



MONASH University

Crude Oil Forecasting via Events and  
Outlook Extraction from Commodity  
News

Lee Mei Sin

Doctor of Philosophy

A Thesis Submitted for the Degree of Doctor of Philosophy at  
**Monash University** in 2022  
School of Information Technology

# Copyright notice

©[Lee Mei Sin](#) (2022).

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

# Abstract

Natural language-based financial forecasting is an active area of research, with the majority of publications centering around company stock prediction. There is, however, a huge potential in under-researched financial assets such as commodities. Not only that there is a lack of annotated dataset for commodities but also solutions developed for company stock predictions are not fit-for-purpose for commodities that are influenced by a totally different set of events. The main objective of this research is therefore to contribute towards natural language-based financial forecasting for crude oil, one of the major commodities traded in financial markets.

The first research contribution is the *CrudeOilNews* corpus, an annotated dataset of English Crude Oil news for event extraction. It is the first of its kind for commodity news and serves to contribute towards resource building for economic and financial text mining. First, the corpus was manually annotated and then was expanded through (1) data augmentation and (2) Human-in-the-loop Active Learning. The annotation closely follows the ACE/ERE standard. Apart from event extraction, equal emphasis was placed on event properties (Polarity, Modality, and Intensity) classification to determine the factual certainty of each event. The resulting corpus has 425 news articles with approximately 11k events annotated.

Secondly, this work presents a complete framework suitable for extracting and processing crude oil-related events found in *CrudeOilNews* corpus. These events are distinctly different from generic events and company-related events and as such new solution architecture and training approaches were introduced. In terms of solution architecture, Graph Convolutional Network (GCN) was used to effectively extract event arguments of homogeneous type. In terms of training, rather than training models from scratch, event extraction tasks were fine-tuned from ComBERT, a language model produced by Domain Adaptive Fine-Tuning from BERT on a collection of commodity news. An ensemble of Transfer Learning approaches (Domain Adaptive Pre-training, Multi-task Learning, and Sequential Transfer Learning) was also used to address the issue of class imbalance and to generate models with better performance than those trained via Supervised Learning alone.

Accurate and holistic event extraction from crude oil news is very useful for downstream tasks such as commodity price prediction to support a wide range of business decision-making. Unlike previous financial forecasting methods that used historical price data (time series data), this work introduced a new approach of using ‘market summaries’ as the only source to obtain both signals (semantic information extracted from text) and price information. ‘Market summaries’ is a genre of financial news that contains a good distillation and a retrospective view of global events and how the market reacted in the form of price change. These strong correlations are used to train machine learning models to forecast (1) crude oil directional movement, and (2) returns (percentage of price change) for WTI and Brent, two

of the global oil benchmarks. This solution aims to overcome the drawback of past works where spurious correlation between textual signals and historic price data might be included along with actual, non-spurious ones.

# Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature:

---

Print Name: Lee Mei Sin

---

Date: 29th September 2022

---

# Publications during enrolment

Conference articles:

1. Meisin Lee, Lay-Ki Soon, and Eu-Gene Siew. **Effective use of graph convolution network and contextual sub-tree for commodity news event extraction**. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 69–81, 2021 (Co-located with EMNLP 2021)
2. Lee, M., Soon, L.-K., Siew, E.-G., & Sugianto, L. F. (2022). **CrudeOilNews: An Annotated Crude Oil News Corpus for Event Extraction**. LREC 2022

Journal articles:

1. (Submitted to Expert Systems with Application Journal) Meisin Lee, Lay-Ki Soon, and Eu-Gene Siew. **Crude Oil-related Events Extraction and Processing: A Transfer Learning Approach**
2. (To be submitted to Energy Economics Journal) Meisin Lee, Eu-Gene Siew, Lay-Ki Soon, and Ly Fie, Sugianto. **Crude Oil Price Movement and Return Prediction via Text Analytics on Market Summaries**.

# Acknowledgements

This thesis and the whole PhD journey would not have been possible without the guidance and help of many.

**To my supervisors:** Thank you for your continuous guidance and advice, which has contributed tremendously to the completion of my work.

**To my husband:** Thank you for supporting me in my journey and sharing the load of household responsibilities. To borrow Queen Elizabeth's tribute to Prince Philip, you too are my strength and stay.

**To my children:** The journey would not have been so exciting and challenging at the same time without both of you.

Thank you all.

*To my husband John,  
who has been my strength and stay in this arduous journey.*



# Contents

Copyright notice	i
Abstract	ii
Declaration	iv
Publications during enrolment	v
Acknowledgements	vi
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	2
1.1.1 Impact of news on Crude Oil prices . . . . .	2
1.1.2 A Lack of Language Resource . . . . .	3
1.1.3 Market Summaries (Expert Analysis and Commentaries) . . . . .	4
1.1.4 Factual Event vs Outlook/Forecast . . . . .	4
1.2 Research Overview . . . . .	5
1.2.1 Problem Statement . . . . .	6
1.2.2 Research Questions . . . . .	6
1.2.3 Research Objectives . . . . .	6
1.3 Overview of Proposed Solution . . . . .	7
1.4 Contributions . . . . .	8
1.5 Thesis Organization . . . . .	9
<b>2 Literature Review</b>	<b>11</b>
2.1 Financial News Analytics . . . . .	11
2.1.1 Non-Event-Based . . . . .	12
2.1.1.1 Sentiment Analysis . . . . .	12
Domain-specific Sentiment Analysis . . . . .	13
Sentiment Analysis Libraries . . . . .	13
2.1.1.2 Topic Modeling + Sentiment Analysis . . . . .	13
2.1.1.3 Other Non-event based methods . . . . .	14
2.1.2 Event-Based . . . . .	14
2.1.2.1 Knowledge-Bases, Rules and Domain Ontology . . . . .	15
2.1.2.2 Fully Supervised Learning . . . . .	16
2.1.2.3 Semi - or Distantly Supervised Approach . . . . .	16
2.1.2.4 Other Methods . . . . .	17
2.1.3 News-oriented Crude Oil Forecasting . . . . .	17

2.1.3.1	Text Source . . . . .	17
2.1.3.2	Scope of Input . . . . .	18
2.1.3.3	Prediction Task . . . . .	19
2.2	Event ‘Factuality’ . . . . .	23
2.3	Language Resources for Event Extraction in the Finance and Economics domain	24
2.4	Summary and Discussion . . . . .	25
<b>3</b>	<b>Construction of <i>CrudeOilNews</i> Corpus</b>	<b>27</b>
3.1	Dataset Collection and Pre-processing . . . . .	29
	A Working Example . . . . .	29
3.2	Manual Annotation . . . . .	29
3.2.1	Annotation Guidelines . . . . .	31
3.2.1.1	Entity Mention . . . . .	31
3.2.1.2	Events . . . . .	32
	Event Triggers . . . . .	32
	Event Arguments . . . . .	33
	Event Typology . . . . .	33
3.2.1.3	Event Properties - Modality, Polarity, and Intensity . . . . .	37
	Polarity . . . . .	37
	Modality . . . . .	38
	Intensity . . . . .	38
3.2.2	Inter-Annotator Agreement . . . . .	39
3.2.2.1	Analysis . . . . .	40
3.3	Expanding the dataset . . . . .	41
3.3.1	Data Augmentation . . . . .	41
3.3.1.1	Trigger Word Replacement . . . . .	42
3.3.1.2	Argument Replacement . . . . .	43
3.3.2	Human-in-the-loop Active Learning . . . . .	45
	<i>Least Confidence</i> score . . . . .	45
3.3.2.1	Baseline models . . . . .	46
	Data Preprocessing . . . . .	46
	Entity Mention Detection Model . . . . .	46
	Event Extraction Model . . . . .	47
	Event Properties Classification . . . . .	47
3.3.2.2	Experiments and Analysis . . . . .	47
	<i>Least Confidence (LC)</i> threshold . . . . .	47
	Experiments . . . . .	48
	Analysis . . . . .	48
3.4	Corpus Evaluation and Analysis . . . . .	49
3.4.1	Corpus Statistics and Analysis . . . . .	49
3.4.2	Unique Characteristics . . . . .	49
3.5	Future Enhancements . . . . .	51
3.6	Comparison with other event extraction corpus . . . . .	51
3.7	Summary and Discussion . . . . .	52
<b>4</b>	<b>Event Extraction</b>	<b>53</b>
	Unique Characteristics of <i>CrudeOilNews</i> . . . . .	53

4.1	Definitions . . . . .	55
4.1.1	Terminologies: . . . . .	55
4.1.2	Tasks . . . . .	56
4.2	Related Work . . . . .	57
4.2.1	Event Extraction in Finance and Economics Domain . . . . .	57
4.2.2	Graph Convolutional Networks . . . . .	58
4.3	Preliminary: Domain Adaptive Pre-training - ComBERT . . . . .	58
4.4	Data Pre-processing . . . . .	60
4.5	Subtask 1: EMD and ED . . . . .	61
4.5.1	Experiments . . . . .	62
4.5.2	Results and Analysis of EMD /ED . . . . .	62
4.6	Subtask 2: ARP . . . . .	63
4.6.1	Contextual Sub-tree . . . . .	63
4.6.2	Graph Convolutional Networks over Contextual Sub-tree . . . . .	66
4.6.2.1	ARP with GCN . . . . .	68
	Encoding Trigger-Entity Pair . . . . .	68
4.6.3	Experiments . . . . .	69
	Parameter settings. . . . .	69
	Models Settings. . . . .	69
4.6.3.1	Results and Analysis . . . . .	70
4.6.3.2	Comparing Word Embedding and various Pre-trained Language Models . . . . .	70
4.7	Subtask 3: Event Properties Classification . . . . .	72
4.7.1	Model Architecture . . . . .	72
4.7.2	Measurement for dataset with class imbalance . . . . .	73
	<b>F1-Score</b> . . . . .	73
	<b>MCC</b> . . . . .	73
4.7.3	Experiments . . . . .	74
4.7.3.1	Train-Test Split . . . . .	74
4.7.3.2	Results and Analysis . . . . .	74
4.8	Summary and Discussion . . . . .	76
5	<b>Enhancing Event Extraction Performance with Transfer Learning</b> . . . . .	77
5.1	Transfer Learning . . . . .	79
5.1.1	Negative Transfer . . . . .	80
5.2	Related Work . . . . .	81
5.2.1	Usage of MTL . . . . .	81
5.2.2	Usage of STL . . . . .	82
5.2.3	Other forms of Transfer Learning . . . . .	83
5.3	Event Extraction . . . . .	83
5.3.1	Cross-domain Sequential Transfer Learning . . . . .	83
	<b>ACE2005</b> . . . . .	84
	<b>SENTiVENT</b> . . . . .	84
5.3.2	Inductive Transfer Learning . . . . .	85
5.3.3	Experiments . . . . .	85
	Results and Analysis . . . . .	86
5.4	Event Properties Classification . . . . .	88

5.4.1	Cross-Domain Sequential Transfer Learning . . . . .	89
5.4.1.1	Available Source Domain Corpora . . . . .	90
	Corpora for Negation Detection . . . . .	90
	Corpora for Uncertainty Detection . . . . .	90
	Domain Similarity . . . . .	91
	Task Modification . . . . .	91
5.4.1.2	Experiments . . . . .	92
5.4.1.3	Results and Analysis . . . . .	93
5.5	Final Model Performance . . . . .	94
5.6	Summary and Discussion . . . . .	94
<b>6</b>	<b>Event-based Crude Oil Futures Trend and Returns Prediction</b>	<b>96</b>
6.1	Domain Knowledge, Background Information and Terminologies . . . . .	97
6.1.1	Crude Oil Benchmark . . . . .	98
6.1.2	Terminologies . . . . .	98
6.1.3	What are Market Summaries? . . . . .	98
6.1.4	Market Summaries as single source input . . . . .	99
6.2	Related Work . . . . .	101
6.2.1	Coarse-grained Event . . . . .	101
6.2.2	Fine-grained Event . . . . .	102
6.2.3	Crude oil Price Forecasting with events . . . . .	103
6.3	The End-to-end Framework . . . . .	103
6.3.1	Event Extraction . . . . .	103
6.3.2	Document-Level Information Mining . . . . .	104
	6.3.2.1 Data Pre-processing: Sanity Checking . . . . .	105
6.3.3	Crude Oil Futures Trend and Returns Prediction . . . . .	106
	6.3.3.1 Crude Oil Trend Prediction . . . . .	106
	6.3.3.2 Crude Oil Return Prediction . . . . .	107
6.4	Experiments . . . . .	107
6.4.1	Dataset Construction . . . . .	107
6.4.2	Experimental Setup . . . . .	108
	6.4.2.1 Comparison of text processing techniques . . . . .	108
	‘Labels’ or dependent variable . . . . .	108
	Frequency . . . . .	108
	Results and Analysis . . . . .	109
	6.4.2.2 Contents: News headlines versus News body . . . . .	111
	Results and Analysis . . . . .	111
	6.4.2.3 Ablation Study . . . . .	111
	Results and Analysis . . . . .	112
	6.4.2.4 Solution Robustness . . . . .	112
6.4.3	Overall Analysis . . . . .	114
6.5	Summary and Discussion . . . . .	115
<b>7</b>	<b>Conclusions and Future Work</b>	<b>117</b>
7.1	Summary . . . . .	117
7.2	Limitations and Future Work . . . . .	118
7.2.1	Limitations . . . . .	118

7.2.1.1	Limited Event Coverage . . . . .	118
7.2.1.2	Wrong Correlations . . . . .	118
	‘Market Reactions’ may not be the outcome . . . . .	118
	Contradiction . . . . .	119
7.2.2	Future Work . . . . .	119
7.2.2.1	Other Financial Assets . . . . .	119
7.2.2.2	Other Uses of Events . . . . .	119
7.3	Conclusion . . . . .	120
<b>A</b>	<b>CrudeOilNews Corpus</b>	<b>121</b>
A.1	Event Schema . . . . .	121
A.1.1	Movement-down-loss, Movement-up-gain, Movement-flat . . . . .	121
A.1.2	Caused-movement-down-loss, Caused-movement-up-gain . . . . .	121
A.1.3	Position-high, Position-low . . . . .	122
A.1.4	Slow-weak, Grow-strong . . . . .	123
A.1.5	Prohibiting . . . . .	123
A.1.6	Oversupply . . . . .	123
A.1.7	Shortage . . . . .	125
A.1.8	Civil Unrest . . . . .	125
A.1.9	Embargo . . . . .	125
A.1.10	Geo-political Tension . . . . .	126
A.1.11	Crisis . . . . .	126
A.1.12	Negative Sentiment . . . . .	126
A.2	Event Types, Distribution and Examples . . . . .	126
A.3	RavenPack Event Taxonomy . . . . .	126
<b>B</b>	<b>Source Dataset</b>	<b>128</b>
B.1	ConanDoyle(neg) . . . . .	128
B.2	SOCC(neg) . . . . .	128
B.3	10kFinStatement(unc) . . . . .	129
B.4	Wikipedia-CoNLL2010(unc) . . . . .	129
B.5	Reviews(neg & unc) . . . . .	129
B.6	SENTiVENT . . . . .	129
<b>C</b>	<b>Background Information &amp; Domain Knowledge</b>	<b>130</b>
C.1	General Domain Event Extraction . . . . .	130
C.1.1	Canonical Event Extraction Program . . . . .	130
	ACE2005 Polarity, Tense, Genericity and Modality . . . . .	130
	TAC-KBP: Realis . . . . .	131
C.1.2	Event Extraction . . . . .	132
	Definitions . . . . .	132
	Event Extraction Subtasks . . . . .	132
	A Working Example . . . . .	133
C.2	Crude Oil-related Terminologies . . . . .	134
<b>D</b>	<b>Additional Results</b>	<b>135</b>

---

D.1	Chapter 3	135
D.1.1	Human-in-the-Loop Active Learning	135
D.1.1.1	Uncertainty Sampling	135
D.1.1.2	Model Performance - Active Learning	136
D.1.2	Event Type Distribution	136
D.2	Chapter 4: Baseline results	137
D.2.1	EMD Results (Baseline)	137
D.2.2	ED Results (Baseline)	138
D.3	Chapter 5: Results Post-Transfer Learning	139
D.3.1	EMD Results Post-Transfer Learning	139
D.3.2	ED Results Post-Transfer Learning	140
D.3.3	ARP Results Post-Transfer Learning	140
	References	141
	References	141

# Chapter 1

## Introduction

Information plays a crucial role in financial markets, and market efficiency relies upon the availability of information (Fama, 1965). Natural Language Processing (NLP) has been actively used in the Finance and Economics domain to extract information outside of market-historic-data from sources such as news, tweets, company reports, message boards, corporate disclosures, and financial periodicals. Information extracted from textual data is used to improve the modelling of financial markets for financial forecasting. Within this domain, applications of NLP cover a wide range of areas such as inflation rate prediction, credit scoring, social-economic indicators, market volatility prediction, stock market, foreign exchange rate (FOREX) prediction, and etc. Among these, the majority of the research focus and publication centres around stock market and foreign exchange rate(FOREX) prediction (Xing et al., 2018). There is a huge potential in under-researched financial assets such as commodities. Among them, crude oil is a crucial component of the global economy, where nearly one-third of the world's energy consumption comes from it. Accurate crude oil price forecasting facilitates governmental policy making and decision-making regarding energy resources.

There has been a variety of NLP techniques being used in financial forecasting, at the summary level these includes (1) Semantic Modeling using bag-of-words at the early stages and later using improved vector representation or word embeddings such as GloVe(Pennington et al., 2014) and Word2Vec(Mikolov et al., 2013); (2) Topic Modeling (Blei, 2012) to capture the semantics (composition of multiple topics and corresponding relevance coefficients) of a collection of financial articles; (3) Sentiment Analysis with hand-crafted word lists / dictionary such as Henry Dictionary (Henry, 2008), Loughran & McDonald Dictionary(Loughran

& McDonald, 2011), Harvard-IV dictionary (Stone, 2002); (4) Event Extraction from news (Ding et al., 2014, 2015; D. Chen, Zou, et al., 2019).

## 1.1 Background and Motivation

This section focuses on crude oil in detail and provides the necessary background information to explain the formulation of Research Questions and Research Objectives in this work.

### 1.1.1 Impact of news on Crude Oil prices

The usage of news has drawn much attention in recent years, and there have been many proposed solutions to mine news data for better financial market trend predictions. Structured events from news have proven to produce superior stock prediction results compared to other text-based methods (Ding et al., 2014; D. Chen, Zou, et al., 2019). It is highlighted in (Wu, Wang, Lv, & Zeng, 2021; J. Li et al., 2017) that news is an important source of information for crude oil trading since oil market movements are driven by events such as geopolitics (e.g. war, civil unrest, political instabilities), macro-economic events (e.g. economic development), financial environment, as well as oil market factors (e.g., consumption, inventory, and supply of oil). For example, the Iran revolution in 1979 and the Iran-Iraq war during 1980-1988 both resulted in sharply upward trends in oil prices. As discussed in (Brandt & Gao, 2019), these events are generally at the macro-economic level and are geo-political in nature, and they are found to cause crude oil price fluctuations both in the short-term and long-term. The authors from (Brandt & Gao, 2019) presented a comprehensive list of events known to impact crude oil prices. Here are some main examples of geo-political, macro-economic, and supply-demand events while the complete list is shown in Appendix A:

- **geo-political:** Terrorism, war and conflict, civil unrest and natural disasters such as typhoons or hurricanes.
- **macro-economic:** consumer spending, durable goods orders, housing, economic growth, CPI and exports.
- **supply-demand:** cut production, increase production, economic growth cause boost in oil demand, etc.



While there are past works that used news articles as signals (semantic information extracted from text) via Sentiment Analysis, Topic Modeling, and etc. for crude oil forecasting, none uses events extracted from news articles. Hence there is a huge potential for event-driven crude oil forecasting.

### 1.1.2 A Lack of Language Resource

Figure 1.1 shows an example of a piece of news article on Crude Oil.

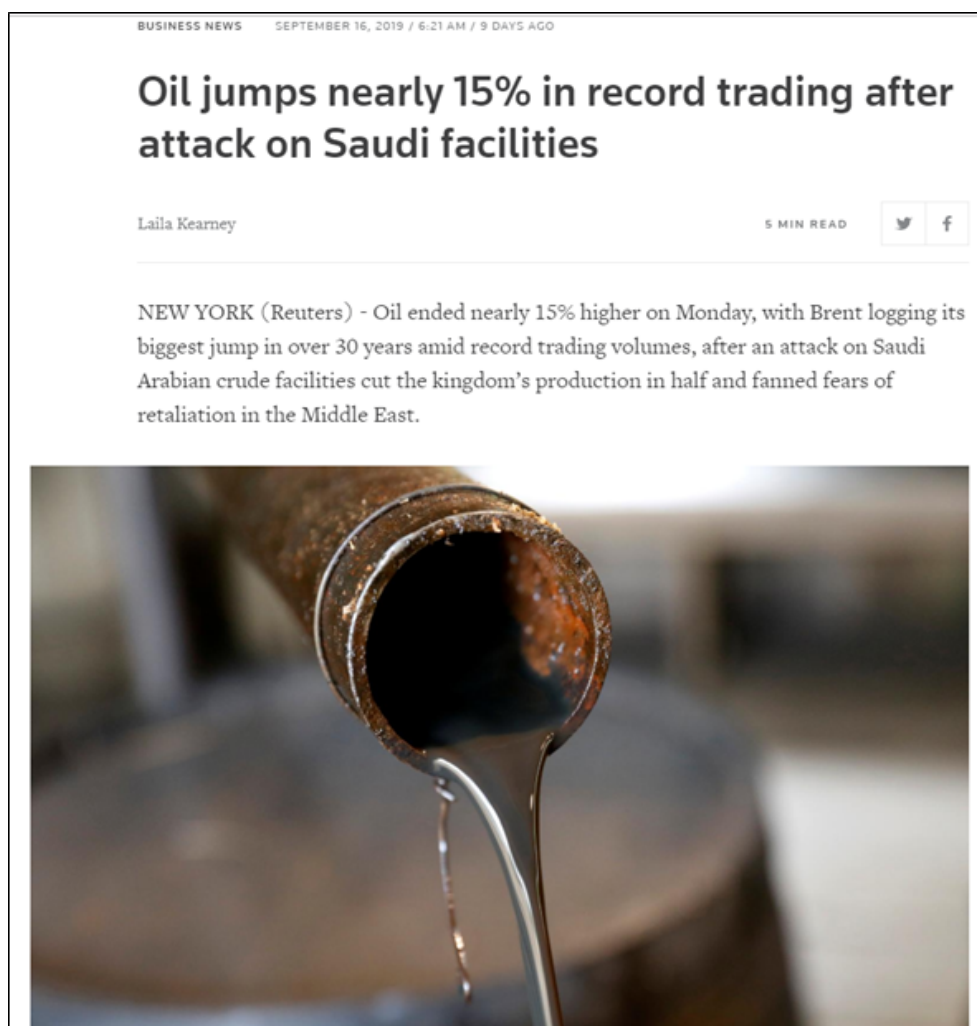


Figure 1.1: An example of news on Crude oil.

There is a small number of corpora in the Finance and Economics domain, such as SENTiVENT in (Jacobs & Hoste, 2021), but all are focused on company-specific events and are used mainly for stock price prediction. As acknowledged by (Jacobs & Hoste, 2021), due to the lack of annotated datasets in the Finance and Economics domain, only a handful of

supervised approaches exist for Financial Information Extraction. To the best of my knowledge, there is no available annotated corpus for crude oil or for any other commodities. To explore the solution of event-driven crude oil forecasting, an annotated corpus for the event extraction task is needed.

### 1.1.3 Market Summaries (Expert Analysis and Commentaries)

Among financial news, there is a special genre of news, termed here as *Commodity market summaries* that contain expert analysis and commentaries analysing the financial market from a retrospective view of what took place and how the commodity market reacted to it. These analyses that are written by financial analysts or journalists contain a good distillation of world events that are truly causal to the movement of crude oil prices. For example, in Figure 1.2, the event of oil price moving higher (highlighted in blue) was led by the event of the US sanctioning Iran (highlighted in red). Commentary-like news is rich in these analyses, which can be fully exploited for predicting oil price movement when another similar event occurs.

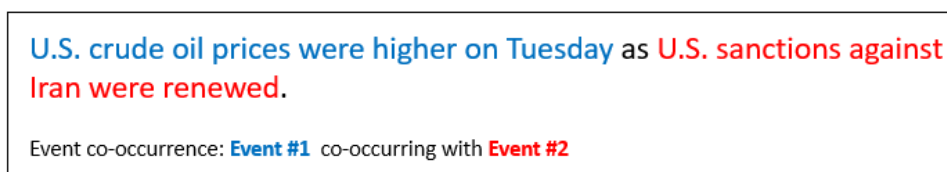


Figure 1.2: A snippet of a news article showing event co-occurrence.

These market summaries are an excellent source to learn the strong correlation between global events and crude oil market reactions.

### 1.1.4 Factual Event vs Outlook/Forecast

Apart from actual events, commodity news also contains anticipated events, expert opinions, analysis, and even financial outlook forecasts. Financial outlook means forward-looking information about prospective financial performance, a financial position that is based on assumption about future economic conditions<sup>1</sup>. In (Brandt & Gao, 2019), the authors made the same observation where apart from event-driven reporting, commodity news also contains potential impact analysis and anticipated event evolution. The forecast of events often exerts

<sup>1</sup>definition according to <https://www.lawinsider.com/dictionary/financial-outlook>

some form of impact on financial markets the same way as actual events do, albeit to a different degree. Take for example, rumours (forecast) of war may impact financial markets the same way as an actual war, but the impact could be of a different magnitude.

To have an accurate interpretation of events, it is important to take event ‘factuality’ into account by distinguishing actual events from forecasted/anticipated events and negated events. In financial news and reports, there are expert opinions, market analyses, and outlook forecasts apart from actual events. This is illustrated in the sentences below:

- (1) Market analysts have forecasted a slow economic recovery in the US.
- (2) China’s annual GDP is expected to grow in-line with the projection of 5.5 point.
- (3) Some market analysts are optimistic about the Syria civil unrest and raised crude oil price forecast to USD53 per barrel.
- (4) Some analysts say it is too early to tell if the latest fall in prices is any different from previous declines , such as in 2012 and 2013.
- (5) EA cautioned that the sanctions are expected to trim Russian demand.

Although Stocks and Crude Oil both fall under the umbrella of Financial Assets, there are some fundamental differences between these two asset classes. One of the major differences is that the set of events impacting company stocks are different from events affecting crude oil prices, which are generally at the macro-economic level and are geo-political in nature. Hence solutions developed for company stock predictions are not fit-for-purpose for crude oil. It is with these differences in mind that this research was formulated.

## 1.2 Research Overview

Given that events that affect crude oil market are distinctly different from generic events and company-related events, existing event-based solutions are not fit-for-purpose for crude oil. Hence a new solution architecture and training approaches are introduced in this research. In a nutshell, the proposed solution is a pipeline approach that starts with resource building, and a machine learning model is then trained to extract fined-grained event details. With this capability, events in crude oil market summaries is then mined for the co-occurrence of

events and market reaction (in the form of price change) to be used in crude oil prediction. The details of the proposed solution are further elaborated in Sections 3, 4 and 6 respectively.

This section gives an overview of the research by presenting the problem statement, research questions, research objectives, and research contribution of the proposed solution.

### 1.2.1 Problem Statement

The following problem statement summarises the problems to be addressed:

*How can events from commodity news articles be extracted and used for crude oil price forecasting?*

### 1.2.2 Research Questions

This research tries to answer the questions listed below:

- **RQ1:** How to define event schemas using available resources, which could be used for labeling corpus with geo-political and macro-economic events?
- **RQ2:** How can extracted events be represented accurately given that there are actual events along with expert opinions, market analyses and outlook forecast?
- **RQ3:** How to improve the accuracy of event extraction models beyond traditional supervised learning approach?
- **RQ4:** How can co-occurring events be mined and used for commodity price movement (a type of event itself) prediction?

### 1.2.3 Research Objectives

With the above research questions in mind, this research aims at the following objectives:

- **RO1:** To build a labeled dataset for crude oil news with the following types of events annotated: (i) macro-economic, (ii) geo-political, and (iii) supply-demand.
- **RO2:** To propose an event extraction model to extract crude oil-related events and market outlook from news corpus.

- **RO3:** To design a classifier for crude oil forecasting using co-occurrence of events.

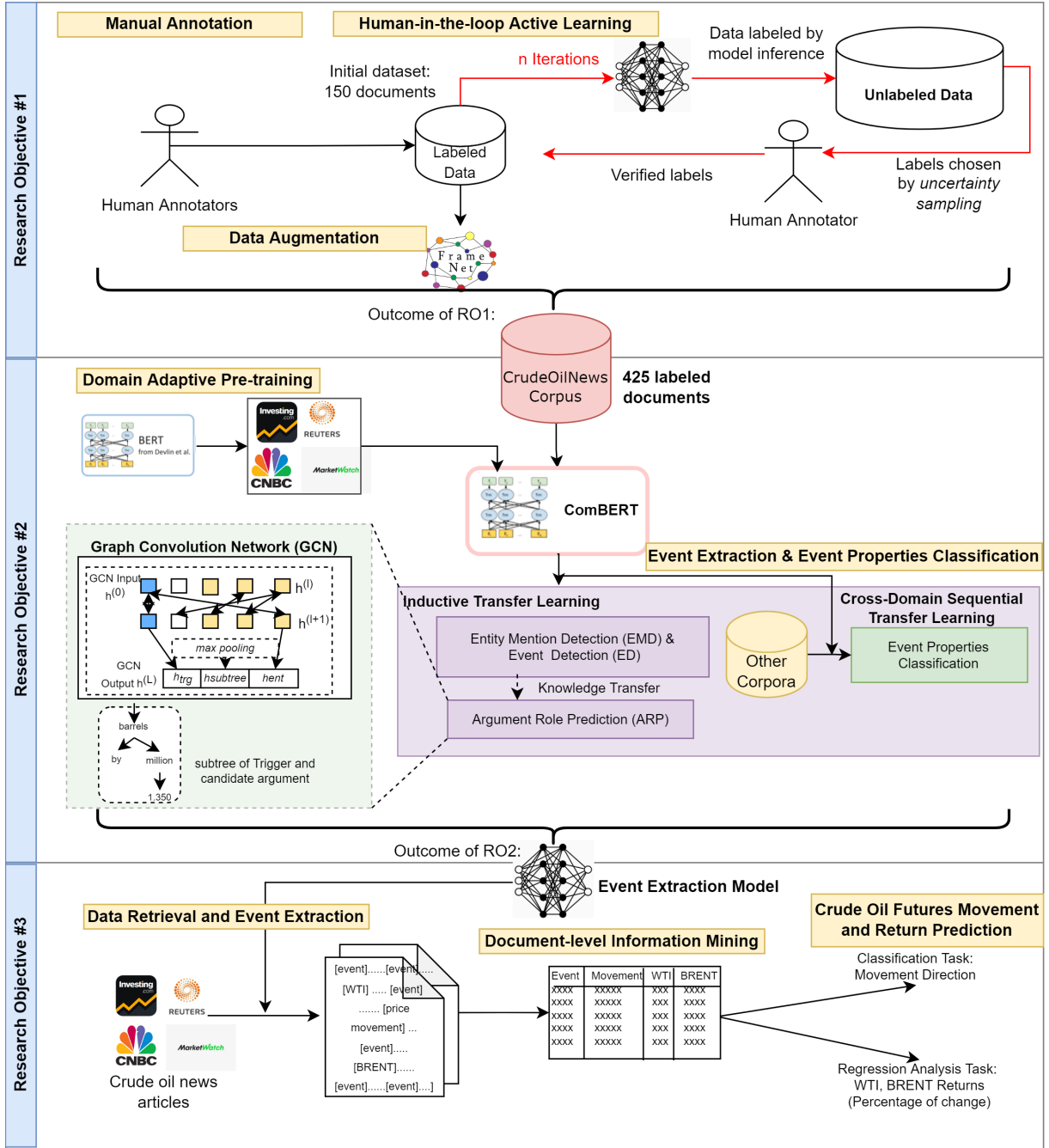


Figure 1.3: Research Objectives in diagrammatic form.

### 1.3 Overview of Proposed Solution

A summarized overview of the proposed solution is shown in Figure 1.3. There are three main components to the proposed solution; each component corresponds to one research objective respectively.

1. The first component is to build an annotated crude oil news corpus for the task of event extraction. First, the corpus was manually annotated and then expanded through (1) data augmentation and (2) Human-in-the-loop Active Learning. Apart from event details, equal emphasis is placed on labeling each event’s properties (Modality, Polarity, and Intensity) to differentiate actual events from financial outlook forecasts or from negated events.
2. With a labeled dataset, the second component is to train an event extraction model for extracting and processing crude oil-related events. A new solution architecture using Graph Convolutional Network (GCN) was introduced to effectively extract event arguments of homogeneous type. Rather than training models from scratch, event extraction tasks were fine-tuned from ComBERT, a language model produced by Domain Adaptive Fine-Tuning from BERT on a collection of commodity news. An ensemble of Transfer Learning approaches (Domain Adaptive Pre-training, Multi-task Learning, and Sequential Transfer Learning) was also used to address the issue of class imbalance and to generate models with better performance than those trained via Supervised Training alone.
3. The third component involves mining events extracted from market summaries’ to obtain both signals (semantic information extracted from text) and price information. ‘Market summaries’ is a genre of financial news that contains a good distillation and a retrospective view of global events and how the market reacted in the form of price change. These strong correlations are used to train machine learning models to forecast (1) crude oil trend, and (2) returns (percentage of price change) for WTI and Brent, two of the global oil benchmarks.

## 1.4 Contributions

The central goal arising from this research is a solution that is able to extract events from crude oil market summaries and utilize the events for crude oil forecasting. This solution is made up of three main components as listed in Section 1.3. Each component contributes to an outcome as listed down below in more detail, with the research objective addressed by each contribution highlighted in parentheses:

1. A crude oil news corpus annotated for the task of event extraction (RO1; Chapter 3).
2. An event extraction model that is able to identify and extract crude oil-related events and as well as classifying each event's event properties (Modality, Polarity, and Intensity) (RO2; Chapters 4, 5).
3. Two crude oil forecasting models: (1) crude oil price trend (UP, DOWN, and STABLE); and (2) returns (percentage of price change) for WTI and Brent, two of the global oil benchmarks (RO3, Chapter 6).

## 1.5 Thesis Organization

The rest of this report are organized as follows:

1. **Chapter 2: Literature Review** - This chapter investigates (1) past works of using text as an input for Financial Asset forecasting with a special focus on the types of text processing techniques involved; (2) Event extraction in general as well as event factuality (factual certainty of an event) and (3) the availability geo-political and macro-economic or similar dataset in the Finance and Economics domain;
2. **Chapter 3: Contruction of *CrudeOilNews* corpus** - This chapter details the outcome of Research Objective #1 - constructing an annotated dataset for event extraction. Apart from manual annotation, the annotated dataset was expanded using data augmentation and Human-in-the-loop active learning, fully optimizing the involvement of human annotators.
3. **Chapter 4: Event Extraction** - This chapter discusses the proposed solution for event extraction, focusing on the architecture, implementation, and experimental results.
4. **Chapter 5: Enhancing Event Extraction Performance with Transfer Learning** - This chapter discusses the usage of Transfer Learning to boost further the performance of models produced via supervised learning as laid out in Chapter 4. Chapters 4 and 5 describe the outcome for Research Objective #2.
5. **Chapter 6: Event-based Crude Oil Futures Movement and Return Prediction** - This chapter lays out the application of events extracted from crude oil market

summaries to build a labeled dataset for the training of models for crude oil futures movement and returns prediction.

6. **Chapter 7: Conclusions and Future Work** - This chapter concludes the thesis by (1) reiterating the contributions of this work, (2) highlighting the limitations of the proposed solution, and (3) presenting a list of potential future work.



