtime/

However, similar to Investing.com the academic variant of RavenPack also did not provide any download functionality. Fortunately, the data was in a tabular order and could be simply marked and copied up to 1000 headlines at a time. However, for obtaining all of the data, this step had to be repeated 46 times which was manageable compared to the retrieval process of the other mentioned sources. Therefore, RavenPack was chosen as the source for the news headlines.

4.3 Overview of collected News Headlines

Through the RavenPack Realtime News Discovery platform, 45'911 news headlines from 1st January 2000 till 1st April 2021 with high relevance to crude oil have been acquired. There were 1034 individual news sources, of which about 400 provided only a single headline. Around half of the news headlines originated from *Dow Jones Newswires*, *Reuters*, *Bloomberg News* and *Platts* (figure 20). A list of the top 40 Publishers can be found in the appendix (table 14).

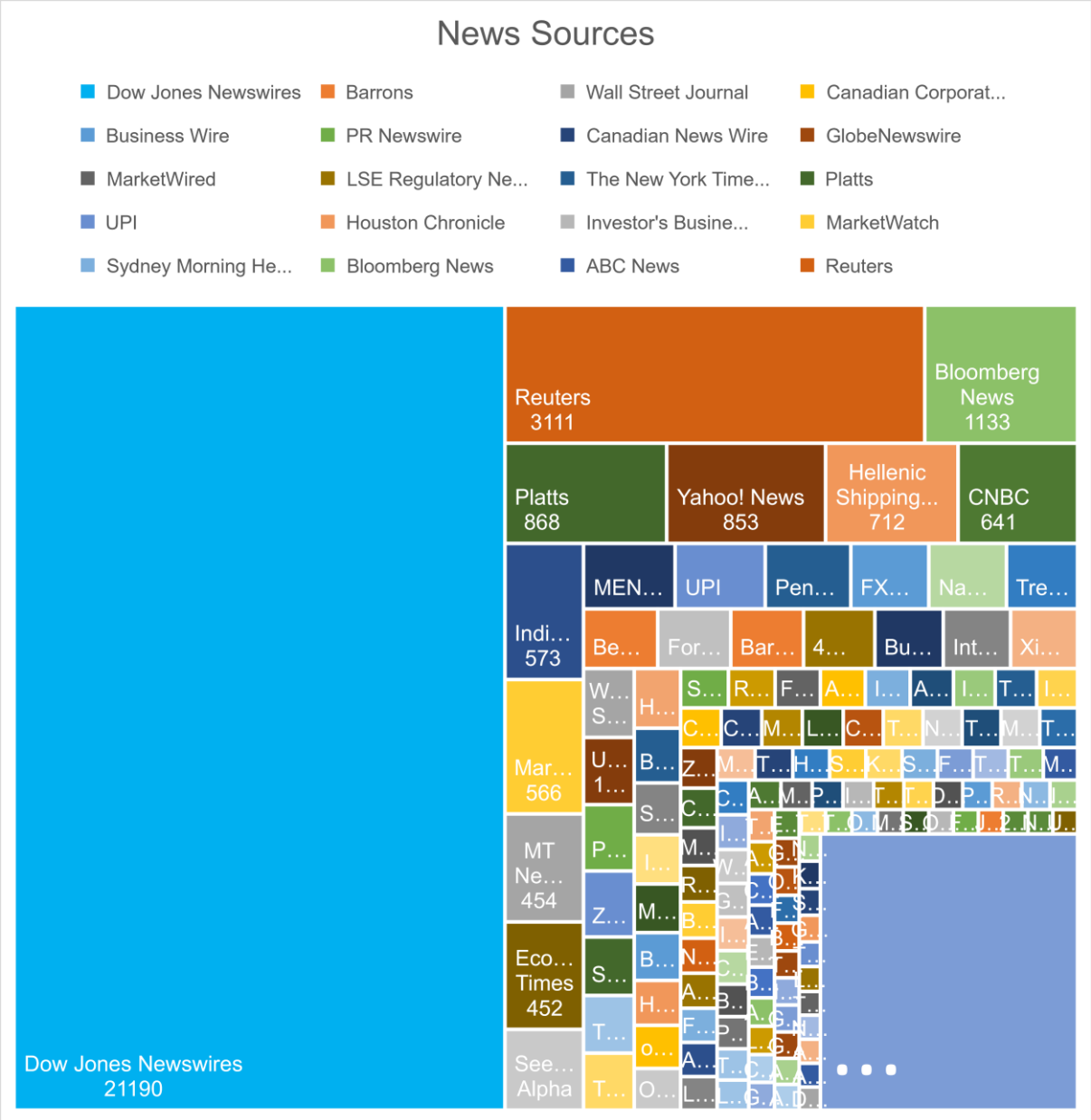

Figure 20: Treemap of the number of news publications of each publisher

Most headlines were published outside the weekends from 1st January 2000 till 1st April 2021 (figure 21). It should be noted that the months January, February and March result from 21 years, compared to the remaining months, which are the accumulation of only 20 years (figure 22). Each year's publications indicate a steady growth until 2012, which could be because RavenPack had possibly added additional publisher feeds over the years (figure 23).

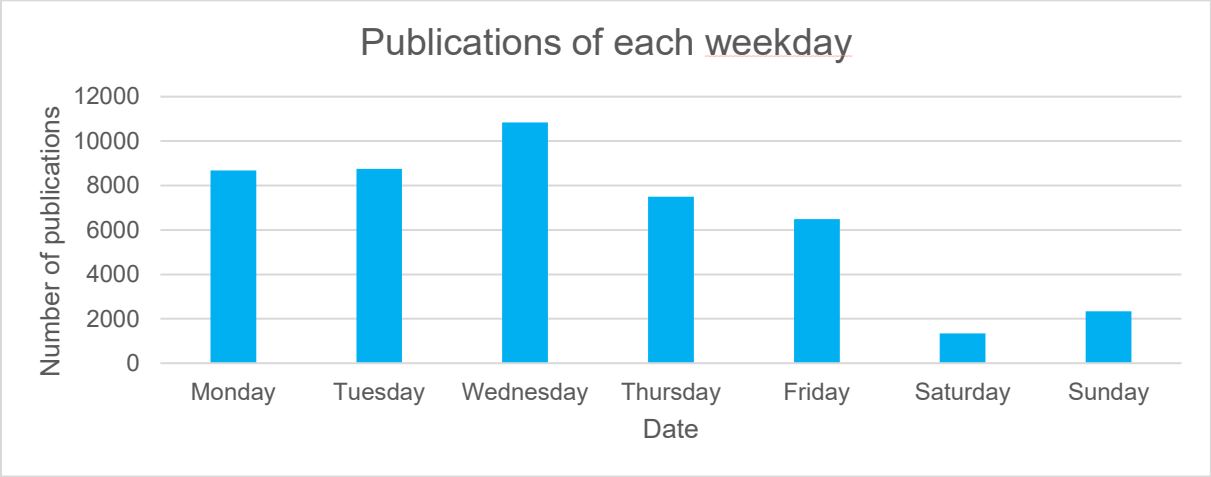

Figure 21: Bar chart of news publications of each weekday

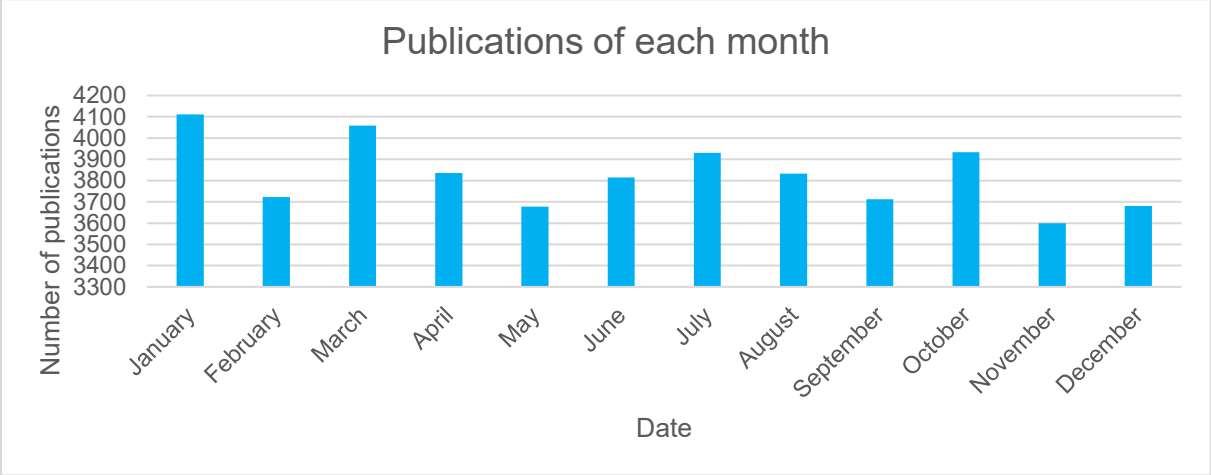

Figure 22: Bar chart of news publications of each month

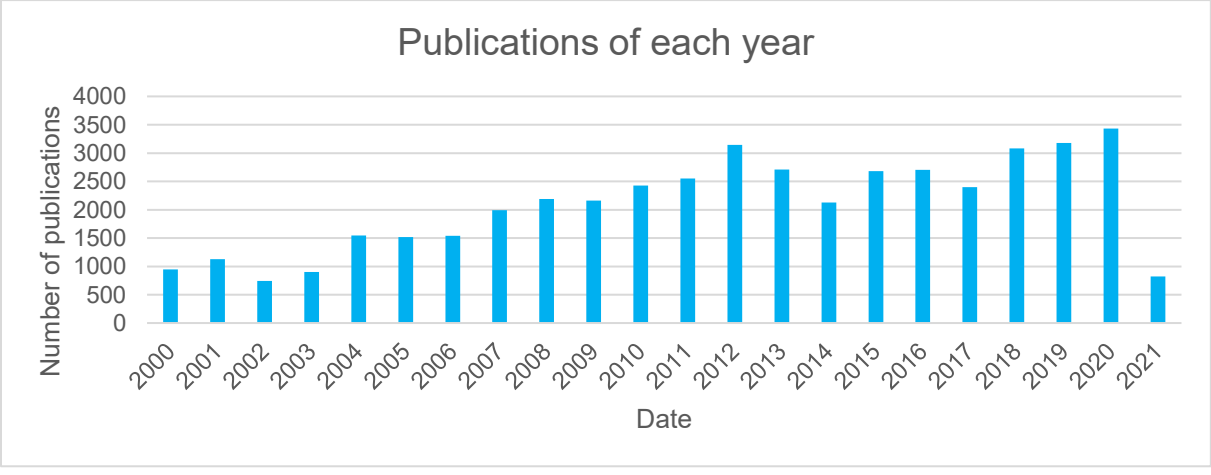

Figure 23: Bar chart of news publications of each year

On average, the headlines contain 10.4 words (median = 10), ranging from at least three to a maximum of 50 words. However, there are two occasions when headlines are made of only a link and are classified as one word (figure 24). From the perspective of character count, the headlines contain an average of 60.2 characters (median = 58) with a minimum of 14 and a maximum of 284 words (figure 25).

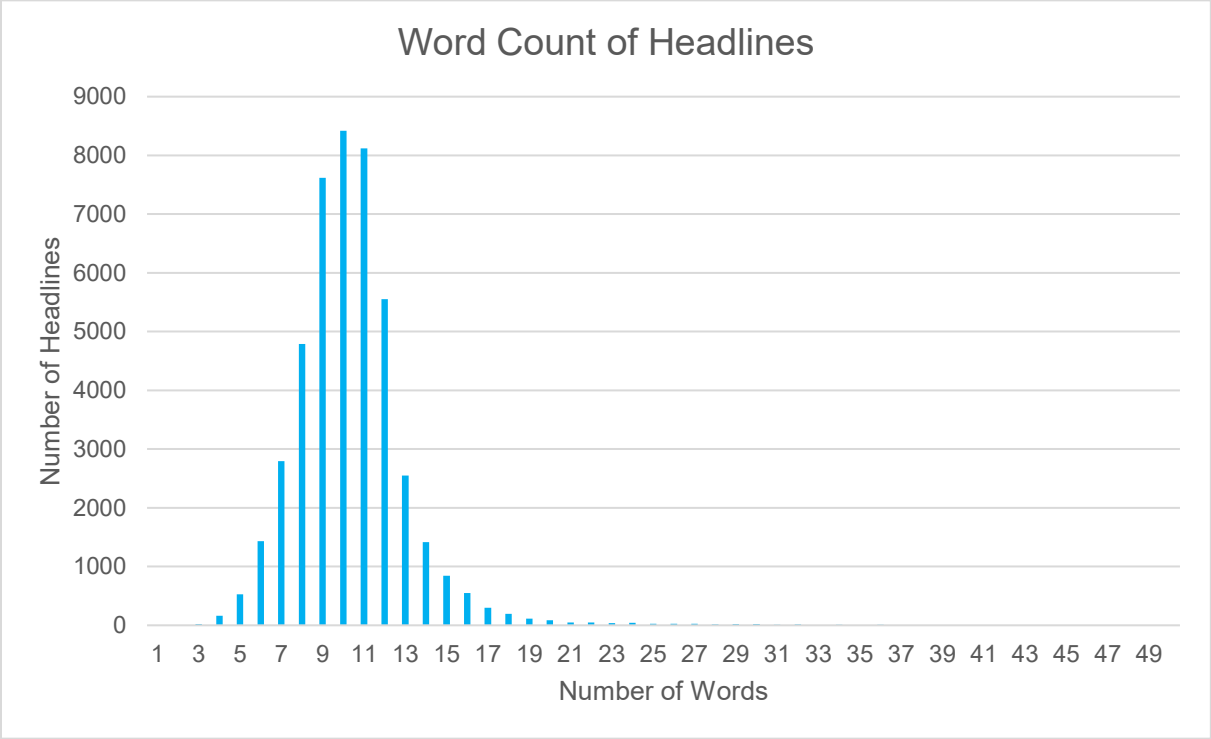

Figure 24: Word count of headlines

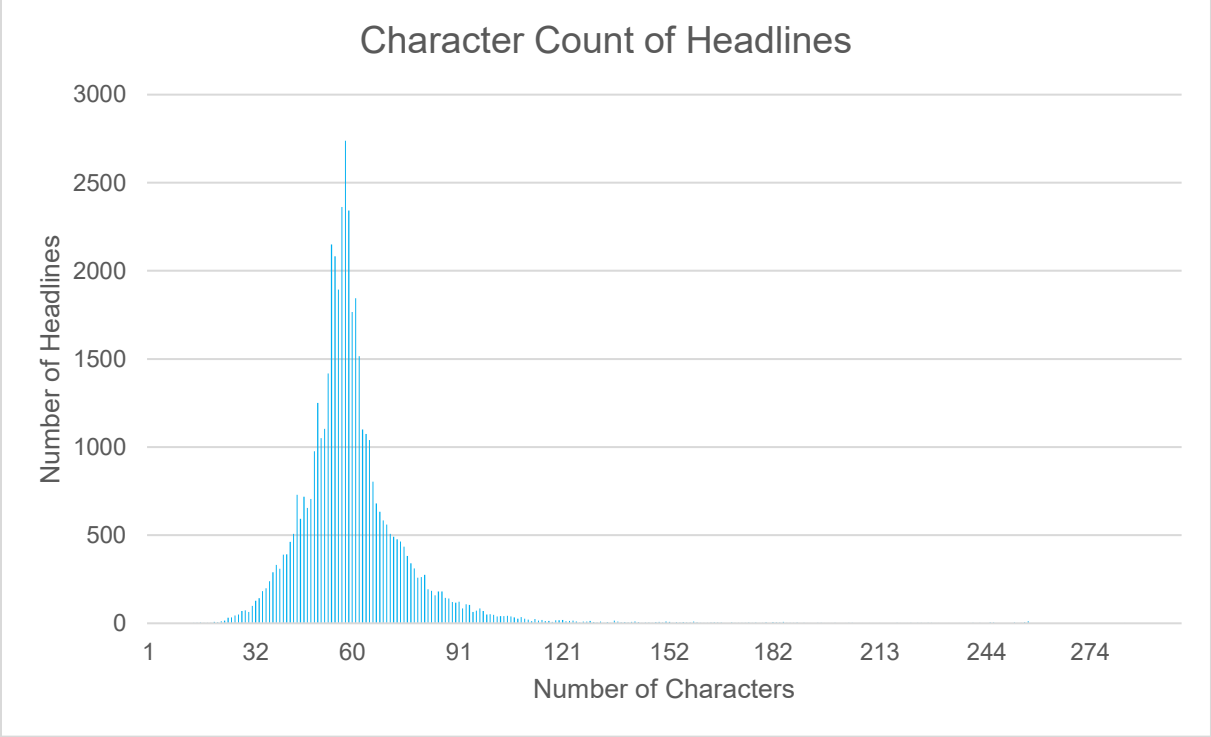

Figure 25: Character count of headlines

4.4 Sources for Oil Prices

Brent Crude and Western Texas Intermediate (WTI) dominate the oil market and dictate its price. Brent crude is the benchmark for crude oil in Africa, Europe, and the Middle East, representing nearly two-thirds of the world's crude oil output. WTI, on the other, is the preferred favoured benchmark by the United States of America. Since the acquired headlines are all in English, the price values of WTI were deemed more appropriate for this thesis.

In contrast to the data collection of news regarding oil, the historical crude oil price was comparatively undemanding to acquire. The first attempt on the already introduced platform Investing.com[14] (chapter XY) produced historical oil prices dating to August 1990. However, it should be noted that Index.com limits the output to 5'000 days. Therefore, to cover the same period of the news headlines acquired from RavenPack (1st January 2000 till 1st April 2021 = 7762 days), the export request has been split into two timeframes and concatenated manually afterwards. The following figure shows the prices of WTI crude oil:

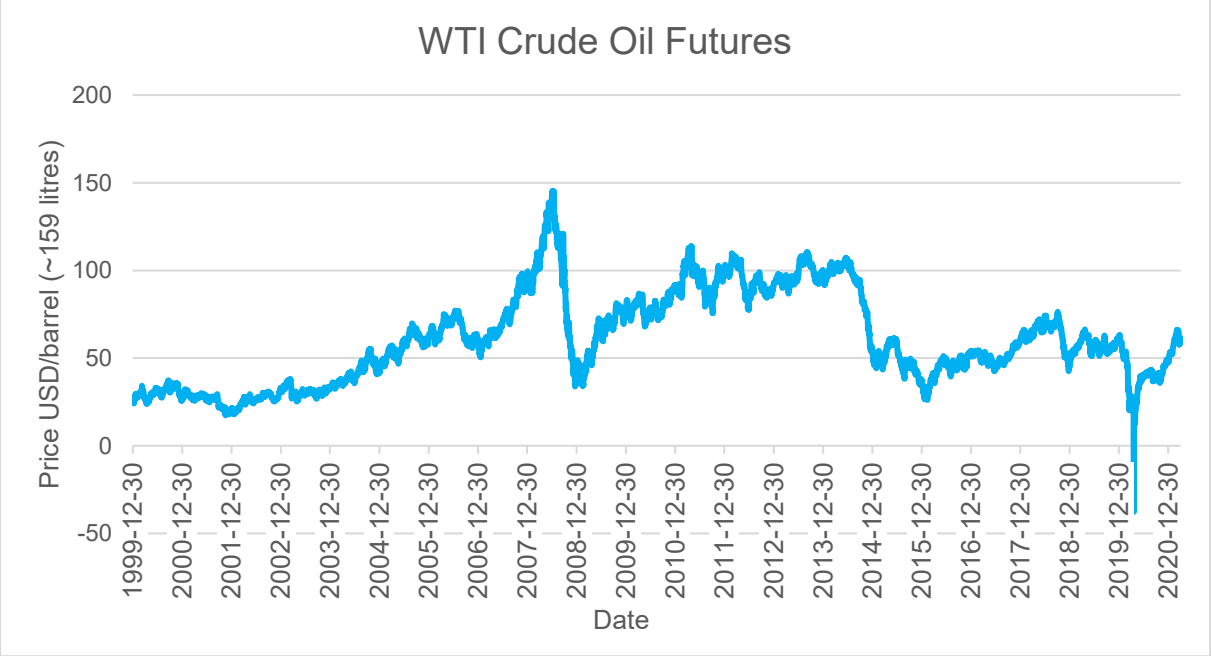

Figure 26: Historical Prices of WTI Crude Oil Futures

Some extreme price movements have been covered in chapter 2.3 (figure 1). However, it did not cover one of the more recent drastic dips, which occurred on April 20, 2020. Due to the outbreak of COVID-19 resulting in a global pandemic, the demand for crude oil decreased dramatically. This decrease was severe enough that crude oil ended up with negative pricing. In other words, the supplier was willing to pay the consumer -40.32 Dollars per barrel for taking his oil since the infrastructure for retrieving oil is hard to slow down and disposal of excess oil is costly. In the entire history of WTI crude oil, this was the first time it ended up with negative pricing. Interestingly, Brent oil futures did not see negative pricing.

[14] https://www.investing.com/commodities/crude-oil-historical-data/

4.5 Preprocessing News Headlines

As mentioned in chapter 2.5, the challenge of representing natural language in a machine-readable yet meaningful format is no trivial task and requires careful text preprocessing. Text preprocessing takes raw text as input and returns tokens that have been cleaned. This process aims to standardise and reduce the noise of the input through tokenization, standardization, stop word removal and stemming steps (Anandarajan et al., 2019, p. 46).