## Chapter 2

# Literature Review

The literature presented in this section serves as a review of state-of-the-art work relating to this research. This literature review is guided by the list of research questions listed in Chapter 1 with the intention of evaluating past works relating to each individual research objective. However, the sections within this chapter are not arranged in the same sequence as RO1-RO3, instead, it is organized as follows:

- Section 2.1 dives into the latest advancement and state-of-the-art methods in financial news analytics; this is then followed by a more specific and focused review of existing news-oriented crude oil forecasting approaches (**RO3**);
- Section 2.2 looks into the subtask of event ‘factuality’ classification;
- Section 2.3 discusses available datasets in the Finance / Economics domain (**RO1**).
- Lastly, the chapter closes with a summary of the analysis of related work, which contain gaps and opportunities in which this research can contribute towards.

Crude oil-related terminologies and their definitions are provided as supplementary information in Appendix C.2 to assist in providing better readers’ understanding.

### 2.1 Financial News Analytics

In recent years, more and more market participants have considered the addition of financial news analytics into their algorithmic trading engine to better predict the direction or

volatility of market movements before making an investment decision. Financial news analytics combines methods from information retrieval, statistical learning, natural language processing, and financial econometrics to collect, categorise, interpret unstructured textual input data and convert this into metric output data, such as a financial sentiment score or into distilled / less-noisy information like event details. The importance of financial news analytics is acknowledged in (Upreti et al., 2019), and the authors provided an overview of existing linguistic resources and methodological approaches to develop knowledge-driven solutions for financial news analytics. For a more targeted discussion, only natural language processing techniques are discussed here, while financial econometrics or time series-related methods are excluded.

Among the financial assets, stocks are the most worked on, while other assets such as commodities (crude oil, natural gas, etc.) are less common. Therefore, apart from discussing crude oil-related work, the scope is extended to include stocks prediction literature to have a more extensive coverage of text processing methods. The section focuses on forecasting methods involving English news articles as input. For a more complete discussion on other text sources (tweets, financial reports, or online forums) and on various types of financial assets other than stocks and crude oil, please refer to surveys (Nassirtoussi et al., 2014; Xing et al., 2018) on Natural Language-based financial and market forecasting. Overall, forecasting methods involving news articles fall into two categories: (i) Non-event-based and (2) Event-based.

### **2.1.1 Non-Event-Based**

#### **2.1.1.1 Sentiment Analysis**

Sentiment Analysis (opinion mining) is one of the most common text processing used in financial forecasting. It offers great value in determining the tone of large amounts of textual data about firms, markets, commodities, and other financial instruments (Feldman, 2013; Loughran & McDonald, 2016). Most existing studies on sentiment analysis in the field of finance look at the aggregate sentiment score assigned to firms after analyzing texts from trade news or social digital media (Tetlock et al., 2008; Bollen et al., 2011; Feldman et al., 2011; Q. Li, Wang, et al., 2014).

**Domain-specific Sentiment Analysis** It has been shown that methods using general sentiment analysis produced rather dismal results because general sentiment words may not carry the same emotional tendency in the finance realm (Loughran & McDonald, 2011). For a more accurate analysis, specific finance-specific dictionaries are created:

- Henry Dictionary(Henry, 2008) - used in stock prediction (Q. Li, Wang, et al., 2014) and in crude oil forecasting(J. Li et al., 2017)
- Loughran & McDonald Dictionary(Loughran & McDonald, 2011) - used in crude oil forecasting(Wex et al., 2013)
- As part of the Stock Sonar Project(Feldman et al., 2011) use domain experts to formulate event rules for rule-based stock sentiment analysis.

Note: Even though (Sadik et al., 2020) uses Sentiment Analysis in crude oil forecasting but strictly speaking, it does not fully utilize sentiment analysis in that the authors used Ravenpack’s sentiment score (available via subscription) without actually generating the sentiment score themselves.

**Sentiment Analysis Libraries** Apart from using domain-specific dictionaries, there are other sentiment analysis resources such as sentiment analysis libraries:

- VADER (Valence Aware Dictionary and Sentiment Resources)library (Hutto & Gilbert, 2014) - used in crude oil forecasting(Zhao, Zeng, et al., 2019)
- TextBlob library - used Polarity and Subjectivity score in TextBlob library for crude oil forecasting(X. Li et al., 2019; Bai et al., 2022)

According to (Ben Ami & Feldman, 2017), irrespective of the accuracy of the applied sentiment analysis method, most systems tend to miss one crucial factor in the analysis, which is context. By counting positive and negative words, or even by performing flawless sentiment analysis, one can, at most, obtain the overall tone of the article, which Ben et al. argue can be a very noisy signal.

#### 2.1.1.2 Topic Modeling + Sentiment Analysis

Apart from standalone sentiment analysis, topic modeling(Blei et al., 2003) is also used jointly with sentiment analysis for better prediction results. Topic models is used to scan a set of

documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. Authors in (T. H. Nguyen & Shirai, 2015) proposed a solution named TSLDA that simultaneously infers the topics and sentiments for stock market prediction. (X. Li et al., 2019), on the other hand, use topic modeling coupled with sentiment analysis for crude oil forecasting. Authors in (Bai et al., 2022) used SeaNMF(Shi et al., 2018), a special topic modeling method developed for doing topic modeling on news headlines for crude oil prediction.

### 2.1.1.3 Other Non-event based methods

Apart from Sentiment Analysis and Topic modeling (the two most common text processing approach), here are other non-event based methods:

1. (X. Li et al., 2019; Wu, Wang, Lv, & Zeng, 2021) use CNN to extract features from news in crude oil forecasting;
2. (Q. Liu et al., 2018) capture complementary information in news headlines and news contents by using Hierarchical Complementary Attention Network (HCAN) for stock price;
3. (Del Corro & Hoffart, 2021) identify financially relevant news using stock price movements and news headlines by re-purposing attention weights initially trained for stock prediction;
4. (D. Chen, Ma, et al., 2019) generate news summaries based on clusters of related news as input features for Forex movement prediction;
5. (Hu et al., 2018) propose the usage of Hybrid Attention Network (HAN) with self-paced learning mechanism based on three principles for news-oriented stock trend prediction, the three principles are: (1) Sequential Context Dependency, (2) Diverse Influence, and (3) Effective and efficient learning by imitating the learning process of human.

### 2.1.2 Event-Based

This sub-section looks at event-driven approaches used in text-based financial forecasting. Apart from sentiments, news events too exert some form of influence on financial markets. The study of events is used in measuring the impact an event has on the value of a firm (MacKinlay, 1997). An automated event extraction enables more data to be processed in

less time and leads to the usage of correlation between an event and stock market movement in stock forecasting, such as in (Ding et al., 2014, 2015; D. Chen, Zou, et al., 2019). Similar to stocks, news events have been shown to exert a significant impact on crude oil price fluctuations (see Section 1.1.1 for analysis).

In event-based approaches, event information is first extracted from news articles and then is used as input in machine learning-based price or trend prediction. This section looks at event-based extraction in the finance and economics domain in detail<sup>1</sup>. A variety of methods are used in financial or economic event extraction, such as hand-crafted rule-sets, ontology knowledge-bases, and using techniques like fully supervised, distant, weakly, or semi-supervised training.

### 2.1.2.1 Knowledge-Bases, Rules and Domain Ontology

The earlier works in Financial/Economics Event Extraction are rule-based and pattern-based. Listed here are several notable examples: (Malik et al., 2011) introduced statistical classifiers aided by rules to extract financial events from company press releases. In (Arendarenko & Kakkonen, 2012), the authors introduced BEECON, an information and event extraction system based on pattern recognition algorithms and hand-written detection rules for business intelligence. (A. Hogenboom et al., 2013) relies on a handcrafted financial event ontology for pattern-based event detection in the economic domain and incorporates lexicons, gazetteers, PoS-tagging, and morphological analysis to extract financial events from news articles. The Stock Sonar project (Feldman et al., 2011) notably uses domain experts to formulate event rules for rule-based stock sentiment analysis. This technology has been successfully used in assessing the impact of events on the stock market (Boudoukh et al., 2019) and in formulating trading strategies (Ben Ami & Feldman, 2017). (D. Chen, Zou, et al., 2019) created the TOPIX Finance Event Dictionary (TFED) and used rule-based data generation methods (distant supervision) to produce ACE/ERE-like event data. In terms of domain-specific ontology, (Lösch & Nikitina, 2009) developed an ontology-based business news event detection system by integrating additional linguistic information by including semantic knowledge from structured resources such as DBpedia. The main drawback of knowledge-based and rule-based approach, as with all approaches using pattern-based, is that it requires intensive human effort and is limited to a narrow variety of sentence structures.

<sup>1</sup>For Generic event extraction-related work, refer to survey papers (Xiang & Wang, 2019) and (F. Hogenboom, Frasincar, Kaymak, De Jong, & Caron, 2016).

### 2.1.2.2 Fully Supervised Learning

There are solutions that use the fully supervised approach. The requisite is a sizeable annotated dataset. Authors in (Jacobs et al., 2018) train a model via supervised training approach on event detection task using the SentiFM dataset released in (Van de Kauter et al., 2015). Subsequently in (Jacobs & Hoste, 2020) train a model via supervised training approach on fine-grained event extraction on SENTiVENT, a new dataset released in (Jacobs & Hoste, 2021).

There are generally very few supervised learning approaches in Financial / Economic event extraction due to the lack of human-annotated datasets (Jacobs & Hoste, 2021; Upreti et al., 2019). Most methods rely on weak or distantly supervision instead.

### 2.1.2.3 Semi - or Distantly Supervised Approach

In Semi / Distantly / Weakly supervised methods, seed sets are manually labeled or labeled by using rules. However, these often lack the granularity and structure of ACE-like events. (Dor et al., 2019) adopted a weakly-supervised training dataset for identifying company-related events at the sentence level in English news articles. They extract weak labels for sentences describing company events to overcome the constraint of using predefined event taxonomies. Their main interest was in the binary task of detecting sentences containing events. Sentences from a company’s Wikipedia article were selected if they appeared in an event section and started with a date-pattern (such as ‘On/In/By/As of’ ? month ?year). They tested their solution on 10 most eventful S&P companies in Wikipedia for 2019. (Rönnqvist & Sarlin, 2017) provide a weakly-supervised event detection solution that successfully identifies banks in financial distress. They identified 243 events of bank distress and government interventions. They utilize a deep learning approach for detecting relevant discussions and extracting these events.

Apart from English news, there are a number of previous works that used semi-supervised or weakly-supervised approaches to extract events from Chinese economic news, such as (Han et al., 2018), (H. Yang et al., 2018), (X. Liu et al., 2019), and (Qian et al., 2019).

#### 2.1.2.4 Other Methods

Other approaches for event-driven stock price prediction conceptualize financial/economic event recognition as the extraction of event tuples (Ding et al., 2014, 2015; Saha et al., 2017) using OpenIE (Open Information Extraction) methods. (Saha et al., 2017) introduced BONIE (Bootstrapping-based Open Numerical Information Extractor), an extension of Open IE to extract numerical arguments in each Open IE tuple. Another approach is formulating financial/economic event extraction as semantic frame parsing in (Xie et al., 2013) using the SEMAFOR semantic parser.

### 2.1.3 News-oriented Crude Oil Forecasting

The literature reviewed in Section 2.1 covers both stocks and crude oil forecasting. Here this section is dedicated to reviewing exclusively news-oriented crude oil forecasting solutions. Purely econometric/financial methods that combine oil price historical data (time series data) with analytical models for price forecasting are excluded here.

#### 2.1.3.1 Text Source

In text-based crude oil price forecasting, the common text source is news articles. According to authors in (Ksiazek et al., 2016), news articles are more persuasive and less noisy; it is considered a more reliable source than other social media source, such as Twitter and blogs. It is also acknowledged in (Wex et al., 2013) that the oil market is deeply affected by extreme events (such as political instabilities and economic development), and text mining algorithms can extract actionable information from online crude oil news. Hence many existing works use news articles in their crude oil forecasting solution.

Apart from news, many utilize other source of textual information such as the ones listed below:

1. (Wang et al., 2018) use internet searches, which the authors termed ‘internet concerns’ (IC) as a way of quantifying investor attention
2. (Wu, Wang, Lv, & Zeng, 2021) use Google Trends along with News headlines (<http://www.google.com/trends>) together with News Headlines from [oilprice.com](http://oilprice.com)

3. (Elshendy et al., 2018) uses four different online media sources: (1) Google Trends, Twitter, Wikipedia and GDELT<sup>2</sup> Dataset for economic significance on WTI crude oil price.
4. (Wex et al., 2013) used online news comments and messages along with news articles. The authors showed that online news messages have powerful oil price predictive capacity by analyzing over 45 million news messages.

All these solutions use multi-channel approaches that combine two or more input features to build the forecasting models. Authors used news text as one of the input features alongside crude oil historical price data as forecast variables. Hence their methods involve some form of time series analysis such as Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional (GARCH), Empirical Model Decomposition (EMD), etc. For a more targeted discussion, only text processing methods are covered here.

### 2.1.3.2 Scope of Input

In terms of the scope of input, some work used only news headlines, arguing that the headlines are succinct and contain just the right amount of useful information, such as in (Ding et al., 2014; Del Corro & Hoffart, 2021; Wex et al., 2013; Wu, Wang, Lv, & Zeng, 2021; X. Li et al., 2019). However, it is noted in (Wu, Wang, Lv, & Zeng, 2021) that their CNN model alone does not achieve high accuracy as expected. They stated that one possible reason is that news headlines contain only partial information that affects oil price and does not reflect the magnitude of these events. On the other hand, these work (Feuerriegel & Neumann, 2013; J. Li et al., 2017; J. Liu & Huang, 2021) used news content alone without news headlines.

(Hu et al., 2018; Q. Li, Wang, et al., 2014; Xie et al., 2013; Q. Liu et al., 2018; D. Chen, Ma, et al., 2019; Zhao, Zeng, et al., 2019) advocate the use of the entire news (headline + content). Among them, (Q. Liu et al., 2018) argued that relying solely on news headlines while ignoring news content degrades prediction accuracy because there are still useful information in the news body. They proposed a hierarchical complementary attention network (HCAN) to capture valuable complementary information in news headline and content for stock trend prediction. Rather than using news individually, some proposed a form of news aggregation or new summarization of related news. Specifically, (Duan et al., 2018) employs the news

---

<sup>2</sup>GDELT stands for Global Data on Events, Language and tone database



abstract as the target to summarize the news content in order to utilize the abundant information of the news body to predict stock returns. (D. Chen, Ma, et al., 2019) generate news summaries based on clusters of related news as input features for Forex trend prediction.

### 2.1.3.3 Prediction Task

In terms of the prediction task, almost all stock forecasting focused on predicting price trend, treating it as a Multi-class classification task with the following labels: DOWN, UP, and STABLE. For crude oil, apart from trend prediction, there is a number of works that attempt to predict crude oil prices, such as in (Wu, Wang, Lv, & Zeng, 2021). For this, they utilize historical crude oil prices along with any textual information as input features. Their scope, however, is mainly limited to WTI benchmark only. Apart from trend and price, (Duan et al., 2018) predicts Cumulative Abnormal Return (CAR) of stock prices, (Feuerriegel & Neumann, 2013) predicts WTI crude **abnormal** returns<sup>3</sup>, and (Zhao, Liu, et al., 2019) predict price risks or price volatility.

Table 2.1 tabulates and analyses all crude oil forecasting solutions based on the following aspects:

1. News input scope: News headlines / content / entire article
2. List of other non-news input used in building the forecasting models
3. Text Processing Methods
4. Data source: *Reuters, investing.com, oilprice.com*
5. Prediction task: usually involves predicting either WTI or Brent crude oil prices. WTI and Brent are global crude oil benchmarks.

---

<sup>3</sup>abnormal returns is *actual return* minus *the expected return*

Table 2.1: Summary of English news-based crude oil prediction solutions and their respective text processing approaches.

References	News	Other Input	Text Processing Method	Dataset	Prediction Task
(Wex et al., 2013)	news headline	WTI historical price	Sentiment analysis using dictionary by (Loughran & McDonald, 2011)	Reuters Jan 2003 - Dec 2010	Regression analysis on daily WTI closing price
(Feuerriegel & Neumann, 2013)	news content	WTI historical price	Sentiment analysis using various Dictionaries (Stone, 2002; Henry, 2008; Loughran & McDonald, 2011), Bi-Normal Separation (BNS)(Forman, 2002) and Tonality (Liebmann et al., 2012)	Reuters Jan 1 2003 - May 31 2012	Regression analysis on WTI abnormal returns <sup>4</sup>
(J. Li et al., 2017)	news content	WTI historical price	Sentiments with Henry's Finance Dictionary(Henry, 2008)	Reuters 2008 to 2016	WTI futures direction

Continued on next page

---

<sup>4</sup>the actual return minus the expected return

Table 2.1 – Continued from previous page

References	News	Other Input	Text Processing	Dataset	Prediction Task
(Zhao, Zeng, et al., 2019)	entire news article	web content including news + Brent Historical price	sentiment analysis using VADER(Hutto & Gilbert, 2014)	Reuters & UPI Jan 1 2013 - 31 August 2018	Daily Brent spot price
(X. Li et al., 2019)	News headlines	WTI historical price data + WTI futures + US Dollar Index + DJIA	Sentiment Analysis using Textblob (Polarity & Subjectivity score) + features extracted from CNN, and then LDA Topic Modeling(Blei et al., 2003) to group features	investing.com September 15th, 2009, to July 20th, 2014	Daily WTI spot price
(Wu, Wang, Lv, & Zeng, 2021)	News headlines	google trends + WTI historical price	use CNN to extract text features and variational mode decomposition (VDM) to construct useful time series indicators	oilprice.com 15 Jun 2011 - 25 August 2019	Weekly WTI spot price

Continued on next page

Table 2.1 – Continued from previous page

References	News	Other Input	Text Processing	Dataset	Prediction Task
(Wu, Wang, Wang, & Zeng, 2021)	News headlines	Oil market data such as production/consumption/inventory data + WTI historical price	use CNN to extract text features (similar to (Wu, Wang, Lv, & Zeng, 2021))	oilprice.com May 2012 - August 2020	Weekly WTI price and oil production, consumption, and inventory during the COVID-19 pandemic
(J. Liu & Huang, 2021)	News content	WTI historical price	Open-domain event extraction (ODEE) (X. Liu et al., 2019) + sentiment analysis using Vader	The guardian & New York Times Jan 2007 - December 2020	Daily WTI closing price
(Bai et al., 2022)	News headlines	WTI historical price	Topic modeling with SeaNMF (Shi et al., 2018) (topic modeling for short text like news headlines) and Sentiment using TextBlob	investing.com 29 March 2011 - 22 March 2019	Daily WTI futures

## 2.2 Event ‘Factuality’

This section focuses on event factuality. Even though ACE2005 dataset is annotated with not just event details but also properties such as *Polarity*, *Tense*, *Genericity*, and *Modality*<sup>5</sup>, previous event extraction work within the ACE and ERE stream focused almost exclusively on event detection and event extraction while under-utilizing the annotation on event properties. Finance and economic corpus such as SENTiVENT in (Jacobs & Hoste, 2021) have Polarity and Modality annotated, but the events’ Polarity and Modality classification are not in scope for event extraction for SENTiVENT (Jacobs & Hoste, 2020). Even in survey papers such as (Xiang & Wang, 2019) and (F. Hogenboom et al., 2016), the focus is solely on the event extraction task. Instead, event properties-related tasks are established separately from event extraction through several shared tasks that are not necessarily event extraction related. These tasks come in a few various variations with a slightly different focus; they are:

1. Event Realis classification (Mitamura et al., 2015) in TAC KBP dataset. There are three types of Realis values: ACTUAL, GENERIC and OTHER. TAC-KBP considers negated (failed) events, future events and conditional statements under the same Realis value, hence losing fine-grained epistemic status of the events.
2. CoNLL-2010 shared task: Hedge detection and scope resolution. The task is to detect hedges and their scope in natural language text. Detailed task description is found in (Farkas et al., 2010).
3. SEM 2012 Shared Task: Negation detection and scope resolution. The task is to detect negation and resolve its scope and focus. Detailed description of task is found in (Morante & Blanco, 2012).
4. Modal sense classification (Marasović & Frank, 2016). This is similar to Uncertainty hedge / Modality cue word detection;
5. Event Factuality Prediction (EFP) (Saurí & Pustejovsky, 2009). This is a combination of Negation and Speculation detection but instead of classification, EFP is a regression task to predict a score between [+3, -3] to quantify the degree to which the current event mention has happened. A +3 score indicates an event that have certainly happened while a -3 score means that an event certainty have not happen.

According to (Rudinger et al., 2018), negation comes in many different forms:

---

<sup>5</sup>Definition of each of these properties are listed in Appendix C.1

- **Negation** - Jo didn't leave.
- **Modal auxiliaries** - Jo might leave.
- **Determiners** - Jo left no trace.
- **Adverbs** - Jo never left.
- **Verbs** - Jo failed to leave.
- **Adjectives** - Jo's leaving was fake.
- **Nouns** - Jo's leaving was a hallucination

According to this tutorial, types of expressions with modal meanings:

- **Modal auxiliaries** - Sandy must/should/might/may/could be home.
- **Semimodal verbs** - Sandy has to/ought to/needs to be home.
- **Adverbs** - Perhaps Sandy is home.
- **Nouns** - There is a slight possibility that Sandy is home.
- **Adjectives** - It is far from necessary that Sandy is home.
- **Conditionals** - If the light is on, Sandy is home.

A more rigorous definition of Modality and Polarity and in-depth analysis of Modality and Polarity can be found in (Morante & Sporleder, 2012) and also in this ACL 2011 conference [http://mirror.aclweb.org/ijcni11/downloads/tutorial/tu3\\_present.pdf](http://mirror.aclweb.org/ijcni11/downloads/tutorial/tu3_present.pdf). These additional resources are helpful reference points for this work in establishing the annotation guidelines for event properties annotation. This is explored in detail in Section 3.2.1.3.

## 2.3 Language Resources for Event Extraction in the Finance and Economics domain

It is highlighted in (Xing et al., 2018; Upreti et al., 2019) that generally, there is a lack of language resources in the Finance and Economics domain. For the task of financial / economic event extraction, there are only a handful of fully annotated datasets:

1. The English and Dutch SentiFM business news corpus (Van de Kauter et al., 2015) contains token-span annotations of 10 event types and 64 subtypes of company-economic events. Some examples of event types are Buy ratings, Debt, Dividend, Merger &

acquisition, Profit, Quarterly results. (Lefever & Hoste, 2016; Jacobs et al., 2018) performed event detection via supervised training on this dataset. As a continuation of the work, the authors introduced SENTiEVENT, a fine-grained ACE/ERE-like dataset in (Jacobs & Hoste, 2021). Just like the earlier work, their focus is mainly on company-related and financial events.

2. As a continuation to (Van de Kauter et al., 2015), the authors introduced SENTiEVENT, a fine-grained ACE/ERE-like dataset in (Jacobs & Hoste, 2021). The authors performed fine-grained event extraction on this corpus via supervised learning in (Jacobs & Hoste, 2020).

The search for commodity news-related resources led us to RavenPack’s<sup>6</sup> crude oil dataset. This dataset is available through subscription at the Wharton Research Data Services (WRDS). It is made up of news headlines and a corresponding sentiment score generated by Ravenpack’s own analytic engine. Unfortunately, this dataset is not suitable for the task of *supervised* event extraction as it contains only sentiment scores without any event annotations. However, Ravenpack’s event taxonomy on crude oil-related events proves to be a useful resource in helping this work define crude oil-related event typology. Details of event typology is covered in Section 3.2.1.2.

## 2.4 Summary and Discussion

This chapter started with a review of financial news analytics methods in Section 2.1. All the solutions, whether event-based or non-event-based, has their pros and cons, and researchers are exploring to find the best approach in terms of extracting key information from news and representing them appropriately for financial asset forecasting. All of the solutions use the multi-channel approach, where they use text input alongside historical price data.

This is then followed by a detailed discussion of previous work in financial/economic event extraction. The methods used consist of (1) Knowledge-based and rule-based, (2) fully supervised, and (3) semi- or distantly supervised approach. In terms of event extraction, more attention is focused on accurately typing the events and extracting event arguments while

---

<sup>6</sup>RavenPack is an analytics provider for financial services. Among their services are finance and economic news sentiment analysis. More information can be found on their page: <https://www.ravenpack.com/>

disregarding important event attributes like Modality and Polarity, which are crucial in accurately interpreting events.

Lastly, an in-depth analysis of existing language resources in the Finance and Economics domain was carried out to find a suitable corpus for crude oil forecasting. The research questions and research objectives are formed based on this in-depth review and analysis of existing works. The gaps or opportunities found in past work are:

- Section 2.3 has highlighted that there is no readily available fully annotated dataset on crude oil news, nor are there any on other commodities. This gap is addressed as part of RO1 in Chapter 3.
- Even though event extraction datasets are annotated with *factuality* features or properties (ACE2005, TAC KBP, and SENTiVENT), none of the existing event extraction work, whether in the generic domain or in the finance & economics domain, includes classifying event factuality into scope. Events extracted are treated as actual events because the events are not differentiated into actual and hypothetical events. This severely impairs the true interpretation of events, especially in commodity news where expert analysis and market outlook forecasts abound. This led to the formulation of RO2.
- The standard approach used in all text-based financial asset forecasting involving text (news, tweet, financial reports, financial periodicals or etc.) built machine learning models based on the correlation between signals (semantic information extracted from text) and historical financial asset prices (time series data) for the task of price prediction. One of the challenges of this approach is that spurious correlations between the input text and time series data are often included together with actual, non-spurious ones. None has considered utilizing market summaries as a single-source (purely text-based) input to mine for correlation between the occurrence of events and price movement. This has helped formulate RO3.

The proposed solution attempts to create a fit-for-purpose solution for crude oil forecasting by taking into consideration the existing state-of-the-art solutions and also gaps and opportunities highlighted in this chapter. Details of the proposed solution are covered in detail in the chapters that follow.



## Chapter 3

# Construction of *CrudeOilNews* Corpus

**RO1:** To build a labeled dataset for crude oil news with the following types of events annotated: (i) macro-economic, (ii) geo-political, and (iii) supply-demand.

Financial markets are sensitive to breaking news on economic events. Specifically for crude oil markets, it is observed in (Brandt & Gao, 2019) that news about macroeconomic fundamentals and geopolitical events affect the price of the commodity. Apart from fundamental market factors, such as supply, demand, and inventory, oil price fluctuation is strongly influenced by economic development, conflicts, wars, and breaking news (Wu, Wang, Lv, & Zeng, 2021). Therefore, accurate and timely automatic identification of events in news items is crucial for making timely trading decisions.

Event extraction has long been investigated in a supervised learning paradigm. For the case of generic event extraction, canonical datasets such as ACE2005<sup>1</sup> and TAC KBP<sup>2</sup> are used to train machine learning models via the supervised learning approach. Supervised learning requires a large amount of training data, however, annotated is hard and expensive to obtain. This challenge is even more apparent in specialised domains such as finance and economics, where only experts with domain knowledge can provide reliable labels (Konyushkova et al., 2017).

---

<sup>1</sup><https://projects.ldc.upenn.edu/ace>

<sup>2</sup><https://tac.nist.gov/2015/KBP/>

It is highlighted in (Jacobs & Hoste, 2021) that there is very few *strictly* supervised approaches to financial event extraction due to the lack of annotated dataset. For company-related financial events, there are only two datasets: (1) in (Van de Kauter et al., 2015), the authors introduced a dataset focused on annotating continuous trigger spans of 10 types and 64 subtypes of company-economic events in a corpus of English and Dutch economic text, some examples of event types are Buy ratings, Debt, Dividend, Merger & acquisition, Profit, Quarterly results and (2) as a continuation of the work, the authors introduced SENTiEVENT, a fine-grained ACE/ERE-like dataset in (Jacobs & Hoste, 2021). As highlighted in Section 2.3, there is no known dataset for crude oil or any other commodities. The contribution of Research Objective 1 is, therefore, to build a crude oil-specific annotated dataset suitable for training event extraction via the supervised learning approach. This chapter describes the data collection process, the annotation methodology, and the event typology used in producing the corpus.

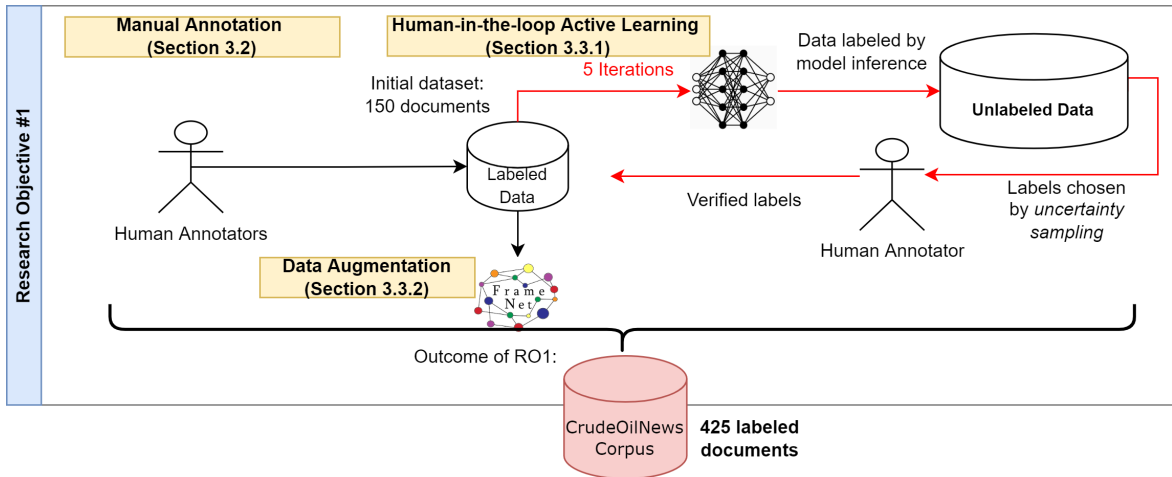


Figure 3.1: The *CrudeOilNews* corpus is built using three components: (1) Manual data annotation, (2) Data Augmentation, and (3) Human-in-the-Loop Active Learning.

Figure 3.1 gives an overview of the components of building the *CrudeOilNews* corpus diagrammatic form. The task of building this labeled dataset is broken down into:

1. Data collection and pre-processing in Section 3.1;
2. Manual annotation in Section 3.2;
3. Expanding the data through:
  - Data Augmentation in Section 3.3.1;
  - Human-in-the-loop Active Learning in Section 3.3.2
4. Corpus Statistics and Analysis in Section 3.4.1;

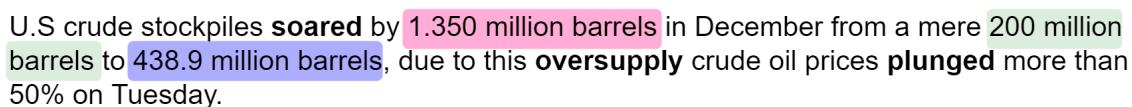
5. Future enhancements in Section 3.5.

### 3.1 Dataset Collection and Pre-processing

First, a crawler is run to extract crude oil news articles from *www.investing.com*<sup>3</sup>, a financial platform and financial/business news aggregator, and is considered one of the top three global financial websites in the world. *www.investing.com* is a notable source for finance-related news and is used as the input source for crude oil prediction in (X. Li et al., 2019; Bai et al., 2022).

News articles dating from Dec 2015 to Jan 2020 (50 months) are extracted. From the pool of crude oil news, 175 pieces of news articles throughout the 50-month period are uniformly sampled to ensure events are evenly represented and not skewed towards a certain topic of a particular time window. These 175 news articles were duly annotated by two annotators. and they form the gold-standard annotation. For the purposes of assessing the inter-annotator agreement and evaluating the annotation guidelines, 25 news articles were selected out of the gold-standard dataset as the adjudicated set (ADJ).

**A Working Example** An example sentence taken from a piece of crude oil news is shown in Figure 3.2. This working example is used throughout this chapter to illustrate how annotation work is carried out.



U.S crude stockpiles **soared** by 1.350 million barrels in December from a mere 200 million barrels to 438.9 million barrels, due to this **oversupply** crude oil prices **plunged** more than 50% on Tuesday.

Figure 3.2: An example of a sentence from a piece of crude oil news, consisting of three events: (1) Crude oil inventory increase, (2) oversupply and (3) Crude Oil price decrease.

### 3.2 Manual Annotation

The dataset is annotated using Brat rapid annotation tool (Stenetorp et al., 2012), a web-based tool for text annotation. This annotated version of the example sentence is shown in Figure 3.3.

<sup>3</sup><https://www.investing.com/commodities/crude-oil-news>

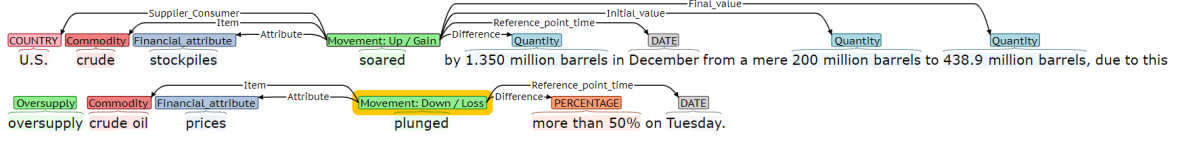


Figure 3.3: An annotation example using Brat annotation tool.

The annotation process is designed to have a high inter-annotator agreement (IAA). One of the criteria is that the annotators should possess domain knowledge in business, finance, and economics. It is imperative for the annotators to understand financial and macro-economic terms and concepts to interpret the text accurately and annotate events accordingly. For instance, sentences containing macro-economic terms such as *contango*, *quantitative easing*, and *backwardation* will require annotators to have finance and economics domain knowledge. To meet these criteria, two annotators were recruited from a pool of undergraduate students from the School of Business of a local university. Annotators were then given annotation training and provided with clear annotation schemas and examples. Every piece of text was duly annotated by two annotators independently.

The annotation was done based on the following layers sentence by sentence:

- Layer 1: Identify and annotate entity mentions (Section 3.2.1.1);
- Layer 2: Annotate events by identifying event triggers (Section 3.2.1.2);
- Layer 3: Using event triggers as anchors, identify and link surrounding entity mentions to their respective events. Annotate the argument roles each entity mention plays with respect to the events identified (Section 3.2.1.2);
- Layer 4: Annotate event properties: modality, polarity and intensity (Section 3.2.1.3).

The example sentence shown in Figure 3.2 is used to illustrate how each category of annotation is carried out.

After each layer, an adjudicator assessed the annotation and evaluated the inter-annotator agreement before finalizing the annotation. For cases where there are annotation discrepancies, the adjudicator will act as the tie-breaker to decide on the final annotation. Once finalized, annotators then proceed with the next layer. This is done to ensure no accumulation of the previous layer's errors in the subsequent layers of annotation.

### 3.2.1 Annotation Guidelines

The annotation schema used is aligned to ACE (Automatic Content Extraction) version 5.4.3 and ERE (Entities, Relations, and Events) version 4.2 standards, so event extraction systems developed for ACE/ERE can be used readily on this corpus. An extensive comparison has been made in (Aguilar et al., 2014), where authors analyzed and provided a summary of the different annotation approaches. Subsequently, there were a number of works that expanded earlier annotation standards, such as in (O’Gorman et al., 2016), authors introduced the Richer Event Description (RED) corpus and methodologies that annotate entities, events, times, entities relations (co-reference and partial co-reference), and events relations (temporal, causal, and sub-events). The aim of this work is to align to ACE/ERE programs as closely as possible, but minor adaptations are made to cater to special characteristics found in crude oil news. For example, *Tense* and *Genericity* defined in ACE2005 are dropped from the annotation scope while the new property - *Intensity* is introduced.

#### 3.2.1.1 Entity Mention

An entity mention is a reference to an object or a set of objects in the world, including named entities, nominal entities, and pronouns. For simplicity and convenience, **values**, and **temporal expressions** are also considered as entity mentions in this work. Nominal entities relating to Finance and Economics are annotated. Apart from crude oil-related terms, below here are some examples of nominal entities found in the corpus and were duly annotated:

- attributes: *price, futures, contract, imports, exports, consumption, inventory, supply, production*
- economic entity: *economic growth, economy, market(s), economic outlook, growth, dollar*

In the example in Figure 3.2, entity mentions are “U.S.”, “crude”, “stockpiles”. Values such as time/date (e.g., “December”, “Tuesday”) and value expressions (e.g., “1.350 million barrels”) are also considered as entity mentions in this work. There are 21 entity types identified and annotated in the dataset, see Table 3.1 for the full list below:

Table 3.1: Entity Types. Those marked with \*\* are identical to Named Entities in NER-tagging.