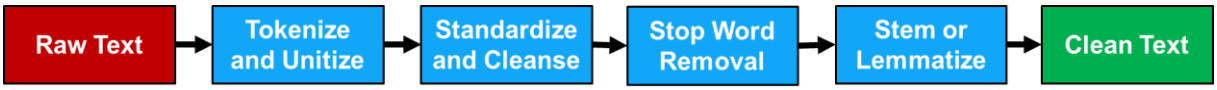

Figure 27: The text data preprocessing process (illustration based on Anandarajan et al., 2019, p. 48)

Finding the right balance between eliminating noise and retaining information is determinantal of the success of the analysis. Any shortcomings in the preprocessing will propagate on to the performance of the target task. Thus, far more time is spent in preprocessing text than in the analysis itself. However, the tokenizing step alone is generally sufficient when working with transformers (chapter 2.5.5) since the attention mechanism can cope with the noise better than traditional NLP algorithms. Tokenization is a process of splitting text into smaller chunks. Multiple ways can achieve this. In the case of transformers, three tokenizing methods prevail Byte-Pair Encoding (BPE), WordPiece (WP), and SentencePiece (SP). (Yi Peng, 2020)

The tokeniser and the corresponding vocabulary are predetermined by the selected model when working with a pretrained transformer model. Depending on the task on hand, the choice of the model might affect the performance. For instance, certain tokenizers such as BPE can recognize Emojis, whereas WP and SP cannot. On the other hand, BPE and WP are not sensitive when numerals are mentioned in a text, whereas SP can react sensitive. The following figure illustrates the variation of numerals with respect to the used tokenizer:

sent1 = *Iraq's August Oil Export +6% Month-on-Month to 2.6 M Barrels per Day.*
sent2 = *Iraq's August Oil Export -6% Month-on-Month to 2.6 M Barrels per Day.*

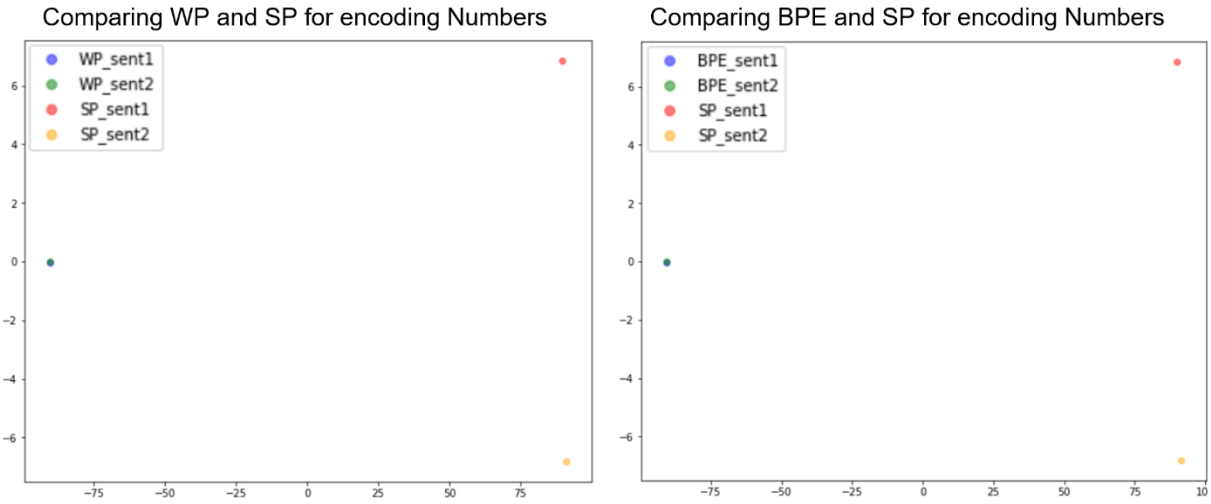

Figure 28: Comparison of the variation of the principal components between BPE, WP with SP

These experiments are based on the works of Yi Peng (2020), who also provides a more detailed treatment of each tokenizer.

4.6 Merging News Headlines and Oil Prices

This chapter covers the process of merging the sentiment scores of the headlines with the crude oil prices. The crude oil price data contained gaps due to market closings on holidays and weekends. Similarly, despite a high number of news headlines (45' 911) for a period consisting of 7762 days, not each day was accompanied by a headline. In contrast, there were also days on which multiple headlines were published (figure 29).

	Date	Time	Headline	Sentiment Score
Gap	01.01.2000	10:08am	India Apr-Nov Oil Imports $6.20B Vs $4.03B; +53.86%	+1
	03.01.2000	07:00am	Texaco and Chevron Missed the Boat on Big Stock Gains	0
Multiple	04.01.2000	07:46am	Taiwan CPC Dec Crude Oil Imports 20.4M Bbl, US$23.79/Bbl	+1
	04.01.2000	07:53pm	Midday US Spot Crude: Prices Sink With Crude Futures	-1
Gap	04.01.2000	11:52pm	API: US Crude Stocks Dn 1.568 Mln Bbl In Week	-1
	06.01.2000	01:30am	Nymex Access Crude Down After 4th Straight Day Of Losses	-1
Multiple	06.01.2000	04:04am	China's CNPC '99 Crude Output 107.08M MT; Dn 0.3%	0
	06.01.2000	10:30pm	Late US Spot Crude: Prices Lower After Choppy Day	-1
Multiple	07.01.2000	03:00am	WSJ(1/7): Texaco Says Oil Find May Hold A Billion Barrels	0
	07.01.2000	03:53pm	Nymex Midmorning:Crude Dn On Report Of Higher OPEC Output	+1
	...	...	...	...

Figure 29: Overview of Data Frame from News Headlines

In order to provide a sentiment score and oil price for the missing days, a quadratic interpolation with the Pandas-function `interpolate(method='quadratic')` has been carried out. In the case of multiple sentiment scores for a given day, one approach would be to take the average of those sentiment scores. However, this would denote that the impact of one news headline would be equal to the impact of multiple headlines on a given day and completely neglect its volume as a factor. Therefore, the works of Hafez et al. (2018) stated that taking the sum instead of the mean should be favoured. Working with the sum can be interpreted as a simple way to take the sentiment score and the sentiment volume simultaneously into consideration, which also yielded better results in their tests (Hafez et al., 2018, p. 1).

It should also be pointed out that the sentiment scores can be positive or negative, whereas the price remains positive (except during the period around 20th April 2020). Therefore, the crude oil returns would be more suitable to compare the sentiment scores instead of the price. For this matter, the daily returns of crude oil have been calculated as follows:

$$Return = \frac{Price_t - Price_{t-1}}{Price_{t-1}} \quad (1.3)$$

However, the price itself could also serve as beneficial for certain analyses since it can be interpreted as the cumulative values of the returns. Thus, it could be compared to the cumulative sentiment scores. The following table shows the first seven rows of the merged data frame, which was achieved by the `pandas.DataFrame.merge function` in which the date was used as the joint index. Note that the sum of multiple sentiment scores occurring on one day was calculated first representing the daily score (blue), followed by an interpolation to fill the price and sentiment scores (orange) gaps. For readability and simplicity, the sentiment scores are listed as fictional integers, whereas in the actual data, they represent float numbers:

Date (index)	Oil Price	Cumulative Sentiment Score	Oil Return	Daily Sentiment Score
January 1, 2000	25.22	1	-0.03908	1
January 2, 2000	25.73	2	2.006354	1
January 3, 2000	25.88	3	0.593844	1
January 4, 2000	25.55	4	-1.27308	1
January 5, 2000	24.91	4	-2.50489	0
January 6, 2000	24.78	3	-0.52188	-1
January 7, 2000	24.22	1	-2.25989	-2
…	…	…	…	…

Table 5: Illustration of the first Rows of the merged Dataframe

It should also be pointed out that the sentiment scores are consisted of floating values between -1 and 1 (when neglecting the volume), whereas the returns have no such limits. Consequently, both values need to be normalized beforehand. Otherwise, the different scaling factors could deliver false and misleading outputs. The normalization has been performed via the function `sklearn.preprocessing.MinMaxScaler(feature_range=(0, 1)`.

In order to reduce the volatility of the data, a Simple Moving Average (SMA) has been applied with the `rolling(window=n).mean()` pandas function. This SMA is calculated by taking the arithmetic mean of a given set of values over a specified period:

$$SMA = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{1.4}$$

Here, X represents the input, whereas n denotes the window (period) of the moving average.

4.7 Frameworks for Sentiment Analysis

This chapter gives an overview of a selection of frameworks suitable for conducting deep learning-based sentiment analysis. Here, emphasis is placed on ensuring that the choices provide transformer implementations and are compatible with the resources stated in chapter 4.1.

4.7.1 Open-Source Library PyTorch

PyTorch was developed by Facebook's AI Research (FAIR) team and released under the modified BSD license as an API for Torch. Torch itself gained much traction after being utilized as the main ML library for renowned organizations such as Facebook, Twitter and Uber. PyTorch aims to simplify the interaction with Torch and is well suited for prototyping with C++ and Python. PyTorch also comes with implementations for transformers available as a library named PyTorch-Transformers, which provides state-of-the-art pretrained models for NLP that utilize various transformer-based models such as BERT. (Wikipedia contributors, 2021)

4.7.2 Open-Source Library AllenNLP

AllenNLP[15] is a PyTorch-based research library for NLP developed and maintained by the Allen Institute for AI (AI2). It is released under the Apache License 2.0. It provides NLP functionalities such as Answer a question, Reading Comprehension, Visual Question Answering, Named Entity Recognition, Semantic Role Labeling, Open Information Extraction, Dependency Parsing, Constituency Parsing and lastly, Sentiment Analysis. The latter can be carried out either based on LSTMs with static word embeddings such as GLoVE or with transformers such as BERT. It is intended for fast prototyping of NLP models. (AllenNLP, n. d.)

4.7.3 Open-Source Library Hugging Face

Hugging Face is a community-driven open-source provider for NLP technologies licensed under the Apache License 2.0. It provides a large repository for training datasets, mainly contributed by the AI community. In addition, it utilizes an intuitive API that exposes to many well-known transformer architectures on its HuggingFace's platform[16]. This platform yields many state-of-the-art pretrained models for NLP that utilize various transformer-based models such as PyTorch-Transformers and pretrained models contributed by the AI community such as FinBERT[17]. (Hugging Face contributors n. d.) Since Hugging Face has an easy-to-use API and provides easy access to the pre-trained FinBERT model, it has been selected as the main framework for implementing and executing the sentiment analysis part in this thesis.

[15] https://allennlp.org/
[16] https://huggingface.co/transformers/v1.2.0/
[17] https://huggingface.co/ProsusAI/finbert

5 Evaluation of the Models

This chapter deals with evaluating the initial output of FinBERT and the subsequent approaches for improvement based on literature review, which leads to CrudeBERT. Additionally, this chapter contains an interview with an expert from the energy sector, which has led to CrudeBERTv2 and its variations. Finally, evaluations in a preliminary manner were carried out to indicate the performance of CrudeBERT, if there is a relationship with the price of crude oil.

5.1 Evaluating FinBERT

Before assessing the forecasting potentials of FinBERT, a quick comparison of the cumulative sentiment scores of FinBERT and the price history of WTI crude oil was performed. According to this comparison (figure 30), it appears that FinBERT shows no recognisable link with the oil price trend.

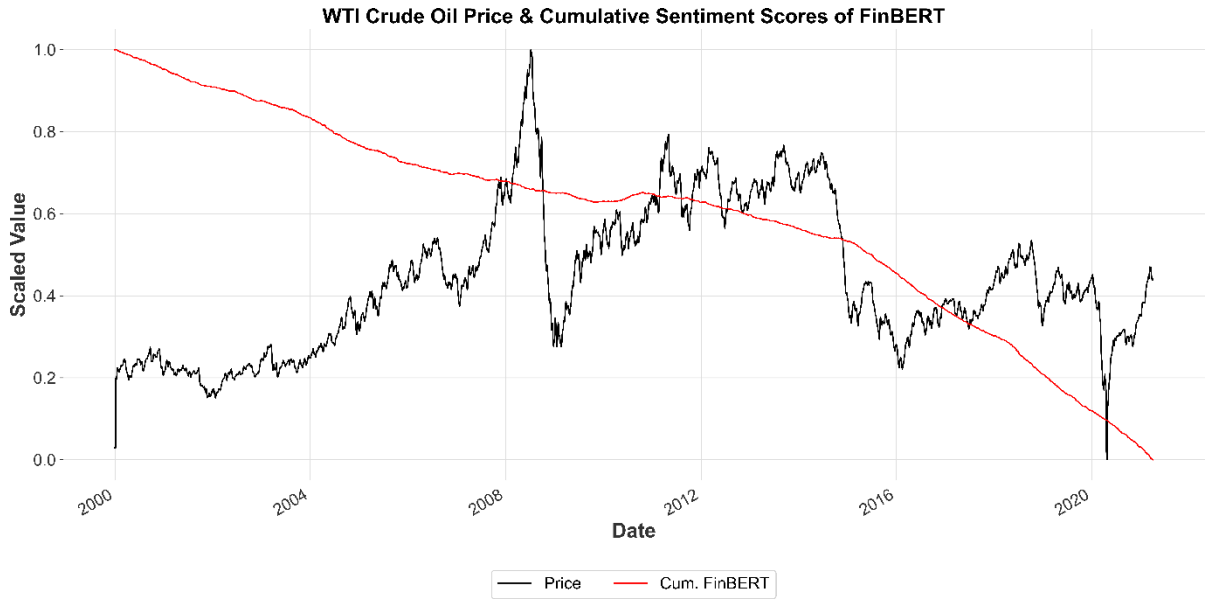

Figure 30: WTI Crude Oil and Cumulative Sentiment Scores of FinBERT

Further analysis of crude oil's price changes (returns) and its associated FinBERT sentiment score in the form of a simple linear regression supports this observation. For this matter, the Pearson correlation value R has been calculated to measure the relationship between them:

$$R = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Where x stands for the returns of WTI crude oil, whereas y denotes the sentiment scores, both input values are non-scaled in this case. The code is available on GitHub (chapter 4.1) as a Jupyter Notebook file (FinBERT_scatterplots_masterthesis.ipynb).

According to the following scatter plot, there is a very weak negative relationship (R = -0.03997) between the sentiment scores of FinBERT and the returns of WTI crude oil (figure 31).

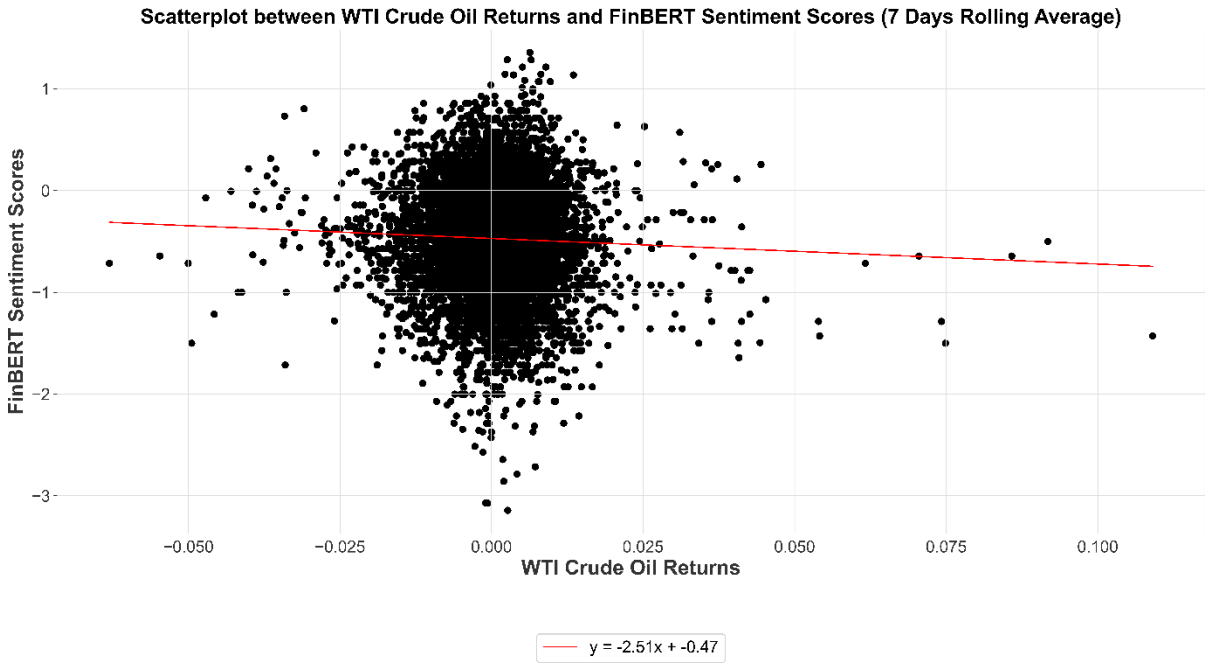

Figure 31: Scatterplot between WTI Crude Oil Returns and FinBERT Sentiment Scores (window=7 days)

Increasing the SMA-window from weakly to quarterly (90 days) slightly emphasised the weak negative relationship (R = -0.1008). However, the linear relationship remains insignificant.

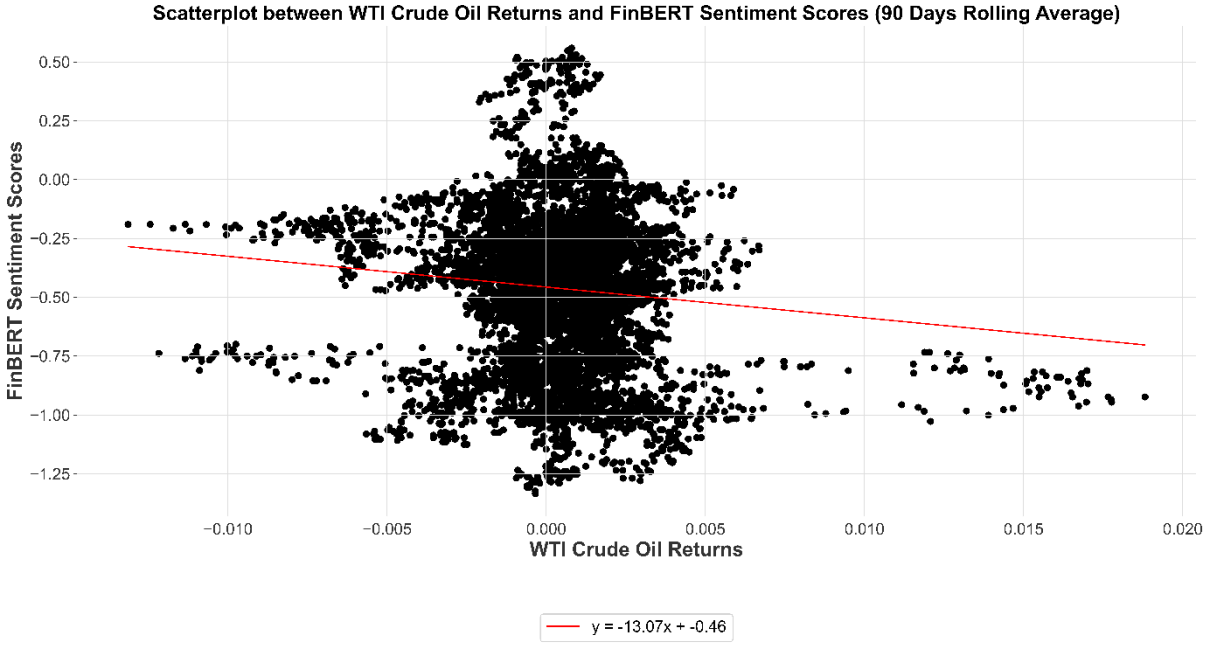

Figure 32: Scatterplot between WTI Crude Oil Returns and FinBERT Sentiment Scores (window=90 days)

The verdict of this arguably simple analysis highlighted the need for further investigation for the potential reasons behind this low relationship and possible solutions for improving it.

5.2 Domain Adaptation

FinBERT is a pre-trained NLP model for analysing sentiment in financial writing. It was created by fine-tuning the BERT language model for financial sentiment classification using a large financial corpus and training it in the finance sector to determine if financial writing expresses optimistic or bearish views on a given text (chapter 2.5.7). Even though crude oil is a publicly tradable asset, financial sentiment analysis as a general domain delivers poor performance and doubtful results. According to Xing et al. (2020), such behaviour is expected when using generic sentiment analysis methods. It is known as the problem of domain adaptation (Xing et al., 2020, S. 978).

For this matter, a manual inspection on a small portion of the output consisting of headlines and its corresponding FinBERT sentiment scores has been conducted. Recapitulating the properties of the oil (chapter 2.3), whose price is mainly determined by supply and demand since it is an essential and limited resource, most of the FinBERT scores appear to contradict this characteristic. In order to better illustrate this observation, the following figure of the supply and demand curve will be used:

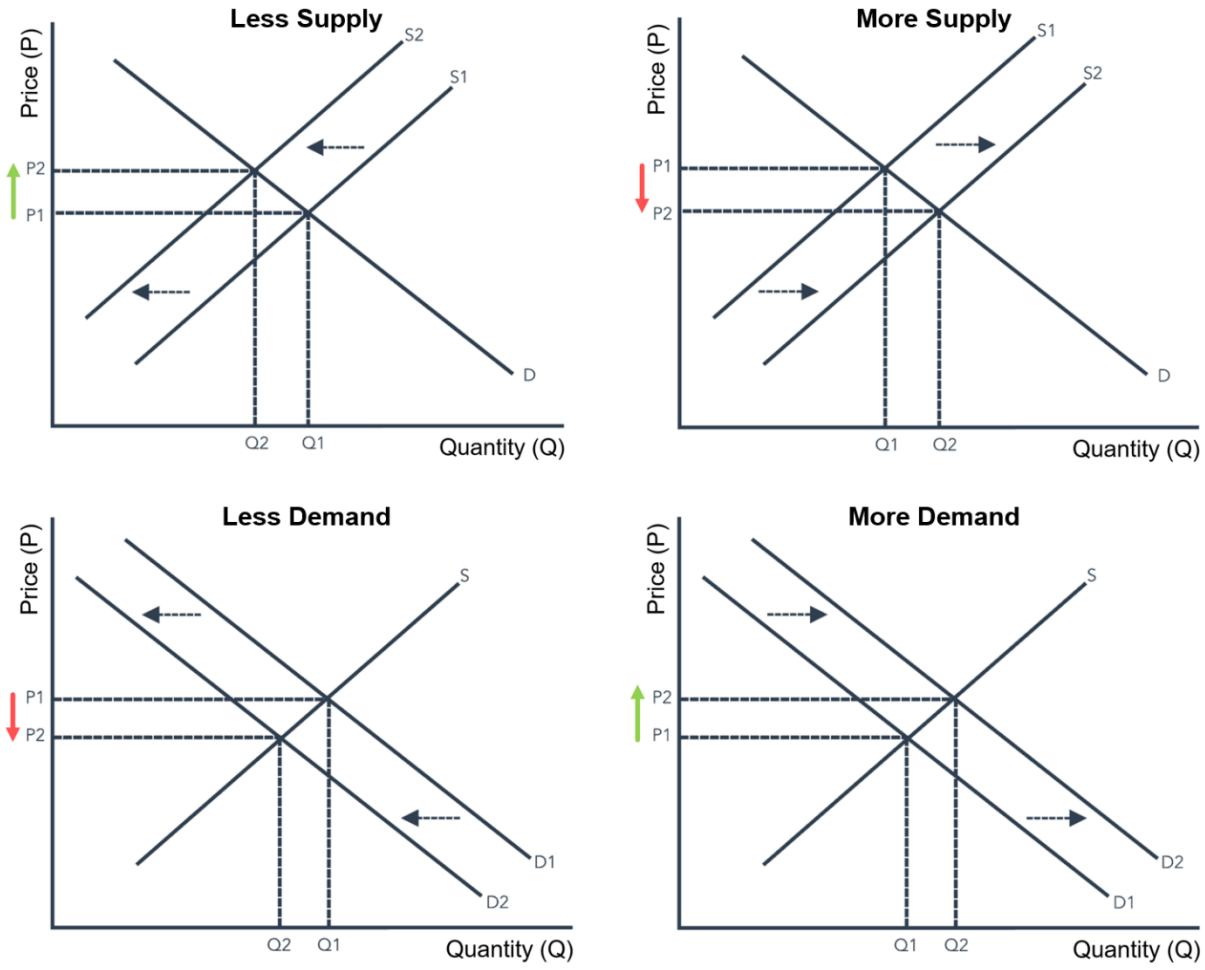

Figure 33: Influence of supply and demand on the price (illustration based on Agarwal, 2018)

Here, “supply” refers to the quantity of a commodity or service that a supplier intends to sell for a given period. The “quantity” of commodity or service a consumer is willing to buy over the same period is described as demand. This interaction between suppliers and consumers results in a competitive market determining the price of the commodity or service. For instance, if the demand remains the same but the supply decreases, it would result in a shortage, thus increasing the price. On the other hand, a shortage can also be caused if the supply remains unchanged instead of the demand increases.

In contrast, increased supply with an unchanged demand would result in a surplus and consequently lowering price. This surplus can also occur if the demand decreases while the supply remains the same. (Agarwal, 2018)

According to the logic behind the supply and demand curve, the influence of each scenario can be summed up as follows:

- Less supply → shortage → higher price
- More supply → surplus → lower price
- Less demand → surplus → lower price
- More demand → shortage → higher price

Under the consideration of the logic behind supply and demand, the following samples consisting of headlines regarding shortage and surplus reveal that numerous FinBERT scores, where 1 = highly positive, -1 = highly negative, and 0 = neutral, appear to be doubtful.

For instance, table 6 contains headlines indicating a supply decrease due to accidents at oil refineries and oil platforms. Accidents like these negatively affect the supply, ultimately leading to a higher price. However, FinBERT associate’s accidents as something negative, which is somewhat unsurprising since accidents in finance would rarely represent good news.

	Headlines indicating a Decrease in Supply	Sentiment Score SHOULD	Sentiment Score FinBERT
Shortage	**Major Explosion, Fire at Oil Refinery in Southwest Philadelphia**	Positive	-0.881674
	Petroleos Mexicanos confirms Gulf of Mexico oil platform accident	Positive	-0.467316
	CASUALTIES FEARED AT OIL ACCIDENT NEAR IRAQ'S BASRA	Positive	-0.754872

Table 6: Sample of Headlines indicating a Decrease in Supply

When looking at examples with headlines indicating an increase in supply due to higher exports or oil discoveries (table 7), FinBERT classified them as positive. One potential reason could be that more export in financial markets is often perceived as indicating a healthy industry. Yet, according to the logic behind the supply and demand of oil, this should not be the case.

	Headlines indicating an Increase in Supply	Sentiment Score SHOULD	Sentiment Score FinBERT
Surplus	**Iraq's Feb Oil Exports +20.9% On Mo at 1.56M B/D - Official**	Negative	0.821053
	Apache announces large petroleum discovery in Texas	Negative	0.713858
	Turkey finds oil near Iran, Iraq border	Negative	0.706201

Table 7: Sample of Headlines indicating an Increase in Supply

On the other hand, when it comes to demand, a decrease caused by lower imports (table 8) should result in a surplus, thus a higher price. Whereas FinBERT interestingly rates them as positive, even though a lower import in finance should, similarly to the case of export, result in negative news for the same reasons. It is probable that FinBERT was trained on datasets in which sustainability was a positively defining feature. Thus, news associated with lower fossil fuel consumption might be recognized as positive for the financial market.

	Headlines indicating a Decrease in Demand	Sentiment Score SHOULD	Sentiment Score FinBERT
Surplus	**Turkey February Crude Imports -16.0% On Year -METI**	Negative	0.725378
	OIL DATA: Japan May Crude Imports -11.0% On Year	Negative	0.724652
	China June Crude Oil Imports -10.9% On Yr,16.60M KL-METI	Negative	0.678462

Table 8: Sample of Headlines indicating a Decrease in Demand

In the opposite case in which news coney an increase in demand as a result of from higher imports (table 9), should lead to a higher price rating of oil. FinBERT rates a higher demand for fossil fuel as positive as well, likely for the same reasons as the earlier example for higher export and neglecting sustainability-related matters instead.

	Headlines indicating an Increase in Demand	Sentiment Score SHOULD	Sentiment Score FinBERT
Shortage	**EIA Chief: Expects Global Oil Demand Growth 1.5M B/D To 2010**	Positive	0.920307
	China Jan-Oct Crude Imports +98.5% To 57.9M MT	Positive	0.882203
	Turkey's crude oil imports up 78.30% in February 2019	Positive	0.926665

Table 9: Sample of Headlines indicating an Increase in Demand

5.3 Developing CrudeBERT

In order to potentially improve the performance of FinBERT, a domain adaptation towards the properties of crude oil has been undertaken. So far, the examples above (tables 5 – 8) covered four topics: accidents, oil discoveries, changes in exports, and **changes in imports**. However, a manual scan of several hundred news headlines revealed six additional frequently reoccurring topics, such as **changes in demand**, **price, supply, pipeline constraints**, **drilling** and **spilling**. Topics associated with changes mainly refers to an increase, decrease or remain steady. As a next step, a query search through all the headlines containing terms representing the ten topics has been conducted. Each match between the headline and the query term has been labelled as the query term in the headline. For instance, the query search for the headline in figure 34 labelled it as a change in import due to the word “import” appearing in the headline. Finally, to determine the direction of the change, a second query search has been conducted consisting of synonyms for increase, decrease and remaining steady. The complete list used for the query search can be found in the appendix (table XY).

Figure 34: Example of Identifying Topics and Polarity in Headlines

The query mentioned above resulted in 30’231 labelled headlines covering almost two thirds (65.85%) of the whole dataset consisting of 45’911 headlines. The following list displays all the detected topics and their amount in brackets:

- Price Increase (1637)
- Price Decrease (1307)
- Supply Increase (5856)
- Supply Same (370)
- Supply Decrease (5654)
- Demand Increase (1265)
- Demand Same (12)
- Demand Decrease (837)
- Export Increase (1956)
- Export Same (127)
- Export Decrease (1536)
- Import Increase (2795)
- Import Same (29)
- Import Decrease (2365)
- Oil Discovery (1633)
- Spill (2304)
- Drilling (113)
- Pipeline Constraint (67)
- Accident (373)

Subsequently, the labelled headlines were examined based on the supply and demand logic covered in chapter 5.2. Headlines conveying news about drilling, oil discoveries, exports, or literally supply increase would lead to more supply and proclaim an oil surplus. Similarly, headlines which state imports decrease or demand decrease should, in theory, also lead to surplus. On the other hand, headlines implying accidents, pipeline constraints, spills or literally supply decrease would cause a shortage since the supply would be affected negatively by such occurrences. This shortage can also be provoked by news stating a demand increase, imports increase, or exports decrease. In short, news headlines indicating a surplus or news stating a price decrease should affect the price negatively. Contrary, news indicating a shortage or news stating a price increase should affect the price positively. However, news headlines also suggested no changes in supply, demand, import and export, resulting in the price remaining the same (figure 35).

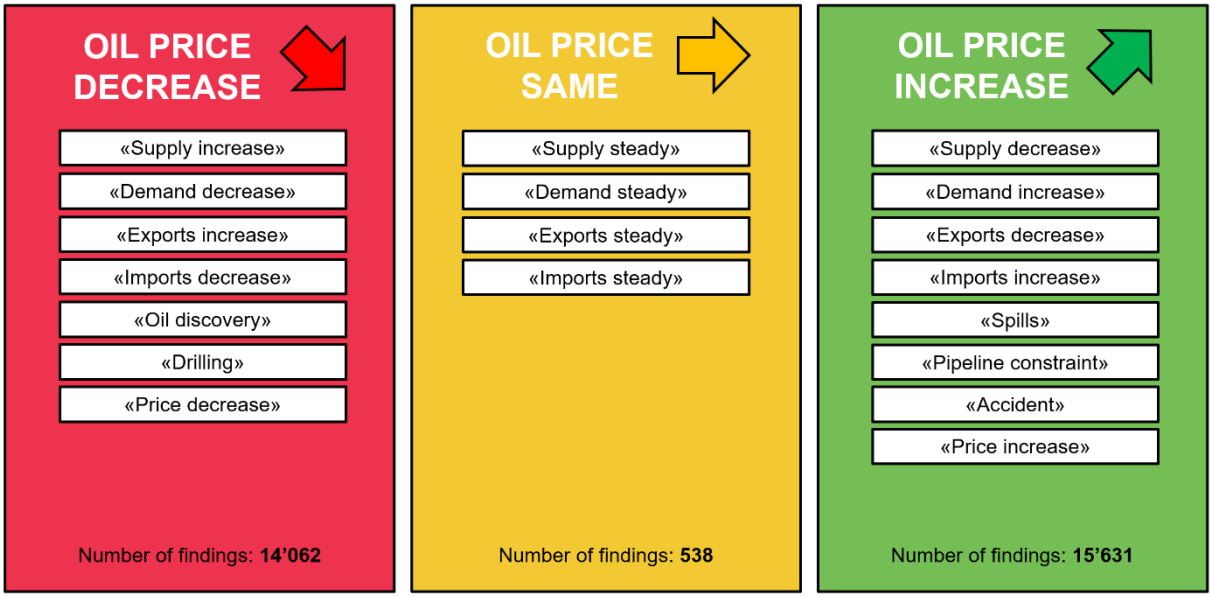

Figure 35: Assignment of the labelled Topics

Every headline, which would negatively influence the price, would be annotated with a negative sentiment score of -1. Conversely, headlines deemed as having a positive impact were annotated with a positive sentiment score of 1. Lastly, the remaining headlines, which indicate an unchanging price, were annotated with a neutral sentiment score of 0. These labelled headlines were ultimately used as a training dataset (S&D-Dataset) to fine-tune FinBERT and create a new model called CrudeBERT (figure 36).

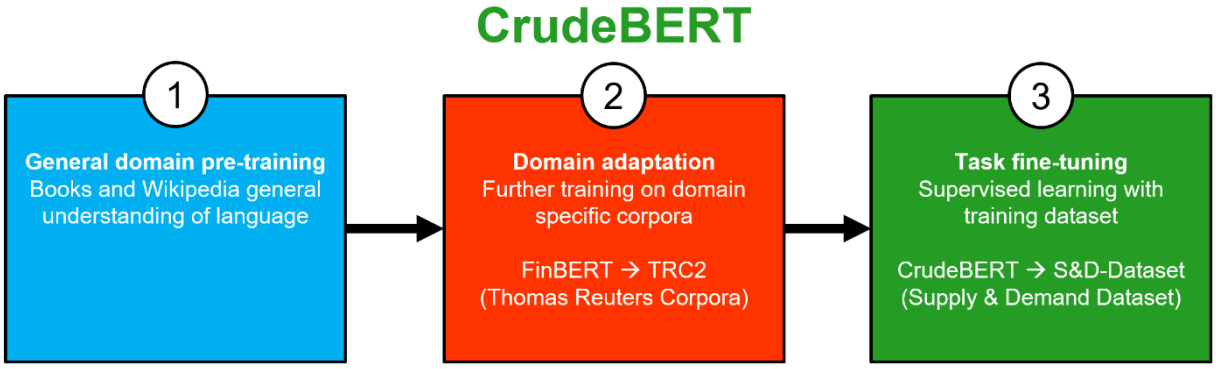

Figure 36: Process of fine-tuning FinBERT and creating CrudeBERT

Since this undertaken segmentation of the S&D-Dataset cannot provide any sensitivity regarding numerals, for instance, the -10 decrease in imports is worse than -5, selecting an appropriate tokenizer (chapter 4.5) was not critical. The selected base models and tokenizer are:

```
finbert.base_model = 'bert-base-uncased'
tokenizer = AutoTokenizer.from_pretrained('bert-base-uncased')
```

The S&D-Dataset consists of 14’062 negative, 538 neutral and 15’631 positive examples. From each class, 20% were preserved for the evaluation of the model. The configuration and the hyperparameters for training the neural network on these three classes can be seen below:

```
config = Config(    data_dir=cl_data_path,
                    bert_model=bertmodel,
                    num_train_epochs=4,
                    model_dir=cl_path,
                    max_seq_length = 48,
                    train_batch_size = 32,
                    learning_rate = 2e-5,
                    output_mode='classification',
                    warm_up_proportion=0.2,
                    local_rank=-1,
                    discriminate=True,
                    gradual_unfreeze=True)
```

The entire code (FineBERT_training_on_oil_headlines_to_generate_CrudeBERT.ipynb) is available on GitHub (chapter 4.1). It should be noted that the S&D-Dataset had to be split into four chunks due to the lack of available RAM capacity. The training itself took nearly two hours and yielded an overall accuracy of 0.97, as can be seen from the following classification report:

	precision	recall	f1-score	support
positive	0.98	0.97	0.98	3126
negative	0.97	0.98	0.97	2812
neutral	0.96	0.96	0.96	107
macro avg	0.97	0.97	0.97	6045
weighted avg	0.97	0.97	0.97	6045
accuracy			0.97	6045

Such a high level of accuracy is not always expected, but in this case not necessarily a surprise when considering the high number of examples available in the training dataset and the fact that the headlines do not contain that many words (chapter 4.3). Meaning, less opportunity for noisy words such as stop words, including negates, thus less potential for misinterpretations.

The confusion matrix (figure 36) illustrates the low number of neutral labels, which is expected since few were detected when creating the S&D-Dataset. Although this kind of skewness can be a problem in certain situations, the main reason it was included in this case was to potentially provide the neural network with a way to assign lower scores when in doubt instead of opting for the two extremes positive, and negative.

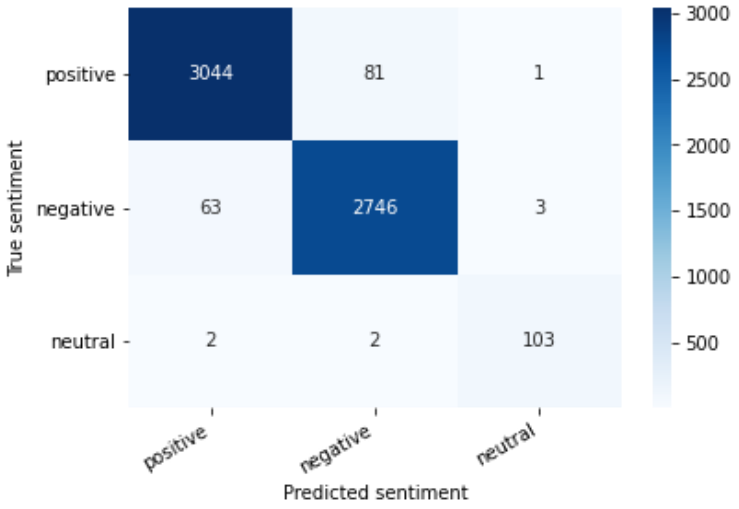

Figure 37: Confusion Matrix of CrudeBERT

Since the training results are quite high, the adjustment of the hyperparameters was not deemed necessary. Even though thousands (20%) of the training data have been extensively used to evaluate the accuracy, the twelve examples from chapter 5.2 (tables 5 – 8) have been classified by CrudeBERT to confirm once again if the output was accurate as desired, which was the case.

	Headlines	Sentiment Score SHOULD	Sentiment Score CrudeBERT
Shortage	**Major Explosion, Fire at Oil Refinery in Southwest Philadelphia**	Positive	0.992527
	Petroleos Mexicanos confirms Gulf of Mexico oil platform accident	Positive	0.995259
	CASUALITIES FEARED AT OIL ACCIDENT NEAR IRAQ'S BASRA	Positive	0.970709
Surplus	**Iraq's Feb Oil Exports +20.9% On Mo At 1.56M B/D - Official**	Negative	-0.995265
	Apache announces large petroleum discovery in Texas	Negative	-0.996137
	Turkey finds oil near Iran, Iraq border	Negative	-0.995416
Surplus	**Turkey February Crude Imports -16.0% On Year -METI**	Negative	-0.985711
	OIL DATA: Japan May Crude Imports -11.0% On Year	Negative	-0.987148
	China June Crude Oil Imports -10.9% On Yr,16.60M KL-METI	Negative	-0.987117
Shortage	**EIA Chief: Expects Global Oil Demand Growth 1.5M B/D To 2010**	Positive	0.988338
	China Jan-Oct Crude Imports +98.5% To 57.9M MT	Positive	0.989277
	Turkey's crude oil imports up 78.30% in February 2019	Positive	0.992335

Table 10: Sample of Headlines to evaluate the Classification of CrudeBERT

5.4 Evaluating CrudeBERT

The same analysis as at the beginning of chapter 5.1 has been conducted for CrudeBERT. Interestingly, the pattern of the cumulative sentiment scores from CrudeBERT feature a much higher resemblance of the WTI crude oil prices compared to FinBERT (figure 37).

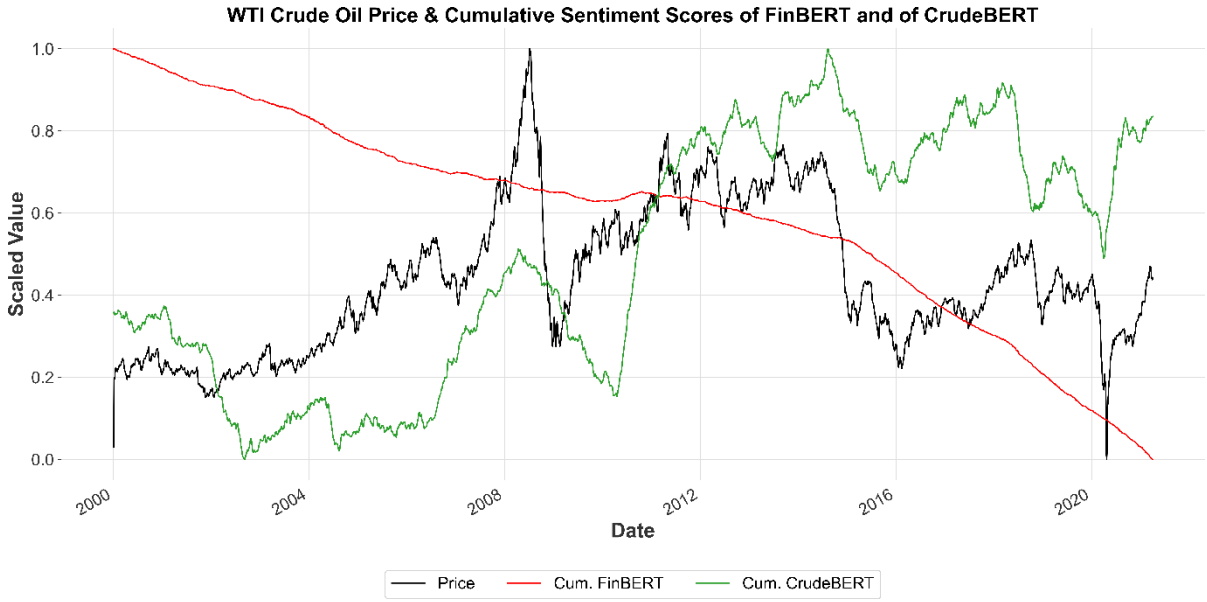

Figure 38: WTI Crude Oil and Cumulative Sentiment Scores of FinBERT and of CrudeBERT

Furthermore, a significant number of the significant changes of direction from CrudeBERT occur prior to the price changes of crude oil, as can be seen, marked green in the following figure:

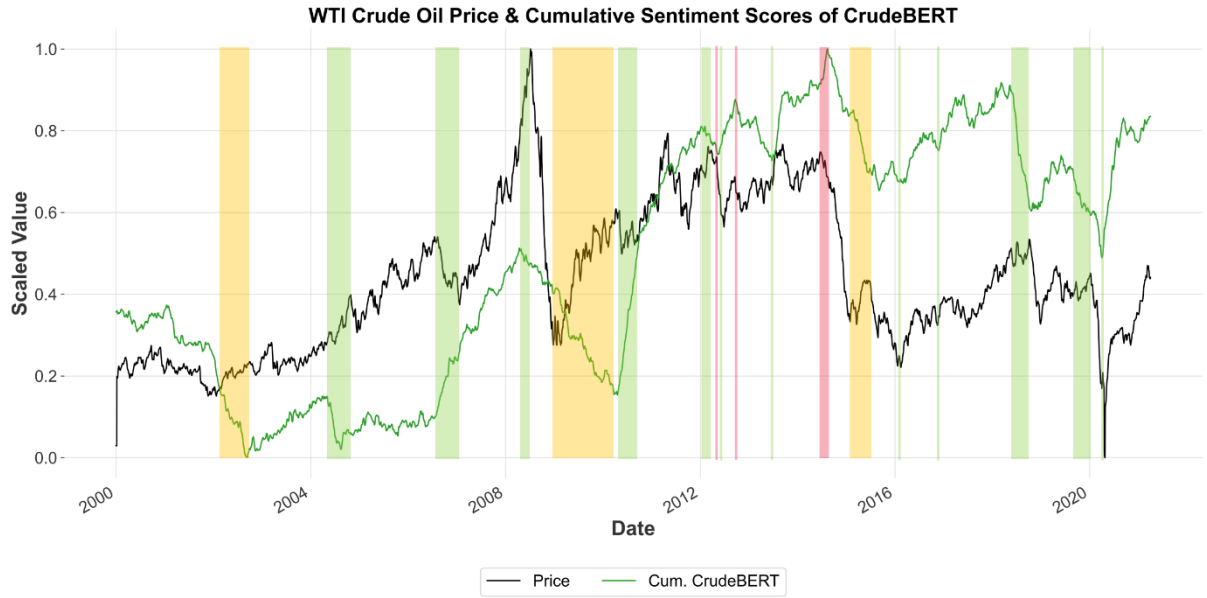

Figure 39: WTI Crude Oil and Cumulative Sentiment Scores of CrudeBERT with highlighted changes of direction

However, there are also instances in which the sentiment of the headlines keeps going down while the price decline has recovered and started rising again (orange). On some occasions, the sentiment lagged slightly behind when the price drastically started falling (red).