Entity Type	Examples
1. COMMODITY	<i>oil, crude oil, Brent, West Texas Intermediate (WTI), fuel, U.S Shale, light sweet crude, natural gas</i>
2. COUNTRY**	<i>Libya, China, U.S, Venezuela, Greece</i>
3. DATE**	<i>1998, Wednesday, Jan. 30, the final quarter of 1991, the end of this year</i>
4. DURATION**	<i>two years, three-week, 5-1/2-year, multiyear, another six months</i>
5. ECONOMIC ITEM	<i>economy, economic growth, market, economic outlook, employment data, currency, commodity-oil</i>
6. FINANCIAL ATTRIBUTE	<i>supply, demand, output, production, price, import, export</i>
7. FORECAST TARGET	<i>forecast, target, estimate, projection, bets</i>
8. GROUP	<i>global producers, oil producers, hedge funds, non-OECD, Gulf oil producers</i>
9. LOCATION**	<i>global, world, domestic, Middle East, Europe</i>
10. MONEY**	<i>\$60, USD 50</i>
11. NATIONALITY**	<i>Chinese, Russian, European, African</i>
12. NUMBER**	<i>(any numerical value that does not have a currency sign)</i>
13. ORGANIZATION**	<i>OPEC, Organization of Petroleum Exporting Countries, European Union, U.S. Energy Information Administration, EIA</i>
14. OTHER ACTIVITIES	<i>(free text)</i>
15. PERCENT**	<i>25%, 1.4 percent</i>
16. PERSON**	<i>Trump, Putin (and other political figures)</i>
17. PHENOMENON	<i>(free text)</i>
18. PRICE UNIT	<i>\$100-a-barrel, \$40 per barrel, USD58 per barrel</i>
19. PRICE UNIT	<i>170,000 bpd, 400,000 barrels per day, 29 million barrels per day</i>
20. QUANTITY	<i>1.3500 million barrels, 1.8 million gallons, 18 million tonnes</i>
21. STATE PROVINCE	<i>Washington, Moscow, Cushing, North America</i>

### 3.2.1.2 Events

Events are defined as ‘specific occurrences’, involving ‘specific participants’. The occurrence of an event is marked by the presence of an event trigger. In addition to identifying triggers, all of the participants of each event are also identified. An event’s participants are entities that play a role in that particular event. Details and rules for identifying event triggers and event Arguments are covered in the subsections below.

**Event Triggers** The annotation of event trigger is aligned to ERE where an event trigger (known as event nugget in the shared task (Mitamura et al., 2015)) can be either a single word (main verb, noun, adjective, adverb) or a continuous multi-word phrase as shown in the examples below:

#### 1. Verb:

- Houthi rebels **attacked** Saudi Arabia..

- US **sanctioned** Iran.

## 2. Noun:

- The government slapped **sanctions** against its petroleum....
- ....supply and demand **uctuations** in the international oil market.

## 3. Adjectives:

- Interest rates were **unchanged**.....
- A fast **growing** economy has....

## 4. Adverb:

- The banks **increasingly** expect oil price to stay low.

## 5. Multi-verb:

- The market **bounced back** ....
- The company **laid** their workers **o** ....

Event trigger is the minimal span of text that most succinctly expresses the occurrence of an event. Annotators are instructed to keep the trigger as small as possible while maintaining the core lexical semantics of the event. For example, for the phrase “Oil price edged lower”, only the trigger word “lower” is annotated.

**Event Arguments** After event triggers and entity mentions are annotated, entities need to be linked up to form events. An event contains an event trigger and a set of event arguments. Referring to Figure 3.4, the event trigger soared is linked to seven entity mentions via arches. The argument role of each entity mention is labeled on each arch, respectively, while entity types are labeled in various colours on top of each entity span. This information is also summarized in tabular format in Table 3.2.

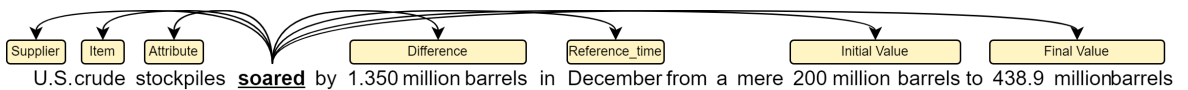


Figure 3.4: Arches link argument roles to the respective event trigger.

**Event Typology** Among the list of event types defined in the SENTiVENT dataset(Jacobs & Hoste, 2021), the only category that overlaps with our work is “Macroeconomics”, an event category that captures a broad range of events that is not company-specific such as economy-wide phenomena, and governmental policy in news. While they choose to remain at a broad

Table 3.2: List of Event Arguments of example in Figure 3.3

Text	Entity Type	Argument Role
U.S.	COUNTRY	SUPPLIER
crude	COMMODITY	ITEM
stockpiles	ATTRIBUTE	FINANCIAL ATTRIBUTE
1,350 million barrels	QUANTITY	DIFFERENCE
December	DATE	REFERENCE_POINT_TIME
200 million barrels	QUANTITY	INITIAL_VALUE
438.9 million barrels	QUANTITY	FINAL_VALUE

level, this work compliments theirs by defining key events in the Macro-economic and Geo-political category at a detailed level. The closest match in terms of language resources, however, is RavenPack data, which contains commodity-related news headlines and their corresponding event category labels. Using Ravenpack's event taxonomy as a reference, a set of 18 oil-related event types are formulated.

1. **Embargo** : Trade or other commercial activity of the commodity is banned.  
*The Trump administration imposed a "strong and swift" economic **sanctions** trigger on Venezuela on Thursday.*
2. **Prohibiting** : Trade or other commercial activity of the commodity is banned.  
*Congress **banned** most U.S. crude oil exports on Friday after price shocks from the 1973 Arab oil embargo.*
3. **Shortage** : Situation where demand is more than supply.  
*Oil reserves are within "acceptable" range in most oil consuming countries and there is no **shortage** in oil supply globally, the minister added.*
4. **Civil Unrest** : Violence or turmoil within the oil producing country.  
*The drop in oil prices to their lowest in two years has caught many observers off guard, coming against a backdrop of the worst **violence** in Iraq this decade.*
5. **Crisis** : (a) A time of intense difficulty, such as other forms of unspecified crisis (grouped under Geo-political event) and (b) Financial / Economic Crisis (which can be grouped under Macro-economic event).  
*Asia 's diesel consumption is expected to recover this year at the second weakest level rate since the 2014 Asian **financial crisis**.*
6. **Geo-political Tension** : Political tension between oil-producing nation with other nations.  
*Deteriorating **relations** between Iraq and Russia first half of 2016 ignited new fears of supply restrictions in the market.*



7. **Oversupply** : Situation where production goes into surplus.  
*Forecasts for an crude **oversupply** in West African and European markets early June help to push the Brent benchmark down more than 20% January.*
8. **Caused-movement-down-loss** : Action taken to reduce / cut production or forecast target, etc.  
*The IMF earlier said it **reduced** its 2018 global economic growth forecast to 3.30% from a July forecast of 4.10%.*
9. **Caused-movement-up-gain**: Action taken to increase production or forecast target, etc.  
*OPEC countries agreed to **boost** production by 1 million barrels per day.*
10. **Movement-down-loss** : Situation where commodity price falls.  
*Globally crude oil futures **fell** \$2.50 to \$59 per barrel on Tuesday.*
11. **Movement-up-gain**: Situation where commodity price rises or trends up.  
*WTI price **surged** to a new high in today's trade.*
12. **Movement- at**: Situation where there is no change to commodity price.  
*In the ICE market, there is little change to BRENT price.*
13. **Slow-weak**: Describes the economic / GDP / Demand of crude oil / Employment condition of a country.  
*Faced with the worst **downturn** in the oil sector in at least three decades , BP reduced its capital spending three times in 2015 to \$19 billion.*
14. **Grow-strong**: Describes the economic / GDP / Demand of crude oil / Employment condition of a country.  
*Post-recession bounces in Europe and **strong** demand in India pushed global oil demand to a ve-year high in the third quarter of 2015.*
15. **Position-high**: Describes the position of the current commodity price.  
*The IEA estimates that U.S. crude oil is expected to seek higher ground until reaching a 5-year **peak** in late April of about 17 million bpd.*
16. **Position-low**: Describes the position of the current commodity price.  
*The BoJ has kept long-term interest rates at **record lows**, but also reduced liquidity in the money market.*
17. **Trade tensions**: An economic conflict often resulting from extreme protectionism in which states raise or create tariffs or other trade barriers against each other in response to trade barriers created by the other party.

*The economic slowdown and threat of trade wars has not just scared oil financial traders.*

18. **Negative Sentiment:** The general feeling towards a situation is pessimistic.

*Energy markets are being battered across the board as negative data fuels concerns that consumption of oil, natural gas and coal will be hit hard by a global economic slowdown.*

Event types and the corresponding list of example trigger words and example key arguments are listed out in Table 3.3 below. Event schemas for all 18 event types, on the other, are listed in Appendix A.1. It is worth highlighting that these 18 event types are the most common ones being reported, there are however other less common ones which are not in scope at the moment, they are events such as ‘positive sentiments’, ‘demand side shocks’ such as demand surge or demand collapse. To have a more complete representation of events, these extra event types may be expanded in future research.

Table 3.3: List of Event types with example trigger words and example key arguments.

Event Type	Example Trigger Word(s)	Example key arguments
1. CAUSED-MOVEMENT-DOWN-LOSS	<i>cut, trim, reduce, disrupt, curb, squeeze, choked off</i>	oil production, oil supplies, interest rate, growth forecast
2. CAUSED-MOVEMENT-UP-GAIN	<i>boost, revive, ramp up, prop up, raise</i>	oil production, oil supplies, growth forecast
3. CIVIL-UNREST	<i>violence, turmoil, fighting, civil war, conflicts</i>	Libya, Iraq
4. CRISIS	<i>crisis, crises</i>	debt, financial
5. EMBARGO	<i>embargo, sanction</i>	Iraq, Russia
6. GEOPOLITICAL-TENSION	<i>war, tensions, deteriorating relationship</i>	Iraq-Iran
7. GROW-STRONG	<i>grow, picking up, boom, recover, expand, strong, rosy, improve, solid</i>	oil production, economic growth, U.S. dollar, crude oil demand
8. MOVEMENT-DOWN-LOSS	<i>fell, down, less, drop, tumble, collapse, plunge, downturn, slump, slide, decline</i>	crude oil price, U.S. dollar, gross domestic product (GDP) growth
9. MOVEMENT-FLAT	<i>unchanged, at, hold, maintained</i>	oil price
10. MOVEMENT-UP-GAIN	<i>up, gain, rise, surge, soar, swell, increase, rebound</i>	oil price, U.S. employment data, gross domestic product (GDP) growth
11. NEGATIVE-SENTIMENT	<i>worries, concern, fears</i>	
12. OVERSUPPLY	<i>glut, bulging stock level, excess supplies</i>	
13. POSITION-HIGH	<i>high, highest, peak, highs</i>	
14. POSITION-LOW	<i>low, lowest, lows, trough</i>	
15. PROHIBITION	<i>ban, bar, prohibit</i>	exports, imports
16. SHORTAGE	<i>shortfall, shortage, under-supplied</i>	oil supply
17. SLOW-WEAK	<i>slow, weak, tight, lackluster, falter, weaken, bearish, slowdown, crumbles</i>	global economy, regional economy, economic outlook, crude oil demand
18. TRADE-TENSIONS	<i>price war, trade war, trade dispute</i>	U.S.-China

### 3.2.1.3 Event Properties - Modality, Polarity, and Intensity

Event extraction (event triggers and event arguments) is simply insufficient to represent the events correctly. In order to have an accurate interpretation of events, it is important to distinguish actual events from speculated/forecasted/anticipated events under the Modality attribute. This is particularly apparent in financial news where expert opinions, analyses, and outlook forecasts are as common as actual events. It is also important to separate positive events from negated ones under the Polarity attribute. Lastly, the Intensity attribute is created to indicate the latest development of an existing event.

Event properties are annotated based on the syntactic or lexical existence of cue words; this is illustrated below (cue words are underlined, and event trigger words are in bold):

- **Syntactic** - using negative cue words or the standard negative syntax (eg: didn't say, don't think)  
did not **cut** supplies, never get to implement **sanctions**.
- **Lexical** - using context  
refused to **sack** him, they backed out of the **purchase**, he denied **killing** the man.

Event properties are formulated as slightly different tasks and has different names, for example:

1. Event Realis classification ([Mitamura et al., 2015](#))
2. Uncertainty detection (CoNLL-2010)
3. Modal sense classification ([Marasović & Frank, 2016](#))
4. Event Factuality prediction (EFP) ([Veyseh et al., 2019](#))

Realis and EFP combined Negation and Speculation into a single task: Realis is classification while EFP is a regression analysis to produce a score between [+3, -3].

#### **Polarity** (POSITIVE, NEGATIVE)

In terms of **Polarity**, an event is NEGATIVE when it is explicitly indicated that the Event did not occur (7). The non-occurrence of the Event must be explicitly and intentionally communicated. All other Events are POSITIVE (6). If event properties are not taken into scope, then event extraction will yield the event cut oil prices for both sentences. However in reality, (6) and (7) mean exactly the opposite.

- (6) OPEC countries **cut** oil supplies. [POSITIVE]  
 (7) OPEC countries *refused* to **cut** oil supplies. [NEGATIVE]

Examples of some of the Polarity-NEGATIVE cue words found in this corpus are *not, not, ease, block, break, lift, refuse, resist, thwart*.

### Modality (ASSERTED, OTHER)

An event is annotated as ASSERTED when the author or speaker refers to it as though it was a real occurrence. All other Events, including *believed events, hypothetical events, commanded / requested event, speculated / forecasted event* are annotated as OTHER. This is explained with (8) and (9).

- (8) Washington issued a statement confirming US's **sanctions** on Iran. [ASSERTED]  
 (9) The market *expects* US to **sanction** Iran. [OTHER]

While the event for (8) and (9) is the same, i.e. US sanctions Iran, but in essence, they are different - (8) is an actual event while (9) is a speculated event that has yet to occur. Example of some of the Modality-OTHER cue words found in this corpus are *forecast, hope, pledge, if, may, might, could, will, would, plan, once, unless, threaten*.

### Intensity (NEUTRAL, INTENSIFIED, and EASED)

Event intensity is a new event property specifically created for this work to better represent events found in this corpus. Often, events reported in Crude Oil News are about the latest development of an existing event, whether the event is further intensified or eased.

Examples of events where one is INTENSIFIED and the other one EASED:

- (10) ...could hit Iraq's output and *deepen* a supply **shortfall**. [INTENSIFIED]  
 (11) Libya's civil **strife** has been *eased* by potential peace talks. [EASED]

The event **strife** (CIVIL UNREST) in (10) and (11) is not an event with negative polarity because the event has actually taken place but with reduced intensity. INTENSITY label is used to capture the interpretation accurately, showing that the civil unrest event has indeed taken place but now with updated 'intensity'. Examples of some of the Intensity cue words found in this corpus are:

1. Intensity-INTENSIFIED: *extended, further, stoking, worsen, deepen, heightened, re-new, prolong*;
2. Intensity-EASED: *pare, lift, o set, ease, reduce, curb, reverse, cap*

With these three event properties, all essential information about an event can be annotated and captured. To further illustrate this point, consider the list of examples of complex events below:

- (12) OPEC *cancelled* a *planned easing* of output **cuts**. [NEGATIVE, OTHER, EASED]
- (13) In order to end the global crisis, OPEC may *hesitate* to implement a *planned loosening* of output **curbs**. [NEGATIVE, OTHER, EASED]
- (14) Oil prices rose to \$110 a barrel on *rumours* of a *renewed* **strife**. [POSITIVE, OTHER, INTENSIFIED]

### 3.2.2 Inter-Annotator Agreement

Inter-annotator agreement (IAA) is a good indicator of how clear the annotation guidelines are, how uniformly annotators understand it, how robust is the event typology and overall, how feasible the annotation task is. IAA on each annotated category is evaluated separately (see Table 3.4 for the list) using the most commonly measurement: Cohen’s Kappa, with the exception of *entity spans* and *trigger spans*. These two annotations are made at the token level, forming spans of a single token or multiple continuous tokens. For the sub-tasks of *entity mention detection* and *trigger detection*, the token-level span annotation were unitized to compute IAA, this approach is similar to unitizing and measuring agreement in Named Entity Recognition(Mathet et al., 2015). According to (Hripcsak & Rothschild, 2005), Cohen’s kappa is not the most appropriate measurement for IAA in Named Entity Recognition. In (Deleger et al., 2012), the authors provided an in-depth analysis of why is the case and proposed the use of a pairwise F1 score as the measurement. Hence for the evaluation of *entity spans* and *trigger spans*, both F1 as well as “token-level” kappa are reported. Both scores were measured without taking into account the un-annotated tokens - labelled “O”.

As for the rest of the annotation category, only Cohen’s Kappa is reported as this is the standard measure of IAA for classification task. The agreement is calculated by comparing

annotation outcomes of the two annotators with each other, arbitrarily treating one as the ‘gold’ reference. Each annotator is also scored separately on the adjudicated (ADJ) set. The ADJ set consists of 25 documents collected through correcting and combining the manual annotations of these documents by the adjudicator. The final scores are calculated by averaging the results across all comparisons. Table 3.4 shows the average agreement scores for all annotation categories.

Event nugget scoring method introduced in (Z. Liu et al., 2015) was not used here because their assessment is rolled up into “Span”, “Type”, and “Realis”, too coarse to show IAA on each annotation category.

Table 3.4: The Inter-Annotator Agreement (IAA) for all annotation categories. For categories involving spans (marked by\*), both Cohen’s kappa (calculated on “token level”) and F1 score measurements are provided.

Task	Cohen’s Kappa	F1 Score
Entity spans*	0.82	0.91
Trigger spans*	0.68	0.75
Entity Type	0.89	-
Event Type	0.79	-
Argument Role	0.78	-
Event Polarity	0.70	-
Event Modality	0.63	-
Event Intensity	0.59	-

### 3.2.2.1 Analysis

These IAA scores are benchmark-ed with the ‘strength of agreement’ of each Kappa range as set out by (Landis & Koch, 1977). Most annotation categories achieved *substantial agreement* with the exception of *Intensity* classification. This is because classifying *Intensity* is more challenging where some of the cue words for determining the event intensity are themselves trigger words. For example:

- (15) **Oversupply** could rise next year when Iraq starts to export more oil.

The word *rise* here is a cue word to indicate that oversupply might be further INTENSIFIED, but it also could be misinterpreted as another separate event. On the other hand, there is very high agreement on identifying entity spans. This is because entities in the news articles are majority Named Entities with very clear span boundaries, and classifying the entities to the correct entity type is also rather straightforward. Even for nominal entities such as *crude oil*, *oil markets*, etc., their span boundaries are clear.

The common mistake in trigger span detection and classification is the different interpretations of the minimum span of an event trigger. Examples of common annotation errors are: (i) the trigger word for “crude oil **inched higher**” should be just “higher”, and (ii) “Oil **pursued an upward trend**” should be just “upward trend”.

The cases where annotators disagree are analyzed and it is found that most of them stem from differences in interpreting special concepts for example:

- If events surrounding US *employment* data are annotated, then what about *unemployment*? Should this be treated as employment data but negated using negative polarity?
- How should double negation be treated? For example, ‘failed attempt to prevent a steep drop in oil prices’, both *failed* and *prevent* are considered negative polarity cue words, creating a double negation situation.

For these non-straight forward cases, each one was handled on a case-by-case basis where the adjudicator discussed each situation with the annotators to seek consensus before finalizing the annotation.

### 3.3 Expanding the dataset

Manual annotations are labour-intensive and time-consuming, as this is seen in our gold-standard manual annotation, where it consists of only 175 documents or news articles. To produce a sufficiently large dataset useful for supervised event extraction, we utilize (1) Data Augmentation and (2) Human-in-the-Loop Active Learning.

#### 3.3.1 Data Augmentation

The main purpose of introducing augmented data is to address the issue of serious class imbalance in Event Properties in the dataset. Using just the gold-standard dev dataset, an initial round of event properties classification model training show rather poor results, as shown in pink-colored cells in Table 3.6. As a strategy to overcome class imbalance, the minority classes are manually over-sampled for data augmentation and introduced into the dataset. The idea is to obtain more data for minority classes through data augmentation.

To this end, data augmentation is carried out through (i) trigger word replacement and (ii) event argument replacement).

### 3.3.1.1 Trigger Word Replacement

The idea is to replace the existing trigger word in an annotated sentence with another valid trigger word while maintaining the same semantic meaning. This is achieved by using FrameNet<sup>4</sup>. Candidate replacements are chosen from the list of lexical units within a particular frame in FrameNet.

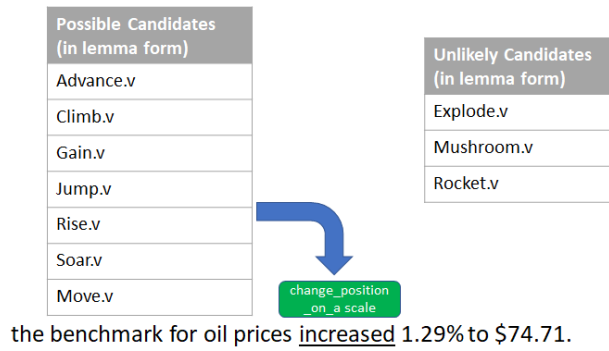


Figure 3.5: Diagrammatic depiction of trigger Word replacement.

In the example shown in figure 3.5, new sentences can be generated by replacing the trigger word. However, if the replacement candidates are purely chosen based on the frame's lexical unit, then invalid sentences might be generated, as illustrated in the examples below:

- (16) The benchmark for oil prices **advanced** 1.29% to \$74.71.
- (17) The benchmark for oil prices **exploded** 1.29% to \$74.71.

While (17) may be grammatically correct but the word *exploded* is generally not used to describe the change in price. It is, however, suitable to be used in a context like "*The population of *ies exploded**". Therefore, FrameNet's lexical units to be used in data augmentation were manually selected to ensure valid and coherent sentences.

<sup>4</sup><https://framenet.icsi.berkeley.edu>



### 3.3.1.2 Argument Replacement

Similar to the approach used in Trigger Replacement, argument replacement aims to generate syntactically diverse but semantically similar sentences. Here event arguments are replaced with entities that have played the same role. In essence, there is no change to the entity type nor argument role labels but only a change in words.

(18) .....after drone [assailant] **attacked** Saudi crude facilities[victim].

(19) ....after an UAV (unmanned aerial vehicle) [assailant] **attacked** major oil installations in Saudi Arabia [victim].

(19) is produced after event arguments in (18) are replaced with semantically similar terms. Care, however, has to be exercised when the argument involves a Geopolitical Entity (GPE). As explained in (Brandt & Gao, 2019), there are differences between news involving oil-producing countries and those involving oil-consuming countries. For example, civil unrest in the major oil-consuming countries often signals economic contraction and decreased oil consumption. Such news is typically positively correlated with oil prices. In contrast, civil unrest in the major oil-producing countries causes concern of supply disruption and is negatively related to oil prices. The authors have classified countries into (1) major oil-consuming and (2) major producing countries, as seen in Table 3.5.

Table 3.5: List of Major Oil-Consuming Countries.

Major oil-consuming countries				
U.S	China	Japan	India	Germany
Eurozone	Korea	France	U.K	Italy
Major oil-producing countries				
Saudi Arabia	Russia	Iran	Mexico	Canada
UAE	Venezuela	Norway	Kuwait	Nigeria
Iraq	Brazil	Algeria	Libya	Angola

Since geo-political news is highly sensitive to the GPE involved, argument replacement for the GPE slot is done only within its category using heuristics, ie. a major-oil consuming country is being replaced by another major-oil consuming country and likewise for major oil-producing countries. For example “America sanctions Iran.” is rewritten to

(20) *Eurozone sanctions Venezuela.*

In replacing words, be it trigger word(s) or arguments, the replacements chosen need to be semantically similar and coherent with the context. The intended result is new sentences

without altering the existing labels nor changing the ground truth of the sentence. Through the above augmentation process, the manually annotated dataset is expanded and enriched by the variety of Trigger Words and Event Arguments. However, this may also increase the risk of generating a dataset too homogeneous which will lead to creating a machine learning model that has a tendency to overfit the data. To overcome this, the percentage of augmented data to be added to the labeled dataset is limited to just 13% of the sentences in the manually annotated dataset.

After adding augmented data into the training set, the green-coloured cells in Table 3.6 show improved F1-scores for minority classes across all three event properties.

Table 3.6: Event Properties Distribution and classification results (F1-score) before and after data augmentation. Prior to data augmentation: the corpus shows obvious class imbalance for all three event properties. Post-Data Augmentation: Class distribution is slightly adjusted and F1-scores for minority classes improved accordingly.

	Gold Dev Set		Before	Augmentation	Updated Count		After
Event Properties	Ratio	# Events	F1	# Events	Ratio	# Events	F1
Polarity: POSITIVE	97.01%	2,855	0.76	965	95.40%	3,820	0.76
Polarity: NEGATIVE	2.99%	88	0.24	96	4.60%	184	0.39
Modality: ASSERTED	82.94%	2,441	0.71	771	80.22%	3,212	0.74
Modality: OTHER	17.06%	502	0.35	290	19.78%	792	0.42
Intensity: NEUTRAL	93.78%	2,760	0.76	745	87.54%	3,505	0.85
Intensity: EASED	3.64%	107	0.36	196	7.57%	303	0.49
Intensity: INTENSIFIED	2.58%	77	0.25	120	4.90%	196	0.37

The bar chart in Figure 3.6 shows the class distribution of Event Polarity, Modality and Intensity ‘before-and-after’ data augmentation (left bar of each group is the original distribution before data augmentation). The minority classes are specifically targeted as candidates for data augmentation. As a result, post-data augmentation saw an increase in minority classes, as shown in Figure 3.6.

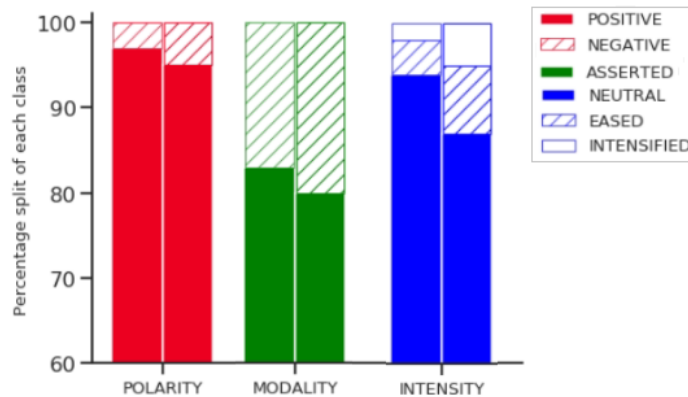


Figure 3.6: Bar chart shows the ‘before-and-after data augmentation’ class distribution for Polarity, Modality and Intensity.

As part of the Augmented Data exercise, a total of **372** sentences containing **1,061** events are added into the dataset.

### 3.3.2 Human-in-the-loop Active Learning

Active learning is well-motivated in many modern machine learning problems where data may be abundant, but labels are scarce or expensive to acquire (Settles, 2009). *Human-in-the-loop Active Learning* is a strategy of utilizing human expertise in data annotation in a more efficient manner. It is a process of training a model with available labeled data and then using the model to predict the labels for unlabeled data. Predictions that are ‘uncertain’ (or of low confidence) is then given to human experts for verification. Verified labels are then added to the pool of labeled dataset for training. These predictions are chosen based on *uncertainty sampling*, a sampling strategy to filter out predictions that the model is least confident with. This way, the scope is narrowed down, and have human experts work specifically on these instances. Rather than blindly adding more training data incurring more cost and time, here instances that are near the model’s decision boundary are targeted as they are valuable when labeled correctly and added to the training data to improve model performance. The whole active learning cycle is shown in Figure 3.7.

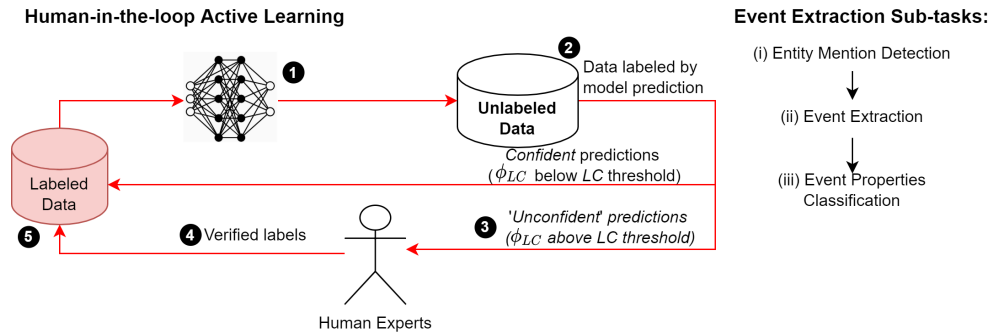


Figure 3.7: Human-in-the-loop active learning cycle: it involves (1) training the model with labeled data, (2) using the model to label new data via model prediction, (3) generating sample instances via uncertainty sampling, (4) validating these sample instance by human experts (relabeling if necessary), and (5) adding checked instances to the pool of training data and re-train the models. Steps 1 - 5 are repeated for each event extraction sub-task.

**Least Confidence score** : *Least confidence score*,  $LC$  captures how un-confident (or uncertain) a model prediction is. For a probability distribution over a set of labels  $y$  for the input  $x$ , the least confidence score is given by the following equation, where  $P(y|x)$  is the highest confidence softmax score:

$$_{LC}(x) = (1 - P(y \neq \hat{y})) \cdot \frac{n}{n-1} \quad (3.1)$$

The equation produces a *Least Confidence (LC)* scores with a 0-1 range, where 1 is the most uncertain score while 0 is the most confidence score,  $n$  is the number of classes for  $y$ . The score is normalized for  $n$  number of classes by multiplying the result by the number of classes, and dividing by  $n - 1$ . Hence it can be used in binary classification as well as multi-class classification. Any model predictions with  $_{LC}$  score above the threshold is sampled as they are most likely to be classified wrongly and need to be relabeled by a human annotator.

### 3.3.2.1 Baseline models

As the baseline for the first round of Active Learning, a number of basic or ‘vanilla’ machine learning models is trained, one for each sub-tasks using the **new development set**, which is made up of gold-standard manually annotated data in Section 3.2 and augmented data in Section 3.3.1, as training data and ADJ set as test data (See Table 3.7 for key statistics). These “vanilla” models also act as the pilot study demonstrating the use of this dataset in event extraction. The following section describes how these models are trained.

**Data Preprocessing** The annotated data consist of original text files (one *.txt* file for each commodity news article) and their corresponding annotation files (*.ann*) generated by Brat rapid annotation tool. As part of data preprocessing, each pair of *.txt* and *.ann* files were processed and converted to a json file. Sentences in each json files were then parsed using Stanford CoreNLP toolkit<sup>5</sup>, which includes sentence splitting, tokenization, POS-tagging (Part-of-Speech tagging), NER-tagging (Named Entity Recognition tagging), and dependency parsing to generate dependency parse trees.

**Entity Mention Detection Model** The Entity Mention Detection task is formulated as multi-class token classification. Similar to the approach used in (T. H. Nguyen et al., 2016), the BIO annotation schema is employed to assign entity type labels to each token in the sentences. For the model architecture, Huggingface’s BERT model with a BERTForTokenClassification head is used to fine-tune on this task.

<sup>5</sup><http://stanfordnlp.github.io/CoreNLP/>

**Event Extraction Model** Event Detection and Argument Role Prediction are jointly trained using JMEE (Joint Multiple Event Extraction), an event extraction solution proposed by (X. Liu et al., 2018). The original version of JMEE uses GloVe word embedding; for this work a modified version of JMEE is used where GloVe is replaced with BERT (Devlin et al., 2019) contextualized word embeddings, original codes (GloVe embedding) are available here: <https://github.com/nlpcl-lab/bert-event-extraction>.

**Event Properties Classification** The event properties classification is formulated as a classification task. The BERT architecture with a BERTForSequenceClassification head is used to fine-tune on this task. For every event identified in the earlier model, the event ‘scope’ is extracted as input for the training. This ‘scope’ is made up of the trigger word(s) being the anchor plus  $n$  tokens surrounding it. For the training,  $n = 8$  is used. Using the example sentence presented in Figure 3.3, the ‘scope’ for the second event is “*oversupply crude oil prices plunged more than 50% on*”. This sequence of text is fed into the model for event property classification.

### 3.3.2.2 Experiments and Analysis

**Least Confidence (LC) threshold** : In order to find the optimum sample size for human relabeling, the suitable *LC* threshold needs to be determined. The uncertainty sampling exercise is designed as a Binary Classification task with two outcomes: *sampled* and *not-sampled*. Different threshold values are experimented with to find the optimum sample size for human validation. Apart from being used in the IAA study, the adjudicated (ADJ) set is also used here as the hold-out set to determine the best *LC* threshold. The *sampled* and *not-sampled* instances are checked against the ground-truth in ADJ, and were able to construct the confusion matrix and obtain *Precision*, *Recall* and *F1* scores. Ideally, the threshold should produce a high *Recall* score (sample as many erroneous cases as possible for human relabeling) and a high *Precision* score as well (identify only relevant instances for correction by keeping correct ones away from being sampled). Different *LC* threshold value ranging from 0 to 1 is experimented to find the best threshold that produces *sampled* and *not-sampled* split with the best F1 score (the highest precision-recall pair). All five iterations of active learning (described next) is run using the following *LC* thresholds: Entity Mention Detection - 0.60, Trigger Detection - 0.55, Argument Roles Prediction - 0.50, Event Polarity - 0.40, Modality - 0.30, and Intensity - 0.45.

## Experiments :

Five iterations of active learning are run; each iteration adds 50 unlabeled crude oil news being labeled through model prediction. Then uncertainty sampling is run. Two annotators are assigned to validate the samples and relabel them if needed. For sentences not sampled, they are deemed 'confident' and therefore being validated/checked by just a single annotator.

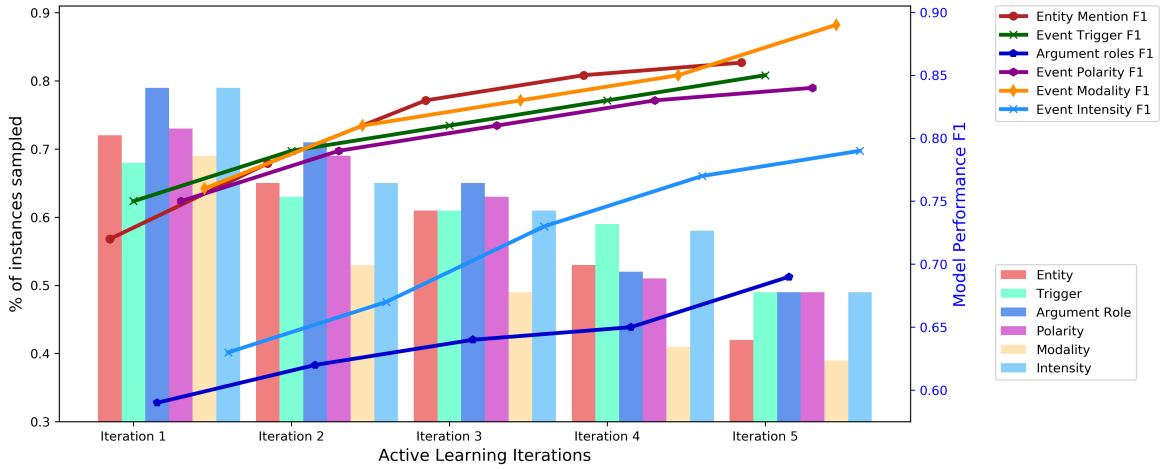


Figure 3.8: Results of Active Learning of 5 iterations of Human-in-the-loop Active Learning: (i) the bar chart captures the percentage of data sampled as part of uncertainty sampling; (ii) the line graph shows the model performance (Micro F1 measure) for each sub-tasks. There is an inverse relationship between model performance and percentage of data sampled through uncertainty sampling. See Tables D.1 and D.2 for results in tabular form.