The following scatter plots reveal improvements compared to FinBERT. Here, the linear relationship is positive and slightly stronger (R = 0.0529):

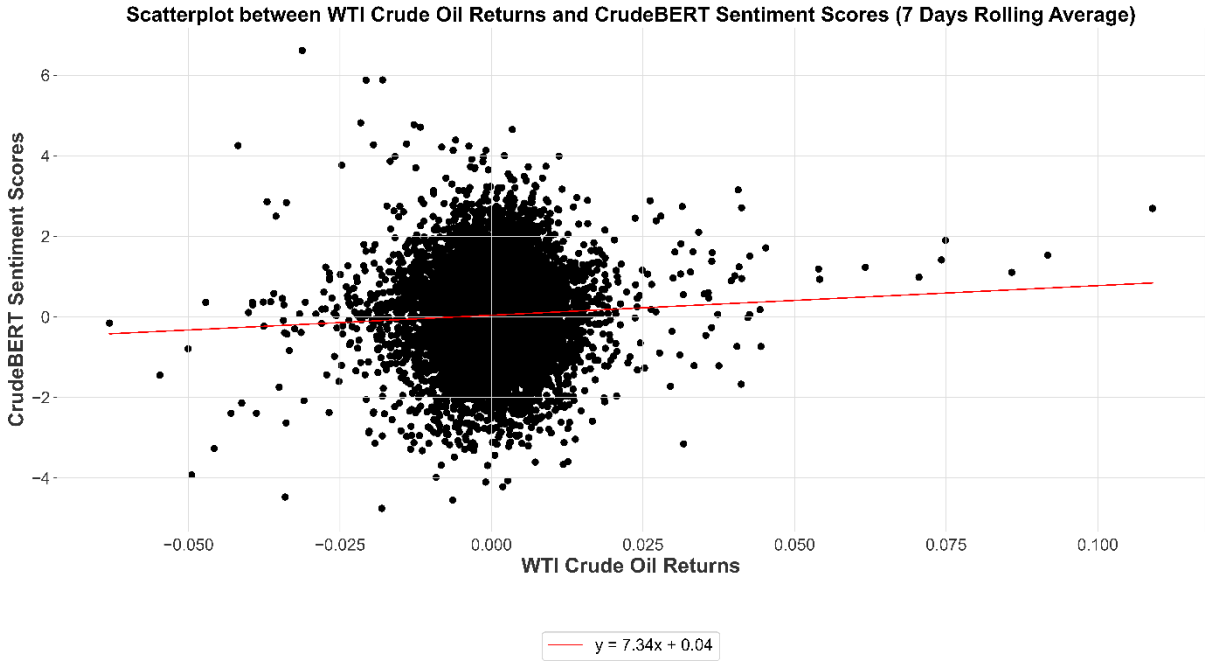

Figure 40: Scatterplot between WTI Crude Oil Returns and CudeBERT Sentiment Scores (window=7 days)

In contrast to FinBERT, increasing the SMA-window to 90 days improves the linear relationship (R = 0.2405). Moreover, when looking at the scatter plot, it visualizes that higher sentiment scores and higher returns co-occur slightly more often. The same goes for lower value pairs.

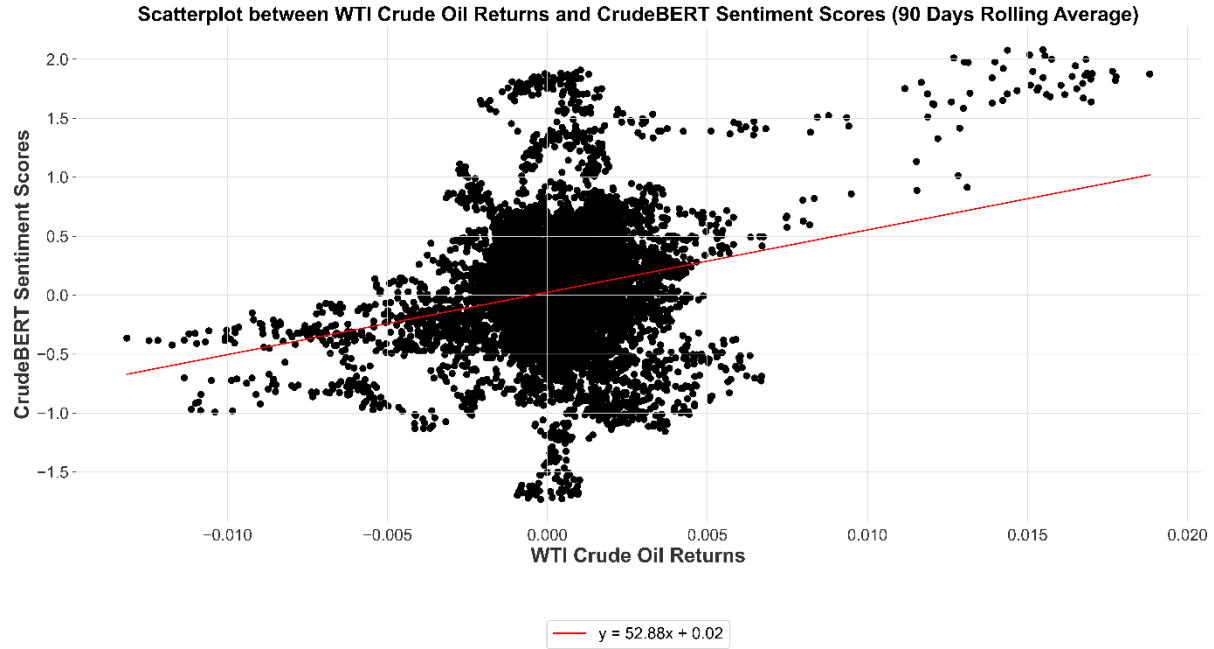

Figure 41: Scatterplot between WTI Crude Oil Returns and CudeBERT Sentiment Scores (window=90 days)

According to these observations, the application of domain adaptation appears to lead to substantial improvements. However, the undertaken domain adaptation is subject to improvement since the S&D-Dataset is based on a fundamental theory regarding supply and demand.

5.4.1 Consulting an Expert

In order to assess and potentially improve the domain adaptation behind CrudeBERT, an interview on 28th and 29th June 2021 with Dr Karin Kneissl has been consulted. This chapter covers a brief introduction and a summary of the interview between her and the author.

Dr Kneissl has published several books and numerous papers and articles about geopolitics, energy and international relations. She joined the Austrian Ministry for Foreign Affairs in 1990, served in Paris and Madrid, as well as in the Legal office. Later, she served her country as Foreign Minister in 2017-19. In addition, she contributes guest comments on middle eastern and energy matters to RT (Russia Today) and several other media sources, including Cicero. Furthermore, she has been doing online briefings, primarily for corporations outside of Austria. In June 2021, she got elected as an independent director of the board of Rosneft, one of the world's largest oil companies. (Kneissl, n. d.)

After a brief introduction regarding the research objectives of this work, Dr Kneissl shared some of her first-hand experiences regarding the price development of oil.

For instance, on a Friday afternoon in May 2008, the price of crude oil increased within seconds by 10 dollars. Reasons behind this were the speculative and hysterical expressions of specific stakeholders such as chief analysts from big banks and other political people. The expressions usually started as so-called if-statements, such as "if problems with X occur, then the oil price will go direction Y!". According to her, this is one of the many prime examples asserting that the oil market price is mainly influenced by hysterical actions, making it hard to estimate. She mentioned that even the general secretary of OPEC stated in an emergency meeting with the OPEC ministers held on 23rd July 2008 that they do not understand the price of oil anymore. She explained that the origins of this volatile oscillation of oil price go back to the early 80s when the separation between paper and physical markets occurred. These allowed the trading of commodities without physical exchange, for instance, in the forms of a futures contract (also called futures). She stated that futures were initially supposed to be smart instruments for stabilizing the market. However, the involvement of a high number of participants caused dramatic distortion in prices and disrupted the oil market. She reasons that before the market separation, the price of crude oil remained largely stable for decades because the market was only physical. Back then, the price of crude oil did not experience a sudden 100% change in value. At the same time, the oil itself physically (like extraction, production, transport) saw no such radical changes. Therefore, Dr Kneissl states that crude oil pricing is, since the separation of the physical market, mainly driven by financial assessments and not by physical supply and demand. She gave one example for the year 2014, in which Syria was affected by the Islamic state crisis. Syria is an oil-exporting country, but due to this crisis, its oil output diminished substantially. However, despite this reduction in supply, the price kept falling sharply during this period. So, the logic behind less supply resulting in an increased price is not always the case. Furthermore, she adds that the entire energy market is rarely rational instead mainly emotionally driven. According to her, energy politics is tightly entangled with climate politics, which is a concerning dilemma since it leads to significant underinvestment in oil-related infrastructures. (K. Kneissl, personal communication, 2021)

However, despite her firm conviction that the physical oil supply and demand do not primarily determine oil prices, she did not entirely dismiss the potentials of analyzing these physical properties. She could imagine that such an analysis could potentially lead to short-term profits. For this matter, she made the following recommendations in order to improve the S&D-Dataset, thus the domain adaptation of CrudeBERT:

1. She suggested including futures as a topic since the futures market has many participants and can convey lots of additional information.

2. She also stated that journalists are rarely keen on matters regarding energy, therefore, are not truly competent to deliver accurate intel about the oil market. Therefore, it could be beneficial to reduce sources to a few but highly reputable news agencies.

3. According to her, another critical factor could be incorporating the technological aspects of crude oil into the domain adaptation. For instance, through the technological advancements in the shale industry, the USA transformed from the most prominent importer to an exporter of oil.

4. She is also a firm supporter of involving human experts when interpreting a text since they could read between the lines or get more out of it than machines.

(K. Kneissl, personal communication, 2021)

5.4.2 Development and Evaluation of CrudeBERTv2

Her first advice, introducing and including changes in futures, did bring a considerable amount (3'964) of additional findings. Originally the S&D-Dataset represented around two-thirds of all the news (chapter 5.3). With the inclusion of futures, the ratio of coverage increased to three-quarters (74.48%):

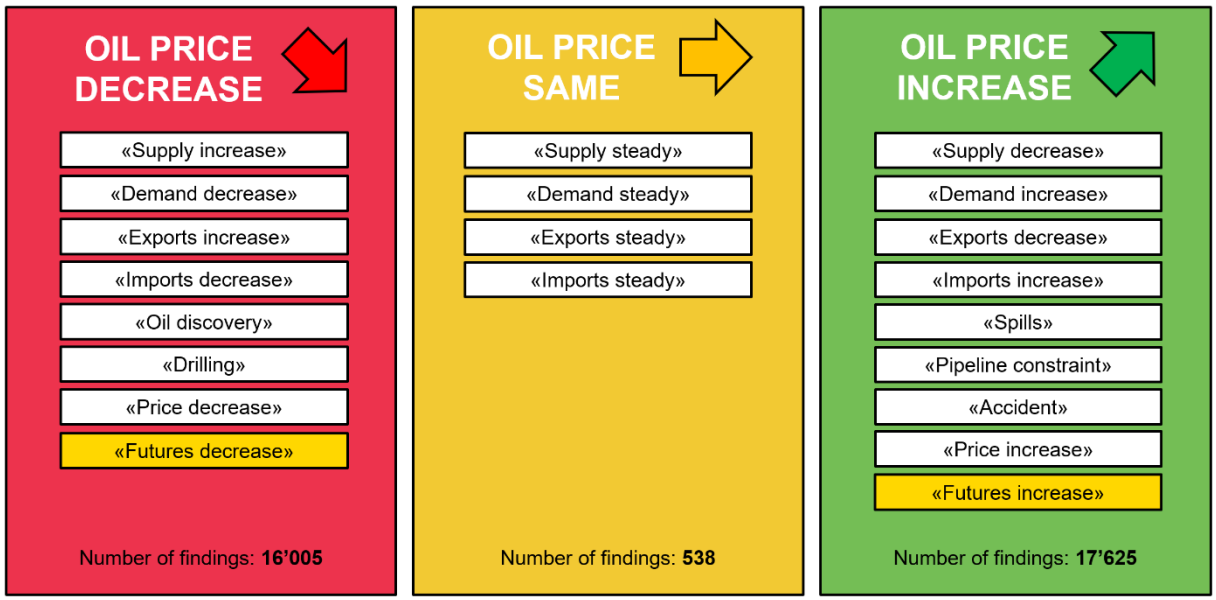

Figure 42: S&D-Dataset including Futures

For this matter, headlines containing the words "futures" and "nymex" were labelled as "futures". Its direction of change has been determined the same way as the other topics. The exact training process (chapter 5.3) has been repeated on the new enhanced S&D-Dataset, leading to the creation of CrudeBERTv2. The accuracy of this dataset was also very high. The classification report and confusion matrix of CrudeBERTv2 can be seen in Appendix C.

The comparison between the cumulative sentiment scores of CrudeBERTv2 and WTI crude oil price revealed similar characteristics as the original CrudeBERT (figure 43).

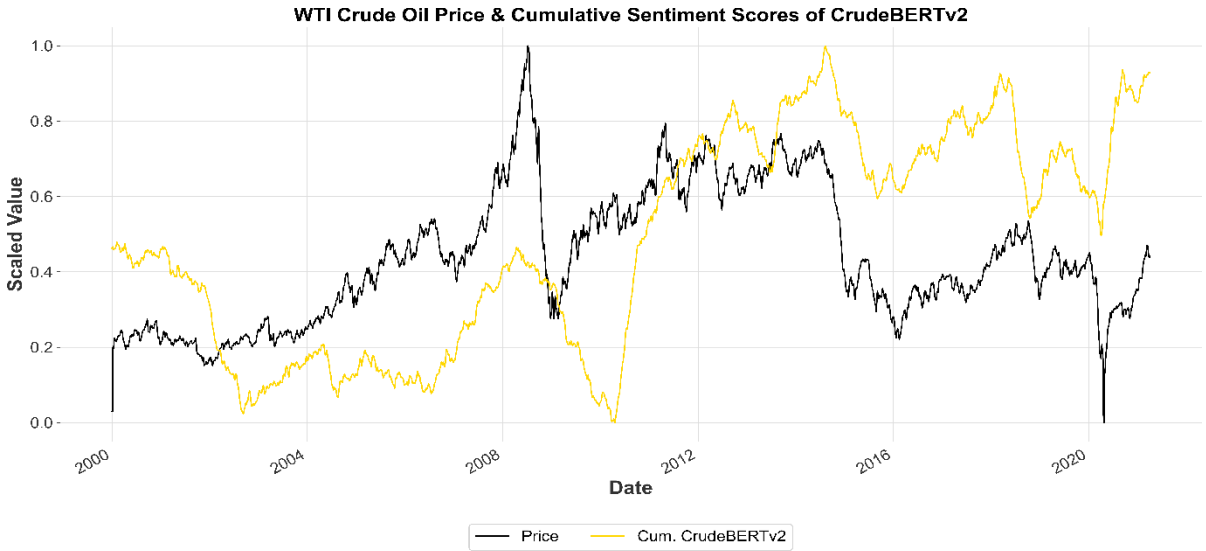

Figure 43: WTI Crude Oil and Cumulative Sentiment Scores of CrudeBERTv2

However, the regression analysis revealed a slight increase in the linear relationship between the sentiment scores of CrudeBERTv2 and the returns of WTI crude oil (R = 0.0590).

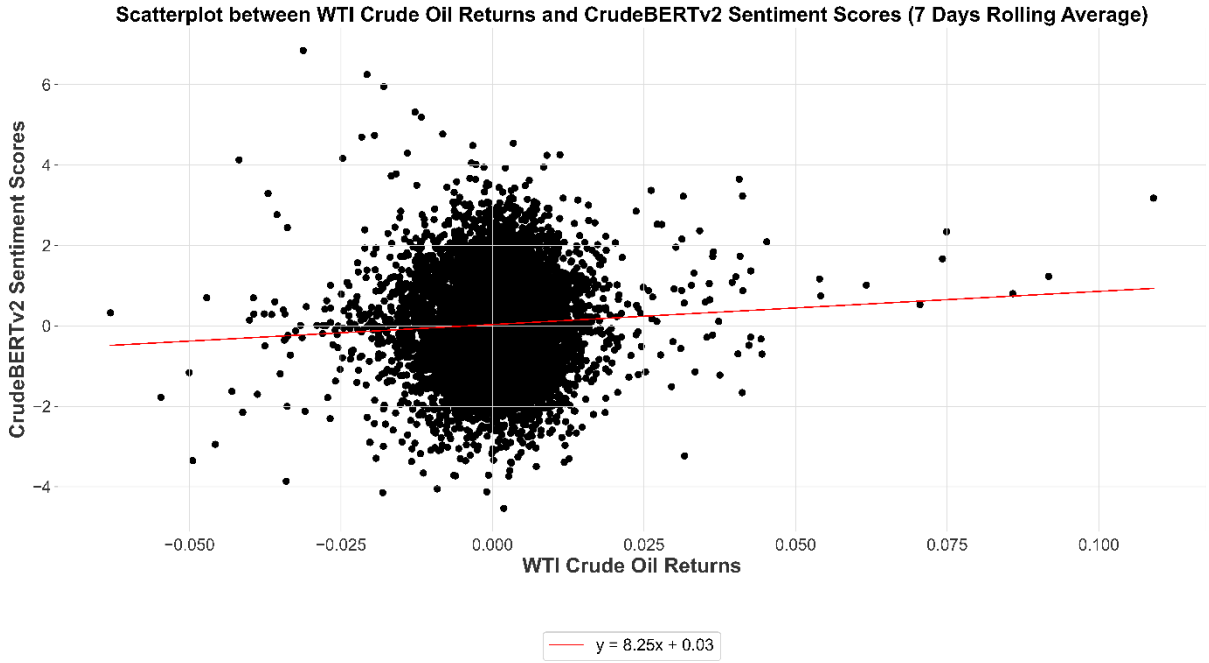

Figure 44: Scatterplot between WTI Crude Oil Returns and CudeBERTv2 Sentiment Scores (window=7 days)

This increase is also true when a longer window of 90 days for the moving average is selected. The Pearson correlation coefficient did increase marginally from 0.2405 to 0.2413 compared to CrudeBERT (figure 45).

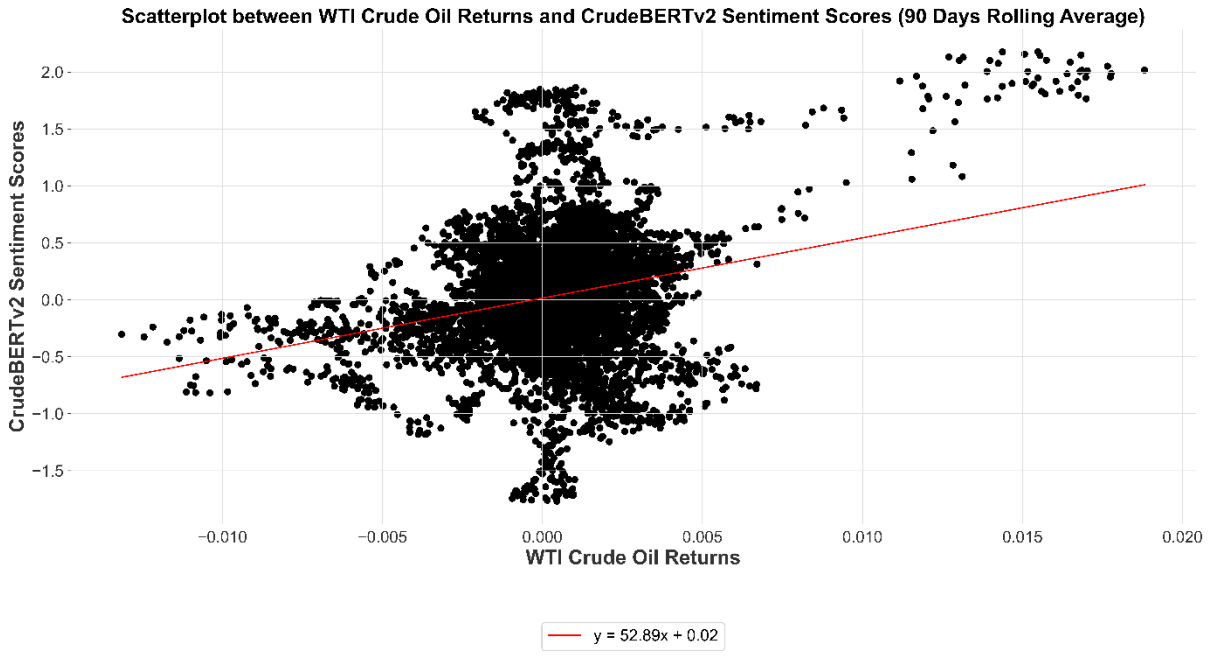

Figure 45: Scatterplot between WTI Crude Oil Returns and CudeBERTv2 Sentiment Scores (window=90 days)

Including futures as a topic in the S&D-Dataset, seems to improve, albeit very slightly. The model.