**Analysis :** Overall there are improvements in model performance across all sub-tasks. As shown in Figure 3.8, models performance progressively improved after each iteration. This is because as more annotated training data are added to the training, the more “confident” the model gets, the fewer instances are sampled under *uncertainty sampling* in each iteration. This inverse relationship is shown in Figure 3.8. It is clear that as model performance (Micro F1 measure) improves, the percentage of sampled data decreases.

The least confidence sampling approach is very effective in identifying data points that are near the model’s decision boundary. In the case of event type, typically, these are events types that can easily be confused with other types. For example, the model erroneously classify **trade tension** as Geopol i t i c a l -Tensi on when the right class should be Trade-Tensi ons. As the word ‘tension’ exist in both event types, it is understandable why the model makes such a mistake. Least confidence sampling is also able to pick up instances of minority classes. Due to the fact that the model has significantly fewer data from minority classes to learn from, this caused the model to generate predictions that are less ‘confident’.

After five iterations of Human-in-the-Loop Active Learning cycle, a total of **250 documents** containing **approximately 4000 sentences** and **approximately 6000 events** are added into the dataset.

### 3.4 Corpus Evaluation and Analysis

#### 3.4.1 Corpus Statistics and Analysis

In the end, a final dataset consisting of **425 documents**, which consist of **7,059 sentences**, **10,578 events**, **22,267 arguments** is produced. The breakdown is shown in Table 3.7.

Table 3.7: *CrudeOilNews* Corpus Statistics

	Gold-standard		Augmented Data	5-Iterations of Active Learning
	Development	Test/ADJ		
# documents	150	25	-	250
# sentences	2,557	377	372	3,753
# tokens	68,219	9,754	12,695	99,884
# Entities	7,120	1,970	1,838	19,417
# Events	2,943	577	1,061	5,997
# Arguments	5,716	1,276	1,693	13,582

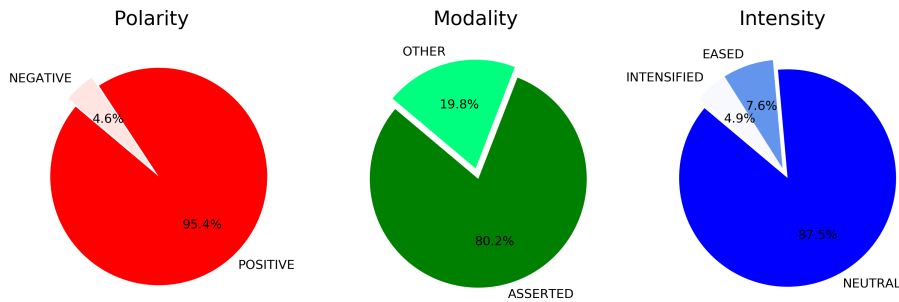


Figure 3.9: Event Polarity, Modality and Intensity Class Distribution.

#### 3.4.2 Unique Characteristics

As part of the analysis of the corpus, it is observed that the *CrudeOilNews* exhibits a set of unique characteristics:

1. **Number intensity** - There is an abundance of Numbers (e.g.: price, difference, percentage) and date (including Day, Time, duration) in the *CrudeOilNews* corpus as they are widely used to express financial information.

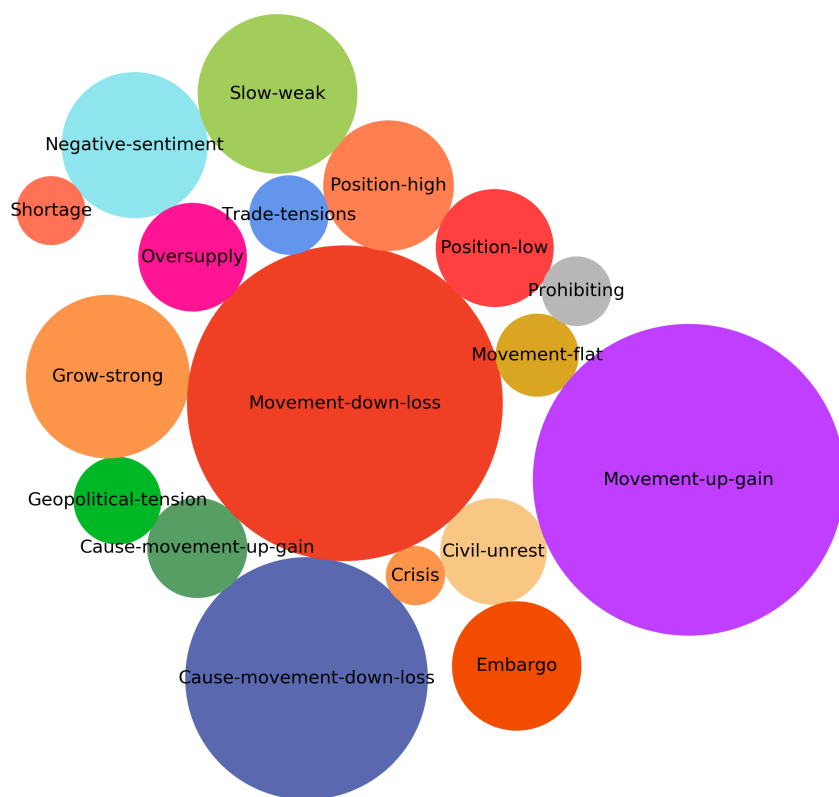


Figure 3.10: Event Type Distribution. See Table D.3 for event type distribution in tabular form.

2. **Arguments homogeneity** - argument roles are non-differentiable from their entity-mention type. Many arguments in the example sentence in Figure 3.2 have the same entity type but actually play different roles to different events. Figure 3.4 shows that *1.350 million barrels*, *200 million barrels*, *438.0 million barrels* are tagged as *QUANTITY*, however all three arguments play a different role in relation to the event (see Argument roles at the top row).
3. **Class imbalance / topic bias** - serious class imbalance in event Properties in Figure 3.9 and as well as event type distribution in Figure 3.10 where the majority class outnumbers the minority classes by a large margin;
4. **Numerous events in a sentence** - as seen in the example sentence above that contains, on average, three events in every sentence in the annotated dataset contains three events. It is therefore challenging to link the arguments to the right event correctly.
5. **Single-type event triggers** - overall, trigger type classification is fairly simple and straightforward because, unlike the ACE2005 dataset, event triggers are associated with only one type of event; there are no multi-type triggers.



6. **Expert Opinions, Analyses and Outlook forecasts** - apart from actual events, it is common to find financial and economic forecasts being reported as well. Hence there is a strong need to differentiate them to have an accurate and complete interpretation of the events.

It is with these characteristics in mind that event extraction model is designed. This is further elaborated in Chapter 4.

### 3.5 Future Enhancements

Even though the event schema used here follows ACE/ERE guidelines closely, there are complicated events that are not specified in the guides and hence may not have the best annotation schema at this stage. These complicated events are considered out of scope for this work. As future enhancement, a new set of event schemas can be designed to cater to these special types of events. Below here is an example:

- (21) Then spent the rest of the week trying to defend those gains as market optimism over the vaccine gave way to concerns over the logistics of its eventual roll-out, though other factors contributed to the two-sided trade.

(21) is an example of a sentence written differently than a standard news article; the information is conveyed in an indirect way that makes it challenging to pinpoint clear-cut events. Hence more thought need to go into determining the best way to annotate events like these so that the events can be accurately represented.

### 3.6 Comparison with other event extraction corpus

Here the *CrudeOilNews* corpus is compared with the canonical ACE2005 dataset and a similar corpus - SENTiVENT. Table 3.8 shows the comparison between these three datasets in terms of their statistics as well as annotation details.

Table 3.8: Comparison of *CrudeOilNews* with other datasets.

	ACE2005 (English corpus)	SENTiVENT	CrudeOilNews
# documents	599	288	425
# Events	6,000	6,194	10,578
# Event types	8 main and 33 subtypes	18 main and 42 subtypes	18 types
# Entity types	4 main entity type, values, temporal expressions	-	21 entity types
Event Properties	Polarity, Tense, Genericity, and Modality	Polarity and Modality	Polarity, Modality and Intensity
Other annotations	Entity Relation	Event Co-reference and Canonical referent of a pronominal and anaphoric noun phrase ('it', 'the company')	-

### 3.7 Summary and Discussion

The contributions to Research Objective #1 are as follows:

- Introduced *CrudeOilNews* corpus, the first annotated corpus for crude oil consisting of 425 crude oil news articles. It is an ACE/ERE-like corpus with the following annotated: (i) Entity mentions, (ii) Events (triggers and argument roles), and (iii) Event Properties (Polarity, Modality, and Intensity);
- Introduced a new event property to capture a complete representation of events. The new property -INTENSITY captures the state of an existing event, whether it further intensifies or eased;
- Addressed the obvious class imbalance in event properties by over-sampling minority classes and adding them into the corpus through data augmentation;
- Used Human-in-the-Loop Active Learning to expand the corpus through model inference while optimizing human annotation effort to focus on just less confident (and likely less accurate) predictions.

RO1 is met, and the deliverable for RO1 is the *CrudeOilNews* corpus, an annotated dataset with macro-economic, geo-political, crude oil supply-demand events identified and annotated. This corpus is used in subsequent chapters to train for event extraction.

## Chapter 4

# Event Extraction

**RO2:** To propose an event extraction model to extract crude oil-related events and market outlook from news corpus.

Event extraction is an important task in Information Extraction. It is the process of gathering knowledge about incidents found in texts, automatically identifying information about what happened, when it happened, and other details. Event extraction has long been a challenging task, addressed mostly with supervised methods<sup>1</sup> that require massive amounts of annotated data. Here, the task of event extraction on the *CrudeOilNews* corpus is investigated.

**Unique Characteristics of *CrudeOilNews*** Crude oil-related events are distinctly different from generic events and even company-related events. As a result, existing solutions may not be effective for the *CrudeOilNews* corpus. The full list of unique characteristics of this corpus is reported in Section 3.4.2. A portion of them is highlighted here using the example sentence in Figure 4.1 to justify the need for a new fit-for-purpose solution.

### A working example

Below is a list of unique characteristics explained using the example sentence in Figure 4.1:

1. **Number intensity** - numbers (e.g., price, difference, percentage of change) and dates (including date of the opening price, dates of closing price) are abundant in crude oil

---

<sup>1</sup>Apart from supervised methods, there is fewer who use some form of Weak Supervision, Distant Supervision, etc. (see (Xiang & Wang, 2019) for a survey of existing event extraction methods).

U.S crude stockpiles **soared** by 1.350 million barrels in December from a mere 200 million barrels to 438.9 million barrels, due to this **oversupply** crude oil prices **plunged** more than 50% on Tuesday.

Figure 4.1: The same example as shown in Section 3.2.1 but reproduced here to highlight the unique characteristics of events found in the *CrudeOilNews* corpus.

news. These numerical data is critical in expressing financial information. In Figure 4.1, the numerical values are: *1.350 million*, *200 million*, *438.9 million* and date information are *December*, *Tuesday*;

2. **Arguments homogeneity** - argument roles are non-differentiable from their entity-mention type. Many arguments in the example above have the same entity type but actually play different roles in different events. In the example, all numerical values are of the same entity type, ie. *1.350 million barrels*, *200 million barrels*, *438.0 million barrels* are tagged as QUANTITY, however all three arguments play a different role in relation to the event;
3. **Numerous events in a sentence** - the example sentence above contains 3 events: **soared**, **oversupply** and **plunged**. Similarly, the majority of the sentences *CrudeOilNews* corpus contain a few events. In fact, on average, every sentence contains about three events. It is therefore challenging to link the arguments to the right event correctly;
4. **Event factuality**: apart from actual events, it is common to find financial and economic forecasts being reported as well. Hence there is a need to differentiate them in order to have an accurate and complete interpretation of the events.

Given the unique characteristics of the *CrudeOilNews* corpus listed above (complete list if found in Section 3.4.2), this work proposes a solution suitable for event extraction from the *CrudeOilNews* corpus. Rather than training models from scratch using Supervised Learning, here the proposed approach leverages the power of transfer learning where event extraction tasks are fine-tuned from a pre-trained language model.

Figure 4.2 gives an overview of the proposed event extraction solution in diagrammatic form. The task of event extraction is broken down into:

1. Preliminary: Domain adaptive pre-training on in-domain text to produce ComBERT in Section 4.3;

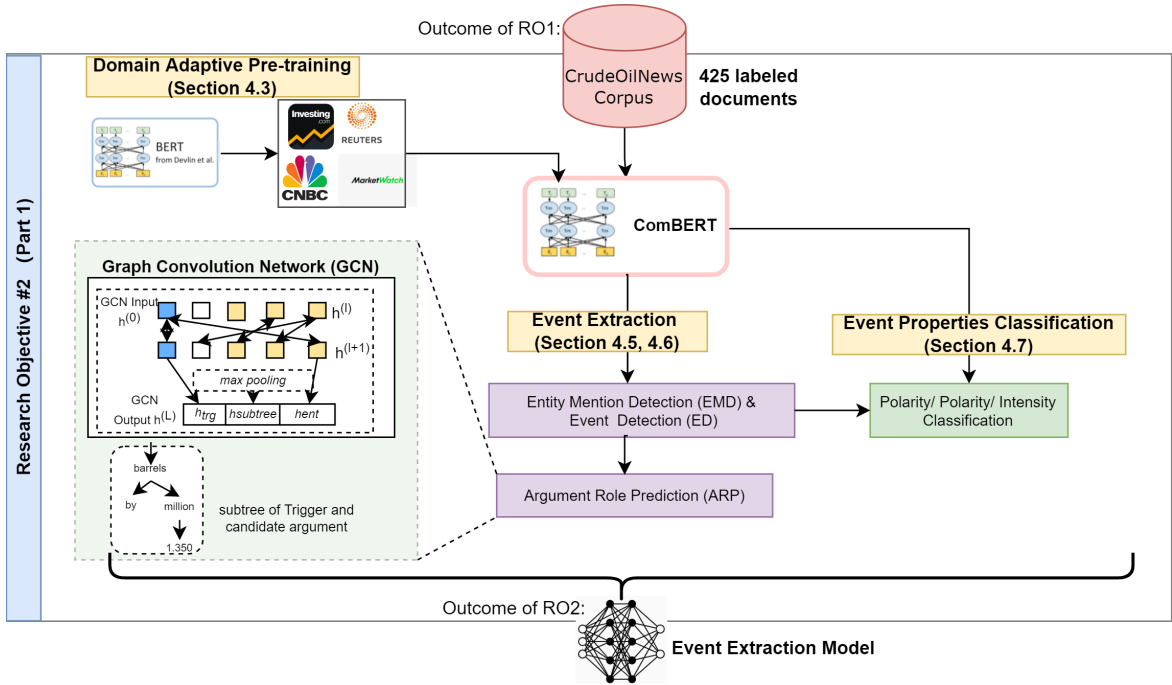


Figure 4.2: Event Extraction solution is made up of these components: (1) Domain Adaptive Pre-training, (2) Subtask 1: EMD and ED, (3) Subtask 2: ARP and (3) Subtask 3: Event Properties Classification

2. Subtask 1: Entity mention detection (EMD) and Event Detection (ED) in Section 4.5. Both are standalone and separate task but use the same solution and hence shares the same solution architecture;
3. Subtask 2: argument role prediction (ARP) in Section 4.6 using Graph Convolutional Network (GCN) with Contextual Sub-tree;
4. Subtask 3: Event properties (Polarity, Modality, and Intensity) classification in Section 4.7.

## 4.1 Definitions

Before diving into the technical details of the proposed solution, this section is dedicated to laying out the terminologies and task descriptions to aid the readability of this chapter.

### 4.1.1 Terminologies:

1. An **entity mention** is an explicit mention of an entity in a text that has an entity type.

2. An **event trigger** is the main word(s) that most clearly expresses the occurrence of an event, usually a word or a multi-word phrase. It can come in the form of verb, noun, adjective or adverb).
3. An **event argument** is an argument filler that plays a certain role in an event.
4. **Polarity** also known as negation in (Morante & Blanco, 2012) denotes whether an event actually happened or was negated (did not happen), not to be confused with the polarity in sentiment analysis. Its value can be POSITIVE or NEGATIVE.
5. **Modality** also known as hedge in (Farkas et al., 2010), denotes whether an event actually happened or will happen in the future. Its value can be ASSERTED or OTHER.
6. **Intensity** denotes if an event further intensified and lessen. Its value can be INTENSIFIED, NEUTRAL and EASED.

The definition of Event Polarity and Modality in *CrudeOilNews* corpus are aligned with *ACE2005* Version 5.4.3. In contrast, Event Intensity is a newly defined property specially crafted for how events are reported in commodity news.

Here a solution for sentence-level Event Extraction is proposed; it is made up of a few sub-tasks. This work uses the same event extraction naming convention in (T. M. Nguyen & Nguyen, 2019). These tasks are described below:

#### 4.1.2 Tasks

1. Entity Mention Detection (**EMD**): a task to detect entity mentions (named or nominal) and assign each token an entity type or NONE for tokens that is not an entity mention.
2. Event Extraction :
  - (a) Event Detection (**ED**): similar to EMD, it is a task to detect event trigger word(s) and assign it to an event type or NONE for tokens that is not an event trigger.
  - (b) Argument Role Prediction (**ARP**): a task aims to assign an argument role label or NONE to a candidate entity mention.
3. Event Properties Classification: a task to classify each event in terms of its Polarity, Modality and Intensity classes.

## 4.2 Related Work

### 4.2.1 Event Extraction in Finance and Economics Domain

An in-depth analysis of financial / economic event extraction is found in Section 2.1.2. Here, a concise summary is presented to provide background information for a better appreciation of the proposed solution architecture in this Chapter.

1. Rule-based/semantic-based/ontology-based approaches: the earliest work proposed solutions that are rule-based, semantic-based or domain-specific ontology knowledge-based or a combination of these components;
2. Supervised Learning - (Lefever & Hoste, 2016; Jacobs et al., 2018; Jacobs & Hoste, 2020);
3. Semi / Distantly / Weakly Supervised: (Dor et al., 2019) use Wikipedia in financial/economic event extraction via weak supervision;
4. Usage of generic text processing methods such as semantic frame parsing (Xie et al., 2013) and as event tuple extraction via OpenIE (Ding et al., 2014, 2015) or a variation of OpenIE (Saha et al., 2017).

So far, all existing financial/economic event extractions are for extracting company events such as mergers & acquisitions, dividend payout, quarterly results, etc. Although company financial events and commodity news fall under the same domain, and both may involve numerical data as event arguments, existing methods for company financial event extractions are insufficient to cater to the unique characteristics shown in *CrudeOilNews* corpus. For example, the solution in (Saha et al., 2017) caters to extracting only one numerical argument for each Open IE tuple. In comparison, even though the solution proposed in (H. Yang et al., 2018) extracts numerical information from company-related information with the help of a financial event knowledge database consisting of labeled event trigger and argument samples, however, the knowledge dataset has limited coverage of only nine company-related pre-determined events. Furthermore, it is a solution for Chinese text and focuses on document-level event extraction. Hence it is simply not feasible to apply any existing event extraction methods as-is without first modifying them to be suitable for extracting events in the *CrudeOilNews* corpus.

### 4.2.2 Graph Convolutional Networks

The usage of Graph Convolutional Networks (GCN) coupled with syntactic information from dependency parse tree has been used for event extraction in (T. H. Nguyen & Grishman, 2018) and in (X. Liu et al., 2018). The combination has also proven to be effective in relation extraction in (Y. Zhang et al., 2018). In (T. H. Nguyen & Grishman, 2018), Convolution Neural Network coupled with dependency parse tree to perform event detection produced SOTA results at the point of publication. The authors utilized syntactic representation from dependency parse tree to link words directly to their informative context in event extraction in sentences.

In (T. H. Nguyen & Grishman, 2018), the authors proposed using GCN over syntactic dependency graphs of sentences to produce non-consecutive  $k$ -grams as an effective mechanism to link words to their informative content directly for event detection. On the other hand, authors in (X. Liu et al., 2018) used attention-based GCN to model graph information to extract multiple event triggers and arguments jointly. Their proposed solution, Joint Multiple Events Extraction (JMEE) framework, focuses on modeling the association between events to enhance the accuracy of event extraction. Both of these solutions use the shortest dependency path.

Apart from event extraction, GCN has been used successfully for relation extraction in (Y. Zhang et al., 2018). Instead of obtaining tokens strictly from the shortest dependency path, authors in (Y. Zhang et al., 2018) made modifications to produce pruned a sub-dependency tree to include off-path information, such as negation cue words. Among the related work listed here, the one that is closest in terms of the task (event extraction) and scope (sentence level) is JMEE by (X. Liu et al., 2018). Inspired by the effectiveness, this work looks at using GCN and contextual sub-tree to overcome the challenge of classifying event arguments of homogenous type.

## 4.3 Preliminary: Domain Adaptive Pre-training - ComBERT

It is shown in (Brown et al., 2020) that extremely large language models can perform competitively on downstream tasks with far less task-specific data than would be required by smaller



models. Pre-trained language models such as BERT (Devlin et al., 2019) can be rapidly fine-tuned on downstream GLUE tasks to produce state-of-the-art results. Apart from GLUE tasks, BERT has also been fine-tuned on event extraction using the ACE2005 dataset in (S. Yang et al., 2019). Instead of using the ‘vanilla-version’ of BERT, this work uses domain adaptive pre-training to further pre-train BERT on a large collection of commodity news. Further pre-training in a specific domain is essential in creating a contextualized language model for tasks that involve a domain-specific corpus. This is evident in SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2020).

Apart from the famous ‘bank’ example, where it could mean (1) a financial institution or (2) terrain that is part of the river, there are some commodity-specific polysemous words in the *CrudeOilNews* Corpus which can be better represented with further pre-training with in-domain data, for example:

- stocks: (1) Inventory and (2) Shares
- tank: (1) storage vessel (noun), (2) market / price drop (verb)

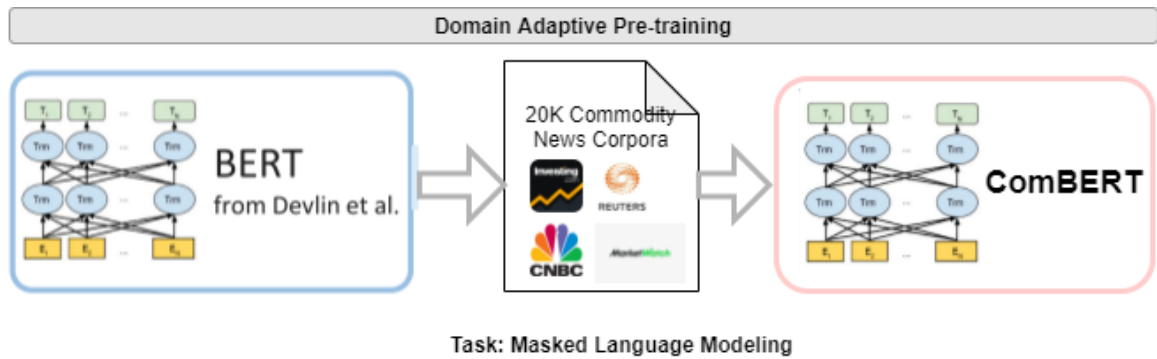


Figure 4.3: Domain Adaptive Pre-training: Using BERT as baseline, the language model is further pre-trained BERT on a commodity news corpus, adapting the model to the finance and economic news domain.

The resulting model is referred to here as **ComBERT**. The process of domain adaptive pre-training is shown in Figure 4.3. Besides being able to produce even more contextualized word embedding, ComBERT is also ‘domain adapted’ to commodity news, which contains opinion, expert analysis and financial forecasts apart from factual information<sup>2</sup>.

The collection of commodity news from which ComBERT is fine-tuned on is made up of about 20k news articles extracted from <https://www.investing.com/news/commodities-news>, with publishing dates ranging from 2013 to 2019. ComBERT was initialized with

<sup>2</sup>BERT was trained on ‘factual’ text like English Wikipedia and Brown Corpus

bert-base-cased and with the same model settings and hyperparameters as BERT. ComBERT is fine-tuned on masked language modeling with 15% of tokens masked, of which 80% are replaced with the token “MASK”, 10% with a random token from the corpus, and 10% with the original token. The masked language modeling task is trained on Cross Entropy Loss. ‘Cased’ vocabulary (bert-base-cased) is used instead of ‘uncased’ because the event extraction task involves extracting event arguments that are made up of named entities and nominal entities. Named entities, such as are countries, organizations, and specific commodities-related terms such as WTI, ICE, NYMEX, Brent and etc, are better represented by the bert-base-case vocabulary. A case-sensitive model yielded slightly better performance for the downstream event extraction task.

Subsequent sections show how ComBERT is used in all event extraction subtasks (EMD and ED in Section 4.5, ARP in Section 4.6 and event properties classification in Section 4.7) to produce superior results.

## 4.4 Data Pre-processing

The *CrudeOilNews* corpus is made up of annotation files in *.json* format. Each json files were pre-processed using Stanford CoreNLP toolkit (see Section 3.3.2.1 for more details) and contain the following information apart from annotations:

1. POS tags for each word token;
2. NER tags for each word token;
3. dependency annotation for each word token generated by dependency parsing.

For input to the model, this work adopts the “multichannel” strategy, which concatenates three components listed above along with ComBERT word embeddings. The input is defined as follows: Let  $W = w_1; w_2; \dots; w_n$  be a sentence of length  $n$  where  $w_i$  is the  $i$ -th token:

1. The word embedding vector of  $w_i$ : this is the feature representation from a word embedding. Various word embeddings were experimented including GloVe and contextulized word embedding such as BERT, RoBERTa and ComBERT. Details and experimental results are found in Section 4.6.3.2.
2. The POS-tagging label embedding vector of  $w_i$ : This is generated by looking up the POS tagging label embedding.

3. The entity type label embedding vector of  $w_i$ : Similar to the POS-tagging label embedding vector of  $w_i$ , entity mentions in a sentence were annotated using BIO annotation schema and the entity type labels were transformed to real-valued vectors by looking up the embedding table.

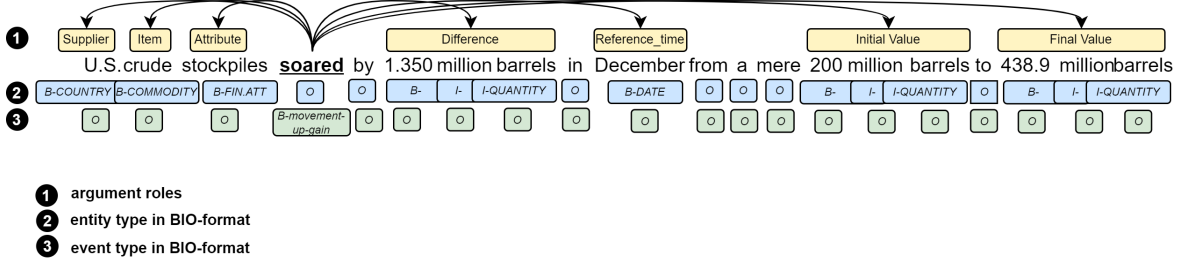


Figure 4.4: Example in Figure 4.1 is shown here with the following annotations: (1) argument role each entity plays (in light yellow), (2) entity mention tags in BIO format (in blue below), and (3) event trigger in BIO format (in light green). Event trigger word is underlined and in bold; arches link arguments to their respective trigger word.

Figure 4.4 shows the example sentence (original sentence in Figure 4.1) with its corresponding (1) entity mention tags, (2) event trigger tags and (3) argument roles each entity plays.

## 4.5 Subtask 1: EMD and ED

Entity Mention Detection (EMD) and Event Detection (ED) are both formulated as a multi-class token classification (also commonly known as *sequence labeling* in NER and POS tagging) problem. The input is tokens in BIO-tagging (Begin, Inside, Outside) format that is able to cater to both single-word and multi-word event triggers. A token is labeled as *B-EVENT* if the token is the “beginning” of trigger for *EVENT*, or *I-EVENT* if the token is part of the event span or *O* otherwise. There are 18 types<sup>3</sup> where the classifier will predict one of 37 classes (B-tag, I-tag for each event plus an additional tag “NONE”) for each token. The outcome is  $T = t_1; t_2; \dots; t_j$ , the list of triggers identified in the sentence where  $j$  is the number of triggers identified (note: multi-word trigger is counted as one).

As for EMD, the same approach is used as it is also multi-class token classification with input in BIO-tagging format. There are 21 entity types<sup>4</sup> and the classifier will predict one of 43 classes (B-tag, I-tag for each event plus an additional tag “NONE”) for each token. The

<sup>3</sup>see Section 3.2.1.2 for complete list of event types

<sup>4</sup>see Section 3.2.1.1 for complete list of entity types

outcome is  $E = e_1; e_2; \dots; e_k$ , the list of entity mentions identified in the sentence where  $k$  is the number of the entity mentions.

#### 4.5.1 Experiments

Both tasks are standalone tasks and are trained separately. They are discussed here together because both tasks are identical, which is essentially a multi-class token classification task.

The following neural network models have been proven to be effective in NER tagging and are experimented for the EMD and ED tasks:

1. BiLSTM-CRF (Lample et al., 2016) with BERT embeddings<sup>5</sup>
2. Flair embeddings (Akbi et al., 2018)<sup>6</sup>
3. BERTForTokenClassification head on BERT architecture (Devlin et al., 2019)

The data is split into 70% for training and 30% for testing. Each model is trained with a batch size of 16 and with the Cross-entropy loss function. The Adam optimizer is used.

#### 4.5.2 Results and Analysis of EMD /ED

As shown in Table 4.1, Entity Mention Classification (EMD) achieve rather high F1 scores.

Table 4.1: EMD and ED results across various methods.

Methods	EMD Task			ED Task		
	Precision	Recall	F1	Precision	Recall	F1
BiLSTM-CRF	0.782	0.691	0.734	0.761	0.692	0.725
Flair embeddings	0.825	0.832	0.828	0.792	0.808	0.800
BERTForTokenClassification + BERT	0.822	0.812	0.817	0.809	0.831	0.820
BERTForTokenClassification + ComBERT	<b>0.903</b>	<b>0.912</b>	<b>0.907</b>	<b>0.915</b>	<b>0.899</b>	<b>0.907</b>

As described in Section 3.4.2, this dataset is considered simple and straightforward where it does not contain multi-type event triggers. In other words, each trigger is associated with only one event type. The model with BERTForTokenClassification coupled with ComBERT contextualized word embeddings produced the best result. In the mean time, BERTForTokenClassification with BERT embeddings produced almost similar results as the model that uses Flair. Flair embeddings have shown to produce competitive results to the BERT model especially on tasks related to syntax and morphology.

<sup>5</sup>codes here: <https://github.com/hertz-pj/BERT-BiLSTM-CRF-NER-pytorch>.

<sup>6</sup>codes here: <https://github.com/flairNLP/flair>.

## 4.6 Subtask 2: ARP

Argument Role Prediction (ARP) is the task of classifying the argument role each entity plays in an event. With the list of predicted candidate triggers  $T$  from ED, and entity mentions  $E$  from EMD, the next task is to predict the argument roles (ARP) each entity mention  $e$  plays in its respective event. If the entity does not belong to the event, then the argument role is “NONE”. On the other hand, if the entity is linked to the trigger, then the classifier will predict the argument role the entity plays. This work proposes a solution using Graph Convolutional Networks (GCN) with contextual sub-tree for effective argument role prediction.

### 4.6.1 Contextual Sub-tree

A syntactic dependency parse tree is a special form of graph; it represents sentences as directed trees with head-modifier dependency arcs between related words. The combination of Graph Convolution Network (GCN) and dependency parse tree has been shown to be useful in Event Detection (T. H. Nguyen & Grishman, 2018; X. Liu et al., 2018) and also in Relation Extraction (Y. Zhang et al., 2018).

The example sentence in Figure 4.1 and its corresponding parse tree in Figure 4.5 are used throughout this chapter to aid the explanation of the proposed solution involving the Graph Convolution Network (GCN).

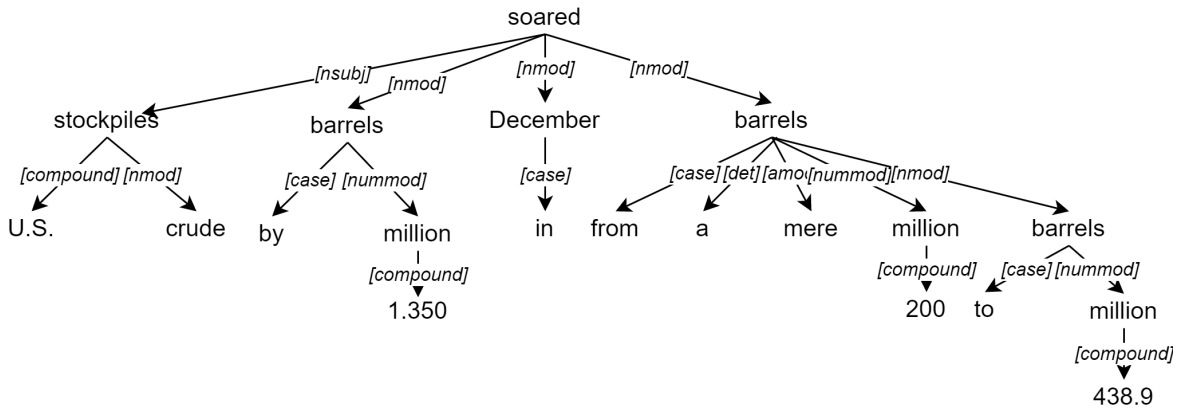


Figure 4.5: A segment of the dependency parse tree for the example sentence in Figure 4.1

Rather than using the full dependency parse tree as input into a GCN, this work proposes to use a uniquely pruned dependency tree that is made up of the shortest path between two

nodes (in this case - the **trigger candidate** and **entity mention**) and additional **off-path nodes**. The usage of a pruned tree in the proposed solution is inspired by (T. H. Nguyen & Grishman, 2018) where authors have used a pruned dependency tree with the shortest path to maximally remove irrelevant information without omitting crucial contextual information. The *path-centric pruning* technique described in (Y. Zhang et al., 2018) was originally designed for Relation Extraction. It is tweaked here for the ARP task. The path-centric pruning aims to remove irrelevant information from the parse tree while maximally keeping relevant content, which the authors called “off-path information”. This pruned sub-tree is subsequently referred to as **contextual sub-tree**. Figure 4.6 shows the same dependency parse tree with one of the contextual sub-tree highlighted.