5.4.3 Development and Evaluation of CrudeBERTv2_T4

Regarding her second advice, reducing the number of sources, she mentioned four suitable publishers, namely Bloomberg News, Dow Jones Newswire, Reuters and Platts. These news agencies are also the top four publishers who provided the most headlines for this thesis (figure 20 in chapter 4.3).

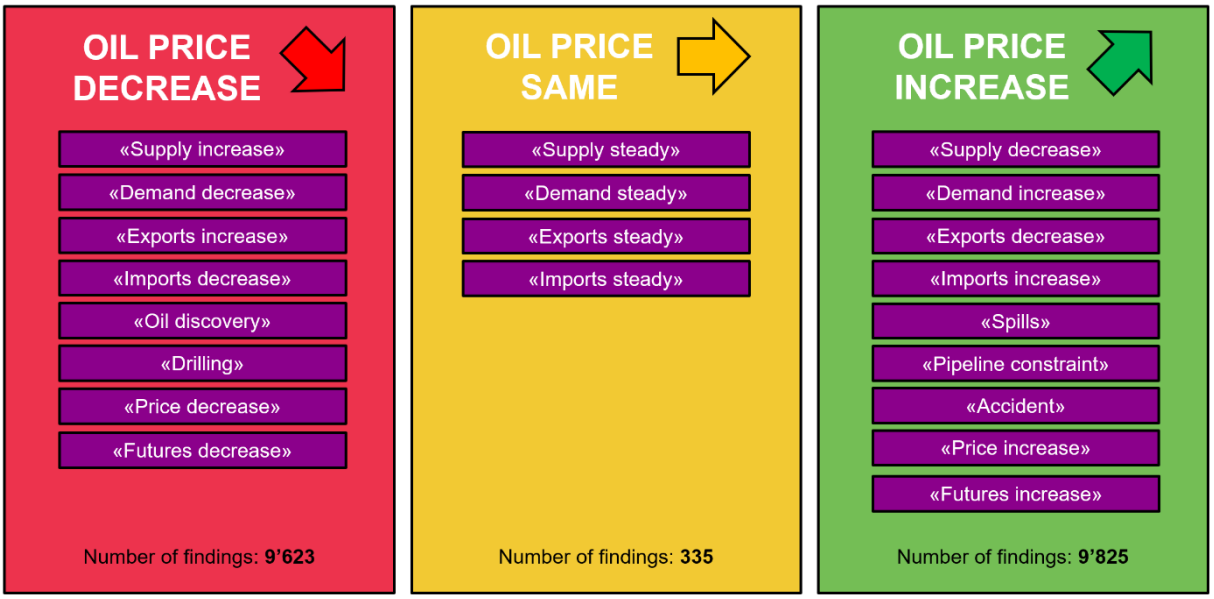

Figure 46: S&D-Dataset containing only headlines from the top four publishers

For this matter, the sources were limited to the above mentioned four publishers, simply via the filter function of MS Excel. This time the size of the new S&D-Dataset was much smaller, down to 19'783 from 34'195 headlines. With this newly training dataset, CrudeBERTv2_T4 was created. Interestingly, this reduction in the dataset did not lead to a major loss in accuracy. The accuracy was the same as the earlier CrudeBERT-Variants (appendix D).

The comparison between the cumulative sentiment scores of CrudeBERTv2_T4 and WTI crude oil also shows a pretty similar characteristic, especially after 2008:

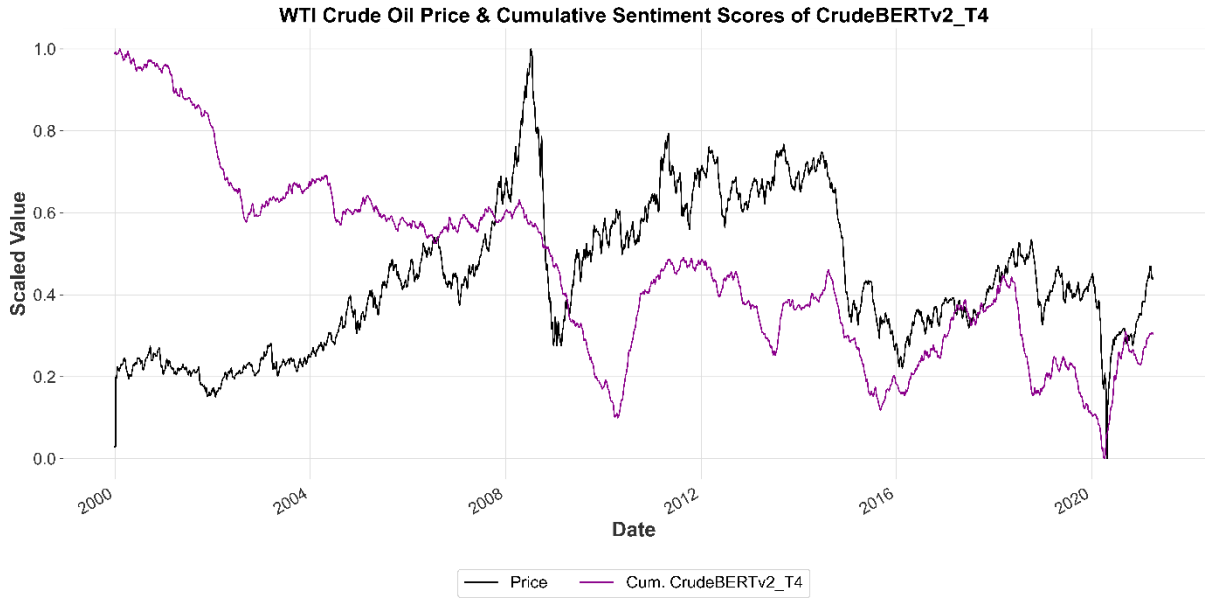

Figure 47: WTI Crude Oil and Cumulative Sentiment Scores of CrudeBERTv2_T4

The distinct deviation in figure 47 could be due to the lack of available headlines from those top four publishers for that period. However, after 2008, despite a much lower number of headlines, the cumulative sentiment scores of CrudeBERTv2_T4 show a similar profile as the earlier CrudeBERT variants. As a result, the regression analysis even yields a higher R-value of 0.06013 (figure 48), despite the initial offset.

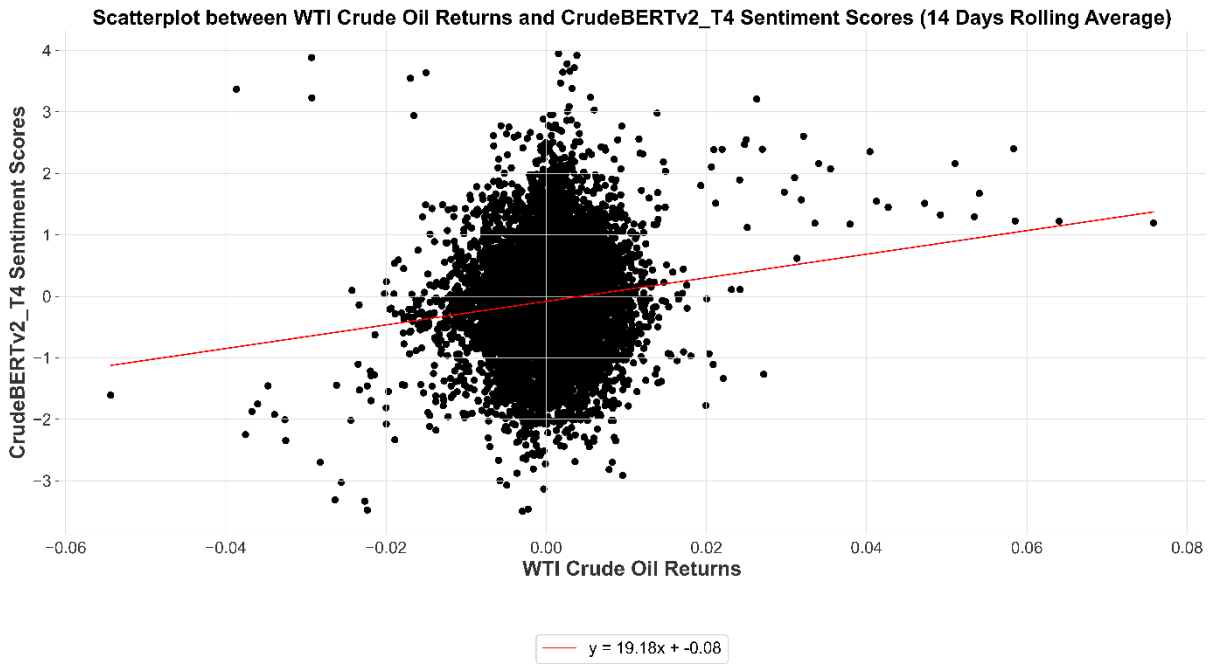

Figure 48: Scatterplot between WTI Crude Oil Returns and CudeBERTv2_T4 Sentiment Scores (window=7 days)

Enlarging the window of the SMA to 90 days also delivers an improved linear relationship between the sentiment scores of CrudeBERTv2_T4 and returns of WTI crude oil (R = 0.2785).

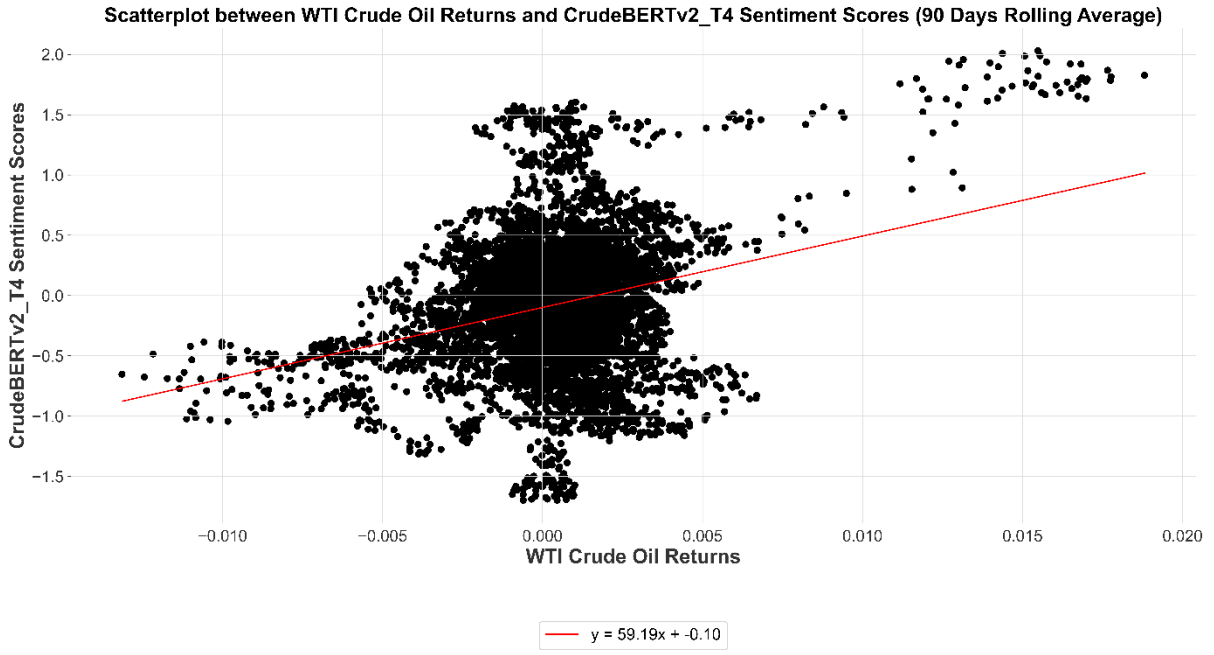

Figure 49: Scatterplot between WTI Crude Oil Returns and CudeBERTv2 _T4 Sentiment Scores (window=907 days)

5.4.4 Development and Evaluation of CrudeBERTv2_GT

Another approach to further improve CrudeBERT would be the incorporation of news volume. The importance of news volume was hinted at in chapter 4.6 when justifying why the sum of multiple sentiment scores should be favoured over the mean on a given day. However, this sum might not represent the actual volume since RavenPack filters out duplicates when high relevance is selected in their filter. In order to provide some insight regarding the news volume concerning crude oil, the traffic activity of specific keywords related to crude oil could be inspected. Google Trends[18] offers keyword statistics for certain words or themes. Figure 50 shows the trends behind the chosen keywords related to crude oil, dating back to 2004.

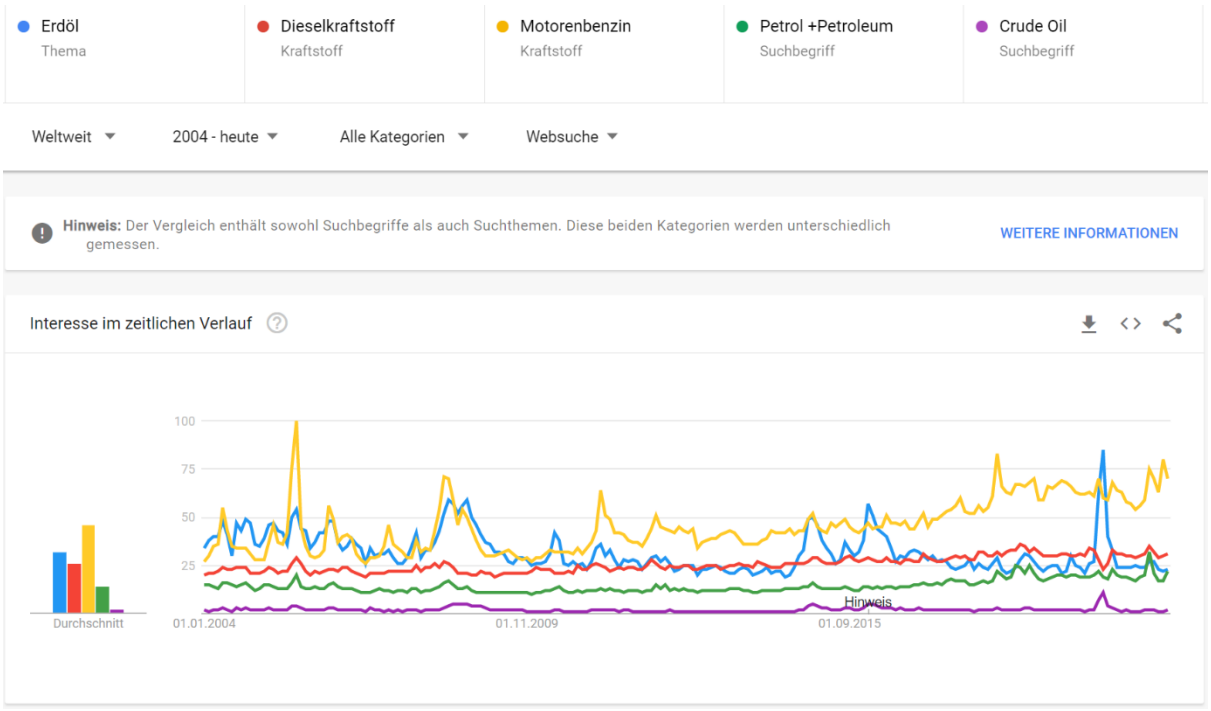

Figure 50: Comparison of various Google Trends Keyword Statistics related to Crude Oil

The statistics behind the word “crude oil” coincided with the WTI crude oil prices' peaks quite well and were therefore selected as an alternative source for news volume.

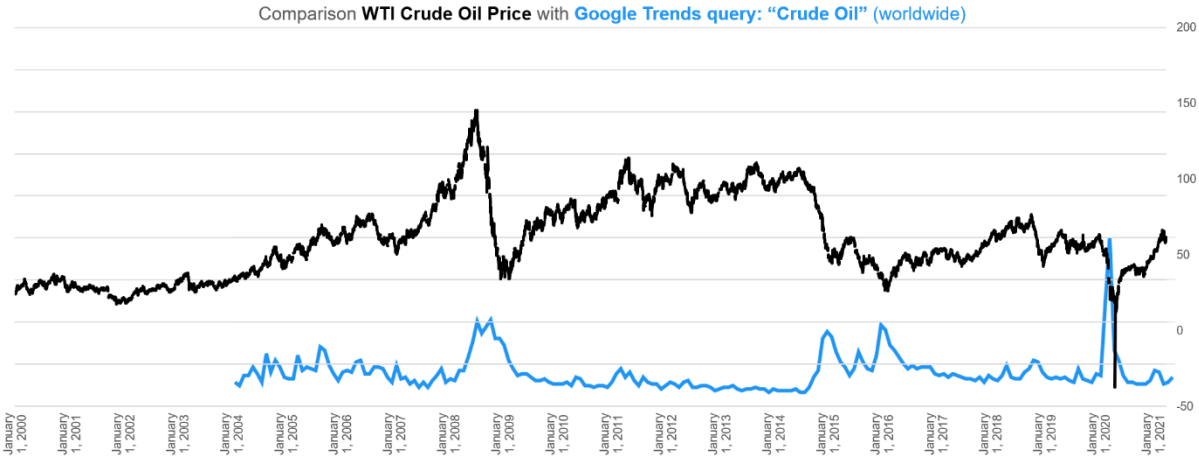

Figure 51: Comparison WTI Crude Oil Price with Google Trends Keyword Statistics of “Crude Oil”

[18] https://trends.google.com/trends/

In theory, these keyword statistics could provide information about global online traffic of the word “crude oil”. One way to utilize this keyword statistics as a volume would be by multiplying it with the sentiment scores of CrudeBERTv2. Figure 52 illustrates the result of this operation:

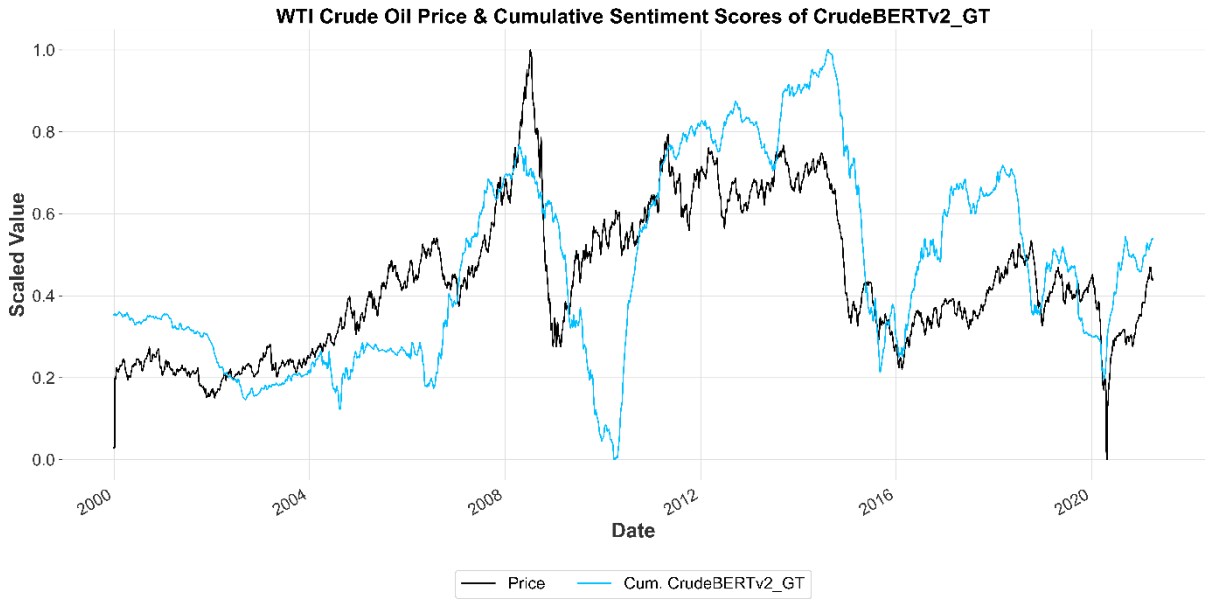

Figure 52: WTI Crude Oil and Cumulative Sentiment Scores of CrudeBERTv2 including the Google Trends Factor

At first glance, it appears that the slopes around the years 2008 and 2014 are better tracked. The regression analysis with the weekly SMA further backs this appearance with the highest R-value of 0.0649 achieved so far.

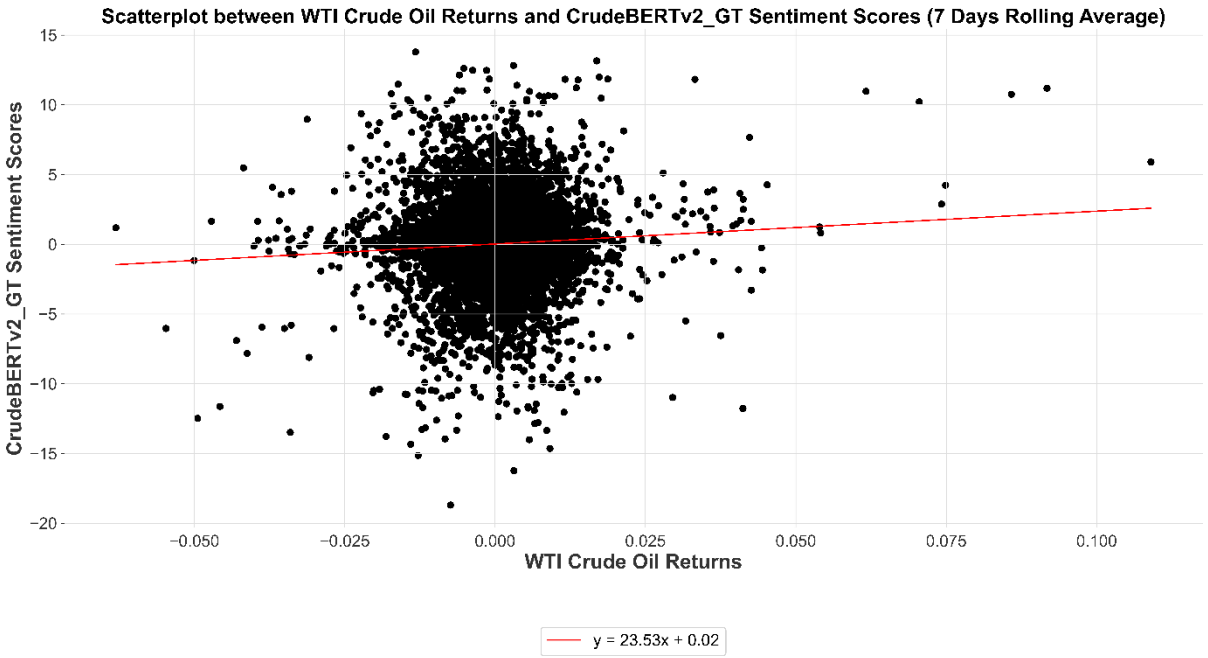

Figure 53: Scatterplot between WTI Crude Oil Returns and CudeBERTv2_GT Sentiment Scores (window=7 days)

However, the R-value is the lowest when taking a longer SMA window of 90 days (R = 0.1983). This result indicates that additional volume acquired from an external source could benefit short-term forecasts but worsen the long-term ones (figure 54).