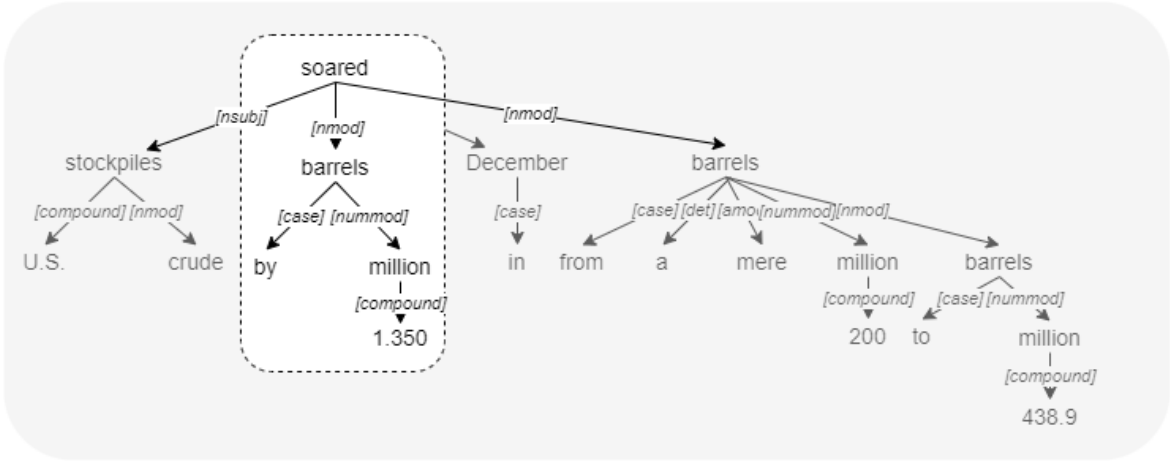


Figure 4.6: The same dependency parse tree as shown in Figure 4.5 with a sub-tree highlighted

The reason for using a contextual sub-tree instead of the entire dependency parse tree is so that only convolution operations from the GCN can be performed on the most relevant words and avoid the modeling of unrelated words. The size of a contextual sub-tree is between the size of a full parse tree and of a pruned-shortest path tree. The dependency tree is pruned to obtain the sub-tree rooted at the Least Common Ancestor (LCA) between the trigger candidate and the entity mention candidate while also containing off-path nodes. These off-path nodes provide additional and crucial contexts that enable better results in argument role classification. Off-path information is made up of tokens that are up to distance  $DIST$  away from the dependency path.

Algorithm 1 shows the steps of how to build the contextual sub-tree. As shown in (Y. Zhang et al., 2018),  $DIST = 1$  achieves the best balance between including contextual information and keeping irrelevant ones out of the resulting sub-tree as much as possible. In the ARP

**Algorithm 1:** Build sub-parse tree from dependency head indexes**Result:** sub-tree structure

convert head indexes to tree object;

**if**  $DIST < 0$  **then**

| build the whole tree;

**else**

| find all ancestor nodes of trigger;

| find all ancestor nodes of entity;

| find lowest common ancestor;

| generate PathNodes (common nodes between trigger and entity);

| insert more nodes based on  $DIST$  away from PathNodes;**end**

task, candidates are classified into one of the 19 argument roles. Figure 4.7 (Left) shows the sub-tree with the LCA path between event trigger and argument, while Figure 4.7 (Right) shows a slightly ‘bigger’ contextual sub-tree produced by Algorithm 1 with  $DIST = 1$ . The off-path information included in this sub-tree is the word ‘by’, which provides additional and crucial contexts to help classify the role ‘1.350 million barrels’ plays.

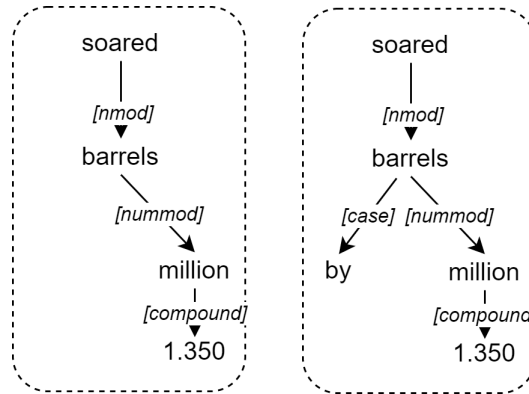


Figure 4.7: Left: Sub-tree with shortest path, Right: Contextual sub-tree with off-path information

Table 4.2 below shows more examples of sub-dependency parse tree, in the form of a list of words, between **trigger word** and **entity**. In the examples below, trigger word(s) is bolded, entity mention in square brackets [ ], and additional off-path information underlined>.

Table 4.2: Event Arguments for the event **soared**.

Words in sub-dependency parse tree	Entity Type	Argument role
(a) [stockpiles] <b>soared</b>	FINANCIAL ATTRIBUTE	ATTRIBUTE
(b) <b>soared</b> <u>by</u> [1.350 million barrels]	QUANTITY	DIFFERENCE
(c) <b>soared</b> <u>in</u> [December]	DATE	REFERENCE TIME
(d) <b>soared</b> <u>from</u> a mere [200 million barrels]	QUANTITY	INITIAL VALUE
(e) <b>soared</b> <u>to</u> [438.9 million barrels]	QUANTITY	FINAL VALUE

For examples (b), (d), and (e), all three of their entities are of “Quantity” type, which can make classifying their argument role a challenge given they are all non-differentiable in terms of the entity type. By using a sub-dependency parse tree, not only are words in the shortest is used, but other off-path information such as the conjunctions by, from and to are included. These off-path information are crucial in providing more context and clues for the accurate classification of argument roles.

The ARP task is conceived as sentence-level multilabel classification task. The overall architecture is shown in Figure 4.8.

#### 4.6.2 Graph Convolutional Networks over Contextual Sub-tree

The ARP task is set up as a sequence classification task. Candidate arguments are selected from the pool of entity mentions within the sentence. Each candidate argument will be paired with a candidate trigger for argument role classification. The classifier will classify each trigger-entity pair into 20 classes (19 argument roles and ‘NONE’ for entities with no links to the candidate trigger).

The candidate entity mentions  $E$  and candidate event triggers  $T$  produced by EMD and ED models respectively in Section 4.5 is used as input in the ARP task.  $E = e_1; e_2; \dots; e_k$  and  $T = t_1; t_2; \dots; t_j$  where  $k$  is the number of entity mentions while  $j$  is the number of triggers in a sentence. Each candidate trigger is paired with an entity, resulting in  $j \times k$  number of pairs. For the pair  $t_x e_y$ , the task is to classify the argument roles entity  $e_y$  plays in event  $t_x$ .

A sentence’s syntactic parse tree can be seen as a directed graph. Let  $G = fV; Eg$  be the dependency parse tree for the sentence  $w$  with  $V$  and  $E$  as the sets of nodes and edges of  $G$  respectively.  $V$  contains  $n$  nodes corresponding to the  $n$  tokens  $w_1; w_2; \dots; w_n$  in  $w$ . Each edge  $(v_i; v_j) \in E$  is directed from the head word  $w_i$  to the dependent word  $w_j$  with the Universal Dependency (UD) relation tags. Given that a sentence’s dependency parse tree with  $n$  nodes, each tree is converted into its corresponding  $n \times n$  adjacency matrix  $A$  with the following modifications:

1. Treating the dependency graph as undirected, i.e  $\delta(i; j); A_{i;j} = A_{j;i}$ , where  $A_{i;j} = A_{j;i} = 1$  if there is a dependency edge between tokens  $i$  and  $j$ ;
2. Adding self-loops to the each node in the graph, following (Kipf & Welling, 2017):  
 $\hat{A} = A + I$  with  $I$  being the  $n \times n$  identity matrix



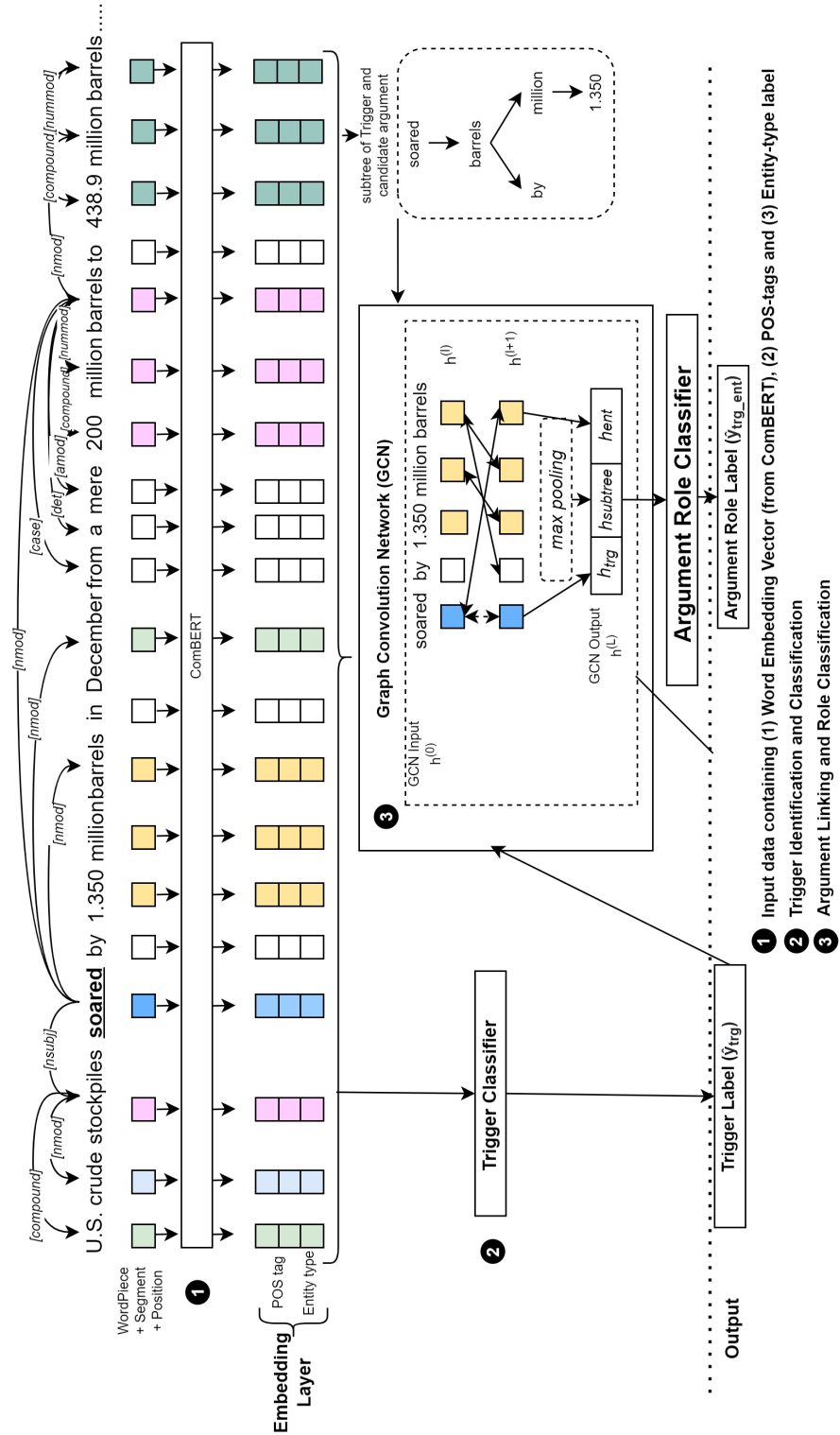


Figure 4.8: Proposed solution architecture for Argument Role Prediction (ARP) task: Graph Convolution Network (GCN) + Contextual Subtree

Stacking a GCN layer  $L$  times gives us a  $L$ -layer GCN where  $L$  is a hyperparameter of the model. During graph convolution at each layer  $l$ , each node gathers and summarizes information from its connected nodes ( $\mathcal{A}_{i,j} = 1$ ) in the graph.  $h^{(0)}$  is set as the input word vectors for an  $L$ -level GCN network and  $h^{(L)}$  as the output word representations. The graph convolution operation of a single node, node  $i$  at level  $l$  of the GCN is as follows:

$$h_i^{(l)} = \left( \sum_{j=1}^{\mathcal{N}} \mathcal{A}_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)} \right) \quad (4.1)$$

where  $h_i^{(l-1)}$  is the input vector,  $h_i^{(l)}$  denotes the collective hidden representations,  $W^{(l)}$  is the weight matrix,  $b^{(l)}$  is a bias term,  $\sigma$  is the sigmoid activation function and  $d_i = \sum_{j=1}^n \mathcal{A}_{ij}$ ,  $n$  is the number of arches in the resulting graph.

#### 4.6.2.1 ARP with GCN

This section describes the operations shown as “3” in Figure 4.8.

**Encoding Trigger-Entity Pair** Given a trigger-entity pair  $t_x e_y$ , the dependency tree is pruned to obtain the contextual sub-tree between trigger  $t_x$  and entity  $e_y$  based on Algorithm 1. The nodes of this sub-tree form the input word vectors  $h^{(0)}$  to the  $L$ -layer GCN network.

The subtree representation after  $L$  times of graph convolution is obtained as follows:

$$h_{subtree} = f(h^{(L)}) = f(GCN(h^{(0)})) \quad (4.2)$$

where  $h^{(L)}$  is the output word representations produced by the  $L$ -layer GCN network and  $f$  is a max-pooling function that maps the input to the subtree vector,  $h_{subtree}$ . Besides the subtree representation, a representation  $h_{trg}$  for trigger and  $h_{ent}$  for entity is also obtained:

$$h_{trg} = f(h_t^{(L)}); h_{ent} = f(h_e^{(L)}) \quad (4.3)$$

Besides max-pooling, average-pooling and sum-pooling are also experimented to obtain the final vector for all three vectors (subtree, trigger and entity). All three vectors are then concatenated into a vector which is then propagated through a fully-connected layer to

classify the argument role:

$$\mathbf{y}_{trg\_ent} = g(W_a[h_{subtree}; h_{trg}; h_{ent}] + b_a) \quad (4.4)$$

where  $g$  is the *softmax* operation to obtain a probability distribution over argument roles.  $\mathbf{y}_{trg\_ent}$  is the final output of the role the entity *ent* plays in the event triggered by the trigger candidate *trg*.

### 4.6.3 Experiments

**Parameter settings.** The data is split into 70% for training and 30% for testing. For all the experiments, the word embedding is of size 768 dimensions (same as bert-base-cased) while 50 dimensions for the other two embeddings - POS-tag embedding and entity-type embedding. A two-layer GCN ( $L = 2$ ) with a batch size of 4 is used for the GCN module. The model is trained using the Cross-entropy loss function and Adam optimizer.

**Models Settings.** The architecture and setup for models listed in Table 4.3 are as follows:

1. **Model A** - The embedding of trigger and candidate argument (from ComBERT) are concatenated and fed into a Bi-LSTM, which is then fed into a classifier with one fully connected (FC) layer.
2. **Model B** - Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation (JMEE) (X. Liu et al., 2018)<sup>7</sup>.
3. **Model C** - GCN with Full Tree uses the full dependency tree,  $h_{fulltree}$ . The same convolution operations are done on  $h_{fulltree}$  in the place of  $h_{subtree}$ .
4. **Model D** - GCN with LCA sub-tree with shortest dependency path between trigger candidate and entity candidate.
5. **Model E** (Proposed solution) - GCN with contextual sub-tree, this setup is described in the Proposed Solution section.

---

<sup>7</sup>This was developed for the ACE2005 dataset.

#### 4.6.3.1 Results and Analysis

From the results shown in Table 4.3, it can be concluded that syntactic representation (**Model C, D, E**) of a sentence yields better event extraction results. The results of Model B and C are not as good as those using sub-tree because the full dependency tree contains unnecessary and noisy information that is not helpful in argument role classification. As for (X. Liu et al., 2018) (Model B), it did not produce the best results because it was designed for capturing the association between multiple events within a sentence via the attention mechanism. The events in the commodity news dataset do not exhibit the same strong association as the events in the ACE2005 dataset. Model D uses the LCA sub-tree that has only the “bare minimum” information. In contrast, Model E contains additional crucial context information that has proved to be useful in argument role classification.

Table 4.3: Comparing ARP results across various methods.

Method	ARP Task		
	Precision	Recall	F1
<b>A</b>	0.701	0.559	0.622
<b>B</b>	0.751	0.801	0.775
<b>C</b>	0.740	0.722	0.731
<b>D</b>	0.812	0.700	0.752
<b>E</b>	<b>0.790</b>	<b>0.814</b>	<b>0.802</b>

Table 4.4 presents the breakdown of Argument Classification by Argument Types to fully provide evidence of the effectiveness of the proposed solution on the corpus, which exhibits the characteristic of arguments homogeneity. It is shown clearly that arguments of the same entity type, for example, FINAL\_VALUE, INITIAL\_VALUE and DIFFERENCE can be better differentiated and classified using a contextual sub-tree that contains the shortest path between an event trigger and its event argument as well as crucial off-path information. Symbols ( / , } , • ) in Table 4.4 indicate the grouping of arguments by entity type.

As for the characteristic of having multiple events in a sentence, the proposed solution can detect and classify the events and link arguments to their rightful event, as shown in both Table 4.3 and Table 4.4.

#### 4.6.3.2 Comparing Word Embedding and various Pre-trained Language Models

In order to test out the most suitable word embeddings for ARP, the following word embeddings are used:

Table 4.4: F1-scores for each argument type.

Argument Roles	Entity Type	ARP Task				
		A	B	C	D	E
NONE	-	0.84	0.84	0.90	0.91	<b>0.92</b>
Attribute	FINANCIAL ATTRIBUTE	0.40	0.65	0.79	0.75	<b>0.83</b>
Item	ECONOMIC ITEM	0.64	0.85	<b>0.88</b>	0.85	<b>0.88</b>
Final_value /	MONEY / PRODUCTION UNIT / PRICE UNIT / PERCENTAGE / MONEY / QUANTITY	0.43	0.39	0.71	0.75	<b>0.79</b>
Initial_value /		0.56	0.56	0.73	0.69	<b>0.77</b>
Difference /		0.58	0.69	0.74	<b>0.79</b>	<b>0.79</b>
Reference_point }	DATE	0.54	0.69	<b>0.79</b>	0.71	<b>0.79</b>
Initial_reference_point }		0.40	0.63	0.63	0.60	<b>0.66</b>
Contract_date }		0.52	0.54	0.70	0.66	<b>0.80</b>
Duration	DURATION	0.55	0.55	0.75	0.82	<b>0.84</b>
Type	LOCATION	0.52	0.59	0.70	0.68	<b>0.76</b>
Imposer •	Country / State or province	0.71	0.69	<b>0.81</b>	0.79	<b>0.81</b>
Imposee •	Country / State or province	0.50	0.49	0.60	<b>0.64</b>	<b>0.64</b>
Place •	Country / State or province	0.58	0.69	<b>0.74</b>	0.60	<b>0.74</b>
Supplier_consumer •	Country / State or province / Nationality / Group	0.49	0.71	0.73	0.73	<b>0.79</b>
Impacted_countries •	Country	0.42	0.69	0.72	0.70	<b>0.76</b>
Participating_countries •	Country	0.65	0.75	0.78	0.83	<b>0.89</b>
Forecaster	ORGANIZATION / GROUP	0.62	0.75	0.78	0.80	<b>0.82</b>
Forecast	FORECAST TARGET	0.61	0.61	0.83	0.67	<b>0.81</b>
Situation	PHENOMENON / OTHER ACTIVITIES	0.52	0.59	<b>0.68</b>	0.67	0.56

1. GloVe(Pennington et al., 2014)
2. BERT(Devlin et al., 2019)
3. RoBERTa (Y. Liu et al., 2019)
4. ComBERT

**Model E** in Table 4.5 was further experimented using GloVe (Pennington et al., 2014) and other pre-trained language models namely BERT (Devlin et al., 2019) and RoBERTa (Y. Liu et al., 2019). These were compared against ComBERT.

Table 4.5: Comparing Word Embedding and Pre-trained Language Models for Model E.

Method	ARP Task		
	Precision	Recall	F1
GloVe	0.650	0.691	0.670
BERT	0.750	<b>0.817</b>	0.782
RoBERTa	0.761	0.769	0.765
<b>ComBERT</b>	<b>0.790</b>	0.814	<b>0.802</b>

The results in Table 4.5 show that ComBERT produced the best F1 result, outperforming GloVe by 1%, and RoBERTa by 2% in terms of argument roles prediction. As shown in Table 4.5, word embedding using GloVe produced the worst result, while the contextualized word embeddings produced by pre-trained Language Models like BERT and RoBERTa produced slightly better results. This is not unsurprising given that pretrained language

models like BERT and RoBERTa have shown to produce SOTA results on several benchmark NLP Tasks. The results prove that a contextualized token representation helps boost the performance of event extraction. ComBERT is used in all models listed in Table 4.4.

## 4.7 Subtask 3: Event Properties Classification

One of the challenges faced when training a model for event properties classification in sentences containing multiple events is to classify each event accurately. According to the list of characteristics of the *CrudeOilNews* corpus in Section 3.4.2, on average, the sentences in the *CrudeOilNews* corpus has about 2 to 3 events. Accurate classification of event properties requires identifying cue words at the individual event scope level and not at the sentence level. Classifying at the sentence level will produce incorrect results. Therefore to accurately classify event properties of several events within a sentence, the scope needs to be narrowed down to use only the event scope and not the entire sentence.

### 4.7.1 Model Architecture

Experiments with different input ‘scope’ are carried out to find the one that produces the best results, the following techniques are experimented:

1. Fixed window of words surrounding event trigger word(s)  $(x_{i-r} \dots x_{i-1} \ x_i \ x_{i+1} \dots x_{i+r})$  where  $\dots$  is the concatenation operation,  $r$  represents the length from trigger word  $x_i$ . The sequential word representation is fed into an MLP to generate a vector and then through a *softmax* activation function.
2. GCN over Contextual Sub-tree described in Section 4.6.2
3. SelfAttentiveSpanExtractor (Gardner et al., 2018) (part of the AllenNLP library) to weightedly combine the representations of multiple tokens and create a single vector for the original event span. The span vectors are fed into a two-layer feed-forward network with a *softmax* activation function.

#### 4.7.2 Measurement for dataset with class imbalance

**F1-Score** F1-score reported here is the macro-average F1-score averaged across  $k$  experiments. F1-score for each fold (iteration) is computed; then, the average F1 score from these individual F1 scores is calculated.

$$F1_{avg} = \frac{1}{k} \sum_{i=1}^k F1_i \quad (4.5)$$

**MCC** In (Xie et al., 2013), apart from the familiar F1-measurement, the authors used an additional evaluation metric known as the Matthew Correlation Coefficient (MCC) to avoid bias due to the skewness of data. It takes into account true and false positives and negatives and is generally regarded as a balanced measure, which can be used even if the classes are of very different sizes. MCC is a single summary value that incorporates all four cells of a 2 × 2 confusion matrix<sup>8</sup>.

The equation for Binary Classification:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4.6)$$

and for Multi-class Classification:

$$MCC = \frac{c - \frac{s^2}{\sum_k p_k}}{\sqrt{(s^2 - \frac{s^2}{\sum_k p_k})(s^2 - \frac{s^2}{\sum_k t_k})}} \quad (4.7)$$

with the following intermediate variables:

- $t_k = \sum_i C_{ik}$  is the number of times class  $k$  truly occurred,
- $p_k = \sum_i C_{ki}$  is the number of times class  $k$  was predicted,
- $c = \sum_k C_{kk}$  is the total number of samples correctly predicted,
- $s = \sum_i \sum_j C_{ij}$  is the total number of samples,
- $TP$  is True Positive,  $FP$  is False Positive,  $TN$  is True Negative and  $FN$  is False Negative.

---

<sup>8</sup>For more information on Matthews Correlation Coefficient (MCC), visit [https://scikit-learn.org/stable/modules/model\\_evaluation.html](https://scikit-learn.org/stable/modules/model_evaluation.html)

### 4.7.3 Experiments

#### 4.7.3.1 Train-Test Split

The main challenge in Event Property classification with the *CrudeOilNews* corpus is class imbalance. To address this challenge, **stratified** k-fold cross-validation is used instead of random sampling. This is done to ensure that both the training and testing set in each cross-validation maintain the same class distribution (label ratio) of the original dataset as shown in Figure 3.9.

#### 4.7.3.2 Results and Analysis

All three sub-tasks (Polarity, Modality, and Intensity classification) are standalone and independent tasks where the outcome of one does not influence the outcome of others. Therefore all three classification model were trained independently of each other. Event Modality / Polarity classification is a binary classification task, where the labels for Modality are: ASSERTED and OTHER, and for Polarity are: POSITIVE and NEGATIVE. Both classification tasks are trained on Binary Cross-entropy Loss. As for Event Intensity, it is a multi-class classification task, the labels are: NEUTRAL, EASED, and INTENSIFIED. It is trained on multi-class Cross-entropy Loss. Experiments are conducted to determine the most suitable text span for the classification of Event Property Classification by the text processing methods listed in Section 4.7.1.

Table 4.6: Experiment results of different Input Text Span.

Text Span Generation Methods	Polarity		Modality		Intensity	
	F1	MCC	F1	MCC	F1	MCC
4-grams <sup>9</sup> xed window centered around event trigger	0.485	0.285	0.599	0.305	0.601	0.320
GCN with Contextual Sub-tree (Section 4.6.2)	0.659	0.305	0.683	0.298	0.699	0.398
SelfAttentiveSpanExtractor (Jiang & de Marne e, 2021)	<b>0.729</b>	<b>0.478</b>	<b>0.795</b>	<b>0.498</b>	<b>0.721</b>	<b>0.595</b>

Based on the results in Table 4.6, it can be concluded that the best text span is the ones generated by SelfAttentiveSpanExtractor, then followed by using a dependency parse tree. The dependency parse tree utilizes the syntactic structures of the input sentence and works well for identifying modifiers and negations such as WILL and NOT that is linked to the event trigger’s sub-parse tree. However, it does not work for cases where event trigger is

<sup>9</sup>Experiments using  $k = 2, 3, 4, 5$  is ran. 4-gram produced the best results.



not a verb that forms its sub-tree. This is illustrated with an example below and a portion of its dependency tree in Figure 4.9.

- (22) The Trump administration *will not* consider reimposing **sanctions** on the OPEC member nation.

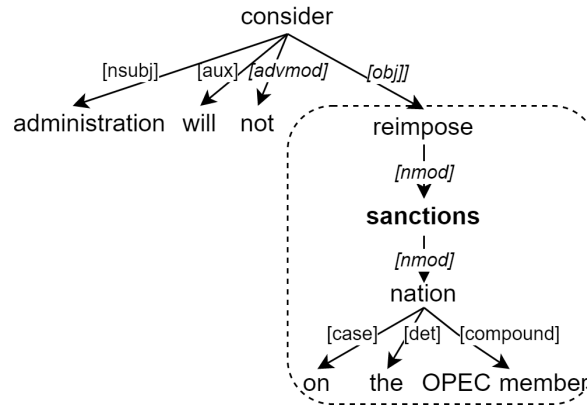


Figure 4.9: An example of a pruned dependency parse tree (enclosed in dotted lines) that did not generate a good text input for modality and polarity classification. Based on the event trigger word - **sanctions**, this pruned dependency parse tree does not contain the modality and polarity cue words: *will* and *not*.

The worst performing text span is the  $k$ -gram fixed window approach because it does not capture words far away from the trigger (i.e.: words are located outside of the  $k$ -grams window). Based on the example above, the word cue word ‘will’ is not extracted as part of the text span for **sanctions** because it is located outside the fixed window centered around the event trigger. Apart from this,  $k$ -grams window will also not work for compound sentences such as the example below:

- (23) Such accommodative policies *tend to* **weaken** the dollar by design and send commodities prices **rising**.

In the  $k$ -grams approach, the cue words *tend to* will not be picked up for the event **rising** leading to an error in Modality classification.

The F1-score for both Polarity and Modality classification is high in contrast to much lower MCC scores. This situation is caused by class imbalance in these two classifications. Error analysis on these two tasks showed the errors are primarily False Positives. The models tend to classify everything to the majority class (POSITIVE for Polarity classification and ASSERTED for Modality classification), which result in low precision. Given that the MCC

score takes into account all four values in the confusion matrix, a low MCC score shows that the models are not good at classifying the minority classes. Chapter 5 attempts to address the issue of class imbalance.

The most challenging among the three tasks is Event Intensity classification. Some of the cue words for determining the event intensity (NEUTRAL, EASED, INTENSIFIED) are themselves trigger words.

(24) **Oversupply** could rise next year when Iraq starts to export more oil.

In example (24), the correct interpretation should be: The event **oversupply** might be further INTENSIFIED (cue word: *rise*, but this word is also a trigger word for the event - MOVEMENT-UP-GAIN).

## 4.8 Summary and Discussion

As part of Research Object #2, a new end-to-end event extraction model tailored for the purpose of extracting crude oil-related events in the *CrudeOilNews* corpus is proposed. The event extraction task capitalized the power of ComBERT, a contextualized pre-trained language model produced through domain adaptive pre-training from BERT on in-domain data. The new model architecture addresses specific challenges related to the special characteristics of the dataset, in particular: (1) sentences containing lots of numerical information such as price, percentage of change, and dates, (2) entities of similar type playing distinctly different argument roles, and (3) the need for arguments extraction to disambiguate the identified events. The end-to-end solution is made up of various architectures suitable for each subtasks: EMD, ED, ARP and Property Classification. Both EMD and ED are fine-tuned using the BERTForTokenClassification head (Devlin et al., 2019), while event property classification is trained using SelfAttentiveSpanExtractor from AllenNLP library (Gardner et al., 2018). As for ARP, the proposed solution uses a Graph Convolutional Network with a contextual sub-tree to effectively predict event argument roles. Experimental results for each subtask demonstrate that the proposed solution outperforms existing solutions with higher F1 scores. While this chapter focuses on solution architecture, the next chapter investigates the best training approach leveraging Transfer Learning to achieve better results despite the limited training data and class imbalance.

## Chapter 5

# Enhancing Event Extraction Performance with Transfer Learning

As a continuation from Chapter 4 that specifically looks at solution architecture, here this Chapter dives deeper into the **training approach**. Rather than training models from scratch, an ensemble of Transfer Learning approaches (Domain Adaptive Pre-training, Multi-task Learning, and Sequential Transfer Learning) are used to address the issue of class imbalance and to generate models with better performance than those trained via Supervised Training alone.

Transfer Learning has been proven to be effective for a wide range of applications, especially for low-resourced domains (Meftah et al., 2021). In fact, ComBERT which was created from domain adaptive pre-training from BERT and the idea of fine-tuning the event extraction task from ComBERT (details in Section 4.3) is a form of Transfer Learning. Here in this chapter, transfer learning techniques are explored to formulate the best training setup among event extraction sub-tasks to produce event extraction and event property classification models with the best possible accuracy despite of the limited training size and class imbalance in the *CrudeOilNews* corpus.

Figure 5.1 gives an overview of the proposed training setup in diagrammatic form. The proposed training approach is described in detailed in the following sections:

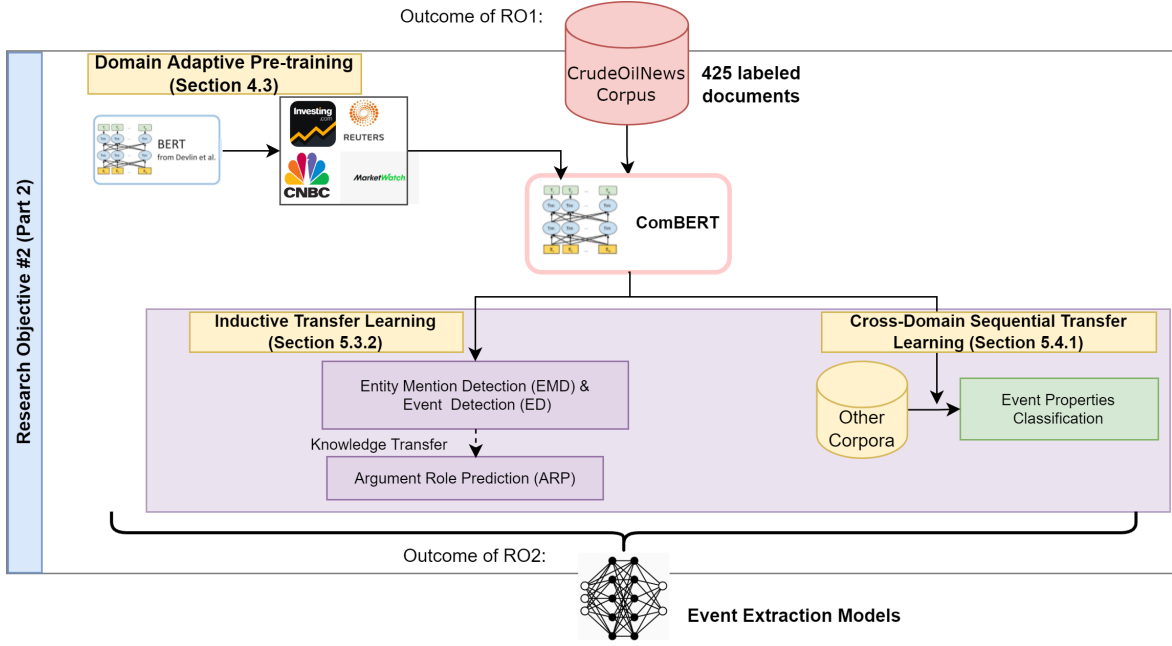


Figure 5.1: Event Extraction leveraging Transfer Learning to achieve best possible performance despite limited annotated data. A number of approaches within the Transfer Learning taxonomy is explored: (1) Inductive Transfer Learning for EMD, ED and ARP sub-tasks; and (2) Cross-Domain Sequential Transfer Learning for Event Properties classification.

1. Inductive transfer learning for the sub-task of (1) Entity Mention (ED) and Event Detection (ED) and (2) Argument Role Prediction (ARP) in Section 5.3.2;
2. Cross-domain Sequential Transfer Learning for event Properties classification in Section 5.4.1;

Transfer learning is a set of methods that leverages resources from other domains or resources intended from other tasks to train a model with better generalization properties. Resources from other domains are known as *source domain*, while resources intended for a different task is known as *source task*. Transfer learning allows the features learned from a source dataset or a source task to be used in, and thus benefiting, the target dataset or target task (Pan & Yang, 2010). In this work, various techniques within the Transfer Learning paradigm are experimented; they are *Domain Adaptive Pre-training*, *Sequential Transfer Learning* and *Multi-task Learning*. Definition of the various types of Transfer Learning techniques are laid out in Section 5.1.

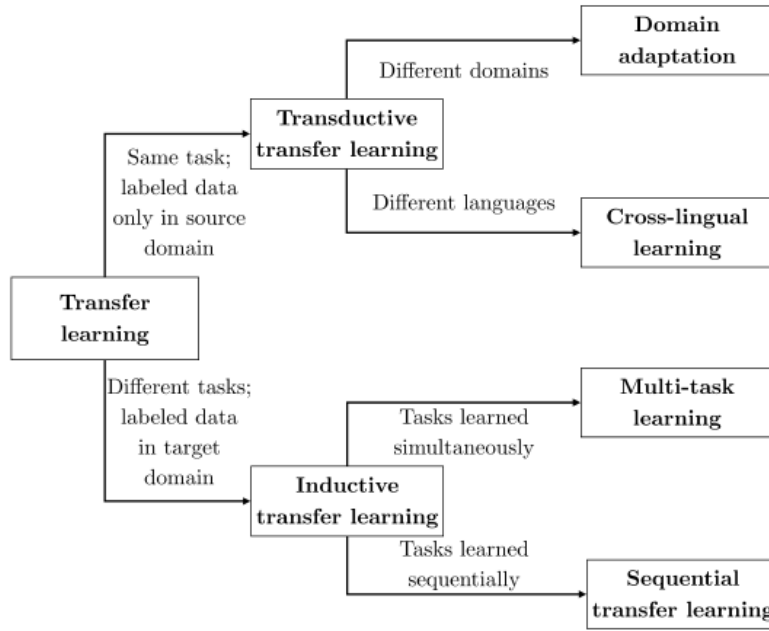


Figure 5.2: A taxonomy of Transfer Learning for NLP (image taken from (Ruder, 2019))

## 5.1 Transfer Learning

Transfer learning is the idea of overcoming the isolated learning paradigm and utilizing knowledge acquired for one task to solve related ones. Transfer Learning allows us to deal with this challenge by leveraging data of related task or domain, known as the *source task* and *source domain*. The knowledge gained in training the source task in the source domain is then applied to the *target task* and *target domain* (Ruder, 2019). The definition of Transfer Learning provided here is based on (Weiss et al., 2016; Pan & Yang, 2010; Alyafeai et al., 2020). Given a source-domain task tuple  $(D_s; T_s)$  and different target domain-task pair  $(D_t; T_t)$ , transfer learning is defined as the process of using the source domain and task in the learning process of the target domain task.

The definition of transfer learning used here is aligned to (Ruder, 2019), and the taxonomy diagram used in (Ruder, 2019) is shown in Figure 5.2. According to this taxonomy, the main branches of Transfer Learning are:

1. **Transductive transfer Learning:** It is the setting where source and target tasks are the same; there is no labeled data at all or there is very few labeled data in the target domain, but sufficient labeled data in the source domain.

2. **Inductive transfer learning:** It is the setting where the source task and the target task are different; labeled data is only available in the target domain.

Based on the above definition and given that the intension is to fully utilize the labels in the *CrudeOilNews* dataset (target domain) to train (or fine-tune) the model, only **Inductive transfer learning** and its sub-types are focused here, namely:

1. **Sequential Transfer Learning (STL):** According to (Ruder, 2019), STL is a type of Inductive Transfer Learning; it is the process of learning multiple tasks ( $T_1; T_2; \dots; T_n$ ). At each step  $t$ , a specific task  $T_t$  is learned. There are two types of STL, each illustrated by an example below:
  - (a) Cross-domain STL: A model is first trained on task  $T$  using a source dataset  $D_s$  and is then transferred to train on the target dataset  $D_t$  on the same task ( $T_s = T_t$ ). In (T. Gui et al., 2018), authors used cross-domain STL to transfer a POS-tagging model trained using News, a resource-rich domain, to train on the same task using Tweets, a lower resource domain;
  - (b) Cross-task STL: A model is first trained on task  $T_s$  and is then transferred to train on a different but related task  $T_t$  in the same domain. This is seen in (Meftah & Semmar, 2018), where the authors first train a Named Entity Recognition (NER) model and then transfer the model to train on POS-tagging task in the same dataset (Meftah & Semmar, 2018).
2. **Multi-task Learning (MTL):** the process of learning multiple tasks ( $T_1; T_2; \dots; T_n$ ) at the same time. All tasks are learned in a parallel fashion. For example, both chunking and POS-tagging are trained concurrently in (Søgaard & Goldberg, 2016; Ruder et al., 2017), as chunking has been shown to benefit from being jointly trained with low-level tasks such as POS tagging. MTL is also used in other tasks such as Toxic Comment Detection in (Vaidya et al., 2020) by jointly learning (1) toxicity score prediction and (2) identity presence detection.

### 5.1.1 Negative Transfer

There are cases when transfer learning can lead to a drop in performance instead of improving it. Negative transfer refers to scenarios where the transfer of knowledge from the source to the target does not lead to improvements, but instead causes a drop in the overall performance

of the target task that might be lower than that with a solely supervised training on in-target data (Torrey & Shavlik, 2010). There can be various reasons or various situation that result in negative transfer, such as:

1. in MTL or STL task-centered transfer learning when the source task is not sufficiently related to the target task or if the transfer method could not leverage the relationship between the source and target tasks very well (Rosenstein et al., 2005);
2. in domain adaptation when the source domain is dissimilar or less related to target domain (Meftah et al., 2021; Blitzer et al., 2007; Ruder, 2019; L. Gui et al., 2018)

Since transfer learning is explored in this work, it is important to be informed about possible causes and pitfalls that could lead to negative transfer. A better understanding of this phenomenon can help us interpret experiment results especially when the use of transfer learning led to a worse off performance.

In this work, the best training setup or configuration of the event extraction subtasks is explored. The focus is to capitalize the effectiveness of transfer learning to improve final model performance through improving embeddings or model representations to overcome issues of labeled data scarcity and class imbalance.

## 5.2 Related Work

The question of how to improve the extraction accuracy from a rather limited set of labeled gold data has become an important one. Many have started exploring transfer learning to improve event extraction through various types of Transfer Learning. This section discusses existing event extraction solutions involving Transfer Learning techniques. There are previous works that explored the usage of MTL and STL in event extraction, they are described below:

### 5.2.1 Usage of MTL

Multi-task Learning (MTL) is also known as *joint learning* or *joint training* in most event extraction literature. Here past works are listed below according to the different combinations of sub-tasks:

1. Partial MTL: jointly training ED + ARP. This is a very common approach in event extraction literature (X. Liu et al., 2018; Q. Li et al., 2013; T. H. Nguyen et al., 2016; Sha et al., 2018). All of them used different deep learning architecture: (X. Liu et al., 2018) uses GCN with Attention Mechanism, (T. H. Nguyen et al., 2016) uses Recurrent Neural Network (RNN), (Sha et al., 2018) uses Dependency-Bridge RNN and Tensor-Based Argument Interaction.
2. Full MTL: Joint modeling of all three sub-tasks : EMD, ED and ARP are trained together. This approach was reported in (Q. Li, Ji, et al., 2014; B. Yang & Mitchell, 2016; Judea & Strube, 2016; T. M. Nguyen & Nguyen, 2019; J. Zhang et al., 2019). The authors in (B. Yang & Mitchell, 2016) consider structural dependencies among sub-tasks, by adopting a two-stage reranking procedure, first by selecting the k-best output of event triggers and entity mentions, then performing joint inference via reranking. (T. M. Nguyen & Nguyen, 2019) build a multi-task model that exploits mutual benefits among the three tasks, by sharing common encoding layers of the input sentence. In this setting, output structures of entity mentions, event triggers and argument semantic roles are decoded separately. (J. Zhang et al., 2019), on the other hand, used neural transition-based framework to predict complex joint structures incrementally in a state-transition process.
3. Hierarchical MTL: training sub-tasks according to a hierarchical fashion. The idea is to utilize a set of low level tasks learned at the bottom layers of the model to create a set of shared semantic representations that will progressively have a more complex representation from the more complex tasks at the higher level. The authors in (Sanh et al., 2019) showed that these low and higher level tasks benefit from models trained in a hierarchical fashion benefit each other. The authors trained Named Entity Recognition (NER), EMD, Entity Coference Resolution and Relation Extraction via a hierarchical fashion. Similarly, authors in (Wadden et al., 2019) also train the same set of sub-tasks (but treating entity co-reference as an auxiliary task) using span representation from BERT and Graph propagation.

### 5.2.2 Usage of STL

In Sequential Transfer Learning (STL), a model is first trained on a task or a dataset, and then it is ‘transferred’ to train another task or to train on another dataset. This means that, as opposed to MTL, STL models are not optimized jointly, but each task is learned sequentially.



(Y. Chen, 2021) is an example of cross-domain STL; the authors used multiple source datasets to help achieve a wider coverage of events in the target dataset using adversarial network-based transfer learning. The authors capitalized on four other corpora with varying degrees of relevance to their target dataset (all within the BioMedical domain) to extract and transfer common features from the related source corpora effectively to boost the performance of event detection in the target dataset.

### 5.2.3 Other forms of Transfer Learning

In (Huang et al., 2018), authors used zero-shot transfer learning to allow their event extraction model to generalize to new unseen event types (events without annotation). They model event extraction as a generic grounding problem and designed a transferable architecture of structural and compositional neural network, that leverages existing event schemas and human annotations for a small set of seen types, and transfers the knowledge from the existing types to the extraction of unseen types. In (Lyu et al., 2021), on the other hand, the authors formulate *zero-shot* event extraction as a set of Textual Entailment (TE) and / or Question Answering (QA) queries, exploiting pretrained (TE/QA) models for direct transfer (transfer learning) to do the new target task of event extraction.

## 5.3 Event Extraction

### 5.3.1 Cross-domain Sequential Transfer Learning

Here, the possibility of utilizing available source datasets on event extraction to improve performance of the same task in the target dataset (*CrudeOilNews*) is investigated. The proposed solution is inspired by the works of (Meftah & Semmar, 2018) who used cross-domain STL in transferring model trained on POS-tagging task from Newswire domain (source domain) to Twitter text (target domain). There are two event extraction datasets annotated according to the ACE/ERE standards: (1) benchmark event extraction dataset ACE2005 in the generic domain, and (2) SENTiVENT (Jacobs & Hoste, 2021) for company-specific events in the finance and economics domain. However, unlike (Y. Chen, 2021) who used source datasets from the same domain (BioMedical Domain), in this case there is no other event extraction corpus from the same domain as *CrudeOilNews*. The two candidate

corpus identified above are somewhat different from the *CrudeOilNews* corpus; analysis for each one is listed down below:

**ACE2005** is a general domain corpus; out of its 33 sub-event types, almost none overlap with the events defined in *CrudeOilNews* corpus. Even though 2 of the events *Conflict - Attack*, *Conflict - Demonstrate* may seem the same as *Civil-unrest*, however upon closer scrutiny, the types of conflict here are different: ACE2005 ones are at a personal level, such as a person attacking another person, while in *CrudeOilNews* the conflicts are geo-political, such as social unrest, and large-scale demonstration.

**SENTiVENT** is a corpus made up of business news with company-related events annotated according to the ACE/ERE standards. Among the event types, there is a ‘placeholder’ event type called ‘Macroeconomic’, a broad category that captures all non-company specific events such as market trends, market-share, competition, regulation issues, etc. This ‘Macroeconomic’ event type is the only event type that overlaps with *CrudeOilNews* corpus. Unfortunately, while they lump non-company events into one category, *CrudeOilNews* corpus focuses on Macro-economic and Geo-political events in a finer-detail. Furthermore, *SENTiVENT* corpus is annotated with discontinuous, multiword triggers, e.g., “upgraded ... to buy”, “cut back ... expenses”, “EPS decline”). This is distinctly different from the way triggers are annotated in *ACE2005* and *CrudeOilNews* where triggers are either single-word or continuous multiwords. The baseline model developed for event detection in *CrudeOilNews* cannot be readily applied to *SENTiVENT* without any modification.

Apart from the fact that there is minimal overlap of event types between the candidate corpora above and the target dataset, an analysis of vocabulary overlap also shows that there is only a mid-range vocabulary overlap (see Figure 5.5). Hence, it can be concluded that it is not feasible to utilize these two candidate corpora for cross-domain STL. This observation is supported by the results in (Y. Chen, 2021), where two out of four of the source dataset has a very low proportion of trigger overlap that produced worse performance in the target dataset. This is the result of Negative Transfer as described in Section 5.1.1.

### 5.3.2 Inductive Transfer Learning

As shown in the section above, there are no suitable datasets to be used in cross-domain STL. Instead, here multi-task learning (MTL) and sequential transfer learning (STL) and the ensemble of the two among event extraction sub-tasks are explored. In the experiment section, **Single Task (baseline) vs Multi-task Learning (MTL) vs Sequential transfer Learning (STL) vs an ensemble of MTL and STL** are investigated.

STL consists of two stages: a pre-training phase in which general representations are learned on a *source* task or domain, followed by an adaptation phase during which the learned knowledge is applied to a target task or domain. The idea here is to optimize the training setup so that the knowledge learned from a particular sub-task will benefit another sub-task.

### 5.3.3 Experiments

Five types of experiments with different combinations of task setups are carried out to determine the best transfer learning configuration with maximum benefits in terms of sharing learned representations of source tasks and target tasks. These five different task setups are:

1. Single Task Learning (Baseline): this is also known as the pipeline approach where the sub-tasks are trained independently one after another, each model is trained from scratch with no transfer learning (EMD, ED, ARP). This is the approach used in Chapter 4.
2. Full Multi-task training: For the experiment, the approach in (J. Zhang et al., 2019) is used, where all sub-tasks are trained jointly using neural transition-based framework, predicting joint output structure as a single task (EMD + ED + ARP).
3. Full Sequential Task Training: in this approach, the EMD task is trained first and upon completion, the model is transferred to train on ED and lastly on ARP (EMD ! ED ! ARP).
4. An ensemble of MTL and STL:
  - (a) Ensemble #1: EMD ! ED + ARP. the approach of jointly training ED + ARP is very common in event extraction literature. This setup is used in many work, see Section 5.2.1. The difference between the setup in (X. Liu et al., 2018; Q. Li et al., 2013; T. H. Nguyen et al., 2016; Sha et al., 2018) and this work is that

they jointly trained ED + ARP together using golden entity mentions, while in this work more realistic setting is used where the input to the ED + ARP task is based on entities predicted from the earlier EMD + ED model (EMD ! ED + ARP). The joint loss for ED + ARP is as follows:

$$joint\_loss = loss\_ED + \alpha (loss\_ARP) \quad (5.1)$$

In the experiment,  $\alpha=2$  is used to give a higher weightage to the loss of the ARP task.

- (b) Ensemble #2: EMD + ED ! ARP as shown in Figure 5.3. The resulting model from training EMD + ED via MTL (in the upper box of Figure 5.3) is then transferred to train for ARP (lower box in the figure). EMD + ED acts as the source task  $T_s$  in the context of cross-task STL to benefit the ARP task, the target task  $T_t$ . The joint loss for EMD + ED is as follows:

$$joint\_loss = loss\_EMD + loss\_ED \quad (5.2)$$

Both loss have equal weightages.

Both Ensemble #1 and Ensemble #2 use Graphical Convolution Network (GCN) + Contextual Sub-tree (described in Section 4.6.2). Figure 5.3 shows the training setup for Ensemble #2: ‘model transfer’ from the *source task* - EMD + ED (top box) to the *target task* - ARP (top box).

**Results and Analysis** Results for these experiments are shown in Table 5.1.

As expected, the worst-performing setup is the individual tasks-pipeline approach, where it not only suffers from error propagation, but each model is trained from scratch for each sub-task (without any interaction between them). Both full MTL and full STL achieved slightly better results. Between these two, the full multi-task training took a few more iterations and took longer to train because the approach is more complex.

The best performing models are those that utilize an ensemble of MTL and STL task setups. Among Ensemble #1 and #2, it is found that Ensemble #2 (jointly training EMD + ED before transferring to train on ARP) brings the best performance. The training of EMD and ED can be done jointly via MTL without much impact on both sub-tasks. This is because

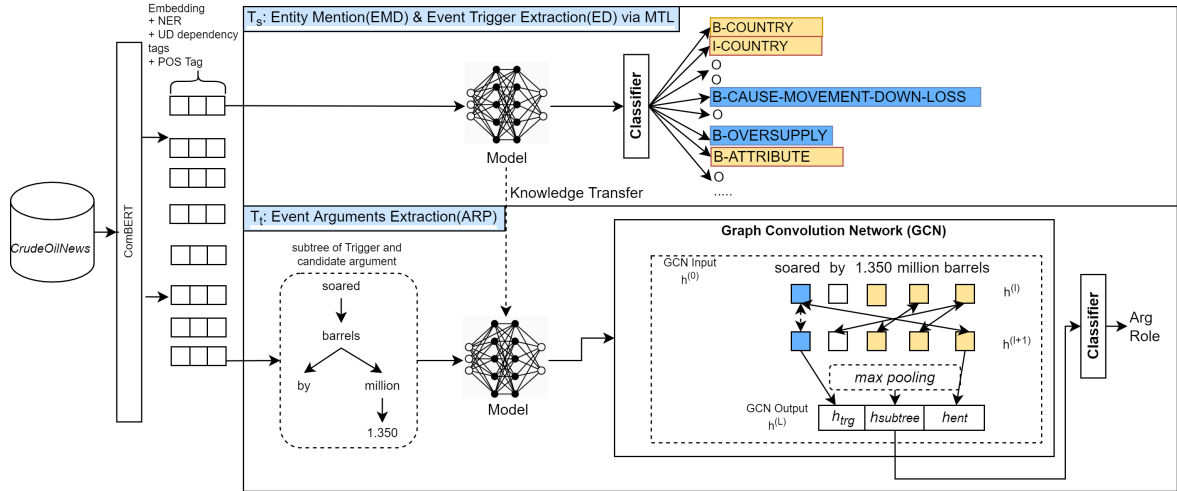


Figure 5.3: Ensemble #2: First, the model is jointly trained on Entity Mention (EMD) and Trigger Extraction (ED) tasks via MTL and then the model is transferred sequentially to train on the task of Entity Arguments Extraction (ARP).

Table 5.1: Experiments with various different task-setups to investigate **Single Task Learning vs Multi-task Learning (MTL) vs Sequential transfer Learning (STL) vs ensemble of MTL & STL**. All setups use the list of entities extracted from the EMD task and not based on Golden annotation.

Task setup			EMD			ED			ARP	
			P	R	F1	P	R	F1	F1	MCC
Individual	sub-task	training	0.903	0.912	0.907	0.915	0.899	0.907	0.802	0.694
(Baseline)										
Task setup			EMD + ED + ARP							
			P	R	F1	P	R	F1	F1	MCC
Full	Multi-task	Training	0.879	0.891	0.885	0.901	0.905	0.903	0.854	0.710
(J. Zhang et al., 2019)										
Task setup			EMD ! ED ! ARP							
			P	R	F1	P	R	F1	F1	MCC
Full	Sequential Task Training		0.903	0.912	0.907	0.911	0.881	0.896	0.854	0.710
Task setup			EMD ! (ED + ARP)							
			P	R	F1	P	R	F1	F1	MCC
Ensemble #1:	EMD !		0.903	0.912	0.907	0.905	0.890	0.897	0.833	0.723
ED+ARP										
Task setup			(EMD + ED) ! ARP							
			P	R	F1	P	R	F1	F1	MCC
Ensemble #2:	EMD + ED		0.926	0.937	0.931	0.916	0.901	0.908	0.888	0.797
! ARP										

it is noticed that entity mentions and trigger words, by definition, are mutually exclusive, e.g. an entity such as *crude oil* is never an event trigger, vice versa an event trigger such as *glut*, though a noun, is never an entity mention. Treating EMD+ED as the source task is useful for the target task. This is related to the fact that the lower embedding and semantic information learned from joint training EMD + ED has a good level of knowledge about entities and triggers, the resulting model has a presentation that is useful for the ARP target sub-task. The detailed results breakdown of Ensemble #2 by entity mention type and event

type are reported in Table D.6 and Table D.7 while results breakdown by event argument roles are reported in Table D.8, in Appendix D.

Ensemble #1 is a common approach in ACE2005 event extraction and this approach have been shown to produce superior results. Unlike ACE2005, *CrudeOilNews* does not exhibit strong interdependence between event type and argument roles as it is in ACE2005. This can be explained using example sentences found in the respective datasets in Table 5.2. As a result, ensemble #1 does not produce the best results in *CrudeOilNews* as was with ACE2005.

Table 5.2: Analysis of events in *ACE2005* and in *CrudeOilNews* respectively. Both datasets exhibit different level of interdependence between event trigger words and its event arguments. The difference in the level of interdependence influenced the selection of the best MTL and STL ensemble for each of the dataset.

Dataset	Analysis
<b>ACE2005:</b> (1) <i>In Baghdad, a cameraman <u>died</u> when an American tank <u>red</u> on Palestine Hotel</i> (2) <i>He has <u>red</u> his air defence chief.</i>	The first occurrence of "\red" is an event trigger of type ATTACK while the second "\red" takes END-POSITION as its event type. The authors in (Q. Li et al., 2013) argues that event arguments play a key role in helping classifying the correct event. For example, the presence of "\tank" plays the role of WEAPON helps determine the right event type. Likewise, in the second sentence, "\defence chief" plays the role of POSITION can help the model classify the second "\red" as END-POSITION. Hence jointly training both ED and ARP helps improve the accuracy of both ED and ARP.
<b>CrudeOilNews:</b> <i>U.S. crude stockpiles <u>soared</u> by 1.350 million barrels in December from a mere 200 million barrels to 438.9 million barrels, due to this oversupply crude oil prices <u>plunged</u> more than 50% on Tuesday.</i>	Events are more straight forward, ie. trigger words are tied to just one type of event, therefore there is no need to utilize arguments to help differentiate the event type.

## 5.4 Event Properties Classification

The main challenge faced with event properties classification is class imbalance. One of the quick and easy ways to overcome this challenge is through oversampling minority classes. However, due to the limited size of the *CrudeOilNews* corpus, oversampling within an already limited pool of annotation is not a good approach. Instead, the usage of other available corpora in all domains was investigated to assist in training a robust model despite the class imbalance.

### 5.4.1 Cross-Domain Sequential Transfer Learning

Here, the usage of other available corpora in all domains for the purpose of cross-domain STL is investigated. The idea is to use resources from different source domains to train a model before fine-tuning the model to adapt to a new domain on the same task. This STL is carried out by first training the corpora from the *source domain* on Event Polarity / Modality classification and then transfer the model to fine-tune on the same task on the *target domain*, i.e., *CrudeOilNews* corpus. Figure 5.4 shows a graphical depiction of the idea of STL. In the top section, a labeled dataset in the source domain is used to train a model on event polarity or modality classification. The model is then transferred and fine-tuned on the same task in the target domain, as shown in the bottom section.

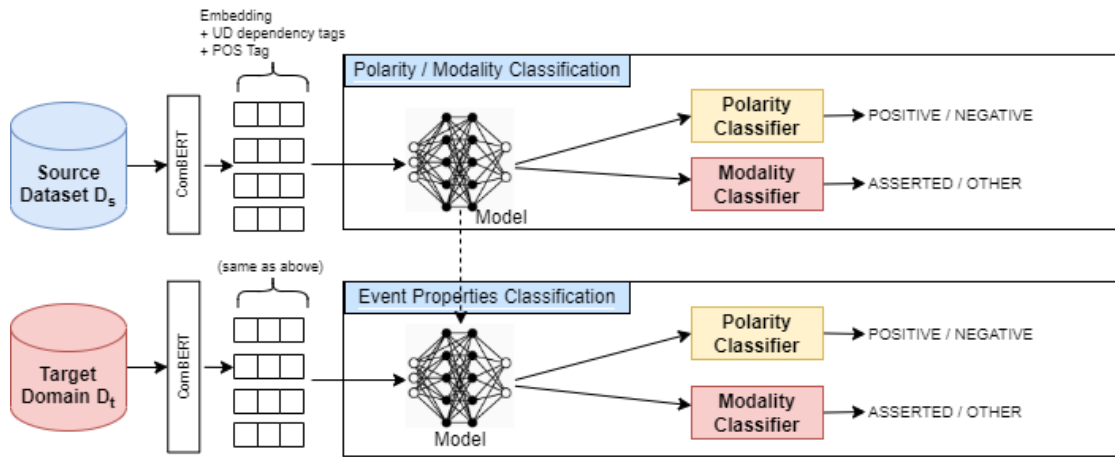


Figure 5.4: *Sequential Transfer Learning (STL)*: The model is first trained on labeled dataset in the source domain (see list of corpora listed in Table 5.3) before being transferred to train on the Target Domain (*CrudeOilNews*). STL is done for Polarity and Modality Classification. Intensity classification is trained from scratch with just the *CrudeOilNews* corpus due to the lack of resources.

Before going into the implementation details of cross-domain STL, it is important that the task definitions are aligned as they may appear as different names:

1. **Event Polarity Classification**<sup>1</sup> can be aligned to **Negation Detection**, the term used in *SEM 2012 Shared Task: Negation Detection and Scope Resolution*.
2. **Event Modality Classification** can be aligned to **Hedge / Uncertainty / Speculation Detection**, the term used in *CoNLL 2010 shared task: Hedge detection and scope resolution*.

<sup>1</sup>Not to be confused with Sentiment Polarity(Positive / Negative sentiment)

In the subsections that follow, the available resources for both negation detection and uncertainty detection are discussed. Resources for Event Factuality Prediction (EFP) (Saurí & Pustejovsky, 2009) is excluded here because EFP combines both negation and speculation detection to determine the ‘factuality’ of an event. Here, only corpora that have Negation and Uncertainty annotated individually are considered.

Event Intensity classification is excluded from cross-domain STL because to the best of our knowledge, there is no available labeled dataset annotated for event intensity classification.

#### 5.4.1.1 Available Source Domain Corpora

##### Corpora for Negation Detection

1. In *SEM2012 Shared Task* (Morante & Blanco, 2012), two corpora in the general text were released for negation scope and focus detection; they are the *Conan Doyle stories* and the *Penn TreeBank* corpus;
2. In the survey paper (Jiménez-Zafra et al., 2020), the authors listed out all English and Spanish corpora annotated with negation (negation cue and its respective scope). According to the list, the available corpora are in the following domains:
  - (a) Bio-related text domain: *BioInfer*, *Genia Event*, *BioScope*, and *DrugDDI*;
  - (b) Consumer reviews: *product review Corpus*, *SFU Review*, *Movie review*;
  - (c) General Domain: *Prop Bank* and *SFU Opinion & Comments (SOCC)*

##### Corpora for Uncertainty Detection

1. The *CoNLL2010 share task* (Farkas et al., 2010) is made up of a collection of corpora include Biology-related publications and general domain factual text from Wikipedia;
2. In the financial domain, (Theil et al., 2018) introduced the *10-k nancial disclosures corpus* for the task of classifying financial statements whether they are *certain* or *uncertain*.
3. Consumer reviews : *SFU Reviews corpus* (Konstantinova et al., 2012) contains both uncertainty and negation cue words and scope annotated.

Corpora for both tasks are summarized into Table 5.3 below.



Table 5.3: List of open source corpora with Negation and Uncertainty Annotation.

Dataset	Domain	Negation		Uncertainty	
		Cue	Scope	Cue	Scope
1. ACE2005	General	only class labels		only class labels	
2. SENTiVENT (Jacobs & Hoste, 2021)	Economic news	only class labels		only class labels	
3. ConanDoyle(neg) (Morante & Blanco, 2012)	Fiction	✓	✓		
4. SFU OCC(neg) (Kolhatkar et al., 2020)	Opinion News & Comments	✓	✓		
5. 10kFinStatement(unc) (Theil et al., 2018)	Corporate Financial Disclosure			only class labels	
6. Wikipedia(unc) (Farkas et al., 2010)	General			✓	
7. Reviews(neg & unc) (Konstantinova et al., 2012)	Product Reviews	✓	✓	✓	✓

**Domain Similarity** In transfer learning, it is observed that the more related the tasks, the easier it is for transfer or cross-utilize the knowledge (Sanh et al., 2019). The same holds true for data where the more related the source domain to the target domain, the easier it is for effective transfer learning (Meftah et al., 2021; Gururangan et al., 2020). Based on this fact, Bio-medical-related corpora are excluded from the experiments because the Bio-medical domain has its specific vocabulary that are deemed different from the Finance and Economics domain. The small vocabulary overlap with crude oil news will potential result in negative transfer. Seven corpora was selected and are listed in Table 5.3; the details of each of these corpora are found in Appendix B. Even though none of the corpora above are related to Economic / Finance domain, they are chosen because their tasks are similar to the task of event property classification and therefore earmarked as potential source datasets in cross-domain STL.

Domain similarity between the source datasets and *CrudeOilNews* are evaluated by obtaining the percentage of vocabulary overlap of each of the source dataset with *CrudeoilNews*. Analysis of vocabulary overlap is shown in Figure 5.5. On a continuum of the proximity between source datasets and target dataset, the source datasets can be ranked as *SENTiVENT* ! *10kFinStatement* ! *ACE2005* ! *Wikipedia-CoNLL2010* ! *SOCC* ! *Reviews* ! *ConanDoyle*.

**Task Modification** The shared task of CoNLL2010 (for Uncertainty Detection) and SEM2010 (for Negation Detection) consist of two sub-tasks: (1) it involves first detecting the cue words at the sentence level and then (2) resolving the scope based on the cue words detected. Event property classification, on the other hand, is slightly different where the main

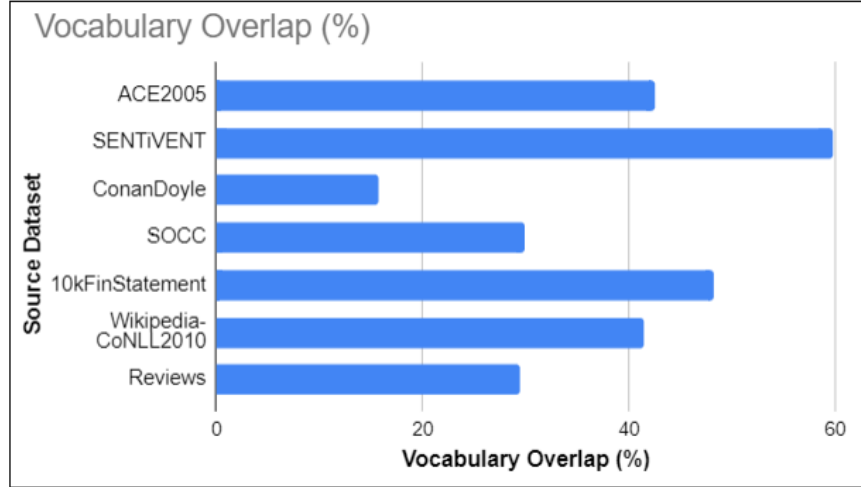


Figure 5.5: Vocabulary overlap (%) between source datasets and *CrudeOilNews* target dataset. Vocabularies of each domain are made up of the top 10K most frequent words (excluding stopwords) in each corpus.

aim is to detect the event first and then determine the event properties based on the event and its scope. Due to the difference in the original source tasks and how the source datasets are annotated, the original tasks are modified to align with the event property classification task:

1. Simplify the shared task to just one task. The original Negation / Uncertainty detection involves two sub-tasks : (1) cue word detection and (2) scope resolution. This is simplified into a sentence level classification task.
2. Next, align class labels:
  - (a) Event Polarity: For sentences that contain Negation cue words, the label NEGATIVE is assigned for the whole sentence. For sentences without, the label POSITIVE is assigned.
  - (b) Event Modality: Similar to polarity classification, sentences with Uncertainty cue words or have the ‘uncertain’ scope annotated are assigned the label OTHER. For sentences without, the label ASSERTED is assigned.

#### 5.4.1.2 Experiments

First, Event Polarity / Modality model is trained using the corpora listed in Table 5.3; these corpora are known as the “source domain”  $D_S$ . Then the model is transferred to train on

the same task on the “target corpus”  $D_t$ , the *CrudeOilNews* corpus. In the experiments, the best model is transferred to train on the same task on *CrudeOilNews*.

#### 5.4.1.3 Results and Analysis

Table 5.4: Results of Polarity and Modality classification of *CrudeOilNews* using various source datasets (as listed in Table 5.3) as source domain  $D_s$  in cross-domain sequential transfer learning. (") indicates an improvement in performance compared to the baseline, while (#) indicates the opposite.

Event Polarity Classification			
Source $D_s$	Target $D_t$	F1	MCC
1a. - ( <i>baseline</i> )	CrudeOilNews	0.729	0.478
1b. ConanDoyle	CrudeOilNews	0.699(#)	0.305(#)
1c. OCC	CrudeOilNews	0.793(")	0.498(")
1d. Reviews	CrudeOilNews	0.713(#)	0.412(#)
1e. <b>SENTiVENT</b>	CrudeOilNews	<b>0.805(")</b>	<b>0.611(")</b>
Event Modality Classification			
Source $D_s$	Target $D_t$	F1	MCC
2a. - ( <i>baseline</i> )	CrudeOilNews	0.795	0.498
2b. 10kFinStatement	CrudeOilNews	0.879(")	0.695(")
2c. Wikipedia	CrudeOilNews	0.841(")	0.481(#)
2d. Reviews	CrudeOilNews	0.723(#)	0.395(#)
2e. <b>SENTiVENT</b>	CrudeOilNews	<b>0.835(")</b>	<b>0.705(")</b>

The results shown in Table 5.4 is analyzed against the list of event properties below:

1. Event Polarity: there is some form of improvement when the model is trained on a source domain first before fine-tuning the model on the target domain. The best “source domain” corpus for Event Polarity is *SENTiVENT*, while models trained on *ConanDoyle* and *Reviews* performed worst than the baseline model. The main reason is that these corpora are somewhat dissimilar to *CrudeOilNews* corpus that resulted in Negative Transfer. Performance deterioration is especially apparent in *ConanDoyle* because it is a corpus made up of dialogues or conversations and has negation cues mainly in a conversational form such as *don't*, *doesn't*, *didn't*, *isn't*, *can't*, *wasn't* that are not found in the target corpus.
2. Event Modality: Due to the similarity between the *CrudeOilNews* and the two finance-related corpora : *SENTiVENT* and *10KFinStatement*, the resulting cross-domain STL models are able to provide a significant boost to model classification performance. Similar to the *ConanDoyle* corpus, the *Reviews* contains conversational-like sentences that have minimal overlap with *CrudeOilNews* in terms of uncertainty cue words that contributed to a worse off model performance.

It is worth highlighting that by comparing between F1-score and MCC-score, MCC-score has a more significant jump in improvement. As highlighted in Section 4.7.3, results of baseline models show a lower MCC-score while a high F1-score, the mismatch is due to class imbalance; the models tend to classify everything to the majority class leading to a high number of False Positives. Upon training the model with source datasets that do not have a serious class imbalance issue, the final models have higher MCC-scores. Based on error analysis, it is shown that the final models have higher True Negatives (higher prediction on minority class) and thus lead to a better MCC-score. An improved MCC score means classification performance improves across all classes, including the minority class; therefore lessening the impact of class imbalance.

It can be concluded from the results above that the more similar the source domain is to the target domain, the easier cross-domain STL can be used to improve the final classifier’s performance. This is also consistent with Ruder’s conclusion in (Ruder, 2019) that the more distant two domains are, the harder it is to adapt from one to the other.

## 5.5 Final Model Performance

The final model performance for each sub-task is tabulated and shown again in Table 5.5 for easier reference. Table 5.5 shows the final model performance after applying transfer learning. The results of *Before* and *After* transfer learning is applied. It is obvious that transfer learning provided a performance boost to all subtasks within Event Extraction and to Event Properties Classification. Based on the results, it is clear that there is marginal improvement in Precision, Recall and F1-scores but there is a significant improvement in MCC score: ARP task showed an increase of 0.103, while Polarity and Modality showed an increase of 0.133 and 0.207 respectively. In other words, the models are no longer always predicting the majority class, but now are more ‘balanced’ in terms of majority-minority class prediction.

## 5.6 Summary and Discussion

It is known that training models via the traditional approach of supervised learning requires a substantial amount of annotated data. This challenge becomes even more apparent for a

Table 5.5: Event Extraction results: **before and after** applying transfer Learning. The models after applying transfer learning show performance improvements of varying degree as compared to the baseline (before applying transfer learning).

	Type	Precision	Recall	F1	MCC
<b>Before</b>					
EE	1. Entity Mention Detection (EMD)	0.903	0.912	0.907	-
	2. Event Detection (ED)	0.915	0.899	0.907	-
	3. Argument Role Prediction (ARP)	0.792	0.821	0.802	0.694
Prop.	1. Event Polarity	0.684	0.780	0.729	0.478
	2. Event Modality	0.763	0.830	0.795	0.498
	3. Event Intensity	0.723	0.719	0.721	0.595
<b>After</b>					
EE	1. Entity Mention Detection (EMD)	0.927(")	0.937(")	0.931(")	-
	2. Event Detection (ED)	negligible difference, same as above			
	3. Argument Role Prediction (ARP)	0.902(")	0.897(")	0.888(")	<b>0.797(")</b>
Prop.	1. Event Polarity	0.697(")	0.917(")	0.805(")	<b>0.611(")</b>
	2. Event Modality	0.803(")	0.842(")	0.835(")	<b>0.705(")</b>
	3. Event Intensity	same as above			

lower-resource domain such as Finance and Economics. Transfer learning is used to improve the performance of event extraction models despite limited training data and dataset suffering from class imbalance.

It has been shown, through vigorous experiments, that the final performance of event extraction model was improved via (1) an ensemble of MTL and STL; (2) Cross-domain STL as a strategy to overcome the issue of class imbalance by leveraging on resources for the event properties classification task but from a different domain. The impact of class imbalance is minimized across several subtasks when MCC scores improved substantially after transfer learning is applied.

## Chapter 6

# Event-based Crude Oil Futures Trend and Returns Prediction

**RO3:** To design a classifier for crude oil forecasting using co-occurrence of events.

In natural language-based financial forecasting, researchers seek information outside historic market data from sources such as news, tweets, company reports, and financial periodicals. Many existing research employed textual data as input features for stock price prediction. In comparison, financial forecasting of other financial assets such as commodities and market risks are less common. The prediction task involves learning the correlations between textual information (unstructured data) and financial asset prices (structured time series data). Various methods were proposed to extract semantic information from textual data. Financial News Analytics has drawn much attention in recent years, and there have been many proposed solutions that mine news data for better market trend predictions. Among the methods are Sentiment Analysis, Topic Modeling, Summarization and Event Extraction in different levels of granularity.