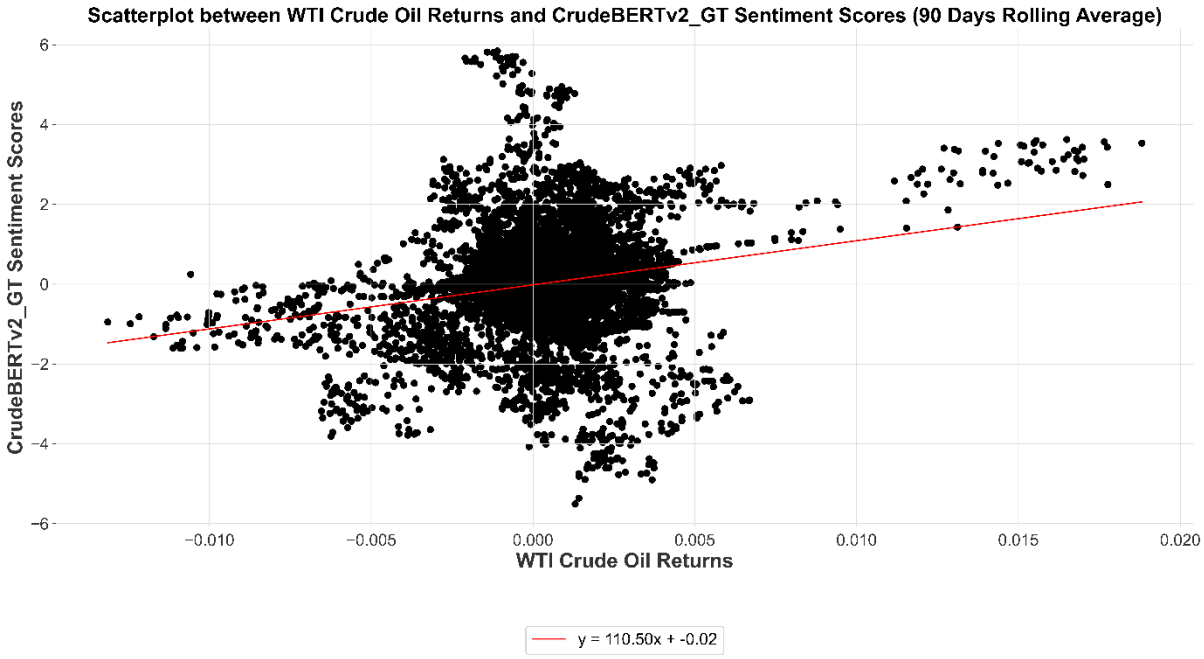

Figure 54: Scatterplot between WTI Crude Oil Returns and CudeBERTv2_GT Sentiment Scores (window=90 days).

The following figure recapitulates this chapter to some extent by illustrating the WTI crude oil price, the cumulative sentiment scores from FinBERT and all of the CrudeBERT-variants.

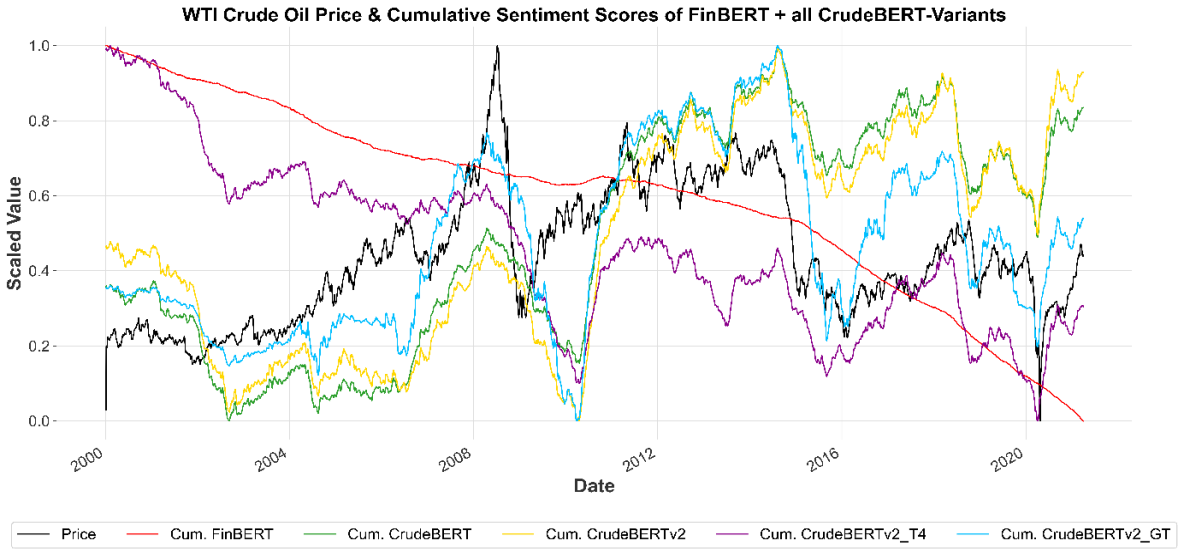

Figure 55: WTI Crude Oil and Cumulative Sentiment Scores of FinBERT and all CrudeBERT-Variants

According to this figure, each iterative step led to an improvement for a particular case. However, there are also trade-offs. For instance, the latest variant, CrudeBERTv2_GT, is superior at tracking extreme price changes, but it requires additional external data. On the other hand, CrudeBERTv2_T4 can work with fewer data and still deliver relatively high results. However, the significant deviation, in the beginning, is concerning. Nevertheless, CrudeBERTv2 seems to provide a balance and is believed the best variant, despite having a slightly lower R-value.

6 Conclusion and further work

"A strong physical constitution can tolerate extremes of hot and cold; people of strong mental health can handle anger, grief, joy and the other emotions." – Epictetus, Greek philosopher

When recapitulating the result of this work, one could add to this quote that people of strong mental health and machines with strong domain adaptation can handle sentiment.

Performing sentiment analysis on the news regarding a specific asset requires domain adaptation. Domain adaptation requires training data made up of examples with text and its associated polarity of sentiment. The experiments show that pre-trained deep learning-based sentiment analysis can be further fine-tuned, and the conclusions of these experiments are as following:

- Deep learning-based sentiment analysis models from the general financial world such as FinBERT are of little or hardly any significance concerning the price development of crude oil. The reason behind this is a lack of domain adaptation of the sentiment. Moreover, the polarity of sentiment cannot be generalized and is highly dependent on the properties of its target.

- The properties of crude oil prices are, according to the literature, determined by changes in supply and demand. News can convey information about these direction changes and can broadly be identified through query searches and serve as a foundation for creating a training dataset to perform domain adaptation. For this purpose, news headlines tend to be rich enough in content to provide insights into supply and demand changes. Even when significantly reducing the number of headlines to more reputable sources.

- Domain adaptation can be achieved to some extend by analyzing the properties of the target through literature review and creating a corresponding training dataset to fine-tune the model. For example, considering supply and demand changes regarding crude oil seems to be a suitable component for a domain adaptation.

In order to advance sentiment analysis applications in the domain of crude oil, this paper presents CrudeBERT. In general, sentiment analysis of headlines from crude oil through CrudeBERT could be a viable source of insight for the price behaviour of WTI crude oil. However, further research is required to see if CrudeBERT can serve as beneficial for predicting oil prices. For this matter, it is made publicly available on the Hugging Face platform[19] for future research.

[19] https://huggingface.co/Captain-1337/CrudeBERT/

7 Limitations and Future Research

Despite the iterative improvements in chapter 5, CrudeBERT has its limitations, and there is still much room for optimization. For instance, the undertaken evaluations using linear regression and utilizing scatter plots did indicate improvements. However, further research on evaluation methods and metrics should be conducted.

Furthermore, the obtained sentiment scores can then be further improved by incorporating factors such as news volume to help determine the magnitude of the sentiment. Additionally, awareness of numerals resulting in a more fine-grained sentiment score could be beneficial. For instance, the sentence “crude oil imports increased by 10 % today!” needs to have a higher sentiment than “crude oil imports increased by 1 % today!”. However, the experiment in Appendix B (table 12 and 13) show that neither FinBERT nor CrudeBERT can differentiate between smaller or larger numbers. Therefore, a training dataset that makes the model sensitive to numerals is required. For this matter, here, the sentiment labels have to be fine-grained.

Furthermore, recognizing and emphasizing the importance of certain countries or companies could serve as determining factor. In addition, news covering technological advancements in the field of crude oil regarding exploration, extraction, refining, production, and transport could also be factored in to improve the impact of the sentiment score towards the crude oil price.

It is also essential that the news headlines, which are being used have high relevance to crude oil. In this thesis, the topics were identified through manual scanning of hundreds of headlines by the author. This step is relatively inefficient and risks missing out on important terms, which was the case concerning futures. Despite being present in about almost every tenth headline, it was overseen. One way to improve this would be using topic modelling algorithms such as Latent Dirichlet Allocation or Contextual Topic Modelling (Tschudy, 2020, p. 2). The query search which was utilized to flag and label topics is also not free from errors. For instance, some headlines include multiple intel regarding supply and demand in the same headline, e.g., "Ivory Coast Jan Crude Oil Exports -1% On Yr, Imports -3%”. Also, despite a low number of words in a headline, it can still contain words that can either contradict or cancel out information. For example, the following headline was assigned in the topic of “Pipeline constraint”: “Canadian pipeline that leaked oil is fixed.” The word fixed cancels out leaked, meaning the supply line has been recovered, which should not express a negative sentiment. Another issue with the query search approach is the potentially incomplete list of synonyms for correctly classifying increases or decreases (Appendix A). However, it could also be assumed that the pretrained BERT embeddings should be able to interpret an increase or decrease in any shape or form when given a few examples.

Another critical potential bottleneck could be the steps involving preprocessing and merging of the data. For example, handling and filling missing values by just quadric interpolation might have shortcomings and are subject to further research. Also, removing seasonality could be a beneficial preprocessing step. For instance, analysis trials with Prophet (Darts) in Appendix E

indicate that both the returns of WTI crude oil (figure 58) and the sentiment scores of CrudeBERT (figure 59) have similar weekly seasonality. Also

Lastly, suitable methods need to be examined for bivariate (also known as multivariate) time-series analysis to unveil any potential in utilizing sentiment scores of the headlines for predicting the price changes of WTI crude oil. In finance, multiple models are employed to make forecasts, but it turns out that their accuracy is generally insufficient. For instance, Mascio et al. (2021, p. 1–2) utilize machine learning as well as logistic regression models to forecast the returns of the S&P 500 a few weeks in advance. Depending on the model, there are different evaluation metrics.

For machine learning-based forecasting, the open-source library Darts[20] could serve as a good starting point. Darts contains many forecasting models for multivariate forecasting models such as RNN's, Temporal Convolutional Networks, Transformer Models and N-Beats. A few forecasting trials with Darts utilizing N-Beats (Oreshkin et al., 2020) have been conducted with Darts to get a mere glimpse of the forecasting potentials of CrudeBERTv2 (Appendix F). It should also be noted that each trial took around 30 minutes to run for this reason. Only a handful of runs for short (1 week) and long term (1 quarter) has been conducted. However, the results of these experiments will not be discussed further, as the experiments were carried out without scientific grounds but only out of curiosity.

For logistic regression, the open-source library PyCaret[21] could be considered promising. It has many models for regression and classification such as Naive Bayes, Decision Tree Classifier, Support Vector Machines, Ridge Classifier, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Light Gradient Boosting Machine, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Extra Trees Classifier, K Neighbors Classifier and Logistic Regression. The results of the trials with PyCaret can be seen without interpretation in Appendix G.

[20] https://github.com/unit8co/darts/
[21] https://github.com/pycaret/pycaret/

8 Literature

About Investing.com. (o. J.). Investing.Com. Abgerufen 6. August 2021, von https://www.investing.com/about-us/

ABOUT ME — English. (o. J.). Abgerufen 11. August 2021, von http://www.kkneissl.com/en/about-me

Agarwal, P. (2018). *Supply and Demand*. INTELLIGENT ECONOMIST. https://www.intelligenteconomist.com/supply-and-demand/

AllenNLP. (o. J.). Abgerufen 8. August 2021, von https://allennlp.org/%PUBLIC_URL%

Anandarajan, M., Hill, C., & Nolan, T. (2019). *Practical Text Analytics: Maximizing the Value of Text Data* (1st ed. 2019). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-95663-3

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv:1908.10063 [cs]*. http://arxiv.org/abs/1908.10063

Archive | Dev Portal. (o. J.). Abgerufen 6. August 2021, von https://developer.nytimes.com/docs/archive-product/1/overview

Baboshkin, P., & Uandykova, M. (2021). MULTI-SOURCE MODEL OF HETEROGENEOUS DATA ANALYSIS FOR OIL PRICE FORECASTING. *International Journal of Energy Economics and Policy*, *11*(2), 384–391. https://doi.org/10.32479/ijeep.10853

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. http://arxiv.org/abs/1409.0473

Bai, Y., Li, X., Yu, H., & Jia, S. (2021). Crude oil price forecasting incorporating news text. *arXiv:2002.02010 [q-fin]*. http://arxiv.org/abs/2002.02010

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Biagioni, R. (2016). *The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts* (Bd. 4). Springer International Publishing. https://doi.org/10.1007/978-3-319-38971-4

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv:1607.04606 [cs]*. http://arxiv.org/abs/1607.04606

Brandstätter, V., Schüler, J., Puca, R. M., & Lozo, L. (2013). *Motivation und Emotion*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30150-6

Brandt, M. W., & Gao, L. (2019). Macro fundamentals or geopolitical events? A textual analysis of news events for crude oil. *Journal of Empirical Finance*, *51*, 64–94. https://doi.org/10.1016/j.jempfin.2019.01.007

Buyuksahin, B., & Harris, J. (2011). Do Speculators Drive Crude Oil Futures Prices? *The Energy Journal*, *Volume 32*(Number 2), 167–202.

Cambria, E., Li, Y., Xing, F. Z., Poira, S., & Kwok, K. (2020). SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis. *Proceedings of The Conference on Information and Knowledge Management*.

Cambria, E., Livingstone, A., & Hussain, A. (2012). The Hourglass of Emotions. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, & V. C. Müller (Hrsg.), *Cognitive Behavioural Systems* (Bd. 7403, S. 144–157). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34584-5_11

Candia, J., & Mazzitello, K. I. (2008). Mass Media Influence Spreading in Social Networks with Community Structure. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(07), P07007. https://doi.org/10.1088/1742-5468/2008/07/P07007

Chollet, F. (2018). *Deep learning with Python*. Manning Publications Co.

Darwin, C. (1872). *The expression of the emotions in man and animal.* John Murray. http://darwin-online.org.uk/converted/pdf/1872_Expression_F1142.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. http://arxiv.org/abs/1810.04805

Evans, H., Hu, M., Kuchembuck, R., & Gervet, E. (2017). *Will You Embrace AI Fast Enough?* A.T. Kearney, Inc.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383. https://doi.org/10.2307/2325486

Feuerriegel, S., & Neumann, D. (2013). News or Noise? How News Drives Commodity Prices. *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design*, *3*, 2695–2714.

Frederking, R. E. (2004). *Machine Translation: From Real Users to Research 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA, September 28 - October 2, 2004. Proceedings*. https://doi.org/10.1007/b100780

Futrzynski, R. (2020, September 1). Getting meaning from text: Self-attention step-by-step video. *Peltarion*. https://peltarion.com/blog/data-science/self-attention-video/

Gan, B., Alexeev, V., Bird, R., & Yeung, D. (2020). Sensitivity to sentiment: News vs social media. *International Review of Financial Analysis*, *67*, 101390. https://doi.org/10.1016/j.irfa.2019.101390

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv:1410.5401 [cs]*. http://arxiv.org/abs/1410.5401

Hafez, P., Matas, R., Lautizi, F., A. Guerrero-Colón, J., Gómez, M., & Gómez, F. (2018). *Effects of Event Sentiment Aggregation: Sum vs. Mean* [White Paper]. RavenPack. https://www.ravenpack.com/research/sum-vs-mean-event-sentiment-aggregation/

Hamilton, J. (2008). *Understanding Crude Oil Prices* (Nr. w14492; S. w14492). National Bureau of Economic Research. https://doi.org/10.3386/w14492

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. *arXiv:1606.02820 [cs]*. http://arxiv.org/abs/1606.02820

Hovy, E. H. (2015). What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis. In N. Gala, R. Rapp, & G. Bel-Enguix (Hrsg.), *Language Production, Cognition, and the Lexicon* (S. 13–24). Springer International Publishing. https://doi.org/10.1007/978-3-319-08043-7_2

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *arXiv:1801.06146 [cs, stat]*. http://arxiv.org/abs/1801.06146

Huggingface (Hugging Face). (o. J.). Abgerufen 9. August 2021, von https://huggingface.co/huggingface

Iworiso, J., & Vrontos, S. (2020). On the directional predictability of equity premium using machine learning techniques. *Journal of Forecasting*, *39*(3), 449–469. https://doi.org/10.1002/for.2632

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T. (2020). SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2177–2190. https://doi.org/10.18653/v1/2020.acl-main.197

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306–324. https://doi.org/10.1016/j.eswa.2014.08.004

Kneissl, K. (2021, Juni 28). *Die Medien und die Preisentwicklung von Rohöl* (H. Kaplan) [Zoom].

Lehmann, J., Mittelbach, M., & Schmeier, S. (2017). *Quantifizierung von Emotionswörtern in Texten.* GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2017-24.pdf

Li, X., Shang, W., & Wang, S. (2019). Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, *35*(4), 1548–1560. https://doi.org/10.1016/j.ijforecast.2018.07.006

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Liew, J. S. Y. (2016). *FINE-GRAINED EMOTION DETECTION IN MICROBLOG TEXT*. 440.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025 [cs]*. http://arxiv.org/abs/1508.04025

Malkiel, B. G. (1989). Efficient market hypothesis. In *Finance* (S. 127–134). Springer.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts: Good Debt or Bad Debt. *Journal of the Association for Information Science and Technology*, *65*(4), 782–796. https://doi.org/10.1002/asi.23062

Mascio, D. A., Fabozzi, F. J., & Zumwalt, J. K. (2021). Market timing using combined forecasts and machine learning. *Journal of Forecasting*, *40*(1), 1–16. https://doi.org/10.1002/for.2690

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2018). Learned in Translation: Contextualized Word Vectors. *arXiv:1708.00107 [cs]*. http://arxiv.org/abs/1708.00107

McCarthy, R. V., McCarthy, M. M., Ceccucci, W., Halawi, L., & SpringerLink (Online service). (2019). *Applying Predictive Analytics Finding Value in Data*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*. http://arxiv.org/abs/1310.4546

Mohammad, S. M. (2020). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *arXiv:2005.11882 [cs]*. http://arxiv.org/abs/2005.11882

Nexis Uni | Designed for Collaborative Research. (o. J.). Abgerufen 6. August 2021, von https://www.lexisnexis.com/en-us/professional/academic/nexis-uni/faculty.page

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv:1905.10437 [cs, stat]*. http://arxiv.org/abs/1905.10437

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of Emotion* (S. 3–33). Elsevier. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25–33.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rasa. (2020, April 20). *Transformers & Attention 1: Self Attention: Bd. Rasa Algorithm Whiteboard*. Rasa. https://www.youtube.com/watch?v=yGTUuEx3GkA

RavenPack. (2016, Juli 1). *Exploiting Themes Derived from Unstructured Content in Trading*. https://www.youtube.com/watch?v=7m6D2TNu7gQ

RavenPack. (2021, Mai 27). *Connect Corporate and Macro Risks*. RavenPack. https://www.ravenpack.com/page/connect-corporate-and-macro-risks/

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). *The Hourglass Model Revisited*.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, *2*(1), 325–347. https://doi.org/10.1146/annurev-linguistics-011415-040518

Tang, D., Qin, B., Feng, X., & Liu, T. (2016). Effective LSTMs for Target-Dependent Sentiment Classification. *Proceedings of COLING 2016, the 26th International*

Conference on Computational Linguistics: Technical Papers, 3298–3307. https://www.aclweb.org/anthology/C16-1311

Tschudy, M. (2020). *Themenerkennung in deutschen Texten* (S. 114) [Masterarbeit]. Fachhochschule Graubünden.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *arXiv:1706.03762 [cs]*. http://arxiv.org/abs/1706.03762

Wei, Y., Liu, J., Lai, X., & Hu, Y. (2017). Which determinant is the most informative in forecasting crude oil market volatility: Fundamental, speculation, or uncertainty? *Energy Economics*, *68*, 141–150. https://doi.org/10.1016/j.eneco.2017.09.016

Weng, L. (2018). Attention? Attention! *lilianweng.github.io/lil-log*. http://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Wex, F., Widder, N., Liebmann, M., & Neumann, D. (2013). Early Warning of Impending Oil Crises Using the Predictive Power of Online News Stories. *2013 46th Hawaii International Conference on System Sciences*, 1512–1521. https://doi.org/10.1109/HICSS.2013.186

Wharton Research Data Services. (o. J.). Abgerufen 6. August 2021, von https://wrds-www.wharton.upenn.edu/pages/grid-items/introduction-capital-iq/

Wikipedia contributors. (2021). *PyTorch—Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=PyTorch&oldid=1037666979

Yenicelik, K. D. (2020). *Understanding and Exploiting Subspace Organization in Contextual Word Embeddings* [Masterthese]. Eidgen¨ossische Technische Hochschule Zürich.