This work focuses on crude oil, one of the major commodities traded in the world. Apart from stock price prediction (Ding et al., 2014; D. Chen, Zou, et al., 2019), news is also an important source of information for the oil market as oil price movements are driven by events such as geopolitics (e.g. war, civil unrest, political instabilities), macro-economic events (e.g. economic development), financial environment, as well as oil market factors (e.g., consumption, inventory, and supply of oil). As discussed in (Brandt & Gao, 2019), these

events are found to cause movement in crude oil prices both in the short-term and long-term. In this chapter, the usage of a special genre of news, known here as **market summaries**, is proposed as an excellent source to learn the correlation between events and crude oil market reactions. The proposed solution uses *Fine-grained Event+*<sup>1</sup> extracted from market summaries to form a dataset for crude oil futures trend and returns prediction.

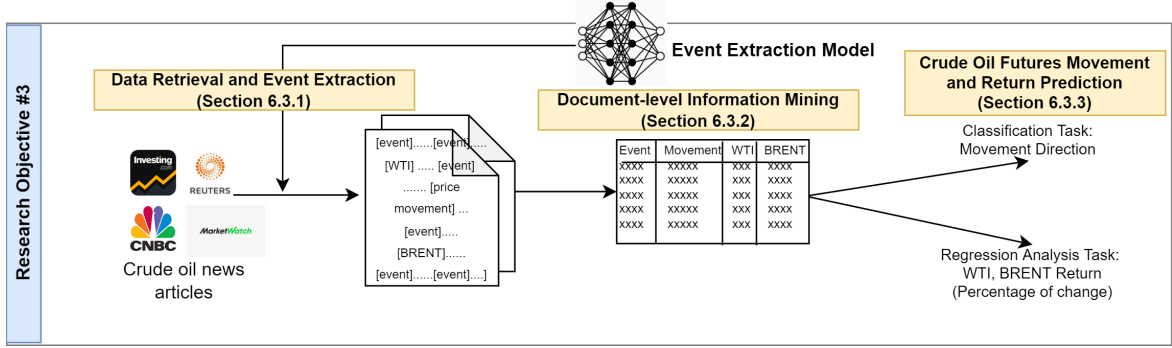


Figure 6.1: The proposed framework is made up of these components: (1) Data Retrieval and Event Extraction, (2) Document-level Information Mining, (3) Crude oil futures trend and returns prediction.

Figure 6.1 gives an overview of the proposed framework in diagrammatic form. The end-to-end framework is broken down into:

1. Data Retrieval and Event Extraction - Extract events from crude oil market summaries using models developed in RO2 (model architecture in Chapter 4, training approach in Chapter 5) in Section 6.3.1;
2. Document-level information Mining - Build a new dataset in Section 6.3.2 based on extracted events;
3. Crude oil futures trend and returns prediction in Section 6.3.3.

## 6.1 Domain Knowledge, Background Information and Terminologies

Here, crude oil / financial market related terminologies and information on WTI and Brent are laid out to assist with the overall readability of this chapter:

<sup>1</sup> *Fine-grained Event+* is defined at the end of this section.

### 6.1.1 Crude Oil Benchmark

For this work, the scope is to predict WTI and Brent **futures**; futures of the front-month contract to be specific. WTI<sup>2</sup> and BRENT<sup>3</sup> is traded in New York Mercantile Exchange (NYMEX) while Brent is traded in Intercontinental Exchange (ICE). Both WTI and Brent<sup>4</sup> are two internationally-recognized types of crude oil that are used as **benchmarks** for prices of crude oil. WTI is also known as light, sweet crude in crude oil news. According to Investopedia, WTI and Brent move somewhat in unison, however, WTI is more sensitive to American economic developments, while Brent responds more to those in other regions.

### 6.1.2 Terminologies

This subsection lays out the important crude oil and financial market-related terminologies<sup>5</sup>.

1. **futures contract** - An oil futures contract is an agreement to buy or sell a certain number of barrels of oil at a predetermined price, on a predetermined date.
2. **futures price** (or more commonly known as just ‘futures’) - Throughout this work, the term price and futures are used interchangeably to mean the same thing. They are not to be confused with spot price which is not used here.
3. **front month contract** - it is also called “near” or “spot” month, refers to the nearest expiration date for a futures contract.
4. **return** - the change in price of an asset, investment, or project over time, which may be represented in terms of price change or percentage change.

### 6.1.3 What are Market Summaries?

Market summaries are a special genre of financial news written by financial analysts or journalists analysing the financial market from a retrospective view of what took place and how the market reacted to it. Hence crude oil market summaries contain an excellent distillation of world events that are truly causal to the movement of crude oil prices.

<sup>2</sup><https://www.investopedia.com/terms/w/wti.asp>

<sup>3</sup><https://www.investopedia.com/terms/n/northseabrentcrude.asp>

<sup>4</sup>Difference between WTI and Brent crude [here](#).

<sup>5</sup>For the complete list of crude oil-related terminologies, refer to Appendix C.2.



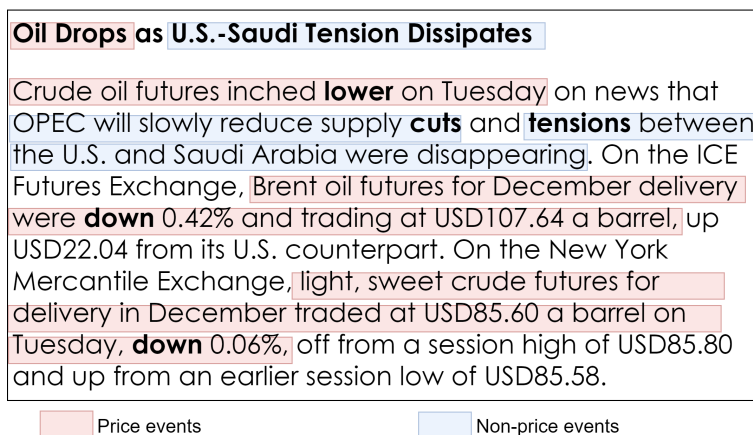


Figure 6.2: An example of Crude Oil Market Summary; Event triggers are in bold, text span of each event is also highlighted: coral colored ones are price events and blue colored ones are non-price events.

The main components of these summaries are (1) a retrospective view of past events, including macro-economic, geo-political, and supply-demand news, and (2) how oil markets reacted to those events. Market reactions here refer to how crude oil market reacts to the news in the form of (1) price trend (UP, DOWN, STABLE) and (2) returns (percentage of price change). Figure 6.2 shows an example of a crude oil market summary. The information found in this news article is extracted and grouped into (1) world events (non-price events highlighted in blue), and (2) its corresponding market reactions (price events highlighted in coral). Each event extracted from Figure 6.2 is organized into ‘Events’ and ‘Market Reaction’, as laid out in Table 6.1.

Table 6.1: Information extracted from Figure 6.2 is re-organized to separate them into (1) events and (2) market reaction in the form of price change and can be further broken down into (1) trend UP, DOWN, STABLE and (2) magnitude of change (in %).

Scope	Event(s)	Market Reaction	
		Trend	WTI and Brent % change
H.line	U.S.-Saudi Tension Dissipates	Oil Drops	
Body	OPEC will slowly reduce supply cuts and tensions between the U.S. and Saudi Arabia were disappearing	Crude oil futures inched lower on Tuesday	Brent oil futures down 0.42% and light, sweet crude down 0.06%

#### 6.1.4 Market Summaries as single source input

The standard approach used in all previous works involving text (news, tweet, financial reports, financial periodicals, etc.) built machine learning models based on the correlation between signals (semantic information extracted from text) and historical financial asset prices for price prediction. One of the challenges of this approach is that spurious correlations

between the input text and time series data are often included with actual, non-spurious ones. The standard practice to overcome the challenge of finding truly correlated news is to find the best lead-lag relationship between a signal and price movement. Authors implement Vector Autoregression (VAR) lag order selection as seen in (X. Li et al., 2019; Bai et al., 2022). An alternative approach is to use Granger causality (Chan & Chong, 2017; J. Li et al., 2017; Feuerriegel & Neumann, 2013) to ascertain that the correlation between text and price is not random. In another approach, authors in (Del Corro & Hoffart, 2021) re-purposed the attention weights of a neural network initially trained for stock price prediction to assign a relevance score to each headline. Another challenge is that not all signals exhibit immediate impact on the asset price, as some take longer to be reflected. The lead-lag relationship between a signal and price movement is investigated in-depth using lag order selection in time series analysis. For simplicity, most works predict price movement for the next day (Ding et al., 2014; Elshendy et al., 2018; Zhao, Zeng, et al., 2019) or other time-interval such as weekly in (J. Li et al., 2017). However, in reality, the impact of news is often shorter as the effects are often intraday, as clearly seen in Figure 6.3 where there is news leading to spikes and trough throughout the same day. Aggregating information to predict a single daily price will muddle out the impact of each piece of news.



Figure 6.3: Brent Crude Oil futures movement in a 24-hour window. (source: <http://www.investing.com>). Each market summaries is indicated by an 'N' indicator, these summaries are released periodically throughout a trading day.

This work proposes to utilize market summaries as a **single-source** input to mine for such signal-price correlation. Fine-grained event extraction is done on market summaries to extract global events and market reactions. The extracted events are categorized into 'price

events’ and ‘non-price events’. A dataset is built by using ‘non-price’ events as input features while ‘price’ events as labels or dependent variable. Apart from event details, (Saurí & Pustejovsky, 2009) pointed out that representation of events in discourse is also incomplete without capturing the veracity or *factuality* (factual certainty) of the events. In other words, it is essential to distinguish events that have actually taken place, events that have not happened, or events that *may* happen. Hence, along with event extraction, the proposed solution captures event properties as well, namely (1) Event Polarity, (2) Event Modality, and (3) Event Intensity.

All these event details make up ‘Fine-grained Event+’ as predictor/independent and market reactions as ‘labels’ or predicted/dependent variables to formulate crude oil price prediction task into two sub-tasks: (1) classification task to determine crude oil price trend (UP, DOWN, STABLE) and (2) Regression analysis task on crude oil price return (percentage of change) for WTI and Brent.

## 6.2 Related Work

An in-depth analysis of financial/economic event extraction techniques is found in Section 2.1.2. Here, the techniques are analysed and evaluated based on the scope and level of event details extracted. The extraction output is categorized into two categories: (1) coarse-grained event and (2) fine-grained event. The definition of ‘coarse-grained’ and ‘fine-grained’ are aligned to the definitions used in (D. Chen, Zou, et al., 2019; Jacobs & Hoste, 2020).

### 6.2.1 Coarse-grained Event

According to the definition used in (D. Chen, Zou, et al., 2019), coarse-grained events are in the form of <subject/ actor, predicate, and object> extracted via Open Information extraction (Etzioni et al., 2008). OpenIE extraction method is used in extracting company events from news headline in (Ding et al., 2014, 2015). With extracted tuples, authors in (Ding et al., 2015) used a neural tensor network to learn effective event embeddings for stock movement prediction. (Saha et al., 2017) introduced BONIE (Bootstrapping-based Open Numerical Information Extractor); it is an extension of Open IE to extract numerical arguments in each OpenIE tuple. (X. Zhang et al., 2018) used HanLP text parser to extract structured events from Chinese news and capture the main verb, object, and subject (similar

to the OpenIE output). Authors in (D. Chen, Zou, et al., 2019) argued that coarse-grained events capture only three components: subject, predicate, and object, therefore losing specific semantic information such as other important arguments such as *time*, *opening price*, *closing price*, *price change*, etc..

### 6.2.2 Fine-grained Event

Fine-grained event extraction focuses on finer event details than coarse-grained event extraction by extracting events' arguments as well as identifying the roles each argument plays with respect to the event.

Authors in (Xie et al., 2013) used SEMAFOR semantic parser (Kshirsagar et al., 2015) in stock price movement prediction. Frames are used to improve sentiment analysis by using (positive or negative) roles that specific companies play in the detected frame to predict the movement of the said companies' share price. However, the authors acknowledged that this approach (1) suffers from inaccuracies in semantic frame parses, which then affects the subsequent classification task of classifying change of stock price and change of polarity, and (2) has the weakness of taking sentences out of context where eventualities of events were not considered, i.e., treating a hypothetical event as a real event.

In (D. Chen, Zou, et al., 2019), the authors built a financial event dictionary and a set of rules and auxiliary information such as POS tags and dependency relations to extract fine-grained events for stock price prediction. On the other hand, (J. Liu & Huang, 2021) use open-domain event extraction (ODEE) for crude oil price forecasting. Event information is used as one of the input features, along with Sentiment Analysis and time-series data modeling for crude oil price prediction.

Comparing all the approaches above, the authors of (D. Chen, Zou, et al., 2019) have shown that using fine-grained events is superior to coarse-grained events because through experiments, it has been shown that fine-grained events and event arguments provide more semantic information and therefore produce better text representation that helps with forecasting performance. Regardless of the granularity and degree of detailed information extracted, all existing work did not consider 'factuality' but merely used event information as-is without detecting any negation or speculation tied to the events.

### 6.2.3 Crude oil Price Forecasting with events

Among the crude oil forecasting works listed in Section 2.1.3, the only one that uses news event as input feature is (J. Liu & Huang, 2021). The solution proposed by authors of (J. Liu & Huang, 2021) use event features (extracted via open domain event extraction) and sentiment features from massive news along with historical crude oil price features to predict future crude oil prices. It is worth highlighting how this work differs from theirs to bring a clearer view of the contribution this work brings. The differences are:

1. Even though events in (J. Liu & Huang, 2021) are extracted from news, the events are not distinguished in terms of factual certainty and are also not differentiated from analysis, expert opinion, and forecast;
2. The open domain event extraction algorithm developed by (X. Liu et al., 2019) was used. This method is able to extract unconstrained types of events and induce universal event schemas from clusters of news articles. In this work, on the other hand, events are extracted via domain-dependent (with pre-defined event typologies) via supervised learning. Rather than clusters of news, a specific genre of financial news: market summaries are used in this work.
3. (J. Liu & Huang, 2021) forecasts WTI daily closing price while this work forecasts both WTI and Brent in terms of trend and returns.

## 6.3 The End-to-end Framework

The end-to-end framework is made up of the following components:

1. Event Extraction - Extract events from crude oil market summaries in Section 6.3.1;
2. Document-level information Mining - Build a new dataset in Section 6.3.2 based on extracted events;
3. Crude oil futures trend and returns prediction in Section 6.3.3.

### 6.3.1 Event Extraction

The event extraction models developed in Chapters 4 and 5 are used here. The models extract the following information: (1) entity mentions and their type as part of the EMD

sub-task; (2) event triggers and event type as part of the ED sub-task; (3) event arguments and the role they play as part of the ARP sub-task; and (4) event properties. To differentiate this from fine-grained events extracted in (D. Chen, Zou, et al., 2019), events extracted here are referred to as *Fine-grained Events+*, which has additional event properties label that other Fine-grained Event does not. Table 6.2 shows the fine-grained event+ components for all events found in the example in Figure 6.2.

Table 6.2: Table below captures all the events identified in Table 6.1 in Fine-grained Event+ form, which consist of *Event Type*, *Event Arguments*, *Polarity(P)*, *Modality(M)* and *Intensity(I)*. Event trigger words are in bold. Abbreviations found in the table: ‘H’ denotes news headlines; ‘S’ indicates sentences in the news body; In Event Trigger column, Lx (eg: L1, L2...) represent ‘Label’ while Ex (eg: E1, E2....) represent ‘Events’ respectively; P (under Polarity) stands for POSITIVE; A and O (under Modality) stand for ASSERTED and OTHER respectively; N and E (under Intensity) stand for NEUTRAL and EASED.

	Event Trigger	Event Type	Argument Role: Text	P	M	I
H	Oil <b>drops</b> (L1)	MOVEMENT-DOWN-LOSS	Item: oil	P	A	N
	U.S.-Saudi <b>tension</b> dissipates (E1)	GEOPOLITICAL-TENSION	Participating country: U.S, Saudi	P	A	E
S1	Crude oil futures inched <b>lower</b> (L2)	MOVEMENT-DOWN-LOSS	Item: crude oil Attribute: futures	P	A	N
	<b>tensions</b> between the U.S. and Saudi Arabia were disappearing (E2)	GEOPOLITICAL-TENSION	Participating country: U.S., Saudi Arabia	P	A	E
	OPEC will slowly reduce supply <b>cuts</b> (E3)	CAUSE-MOVEMENT-DOWN-LOSS	Supplier: OPEC Attribute: supply	P	O	E
S2	Brent oil futures .... <b>down</b> 0.42% (L3)	MOVEMENT-DOWN-LOSS	Item: Brent oil Attribute: futures Difference: 0.42%	P	A	N
S3	light, sweet crude futures ... <b>down</b> 0.06% (L4)	MOVEMENT-DOWN-LOSS	Item: light, sweet crude Attribute: futures Difference: 0.06%	P	A	N

### 6.3.2 Document-Level Information Mining

After event extraction, each piece of news article is organized to form the dataset for training a machine learning model for the prediction tasks. First, the events are separated into *input* and *labels*. As mentioned in Section 6.1.4, market reactions in market summaries come in the form of (1) price trend and (2) how significant is the price movement in terms of the percentage of price change. The trend forms the ‘label’ for the multi-class classification task, while the percentage of price change forms the dependent variable for the regression task. Henceforth, they are referred to as ‘labels’ for easier reference. From Table 6.2, the data

was re-tabulated into Table 6.3 by separating the events into ‘price-related’ and ‘non-price-related’ events based on respective events’ argument. Price-related events  $L1$ ,  $L2$ ,  $L3$ , and  $L4$  form the labels while ‘non-price-related’ events  $E1$ ,  $E2$ , and  $L3$  make up the input features.

Table 6.3: Document-level event mining (Section 6.3.2) and re-tabulating the data for events from Table 6.2. The labels such as E1, E2,...,L4 correspond to the same labels found in Table 6.2.

Input Features / Independent Variable		Label / Dependent Variable		
Header	Events in news body	Trend Label	WTI	Brent
E1	E2 & E3	DOWN (L1, L2)	0.06% (L4)	0.42% (L3)

Unlike any previous work, this approach relies solely on crude oil market summaries to extract price information without using historical price data (structured time series data). In other words, numerical data, such as crude oil returns (percentage of change) or amount of change, initial price, and final price, are all extracted from market summaries alone. There are situations where ‘returns’ are not reported. In contrast, other numerical information such as initial price, final reported price, and price change (not in percentage form but dollars and cents) are available. In these situations, the percentage of change is calculated using one of the two equations below depending on the available information:

$$pt\_change = \frac{p_{t_{final\_reported}} - p_{t_{initial}}}{p_{t_{initial}}} \% \quad (6.1)$$

$$pt\_change = \frac{change\$}{p_{t_{initial}}} \% \quad (6.2)$$

where  $pt\_change$  is the percentage of change,  $p_{t_{initial}}$  is the initial price,  $p_{t_{final\_reported}}$  is the final reported price and  $change\$$  is the price change reported in cents. Price change reported in dollar is converted to cents during data pre-processing.

### 6.3.2.1 Data Pre-processing: Sanity Checking

High quality dataset is vital in producing accurate machine learning models. In this case, accurate event extraction is essential, especially for price-related events. To this end, a sanity check is done to ensure WTI and Brent price-related numerical information extracted from crude oil news summaries against their respective historical price data. Both WTI and Brent historical price data are downloaded from [www.investing.com](http://www.investing.com) for verification. First, each market summary is aligned by date to the historical price data, then the returns (in %),

$pt\_change$ , are checked to make sure the value is within the day's largest jump (in %) and the largest dip (in %) using the equation below:

$$0\% \leq pt\_change \leq \max(\text{flow\_change}, \text{high\_change}) \quad (6.3)$$

where

$$\text{low\_change} = \frac{p_{t\_open} - p_{t\_low}}{p_{t\_open}} \times 100\%; \text{high\_change} = \frac{p_{t\_high} - p_{t\_open}}{p_{t\_open}} \times 100\% \quad (6.4)$$

where  $pt\_change$  is the percentage of change extracted from news,  $p_{t\_low}$  is the low of day  $t$  and  $p_{t\_high}$  is the high of day  $t$  and  $p_{t\_open}$  is the opening price of day  $t$ . Any dataset entry with  $pt\_change$  deviating from the day range is removed from the dataset. The purpose of this sanity checking is to ensure that the labels used in training are as accurate as possible. At the very least, they are bounded by their day range.

### 6.3.3 Crude Oil Futures Trend and Returns Prediction

Two sub-tasks are carried out: (1) crude oil price trend prediction and (2) predict crude oil returns prediction.

#### 6.3.3.1 Crude Oil Trend Prediction

A vanilla BERT-based BERTForSequenceClassification head is used for this classification task. The extracted span vectors are fed into the model, BERT's BERTForSequenceClassification head is equipped with a *sigmoid* activation function to predict one of the three classes: UP, DOWN, STEADY. The model is trained to minimize cross-entropy loss.

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(\hat{y}_{ij}) \quad (6.5)$$

where  $N$  is the number of training instances,  $M$  is the number of classes, which is 3 for UP, DOWN, and STEADY,  $y_{ij}$  is the true label and  $\hat{y}_{ij}$  is the classifier's output.



The measurement for the classification task is F1-score and an additional evaluation metric known as the Matthew Correlation Coefficient (MCC)<sup>6</sup> to avoid bias due to the skewness of data; this is similar to (Ding et al., 2014).

### 6.3.3.2 Crude Oil Return Prediction

Similar to the model architecture for trend classification described above, a BERT-based model is used with BERTForSequenceClassification head for regression task by setting the *num\_class* = 1. The model predicts a single scalar value as output. Two common losses are used for price prediction (an example of regression analysis): root mean squared error (RMSE) and mean absolute percentage error (MAPE), which are defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6.6)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \times 100 \quad (6.7)$$

where  $n$  is the total number of data points,  $y_i$  is the actual returns (percentage of change),  $p_{change}^t$  and  $\hat{y}_i$  is the predicted returns.

## 6.4 Experiments

### 6.4.1 Dataset Construction

The dataset is built by first scraping crude oil news from [www.investing.com](http://www.investing.com)<sup>7</sup> dated between January 2011 and December 2019, which resulted in about 15k pieces of market summaries. Key events that happened during the period between 2011 and 2019 include:

- Trade War: US-China trade war since July 2018
- War and Civil Unrest: Arab spring - Syrian civil war in 2011,
- Greek Debt crisis/Greek exit in the early 2010s, Brexit in 2016,

<sup>6</sup>Refer to Section 4.7.2 for more information, including MCC equation.

<sup>7</sup>investing.com is a notable source for finance-related news and is used as the input source for (X. Li et al., 2019; Bai et al., 2022)

- Sanctions: Sanctions of Russia in 2014, sanction of Iran in 2015, abd
- other more cyclical/regular events such as crude oil shortage/oversupply etc.

The dataset is split into Train, Validation and Test as follows (and also shown diagrammatically in Figure 6.4):

1. Training set: January 2011 - June 2016 (66 months) consisting of 8,763 news;
2. Validation set: July 2016 - March 2018 (21 months) consisting of 3,078 news;
3. Test set: April 2018 - December 2019 (21 months) consisting of 2,979 news.

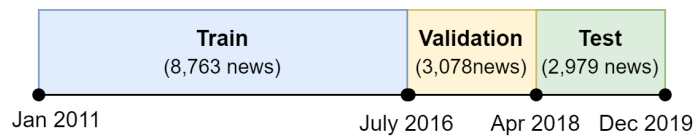


Figure 6.4: Train-test split.

## 6.4.2 Experimental Setup

Three types of experiments are conducted:

1. Comparison of text processing techniques (Section 6.4.2.1)
2. Comparison of content scope: News headlines versus News body (Section 6.4.2.2)
3. Ablation Study (Section 6.4.2.3)

### 6.4.2.1 Comparison of text processing techniques

The proposed solution differs from existing crude oil forecasting in the following aspects:

**Labels' or dependent variable** The proposed solution relies solely on market summaries to obtain 'labels' for the prediction task, while in other solutions, historical price data (time series data) is used.

**Frequency** All existing text-based crude oil forecasting works make price forecasting at the daily or weekly frequency, essentially averaging daily / weekly fluctuations to a single data point. This work, on the other hand, uses a market summary as a single data point.

As highlighted in Figure 6.3, throughout a single day, there are multiple data points. For the training set from the period of Jan 2011 to June 2016, there are only 1,433 daily closing price<sup>8</sup> while here, there are 8,763 pieces of market summaries in this period. Therefore, this solution is able to forecast market reactions at a finer level rather than a daily / weekly average.

Due to these significant differences, conducting a like-for-like comparison with existing solutions is not feasible. Instead, the focus is on implementing existing text processing techniques to determine the best text processing approach for market summaries.

Table 6.4: Examples illustrating the difference in information granularity produced by each event extraction method based on this sentence: *\Brent oil futures for December delivery were down 0.42% and trading at USD107.64 a barrel."*

Methods	Information Granularity
OpenIE / Numerical OpenIE	Arg1: Brent oil futures for December Predicate: were (v) Arg2: down 0.42%
Frame-semantic Parsing	[ <i>Direction</i> frame] Figure: Brent oil futures for December delivery Distance: 0.42%
Fine-grained Event+	Item: Brent oil Attribute: futures Contract: December Difference: 0.42% Final value: USD107.64 a barrel Polarity: POSITIVE Modality: ASSERTED Intensity: NEUTRAL

**Results and Analysis** Two categories of text processing are compared here: Category A- Non-Event-based methods and Category B- Event-Based methods:

1. Category A: Non-event based

- Sentiment Analysis in (Feuerriegel & Neumann, 2013; J. Li et al., 2017; Sadik et al., 2020; X. Li et al., 2019; Zhao, Zeng, et al., 2019; Bai et al., 2022)
- Topic Modeling in (X. Li et al., 2019)
- A combination of Sentiment Analysis and Topic Modeling in (T. H. Nguyen & Shirai, 2015)

2. Category B: Event-based

<sup>8</sup>commodities are not traded over the weekend; hence there is no closing price for the weekends.

- Coarse-grained Event via OpenIE (Mausam, 2016) used in (Ding et al., 2014, 2015) <sup>9</sup>
- Numerical OpenIE (Saha et al., 2017) <sup>10</sup>
- Fine-grained Events via Semantic Frame parsing (SEMAFOR (Das & Smith, 2011, 2012)) in (Xie et al., 2013)) <sup>11</sup>
- Fine-grained Events with event properties classification : Fine-grain events+ (proposed solution)

Table 6.4 shows the different information granularity of event details of Event-based methods.

Table 6.5: Comparing Classification and Regression Analysis Results with other text processing techniques. Category A is non-event methods while Category B is event-based methods and is further divided into B1: coarse-grained events and B2: fine-grained events.

	Text Processing Method	Trend		WTI		BRENT	
		F1	MCC	RMSE	MAPE	RMSE	MAPE
<b>A</b>	Sentiment Analysis using TextBlob (Polarity <sup>12</sup> and Subjectivity Score) (X. Li et al., 2019)	0.456	0.254	14.492	46.59%	13.561	47.05%
	Topic Modeling (Blei et al., 2003)	0.412	0.201	11.104	46.78%	10.12	42.13%
	Sentiment Analysis + Topic Modeling (Blei et al., 2003)	0.501	0.198	8.976	36.17%	7.778	34.54%
<b>B1</b>	OpenIE (Etzioni et al., 2008)	0.421	0.243	8.270	44.40%	7.564	36.58%
	Numerical OpenIE (Saha et al., 2017))	0.491	0.255	12.387	58.92%	12.154	56.55%
<b>B2</b>	Frame-semantic Parsing (SEMAFOR (Das & Smith, 2011, 2012))	0.513	0.345	15.047	50.54%	15.975	51.01%
	Fine-grained Events+ (proposed solution)	<b>0.755</b>	<b>0.554</b>	<b>1.951</b>	<b>10.12%</b>	<b>1.113</b>	<b>9.89%</b>

Based on the results shown in Table 6.5, non-event methods performed worse than event-based methods as expected. This situation is mainly because the non-event approaches do not provide adequate semantic information useful enough for the prediction task. Sentiment Analysis and Topic Modeling individually produced rather dismal results, while a combination of both produced slightly improved results. Among the event-based models, it is clear that fine-grained events provide the best results. This observation is consistent with the conclusion reached by (D. Chen, Zou, et al., 2019) and supported by the fact that fine-grained events contain much more information than coarse-grained events. This is clearly shown in the level of event details generated by each event-based text processing approach in Table 6.4. The input text generated by both OpenIE and Numerical OpenIE is restricted to only

<sup>9</sup>OpenIE functionality in AllenNLP is used, codes here: <https://github.com/allenai/allennlp>

<sup>10</sup>the codes here is used: <https://github.com/dair-iitd/OpenIE-standalone>

<sup>11</sup>the latest version of SEMAFOR using the codes here is used: <https://github.com/swabhs/open-sesame>.

<sup>12</sup>Polarity in Sentiment Analysis is not to be confused with Event Polarity: Sentiment Polarity is Positive or Negative sentiment while Event Polarity Positive indicates event actually happened, while Negative means event did not happen.

extracting just one single numerical argument, while Frame-semantic parsing approach is constrained to frames that do not cater to extracting extra numerical arguments than those already defined in the frame.

#### 6.4.2.2 Contents: News headlines versus News body

Thus far, there is no consensus on whether news headlines or news body is the best input for financial asset prediction task. Some works advocate that headlines are concise enough to capture all needed semantic information (Ding et al., 2014, 2015). In contrast, others strongly argue that headlines are too brief and only the news body captures the right amount of information (Wu, Wang, Lv, & Zeng, 2021). It is hypothesized here that headlines are too brief to capture essential information. This is proved through experiments, and experiment outcomes are captured in Table 6.6.

Table 6.6: Results of using varying contents of market summaries.

Scope of content	Trend		WTI		BRENT	
	F1	MCC	RMSE	MAPE	RMSE	MAPE
News body (proposed solution)	<b>0.755</b>	<b>0.554</b>	<b>1.951</b>	<b>10.12%</b>	<b>1.913</b>	<b>9.89%</b>
Headline	0.454	0.345	7.251	26.97%	6.991	25.54%
Headline + body	0.624	0.475	4.523	13.21%	5.122	10.12%

**Results and Analysis** Based on the experiment results, it is clear that the news body produced the best result in terms of contents. It is the same conclusion made by the authors in (Shi et al., 2018) that news headlines can be arbitrary, noisy, and ambiguous as they contain only a few words. For market summaries, some headlines that are too generic and, therefore not very informative. For instance, *Crude Oil Prices-Weekly Outlook November 19*. In other cases, headlines cover only a portion of breaking events but not all of them. The example headline shown in Table 6.2 (see first two rows under ‘H’) only captures one of the two events (U.S.-Saudi Tension Dissipates) - likely the more prominent event. Based on the results, it is concluded that the best input for market summaries is news content.

#### 6.4.2.3 Ablation Study

An ablation study is conducted to understand the significance each component within *Fine-grained Event+* plays in contributing to the performance of crude oil prediction.

Table 6.7: Ablation Study to understand the significance each component plays in providing semantic richness for crude oil prediction. \* indicates the proposed solution. Performance differences using proposed solution as benchmark is captured in parenthesis.

Information Component	Trend		WTI		BRENT	
	F1	MCC	RMSE	MAPE	RMSE	MAPE
Event Trigger+Arguments+Properties*	<b>0.755</b>	<b>0.554</b>	<b>1.951</b>	<b>10.12%</b>	<b>1.913</b>	<b>9.89%</b>
Event Trigger+Arguments - Properties	(# 0.345)	(# 0.345)	(" 2.421)	(" 12.25)	(" 3.54)	(" 15.09)
Event Trigger+Properties - Arguments	(# 0.042)	(# 0.124)	(" 8.151)	(" 38.54)	(" 8.115)	(" 37.92)
Event Trigger - Properties - Arguments	(# 0.375)	(# 0.405)	(" 8.291)	(" 45.55)	(" 10.552)	(" 40.53)

**Results and Analysis** From the results shown in Table 6.7, the exclusion of Event Properties contributed to the most significant deterioration in the Trend Classification task but had a small impact on WTI and Brent Return Prediction. The description in Section 6.3.1 supports this: event properties are needed to determine the ‘factuality’ or factual certainty of events. For example, ‘**supplies cut**’ correlates to crude oil price moving UP, however ‘**cancellation of supplies cut**’ (with an opposite polarity) correlates to crude oil price moving the opposite direction - DOWN. From this, it is clear that plain event details (sans event properties) are an incomplete and inaccurate presentation of events. On the other hand, the exclusion of Arguments has a noticeable impact on WTI and Brent returns prediction and a negligible impact on trend classification task. It is because event details contribute to the model determining the magnitude of price change. For example **supplies cut of 1 million barrels** versus **supplies cut of 10,000 barrels** will have different levels of impact on crude oil. From this ablation study, it can be concluded that each component within *Fine-grained Event+* is needed for a complete and accurate representation of events for building accurate forecasting models.

#### 6.4.2.4 Solution Robustness

The proposed solution uses an event extraction model trained on the *CrudeOilNews* corpus, a ACE/ERE-like annotated dataset. One of the biggest drawbacks of this supervised approach is that the event type coverage is constrained by predefined event typology, and model performance is constrained by the availability of annotated data. New events not part of the predefined event typology, known as ‘unseen events’, will not be picked up. For example, the *CrudeOilNews* corpus does not contain any Covid-19 pandemic annotations, so any models trained on this dataset via supervised learning will be unable to extract any Covid-19 events. Recently, (Wu, Wang, Wang, & Zeng, 2021) showed that the Covid-19 pandemic had a huge impact on global oil price, oil production, and oil consumption. To test the robustness of

the proposed solution over a period of time with ‘unseen events’, the entire workflow is run on a new collection of market summaries, named *Covid-19* set, from 1st April 2020 to 31 December 2021<sup>13</sup> (same duration as the test set as described in Section 6.4.1). This period (1st April 2020 to 31 December 2021) is chosen to align with the test set which is from the same months, which is from 1st April 2018 to December 2019.

Figure 6.5 shows the event distribution of both the Test set and *Covid-19* set. In the *Covid-19* test set, there is a significant increase in ‘Oversupply’, ‘Slow-Weak’ (Economy, Demand), and a reduction in geo-political events such as ‘Trade tensions’, ‘Geo-political tensions’ and ‘Civil unrest’. It is understandable given that during the pandemic period, most of the world remain in a ‘lockdown’ state.

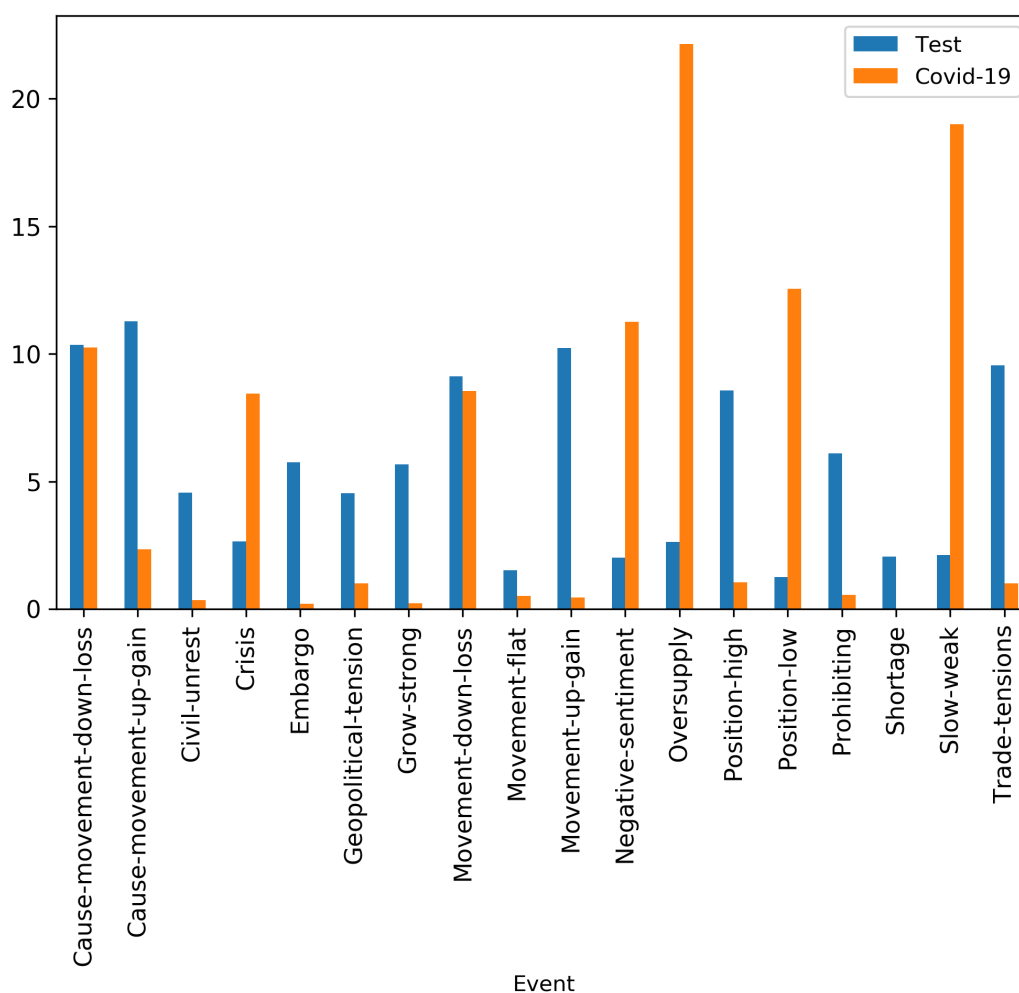


Figure 6.5: Event distribution of the Test set and *Covid-19* set.