Yi Peng, N. (2020, September 20). *Tokenizers: NLP's Building Block*. https://towards-datascience.com/tokenizers-nlps-building-block-9ab17d3e6929

Zhang, Y., & Wallace, B. (2016). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv:1510.03820 [cs]*. http://arxiv.org/abs/1510.03820

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015, Dezember). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

9 Appendix

9.1 Appendix A Terms used for Query Search

Terms used in Query Search to Detect Polarities and Constraints					
INCREASE	**NO CHANGE**	**DECREASE**	**PIPELINE**	**ACCIDENT**	**SPILLS**
“ + ”	FLAT	“ - “	ACCIDENT	ACCIDENT	SPILL
BOOST	FREEZE	BEARISH	BLOW	CASUALTIES	
BULLISH	HOLD	BOTTOM	BOMB	CRASH	
HIGH	SAME	CRUSHING	EXPLODE	DEAD	
INCREASE	STEADY	CUTS	LEAKS	DEATH	
JUMP	UNAFFECT	DECLINE		EXPLODE	
REBOUND	UNCH	DECREASE		FATALITIES	
RISE	UNCHANGED	DN		FIRE	
SOAR		DOWN		INJURED	
UP		DROP		KILLS	
		DWN		LEAKS	
		FALL		OUTAGE	
		FELL			
		LOWER			
		SLIP			

Table 11: Terms used in Query Search to Detect Polarities and Constraints

9.2 Appendix B Experiments to evaluate Sensitivity towards Numerals

Sentiment Classifications with FinBERT to evaluate Sensitivity towards Numerals		
Sentence	**Classification**	**Sentiment Score**
oil exports of iraq +10%	neutral	0.057418
oil exports of iraq -10%	neutral	0.082199

Table 12: Sentiment Classifications with FinBERT to evaluate Sensitivity towards Numerals

Sentiment Classifications with CrudeBERT to evaluate Sensitivity towards Numerals		
Sentence	**Classification**	**Sentiment Score**
oil exports of iraq +10%	negative	-0.995112
oil exports of iraq -10%	positive	0.995444
oil exports of iraq down +10%	positive	0.992829
oil exports of iraq up +10%	negative	-0.995338
oil exports of iraq +0.000001%	negative	-0.994701
oil exports of iraq +100000%	negative	-0.994838
oil exports of iraq 10%	negative	-0.737652
oil exports of iraq 100%	negative	-0.979958

Table 13: Sentiment Classifications with CrudeBERT to evaluate Sensitivity towards Numerals

9.3 Appendix C Classification Report & Confusion Matrix CrudeBERTv2

```
Epoch: 100%|██████████| 4/4 [2:19:35<00:00, 2093.82s/it]


Loss:0.12
Accuracy:0.98

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      3505
           1       0.98      0.98      0.98      3192
           2       0.99      0.94      0.97       107

    accuracy                           0.98      6804
   macro avg       0.98      0.97      0.98      6804
weighted avg       0.98      0.98      0.98      6804
```

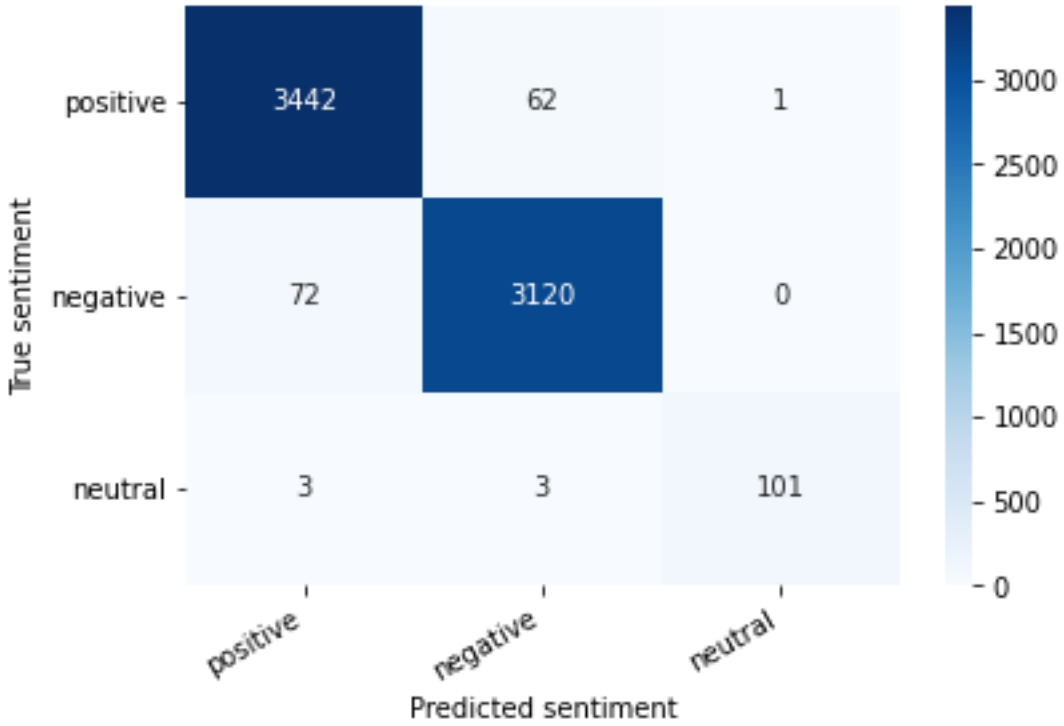

Figure 56: Confusion Matrix of CrudeBERTv2

9.4 Appendix D Classification Report & Confusion Matrix CrudeBERTv2_T4

```
Epoch: 100%|██████████| 4/4 [1:21:59<00:00, 1229.89s/it]


Loss:0.09
Accuracy:0.98

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      1965
           1       0.98      0.98      0.98      1924
           2       0.96      0.97      0.96        67

    accuracy                           0.98      3956
   macro avg       0.97      0.97      0.97      3956
weighted avg       0.98      0.98      0.98      3956
```

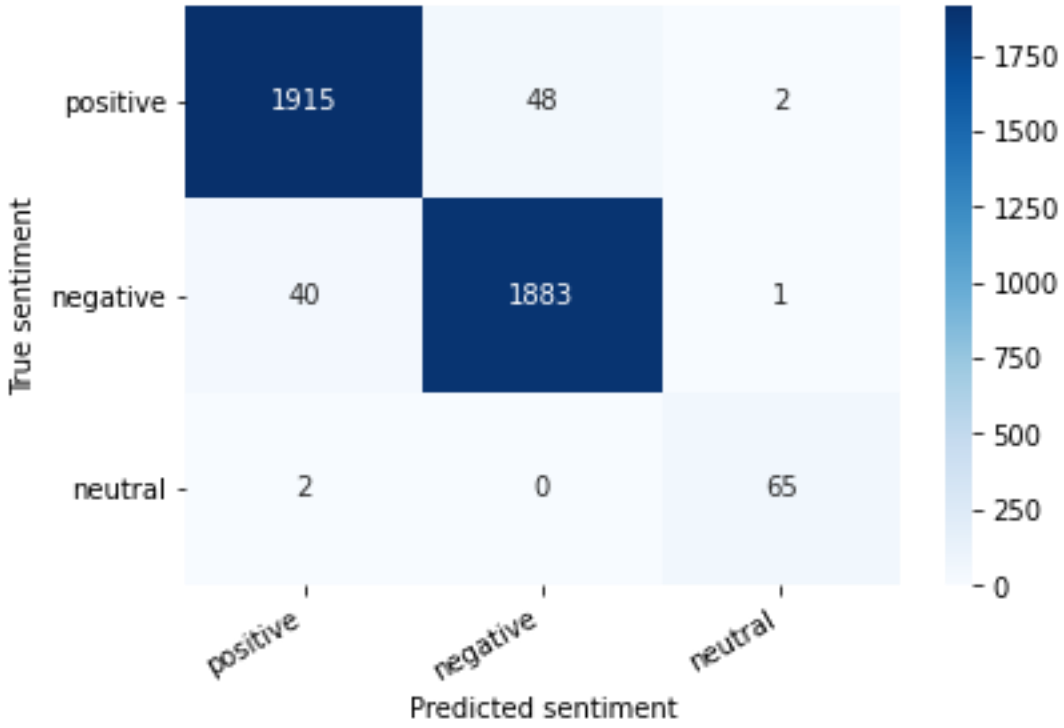

Figure 57: Confusion Matrix of CrudeBERTv2_T4

9.5 Appendix E Results of Prophet

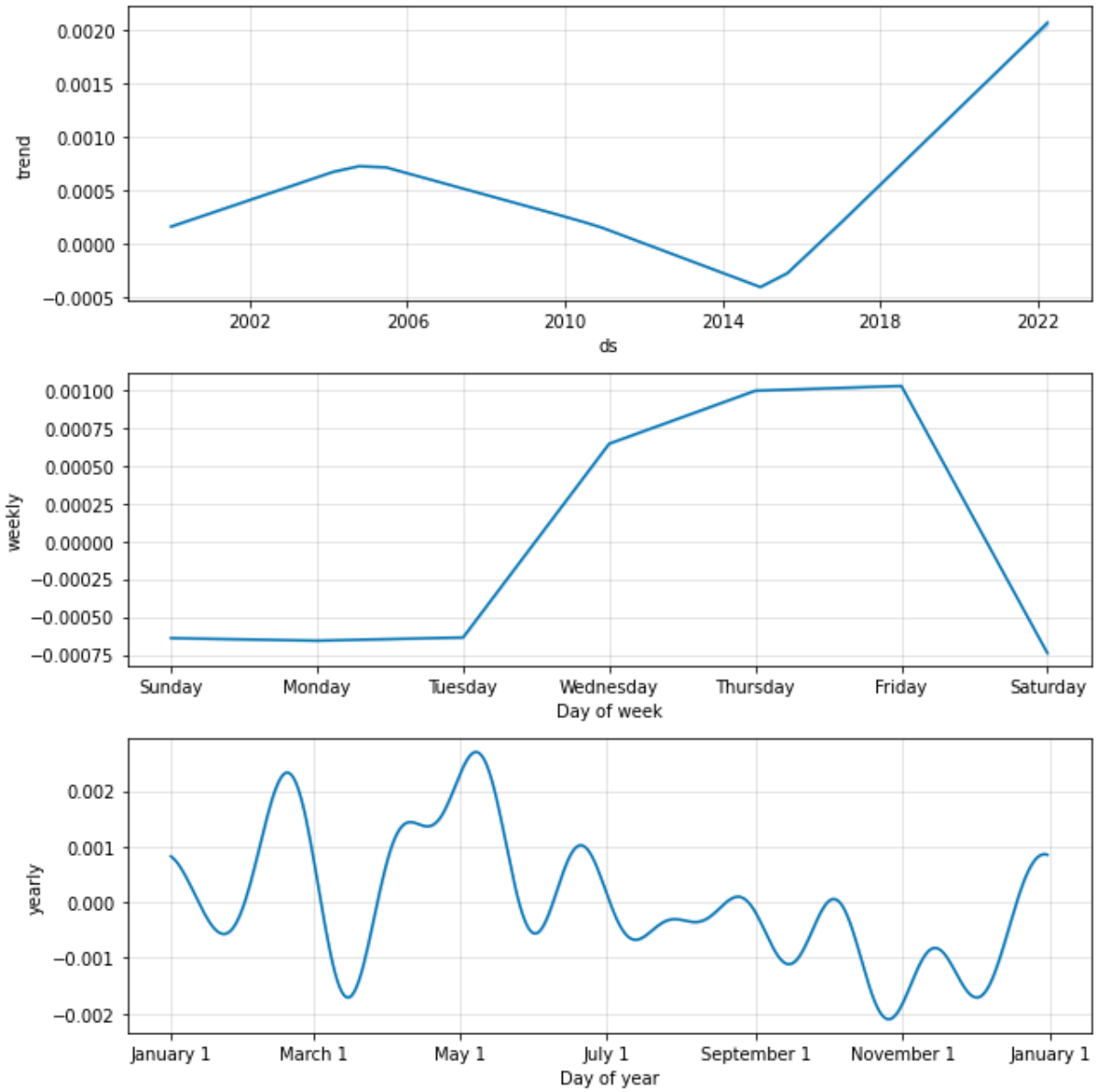

Figure 58: Analyzing Seasonality of WTI Crude Oil Returns with Prophet

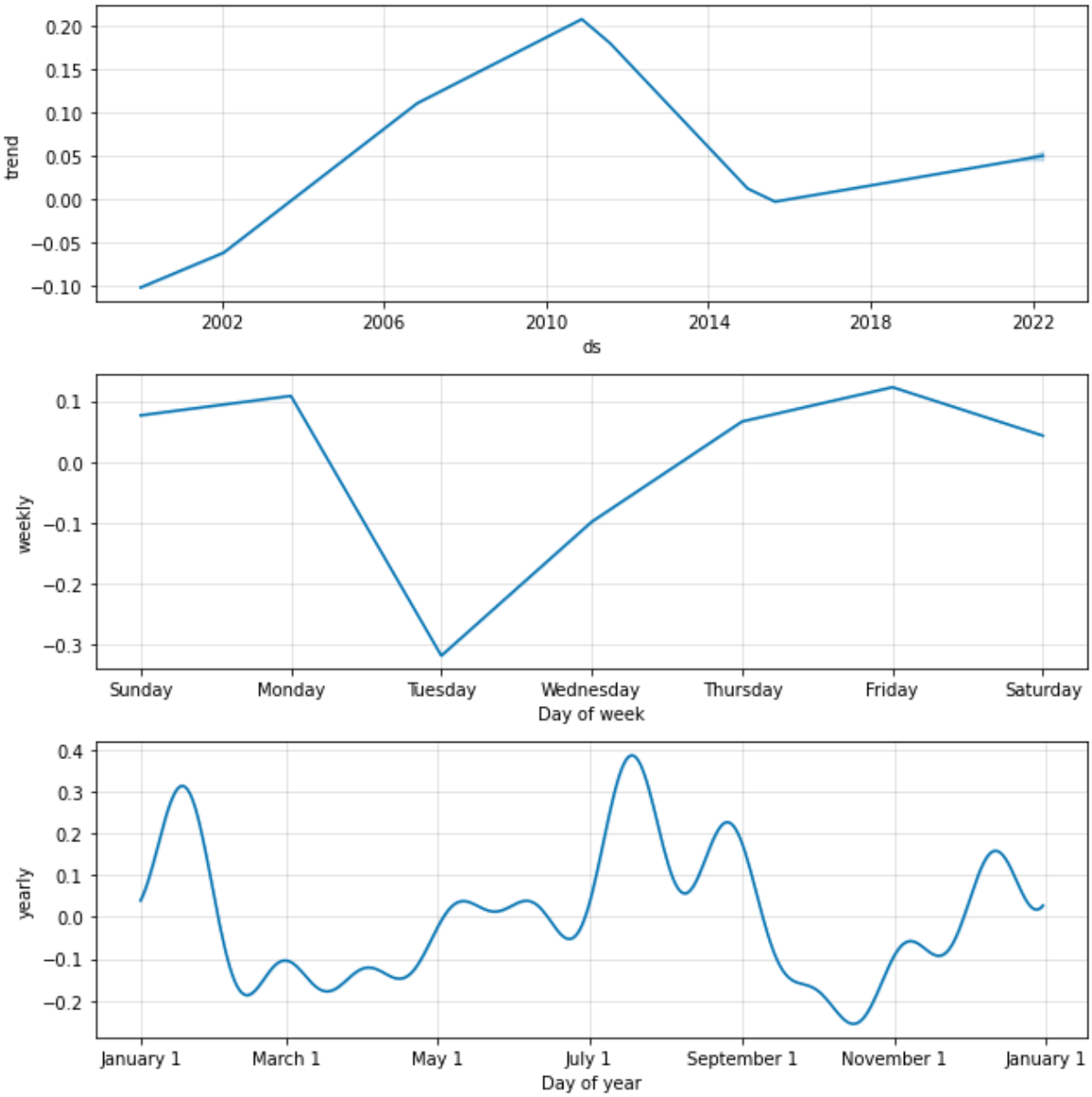

Figure 59: Analyzing Seasonality of CrudeBERT Sentiment Scores with Prophet

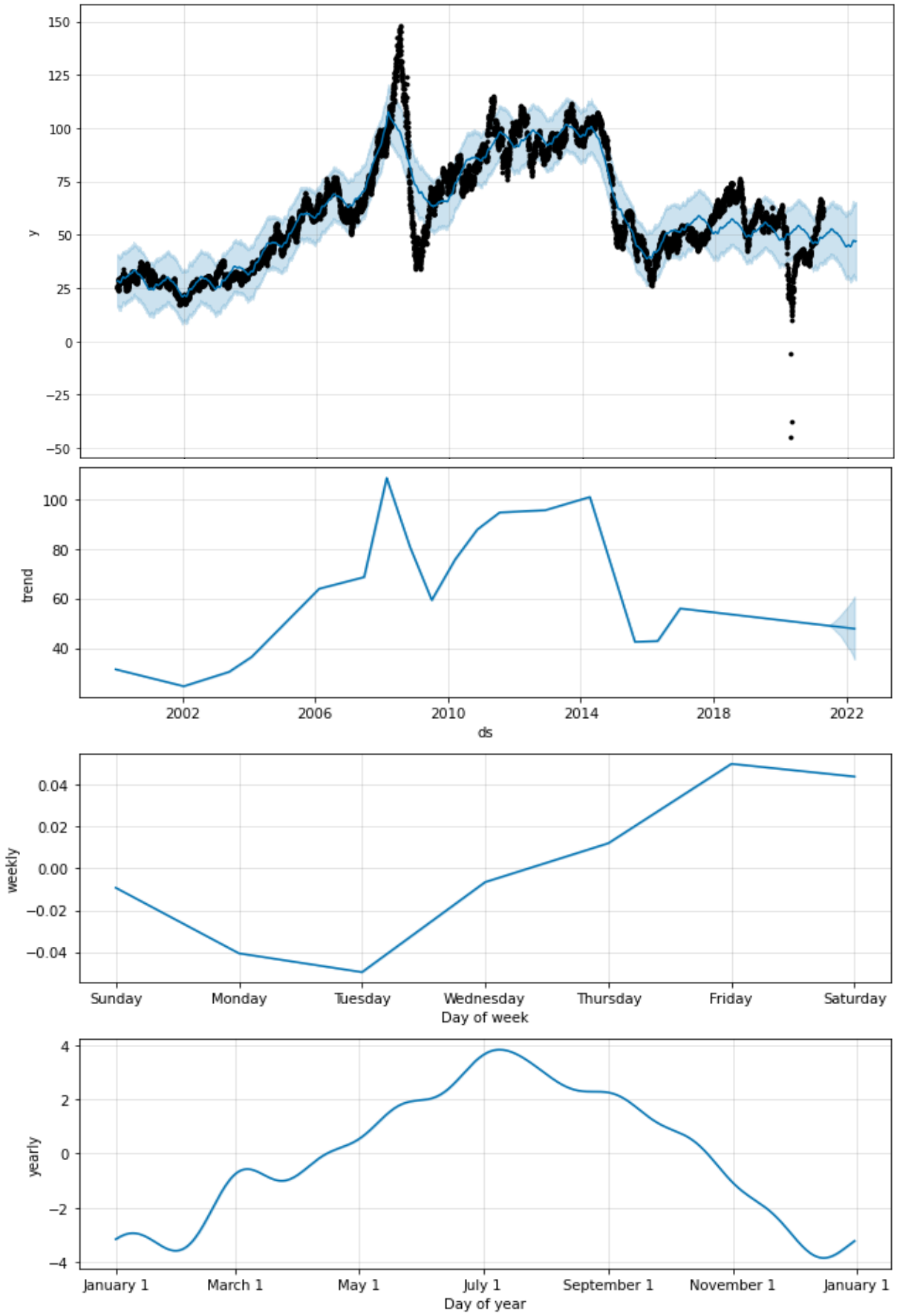

Figure 60: Univariate Prediction and Analyzing Seasonality of WTI Crude Oil Price with Prophet

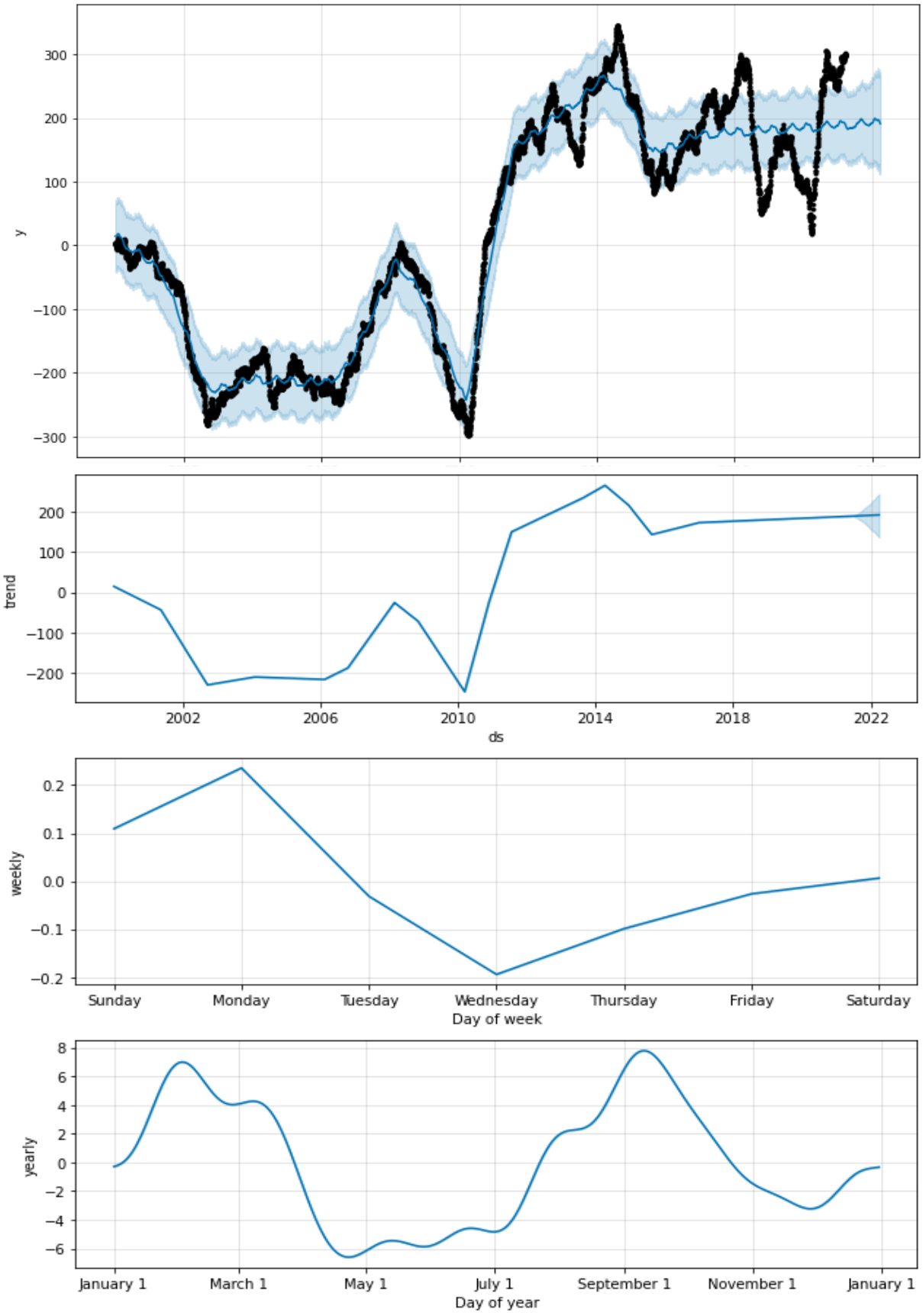

Figure 61: Univariate Prediction and Analyzing Seasonality of Cumulative Sentiment Scores of CrudeBERT with Prophet

9.6 Appendix F Results of Forecasting with NBEATS-Model (u8Darts)

```
input_chunk_length= 365
output_chunk_length= 90
n_epochs= 40
window = 7

MAE = 0.06%
```

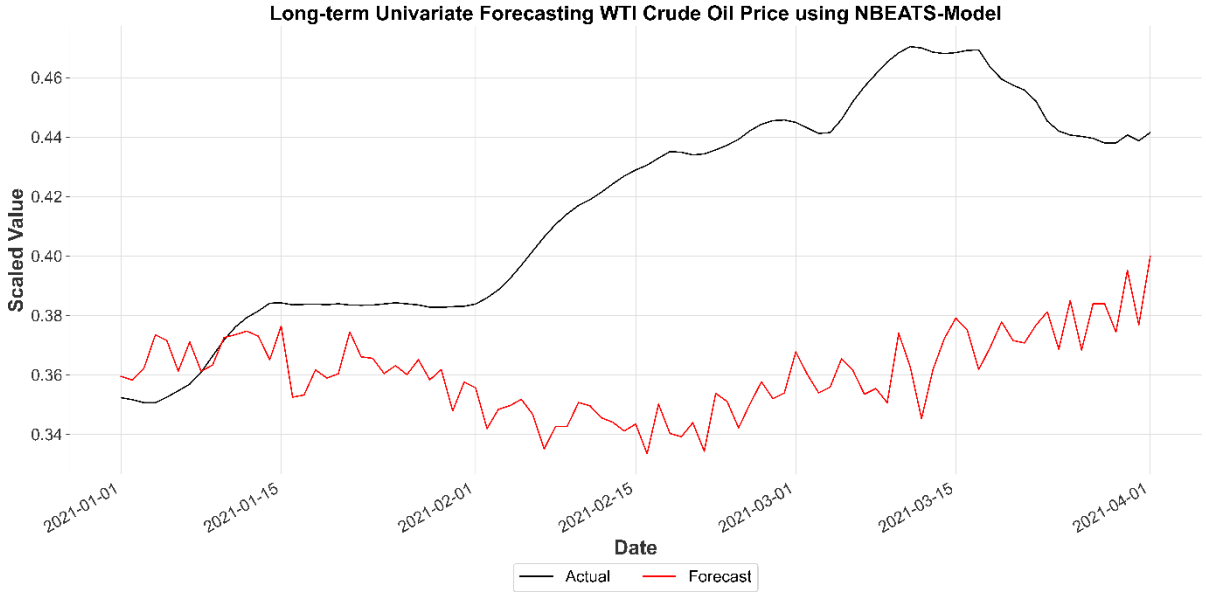

Figure 62: Long-Term Univariate Forecasting WTI Crude Oil Price using NBEATS-Model

```
input_chunk_length= 14
output_chunk_length= 7
n_epochs= 40
window = 7

MAE = 0.01%
```

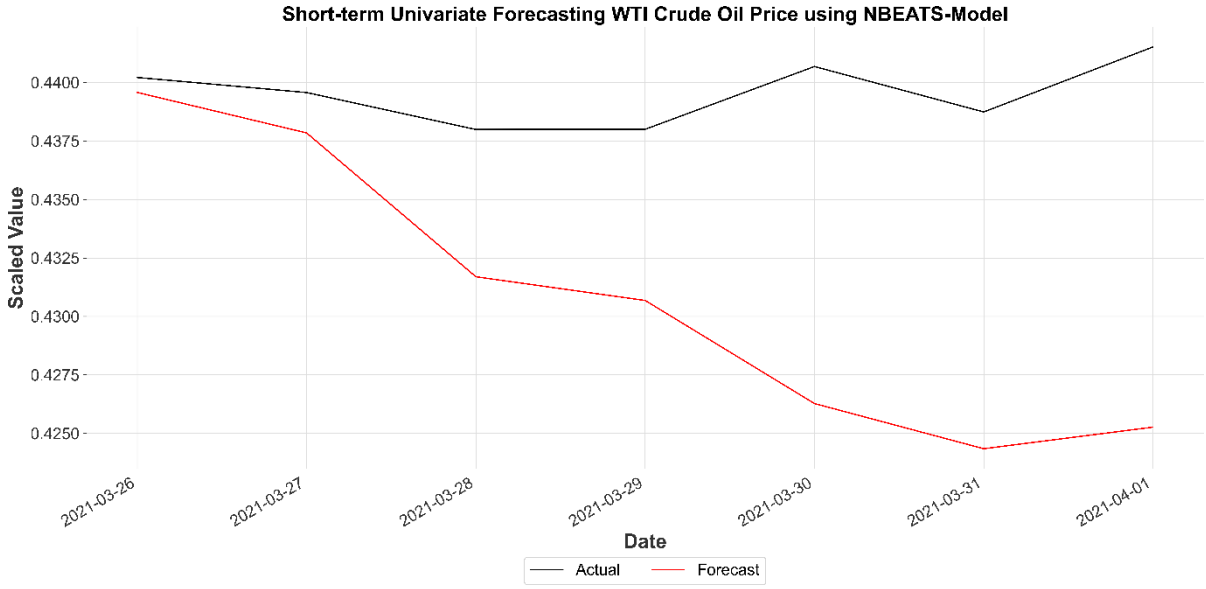

Figure 63: Short-Term Univariate Forecasting WTI Crude Oil Price using NBEATS-Model

```
input_chunk_length= 365
output_chunk_length= 90
n_epochs= 40
window = 7

MAE = 0.06%
```

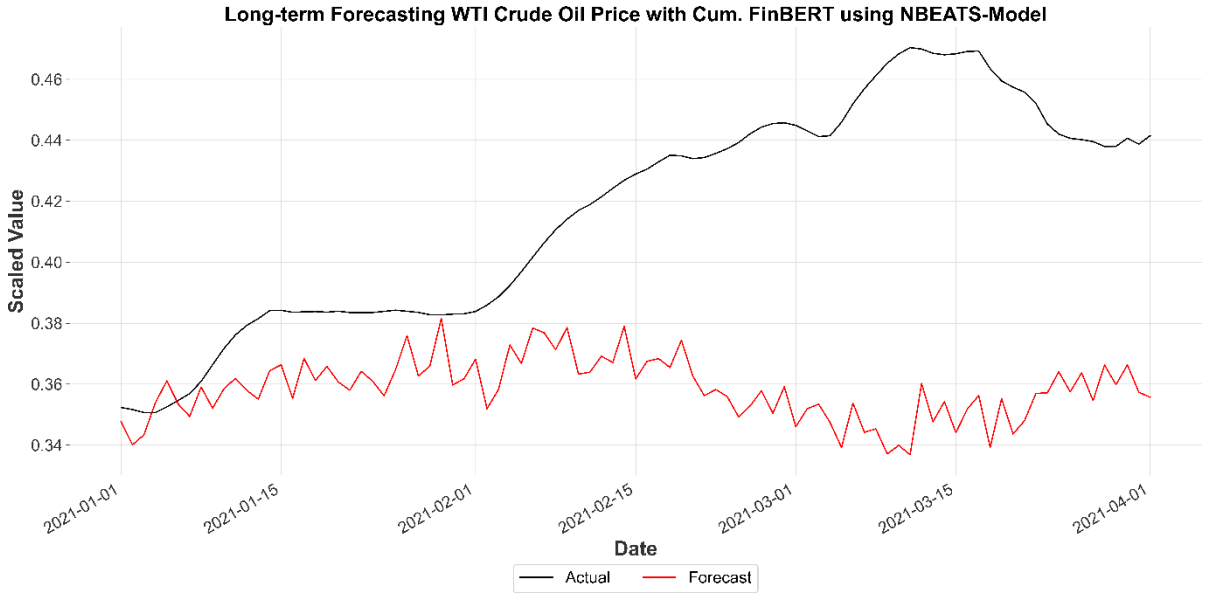

Figure 64: Long-Term Multivariate Forecasting WTI Crude Oil Price with Cum. FinBERT using NBEATS-Model

```
input_chunk_length= 14
output_chunk_length= 7
n_epochs= 40
window = 7

MAE = 0.00%
```

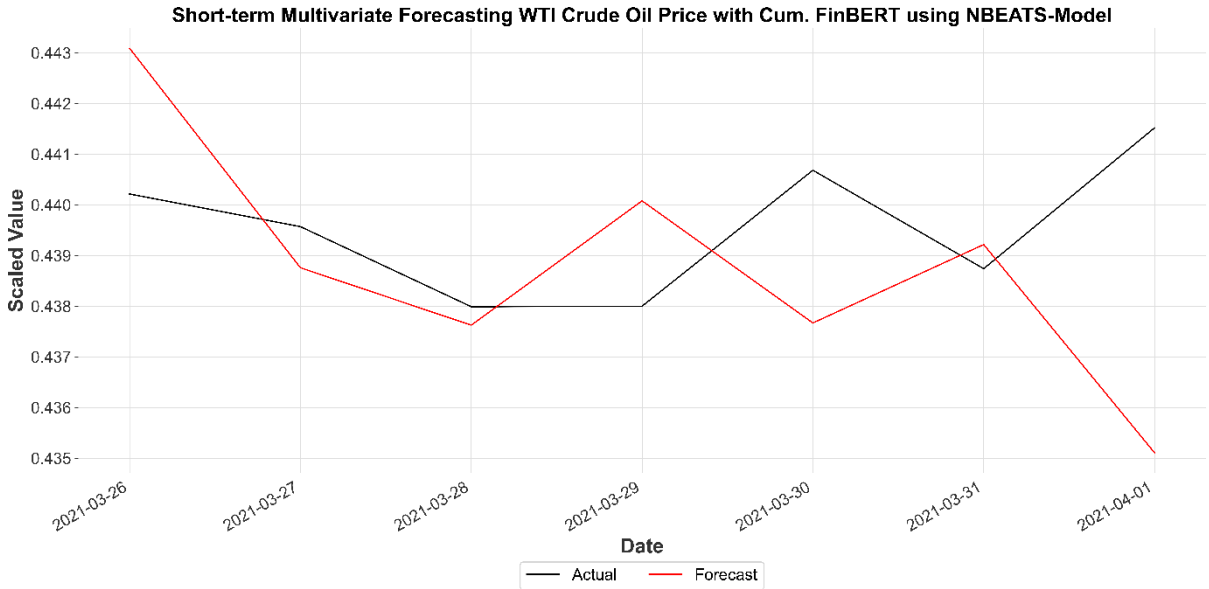

Figure 65: Short-Term Multivariate Forecasting WTI Crude Oil Price with Cum. FinBERT using NBEATS-Model

```
input_chunk_length= 365
output_chunk_length= 90
n_epochs= 40
window = 7

MAE = 0.03%
```

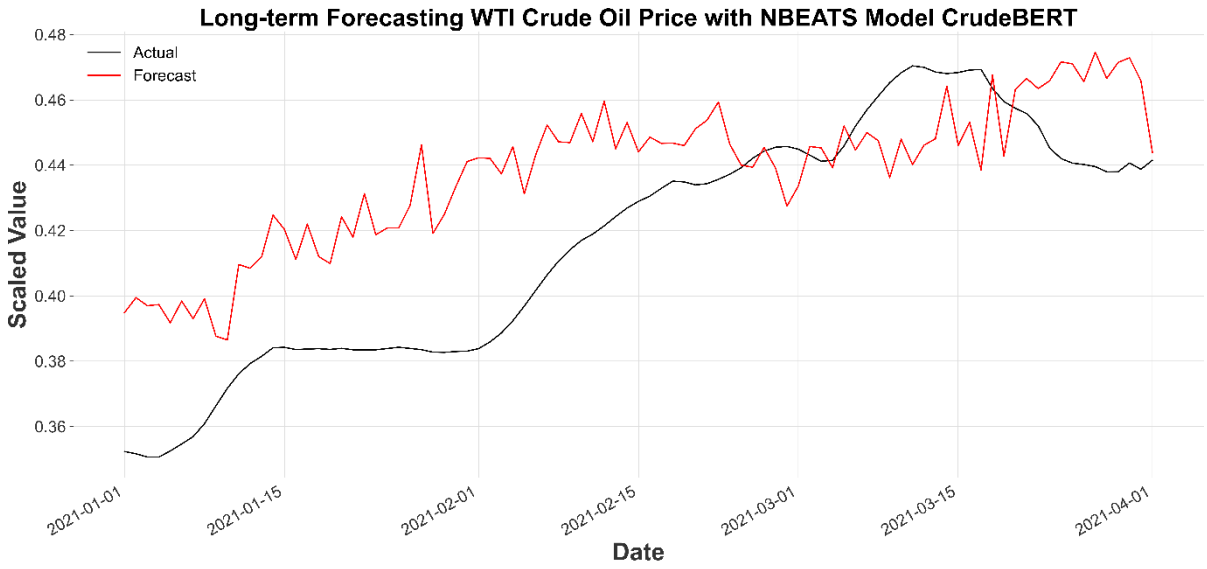

Figure 66: Long-Term Multivariate Forecasting WTI Crude Oil Price with Cum. CrudeBERT using NBEATS-Model

```
input_chunk_length= 14
output_chunk_length= 7
n_epochs= 40
window = 7

MAE = 0.00%
```

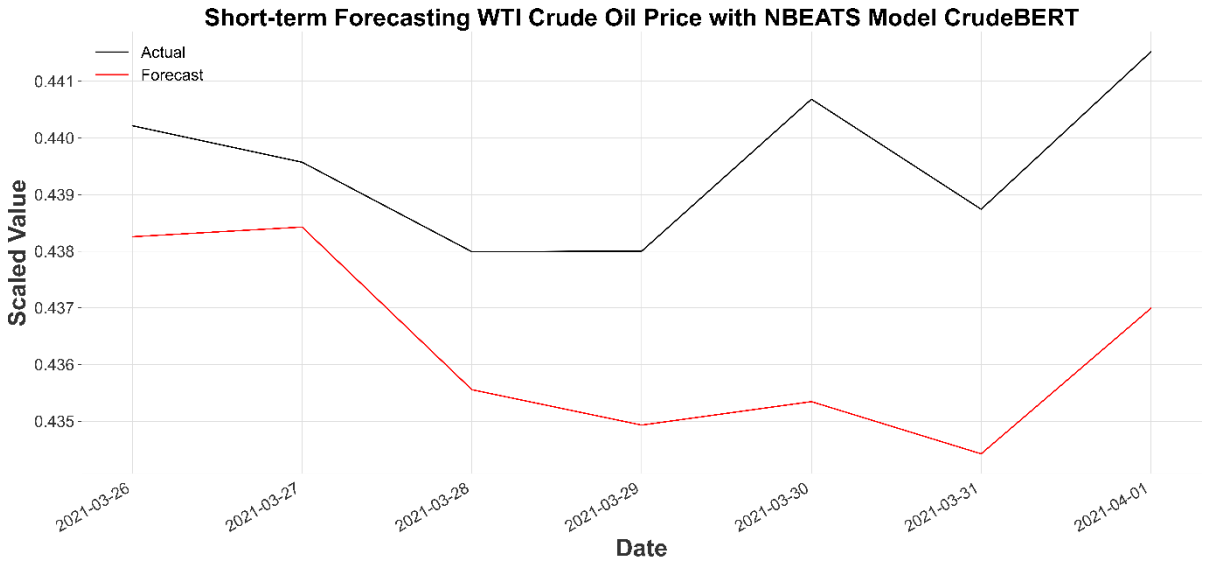

Figure 67: Short-Term Multivariate Forecasting WTI Crude Oil Price with Cum. CrudeBERT using NBEATS-Model

9.7 Appendix G Results of Logistic Regression Classification (PyCaret)

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)	
lr	Logistic Regression	0.5454	0.5449	0.5580	0.5611	0.5592	0.0898	0.0899	0.1890

Figure 68: Calculated Evaluation Metrics of Logistic Regression

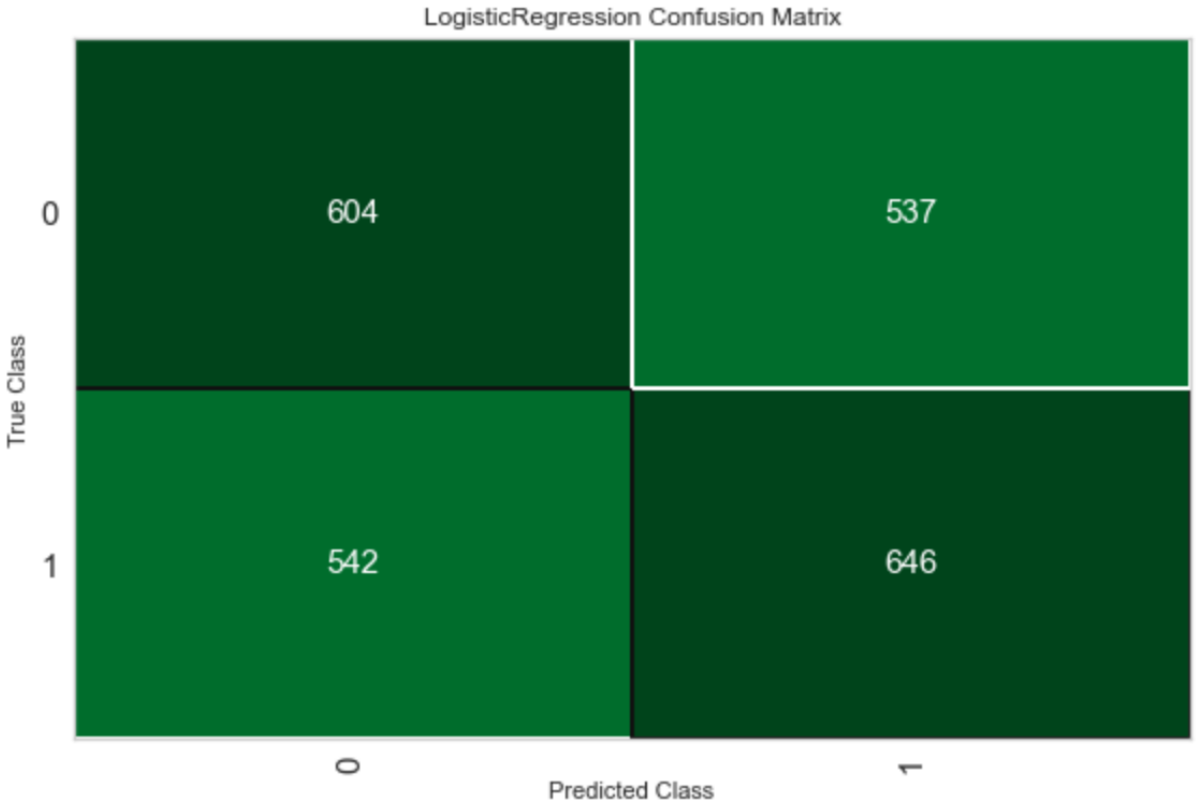

Figure 69: LogisticRegression Confusion Matrix between Sentiment Scores of CrudeBERT and WTI Crude Oil Price

9.8 Appendix H Top 40 Sources for publishing the most Headlines

Source (News Agency)	Number of Headlines
Dow Jones Newswires	21190
Reuters	3111
Bloomberg News	1133
Platts	868
Yahoo! News	853
Hellenic Shipping...	712
CNBC	641
Individual.com	573
MarketWatch	566
MT Newswires	454
Economic Times	452
Seeking Alpha	340
MENAFN	322
UPI	317
Penn Energy	300
FXStreet News	274
Namibia Press Age...	272
Trend News Agency	250
Benzinga	237
Forbes.com	234
Barrons	232
4-Traders	228
Business Recorder	218
Interfax	215
Xinhua News Agency	214
Wall Street Journal	188
UrduPoint	183
PR Newswire	181
Zawya.com	180
Sharenet	160
The Washington Post	157
Trade Arabia	155
Hindu Business Line	148
Business Standard	136
Steel Guru	129
Islamic Republic ...	122
MarketScreener	121
Business Wire	118
Houston Chronicle	112
offshore-technolo...	107

Table 14: Top 40 Sources for Publishing the most Headlines

10 Declaration

I herewith declare that this is my independent work written by me and using only admissible aides and no other sources than those given. I have marked as such, all passages which have been taken literally or analogously from another source. I am aware that if this is not the case, the executive board of the university of applied sciences is entitled to rescind any qualifications awarded or any title bestowed based on this work.

St. Gallen, 13.08.2021

Ort, Datum

Unterschrift