From the results shown in Table 6.8, the overall performance during the pandemic period (April 2020 - Dec 31 2021) has dropped as compared to the test data (April 2018 - Dec

<sup>13</sup>The World Health Organization declared Covid-19 as a pandemic on 11 March 2020.

Table 6.8: Additional experiment to test the robustness of proposed solution using market summaries with ‘unseen events’ during Covid-19 pandemic. The difference in performance between the test set and the *Covid-19* set is in parenthesis.

Dataset	Trend		WTI		BRENT	
	F1	MCC	RMSE	MAPE	RMSE	MAPE
<i>Covid-19</i> set (1st April 2020 - 31 December 2021)	0.590	0.489	14.569	46.04%	10.245	35.95%
Performance difference with test result (1st April 2018 - December 2019)	(# 0.165)	(# 0.065)	(" 12.618)	(" 35.90)	(" 8.332)	(" 26.06)

2019). This drop in performance is expected due to the existence of ‘unseen events’ such as Covid-19-related ones. However, rather than a totally dysfunctional model, the model still performs with a reasonable F1-score for the trend classification task. It can be seen from the results that there is only a slight drop in performance in the trend prediction task but a more significant drop in performance in the returns prediction task. It is because even though ‘unseen’ events are not captured, the majority of crude oil-related events are. This phenomena is illustrated in Figure 6.6. Even though ‘Covid-19 pandemic is not extracted, the majority of resultant or consequential crude-oil-related events are; these include *IEA cuts oil demand forecast*, *air travel outlook darkens*, and other price-related events are extracted.

WTI and Brent returns prediction performs poorly compared to trend prediction because during the pandemic, crude oil markets became extremely volatile with huge fluctuations that were not experienced before. For example, on 20th April 2020, the front-month May 2020 WTI crude contract dropped 306% on the New York Mercantile Exchange. It is the first time in history that any benchmark fell below zero, making U.S. oil not only worthless but a liability <sup>14</sup>. The classification model is able to predict the right direction, but unfortunately, the regression model did poorly in predicting the returns. This massive one-day change (either rise or drop) did not occur during the training period. Therefore such extreme values are out of the model’s prediction range.

### 6.4.3 Overall Analysis

It is shown through rigorous experiments that for better prediction results (1) structured events perform better than other non-event-based methods; (2) among event-based input, *Fine-grain Event+* performs the best followed by fine-grained events, and lastly by coarse-grained events; (3) News body provide more event details and information than headlines for

<sup>14</sup><https://oilprice.com/Energy/Crude-Oil/Could-Brent-Crude-Oil-Prices-Ever-Fall-Into-Negative-Territory.html>



Oil **plunged** as Air Travel Outlook **Darkens**  
 Crude oil futures **plunged** on Friday when IEA **cuts** oil demand forecasts as air travel outlook **darkens** globally caused by the Covid-19 pandemic. On the ICE Futures Exchange, Brent oil futures for May delivery were **down** 5.89% and trading at USD35.54 a barrel. On the New York Mercantile Exchange, light, sweet crude futures for delivery in May traded at USD29.56 a barrel on Tuesday, **down** 8.75%.

Figure 6.6: An example of Crude Oil Market Summary with ‘unseen events’; event triggers that are identified by the model are in bold. ‘Unseen events’ such as ‘Covid-19 pandemic’ is undetected.

crude oil market summaries; (4) Among the components that make up fine-grained event+, event properties contribute significantly to the accuracy of price trend prediction (classification task); and (5) event arguments play an important role in the accuracy of price return prediction (regression analysis task). It is not conclusive whether the models are better at predicting WTI or BRENT futures; both have more or less the same error rate. It could be because WTI and BRENT both move in tandem. It must be acknowledged that the accuracy of prediction is tied to the accuracy of the dataset, which in turn is dependent on the accuracy of the event extraction model. Based on the error analysis, it is discovered that there is still room for improvement in event extraction. The common errors are (i) misclassification of event type, (ii) missed out arguments, and (iii) misclassification of event properties. Therefore, forecasting accuracy can be further improved when event extraction model accuracy is improved. In evaluating solution robustness with ‘unseen events’, the experiment also shows that the proposed solution is robust enough for crude oil trend classification but not as robust in returns prediction.

## 6.5 Summary and Discussion

As part of Research Objective #3, two crude oil forecasting models were trained: (1) a model to predict crude oil price trend; (2) a model to predict the returns (percentage of price change) for Brent and WTI. It is a novel approach where crude oil market summaries are presented as a new and a standalone text source for crude oil prediction. It is different from earlier forecasting work where historical price data is used. Based on the best of our knowledge, this work is the first to use crude oil market summaries to mine for strong correlation between events and market reactions and utilize this information for crude oil prediction tasks. This

work is also the first to predict not just price trend but also the percentage of price change for both crude oil benchmarks (WTI and Brent). For a complete and accurate representation of events, *Fine-grained Events+* is used; it does not just capture event details but also event ‘factuality’ in the form of Event Polarity, Modality, and Intensity. Experiment results show that this is a promising solution.

## Chapter 7

# Conclusions and Future Work

The following sections will recap the work carried out in relation to the research objectives (ROn) and research contributions (RCn) set out in Chapter 1. The final section will outline some ideas for possible further improvements and extensions.

### 7.1 Summary

The main contribution of this research is to build a solution to extract events from crude oil market summaries and utilize these events effectively for crude oil forecasting. The contribution can be broken down into three sub-contributions, namely:

1. *CrudeOilNews*, a new crude oil news corpus annotated for the task of event extraction, is produced. This contribute towards language resource building in the Finance and Economics domain;
2. A new event extraction frame work fit-for-purpose for *CrudeOilNews* corpus is introduced. In terms of solution architecture, event extraction models are fine-tuned from ComBERT rather than trained from scratch for Entity Mention Detection (EMD), Event Detection (ED) and event properties classification. Graph Convolution Network (GCN) is used together with contextual sub-tree to extract event arguments (ARP) effectively. In terms of training approach, an ensemble of Transfer Learning approaches is also used to address the issue of class imbalance and to generate models with better performance than those trained via Supervised Training alone;

3. Two crude oil forecasting models are trained: (1) crude oil price trend (UP, DOWN, and STABLE); and (2) returns (percentage of price change) for WTI and Brent. The models are trained using *Fine-grained Events+*, which are made up of fine-grained events as well as event properties (Polarity, Modality, and Intensity).

## 7.2 Limitations and Future Work

### 7.2.1 Limitations

#### 7.2.1.1 Limited Event Coverage

One of the weaknesses of closed domain event extraction based on pre-defined event typology is that the coverage of events is limited to only pre-defined ones. The performance of models built via Supervised Learning is also constrained by the quality and size of available datasets. Pre-defined event typology is exhaustive and models will not automatically generalize to extract new events. This can be address as part of future work where open domain event extraction, such as using distant supervision in (Dor et al., 2019) or zero-shot learning (Huang et al., 2018) can be used to expand the coverage of events without the need for an expensive human-labeled dataset.

#### 7.2.1.2 Wrong Correlations

The second drawback is that, at the moment, the text mining for crude oil forecasting is based on the correlation of global events and market reaction by way of identifying co-occurring events at the sentence level. The assumption here is that price change (market reaction) is the outcome of a non-price-related event. Most of the time, this assumption holds true, however, there are situations when this assumption is not valid.

**'Market Reactions' may not be the outcome** The first issue is that market reaction in the form of price movement may not always be the outcome of other events. There are counter-examples where price change reported in market summaries is not the outcome but a cause for other events.

- (25) **Rising** oil prices put U.S. driving recovery at risk.

In (25) above, the rise of oil price is not an outcome but the causal event. It would be a wrong interpretation of the sentence if ‘driving recovery at risk’ is treated as the predictor variable and ‘oil price rise’ as the target variable.

**Contradiction** Another issue is Contradictions. Some contradictory words like *despite*, *however*, *even though*, *although*, *in spite of* are used to link two contrasting ideas or to show that one fact makes the other fact surprising.

- (26) Oil prices **climbed** on Monday despite **weaker** fuel demand due to growing coronavirus infections and **higher** production in Libya.

In (26), the rise in crude oil prices is not the expected outcome of weaker demand and higher production; in fact, the logical expectation is the exact opposite. Hence it would be erroneous to treat ‘price rise’ as the target variable for ‘weak demand’ and ‘higher production’ events.

These two issues can be addressed in future work by mining events that are truly causal and not merely at the correlation level. This can be done by including event-event causal relation extraction into the scope.

## 7.2.2 Future Work

### 7.2.2.1 Other Financial Assets

Crude oil forecasting experiment results in Section 6 show that the event-based approach to crude oil forecasting is a promising and robust solution. Even though this work focuses on crude oil, the proposed framework can be extended to other financial assets such as stocks, Forex, and even other commodities. A promising direction is to investigate domain adaptation or zero-shot transfer learning of event extraction from crude oil news to other financial assets.

### 7.2.2.2 Other Uses of Events

Structured events extracted from news have shown to be helpful in stock price prediction (as part of Literature Review in Section 2.1.2) and also in crude oil forecasting (Chapter 6). Apart from these, events have been used in other generic areas such as:

1. Summarization of single documents ([Marujo et al., 2017](#)) or multiple documents ([Glavaš & Šnajder, 2014](#));
2. Financial Risk analysis ([F. Hogenboom et al., 2015](#));
3. Social-economic Indicators prediction ([Chakraborty et al., 2016](#))

As part of future work, other practical uses of commodity events can be explored. This includes (1) financial news summarization, and (2) understanding event chains and learning event sequences, also known as scripts in ([Schank & Abelson, 2013](#)) for more accurate next-event prediction.

## 7.3 Conclusion

The research presented in this thesis is concerned with the building of *CrudeOilNews* corpus (RO1), a language resource for crude oil news event extraction. This resource is used to train an event extraction model (RO2). Extracted events are then used in training crude oil forecast models (RO3): one for trend prediction (multi-class classification) and one for returns prediction (regression analysis and classification). The outcomes of this research are able to meet all research objectives laid out in Chapter 1. Despite its limitations, the proposed solution is proven to be effective in crude oil forecasting. This novel approach adds to the list of applications of Natural Language Processing in financial asset forecasting. Along with the limitations identified in Section 7.2, potential enhancements have also been identified and are to be addressed as future work.

# Appendix A

## *CrudeOilNews* Corpus

### A.1 Event Schema

#### A.1.1 Movement-down-loss, Movement-up-gain, Movement-flat

**Example sentence:** [Globally] [crude oil] [futures] **surged** [\$2.50] to [\$59 per barrel] on [Tuesday].

Role	Entity Type	Argument Text
Type	Nationality, Location	globally
Place	Country, Group, Organization, Location, State or province, Nationality	
Supplier_consumer	Organization, Country, State_or_province, Group, Location	
Reference_point_time	Date	Tuesday
Initial_reference_point	Date	
Final_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	\$59 per barrel
Initial_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	
Item	Commodity, Economic_item	crude oil
Attribute	Financial_attribute	futures
Difference	Percentage, Number, Money, Production_unit, Quantity	\$2.50
Forecast	Forecast_target	
Duration	Duration	
Forecaster	Organization	

#### A.1.2 Caused-movement-down-loss, Caused-movement-up-gain

**Example sentence:** The [IMF] earlier said it **reduced** its [2018] [global] [economic growth] [forecast] to [3.30%] from a [July] forecast of [4.10%].

Role	Entity Type	Argument Text
Type	Nationality, Location	global
Place	Country, Group, Organization, Location, State or province, Nationality	West African, European
Supplier_consumer	Organization, Country, State_or_province, Group, Location	
Reference_point_time	Date	2018
Initial_reference_point	Date	July
Final_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	3.30%
Initial_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	4.10%
Item	Commodity, Economic_item	economic growth
Attribute	Financial_attribute	
Difference	Percentage, Number, Money, Production_unit, Quantity	
Forecast	Forecast_target	forecast
Duration	Duration	
Forecaster	Organization	IMF

### A.1.3 Position-high, Position-low

**Example sentence:** The IEA estimates that U.S. crude oil is expected to seek higher ground until reaching a [5-year] **peak** in [late April] of about [17 million bpd].

Role	Entity Type	Argument Text
Reference_point_time	Date	late April
Initial_reference_point	Date	
Final_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	17 million bpd
Initial_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	
Item	Commodity, Economic_item	
Attribute	Financial_attribute	
Difference	Percentage, Number, Money, Production_unit, Quantity	
Duration	Duration	5-year



### A.1.4 Slow-weak, Grow-strong

**Example sentence:** [U.S.] [employment data] **strengthens** with the euro zone.

Role	Entity Type	Argument Text
Type	Nationality, Location	
Place	Country, Group, Organization, Location, State or province, Nationality	U.S.
Supplier_consumer	Organization, Country, State_or_province, Group, Location	
Reference_point_time	Date	
Initial_reference_point	Date	
Final_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	
Initial_value	Percentage, Number, Money, Price_unit, Production_unit, Quantity	
Item	Commodity, Economic_item	employment data
Attribute	Financial_attribute	
Difference	Percentage, Number, Money, Production_unit, Quantity	
Forecast	Forecast_target	
Duration	Duration	
Forecaster	Organization	

### A.1.5 Prohibiting

**Example sentence:** [Congress] **banned** most [U.S.] [crude oil] [exports] on [Friday] after price shocks from the 1973 Arab oil embargo.

Role	Entity Type	Argument Text
Imposer	Organization, Country, Nationality, State or province, Person, Group, Location	Congress
Imposee	Organization, Country, Nationality, State or province, Group	U.S.
Item	Commodity, Economic_item	crude oil
Attribute	Financial_attribute	exports
Reference_point_time	Date	Friday
Activity	Other_activities	

### A.1.6 Oversupply

**Example sentence:** [Forecasts] for an [crude] **oversupply** in [West African] and [European] [markets] [early June] help to push the Brent benchmark down more than 20% January.

Role	Entity Type	Argument Text
Place	Country, Group, Organization, Location, State or province, Nationality	West African, European
Reference_point_time	Date	this year
Item	Commodity	crude
Attribute	Financial_attribute	markets
Difference	Production_unit	
Forecast	Forecast_target	forecasts

### A.1.7 Shortage

**Example Sentence:** Oil reserves are within “acceptable” range in most oil consuming countries and there is no **shortage** in [oil] [supply] [globally], the minister added.

Role	Entity Type	Argument Text
Place	Country, State or province, Location, Nationality	Congress
Item	Commodity	crude oil
Attribute	Financial_attribute	exports
Type	Location	globally
Reference_point_time	Date	

### A.1.8 Civil Unrest

**Example sentence:** The drop in oil prices to their lowest in two years has caught many observers off guard, coming against a backdrop of the worst **violence** in [Iraq] [this decade].

Role	Entity Type	Argument Text
Place	Country, State or province, Location, Nationality	Iraq
Reference_point_time	Date	this decade

### A.1.9 Embargo

**Example sentence:** The [Trump administration] imposed a “strong and swift” economic **sanctions** on [Venezuela] on [Thursday].

Role	Entity Type	Argument Text
Imposer	Organization, Country, Nationality, State or province, Person, Group, Location	Trump administration
Imposee	Organization, Country, Nationality, State or province, Group	Venezuela
Reference_point_time	Date	Thursday

Note: ‘Imposee’ is not formally a word, but used here as a shorter version of “Party whom the action was imposed on.”

### A.1.10 Geo-political Tension

**Example sentence:** Deteriorating relations between [Iraq] and [Russia] [first half of 2016] ignited new fears of supply restrictions in the market.

Role	Entity Type	Argument Text
Participating_countries	Country, Group, Organization, Location, State or province, Nationality	U.S., China
Reference_point_time	Date	early June

### A.1.11 Crisis

**Example Sentence:** Asia 's diesel consumption is expected to recover this year at the second weakest level rate since the [2014] [Asian] [financial] **crisis**.

Role	Entity Type	Argument Text
Place	Country, State or province, Location, Nationality	Asian
Reference_point_time	Date	this year
Item	Commodity, Economic_item	nancial

### A.1.12 Negative Sentiment

**Example sentence:** Oil futures have dropped due to **concern** about softening demand growth and awash in crude.

Note: **Negative Sentiment** is a special type of event, where majority of the time it contains just the trigger words such as *concerns*, *worries*, *fears* and 0 event arguments.

## A.2 Event Types, Distribution and Examples

### A.3 RavenPack Event Taxonomy

Authors in (Brandt & Gao, 2019) grouped Ravenpack's Event Taxonomy into three main categories.

---

<b><u>Geo-political News:</u></b>		
Terrorism	War & Conflict	Civil unrest
Natural disasters	Government	

---

<b><u>Macro-economic News:</u></b>		
Sovereign Debt	Public finance	Retail sales
Consumer confidence	Housing	Interest rates
Treasury yield	Durable goods orders	Consumer spending
Recession	Economic growth	GDP growth
CPI	PPI	Trade balance
Exports	Foreign exchange	Employment
Private credit		

---

<b><u>Oil supply and demand:</u></b>		
Crude oil supply	Crude oil demand	Price Target
Drilling & pipeline accident		

---

Table A.1: Categories of RavenPack classifications grouped into three broader classes: (1) Geo-political news, (2) Macro-economic news and (3) Oil supply and demand news

# Appendix B

## Source Dataset

Here is the list of Source Datasets used in Cross-domain Sequential Transfer Learning in Section 5.4. Each dataset is accompanied with a short description of the intended task the corpus is designed for and its key statistics.

### B.1 ConanDoyle(neg)

The ConanDoyle-neg (Morante and Daelemans 2012) is a corpus of Conan Doyle stories annotated with negation cues and their scopes, as well as the event or property that is negated. It is composed of 3,640 sentences from The Hound of the Baskervilles story, out of which 850 contain negations, and 783 sentences from The Adventure of Wisteria Lodge story, out of which 145 contain negations. In this case, the three types of negation cues (lexical, syntactic, and morphological) were taken into account.

### B.2 SOCC(neg)

This is the SFU Opinion and Comments Corpus (SOCC) introduced by (Kolhatkar et al., 2020) in 2019. The original corpus contains 10,339 opinion articles (editorials, columns, and op-eds) together with their 663,173 comments from 303,665 comment threads, from the main Canadian daily newspaper in English, The Globe and Mail, for a five-year period (from January 2012 to December 2016). The corpus is organized into three subcorpora: the

articles corpus, the comments corpus, and the comment-threads corpus. Only the articles sub-corpora is used for this work.

### B.3 10kFinStatement(unc)

Dataset introduced in (Theil et al., 2018), contains 1000 sentences taken from 10-Ks<sup>1</sup> each labeled with *certain* and *uncertain* but does not have uncertainty cue words nor scope annotated.

### B.4 Wikipedia-CoNLL2010(unc)

This dataset is provided as part of CoNLL2010 shared task - Learning to Detect Hedges and their Scope in Natural Language Text. 2186 paragraphs collected from Wikipedia archives were also offered as Task1 training data (11111 sentences containing 2484 uncertain ones). The evaluation dataset contained 2346 Wikipedia paragraphs with 9634 sentences, out of which 2234 were uncertain.

### B.5 Reviews(neg & unc)

This is the review dataset introduced by (Konstantinova et al., 2012), it contains annotation of negation and speculation. This corpus consists of 400 documents (50 of each type) of movie, book, and consumer product reviews from the website Epinions.com.

### B.6 SENTiVENT

This dataset is released as part of (Jacobs & Hoste, 2021) and it is a corpus with company financial news annotation in ACE/ERE-like manner.

---

<sup>1</sup>A 10-K is a comprehensive report filed annually by public companies about their financial performance.

## Appendix C

# Background Information & Domain Knowledge

### C.1 General Domain Event Extraction

#### C.1.1 Canonical Event Extraction Program

The main event extraction programs are (1) ACE2005<sup>1</sup> under ACE and (2) TAC-KBP Event track shared task running from 2015-2017<sup>2</sup>. In ACE, there were three primary ACE annotation tasks corresponding to the three research objectives: Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC) and Event Detection and Characterization (EDC). In the mean time, TAC-KBP corpus is annotated based on Rich ERE (Entities, Relations, and Events) standard guidelines and provides an event corpus similar to ACE2005 corpus. According to ACE2005 annotation guidelines, an **event** in ACE is represented by an **event trigger**, an **event type**, and a set of **arguments** with different roles. The goal of event extraction is to identify event triggers with specific types and arguments with specific roles.

**ACE2005 Polarity, Tense, Genericity and Modality** Apart from events and entity relation annotations, each event is annotated with Polarity, Tense, Genericity, and Modality. They are described as follows:

---

<sup>1</sup><https://projects.ldc.upenn.edu/ace>

<sup>2</sup><https://tac.nist.gov/2015/KBP/>



- **Polarity:** An Event has the value *positive* unless there is an explicit indication that the event did not take place, in which case *Negative* is assigned.
- **Tense:** specifies the tense of the event with respect to the author, and can be *Past*, *Present*, *Future* or *Unspecified* (where the tense cannot be determined from the context).
- **Genericity:** has either the *Specific* value, if the event can be understood as describing a singular occurrence at a particular place and time, or a finite set of such occurrences, or *Generic* otherwise.
- **Modality:** Determines whether the event represents a “real” occurrence. There are two possible values: *Asserted* if the author or speaker refers to it as though it were a real occurrence, and *other* otherwise. *Others* could include:

{ Believed Events

Rumors of **arrest**, suspected of **giving** money

{ Hypothetical Events

if..., should he not **pay**

{ Commanded and Requested Events

He was asked to **return**, Iraq was ordered to **cut**

{ Threatened, Proposed, and Discussed Events.

US threatened to **sanction** Iran.

{ Desired Events

China wanted to **increase** production.

{ Promised Events

OPEC agreed to **cut** supplies.

{ Unclear or complicated sentence constructs.

**TAC-KBP: Realis** Rich ERE introduced the Realis attribute in TAC-KBP 2015 Event Nugget Detection task. In the corpus, each event is tagged with one of each of Realis attributes:

- **ACTUAL:** events that actually occurred / attested events.
- **GENERIC:** events that are not specific events with a (known or unknown) time and/or place
- **OTHER:** all other events, including failed events, future events, and conditional statements, and all other non-generic variations.

### C.1.2 Event Extraction

Definitions :

- **Event:** An event is a specific occurrence involving participants, which are entities that are involved in that event.
- **Entity:** an object or a set of objects in one of the semantic categories of interest
- **Entity mention:** a reference to an entity (typically, a noun phrase).
- **Event trigger:** the main word that most clearly expresses an event occurrence.
- **Event argument:** an entity mention, temporal expression or value that is involved in an event (participants).
- **Argument role:** the relationship between an argument and the event in which it participates.
- **Event type:** the semantic type of an event, which has its own set of potential argument roles.
- **Event mention:** a phrase or sentence within which an event is described, including a trigger and arguments.
- **Event Extraction (EE):** is a traditional task in Information Extraction (IE) which aims at extracting event mentions of specific types and their corresponding event arguments with their argument roles (Argument Detection) from unstructured text.

**Event Extraction Subtasks** Event extraction is broken down to the following subtasks:

1. **Entity Mention Detection (EMD):** a task to detect entity mentions (named or nominal) and assign each token an entity type or NONE for tokens that is not an entity mention.
2. **Event Extraction :**
  - (a) **Event Detection (ED):** similar to EMD, it is a task to detect event trigger word(s) and assign it to an event type or NONE for tokens that is not an event trigger.
  - (b) **Argument Role Prediction (ARP):** a task aims to assign an argument role label or NONE to a candidate entity mention.
3. **Event Properties Classification:** Classifying each event's in terms of their Polarity, Modality and Intensity classes.

**A Working Example** The whole end-to-end process of event extraction is explained using the example shown in Figure C.1.

*World oil prices are set to fall further , extending a months-long rout as Saudi Arabia is unlikely to make deep enough production cuts to erase a growing surplus of supply.*

Figure C.1: This is an example taken from *CrudeOilNews* corpus, trigger words are underlined.

The entity mentions extracted via EMD and event triggers extracted through ED are tabulated in Table C.1, while argument roles entity mentions plays are shown in Table C.2. Event properties of each event are shown in Table C.3.

Table C.1: Entity Mention Detection (EMD) and Event Detection (ED) for the example sentence shown in Figure C.1.

Task	Text	Entity Type
EMD	World	LOCATION
	oil	COMMODITY
	prices	FINANCIAL ATTRIBUTE
	months-long	DATE
	Saudi Arabia	COUNTRY
	production	FINANCIAL ATTRIBUTE
	supply	FINANCIAL ATTRIBUTE
Task	Trigger Word(s)	Event Type
ED	fall	MOVEMENT_DOWN_LOSS
	rout	SLOW-WEAK
	cuts	CAUSE MOVEMENT_DOWN_LOSS
	surplus	OVERSUPPLY

Table C.2: Argument Role Prediction (ARP) for the example sentence shown in Figure C.1.

Task	Event	Text	Argument Role
ARP	<u>fall</u>	oil	ITEM
		prices	ATTRIBUTE
	<u>rout</u>	months-long	DURATION
	<u>cuts</u>	Saudi Arabia	SUPPLIER
		production	ATTRIBUTE
	<u>surplus</u>	supply	ATTRIBUTE

Table C.3: Event properties (Polarity, Modality and Intensity) for each of the events from the example sentence in Figure C.1.

Event	Polarity	Modality	Intensity
prices <u>fall</u>	POSITIVE	OTHER	NEUTRAL
<u>rout</u>	POSITIVE	ASSERTED	INTENSIFIED
production <u>cuts</u>	NEGATIVE	OTHER	NEUTRAL
supply <u>surplus</u>	POSITIVE	OTHER	EASED

## C.2 Crude Oil-related Terminologies

Listed below are crude oil-related terminologies and their definitions:

1. **WTI**: West Texas Intermediate crude is a specific grade of crude oil and one of the main three benchmarks in oil pricing, along with Brent and Dubai Crude. WTI is known as a light sweet oil.
2. **Brent**: Brent crude is a specific grade of crude oil and one of the main three benchmarks in oil pricing, along with Brent and Dubai Crude. WTI is known as a light sweet oil.
3. **NYMEX**: New York Mercantile Exchange is where WTI crude is traded.
4. **ICE**: Intercontinental Exchange is where Brent crude is traded.
5. **futures contract** - An oil futures contract is an agreement to buy or sell a certain number of barrels of oil at a predetermined price, on a predetermined date.
6. **futures price** (or mostly just ‘futures’) - Throughout this work, the term price and futures are used interchangeably to mean the same thing. They are not to be confused with spot price which is not used here.
7. **front month contract** - it is also called “near” or “spot” month, refers to the nearest expiration date for a futures contract.
8. **return** - the change in price of an asset, investment, or project over time, which may be represented in terms of price change or percentage change.
9. **technical analysis** - this analysis deals with the time-series historic market data, such as trading price and volume, and make predictions based on that. The main goal of this type of approach is to discover the trading patterns that can be leveraged for future prediction. One of the most widely used model in this direction is the Autoregressive (AR) model for linear and stationary time-series.
10. **fundamental analysis** - is a method of measuring a security’s intrinsic value by examining related economic and financial factors. Fundamental analysts study anything that can affect the security’s value, from macroeconomic factors such as the state of the economy and industry conditions.
11. **bpd**: barrels per day
12. **GDP**: Gross Domestic Product
13. **OPEC**: Organization of Petroleum Exporting Countries
14. **DJIA**: Dow Jones Industrial Average

# Appendix D

## Additional Results

### D.1 Chapter 3

This section captures the additional results of experiments conducted in Chapter 3 - Construction of *CrudeOilNews* Corpus.

#### D.1.1 Human-in-the-Loop Active Learning

The results in this section correspond to experiments reported in Section 3.3.2.

##### D.1.1.1 Uncertainty Sampling

Table D.1: The percentage of instances (not number of sentences) sampled through uncertainty sampling (  $\ell_C$  score above the threshold value). In each active learning iteration, 50 unlabeled crude oil news were randomly selected and labeled through model prediction.

See Figure 3.8 for results in graph form.

	Entity	Trigger	Arguments	Polarity	Modality	Intensity
Threshold	0.6	0.55	0.50	0.40	0.30	0.45
Iter.	# tokens	# tokens	# Trigger-Entity Pair	# events	# events	# events
1	72	68	75	73	69	79
2	65	63	71	69	53	65
3	61	61	65	63	49	61
4	53	59	62	51	41	58
5	42	49	51	49	39	49

### D.1.1.2 Model Performance - Active Learning

Table D.2: Model performance (Micro F1-score) across varying amount of training data. As the amount of training data increases, the performance of each model increases as well. System evaluation is done on Gold-standard Test/ADJ Set. See Figure 3.8 for results in graph form.

Iter.	Training Set	Entity	Trigger	Argument	Polarity	Modality	Intensity
-	Gold Dev	0.71	0.74	0.56	0.74	0.71	0.75
-	Gold Dev + Augmented (New Dev)	0.72	0.75	0.57	0.75	0.73	0.75
1	New Dev + 50 docs	0.72	0.75	0.59	0.75	0.76	0.73
2	New Dev + 100 docs	0.78	0.79	0.62	0.79	0.81	0.77
3	New Dev + 150 docs	0.83	0.81	0.64	0.81	0.83	0.81
4	New Dev + 200 docs	0.85	0.83	0.65	0.83	0.85	0.82
5	New Dev + 250 docs	0.86	0.85	0.69	0.84	0.89	0.83

### D.1.2 Event Type Distribution

Table D.3 shows the detailed breakdown of event type distribution of the *CrudeOilNews* corpus. The same information in diagrammatic form is captured in Figure 3.10.

Table D.3: Event type distribution and sentence level counts. See Figure 3.10 for results in graph form.

Event type	Type ratio
1. Cause-movement-down-loss	14.9%
2. Cause-movement-up-gain	2%
3. Civil-unrest	2.6%
4. Crisis	1.2%
5. Embargo	4.8%
6. Geopolitical-tension	2%
7. Grow-strong	6.0%
8. Movement-down-loss	24%
9. Movement- at	2.6%
10. Movement-up-gain	15%
11. Negative-sentiment	4.07%
12. Oversupply	3.8%
13. Position-high	3.06%
14. Position-low	3.58%
15. Prohibiting	0.9%
16. Shortage	1%
17. Situation-deteriorate	1.1%
18. Slow-weak	5.79%
19. Trade-tensions	1.7%

## D.2 Chapter 4: Baseline results

This section captures the additional results of experiments conducted in Chapter 4 - Event Extraction.

### D.2.1 EMD Results (Baseline)

Table D.4 shows the detailed breakdown of results by entity mention types of the overall EMD performance reported in Table 4.1.

Table D.4: Detailed results of Entity Mention Detection (EMD) - Baseline

	Entity Type	Precision	Recall	F1
Baseline	1. COMMODITY	0.88	0.79	0.83
	2. COUNTRY	0.90	0.84	0.87
	3. DATE	0.91	0.79	0.85
	4. DURATION	0.85	0.90	0.87
	5. ECONOMIC_ITEM	0.79	0.89	0.84
	6. FINANCIAL_ATTRIBUTE	0.91	0.89	0.90
	7. FORECAST_TARGET	0.89	0.85	0.87
	8. GROUP	0.56	0.50	0.53
	9. LOCATION	0.86	0.81	0.83
	10. MONEY	0.92	0.94	0.93
	11. NATIONALITY	0.81	0.82	0.91
	12. NUMBER	0.95	0.92	0.93
	13. ORGANIZATION	0.79	0.86	0.79
	14. OTHER_ACTIVITY	0.33	0.50	0.40
	15. PERCENT	0.86	0.83	0.84
	16. PERSON	0.90	0.87	0.88
	17. PHENOMENON	0.45	0.53	0.49
	18. PRICE_UNIT	0.86	0.79	0.82
	19. PRODUCTION_UNIT	0.80	0.85	0.82
	20. QUANTITY	0.80	0.92	0.86
	21. STATE_OR_PROVINCE	0.77	0.85	0.80

### D.2.2 ED Results (Baseline)

Table D.5 shows the detailed breakdown of results by event types of the overall ED performance reported in Table 4.1. Most of the event types achieve high F1 scores with the exception of *Situation-deteriorate* and *Embargo* having F1 scores below 80%. This could be attributed to data imbalance where these two events are rarer compared to other types of events (see Table D.3) or rare trigger words with low number of occurrence within the specific event type.

Table D.5: Detailed results of Event Detection (ED) - Baseline

	Event Type	Precision	Recall	F1
Baseline	1. CAUSE_MOVEMENT_DOWN_LOSS	0.92	0.93	0.92
	2. CAUSE_MOVEMENT_UP_GAIN	0.87	0.89	0.88
	3. CIVIL_UNREST	1.00	0.89	0.94
	4. CRISES	1.00	1.00	1.00
	5. EMBARGO	0.95	1.00	0.97
	6. GEOPOLITICAL_TENSION	0.75	0.88	0.81
	7. GROW_STRONG	0.79	0.84	0.81
	8. MOVEMENT_DOWN_LOSS	0.92	0.93	0.92
	9. MOVEMENT_FLAT	0.91	0.82	0.86
	10. MOVEMENT_UP_GAIN	0.87	0.89	0.88
	11. NEGATIVE_SENTIMENT	0.93	0.94	0.93
	12. OVERSUPPLY	0.80	0.83	0.82
	13. POSITION_HIGH	0.91	1.00	0.96
	14. POSITION_LOW	0.91	0.97	0.94
	15. PROHIBITING	0.83	0.83	0.83
	16. SHORTAGE	0.91	1.00	0.95
	17. SLOW_WEAK	0.93	0.73	0.82
	18. TRADE_TENSIONS	0.75	0.88	0.81



## D.3 Chapter 5: Results Post-Transfer Learning

This section captures the results of each sub-task trained via Ensemble #2 as described in Section 5.3.3.

### D.3.1 EMD Results Post-Transfer Learning

Table D.6 shows the detailed breakdown of results by entity mention types of the final model performance reported in Table 5.5.

Table D.6: Detailed results of Entity Mention Detection (EMD) Post-Transfer Learning. Model improvements in bold.

	Entity Type	Precision	Recall	F1
Post-Transfer Learning	1. COMMODITY	0.88	0.79	0.83
	2. COUNTRY	0.90	0.84	0.87
	3. DATE	0.91	0.79	0.85
	4. DURATION	0.85	0.90	0.87
	5. ECONOMICITEM	<b>0.83</b>	<b>0.89</b>	<b>0.86</b>
	6. FINANCIALATTRIBUTE	0.91	0.89	0.90
	7. FORECAST_TARGET	0.89	0.85	0.87
	8. GROUP	0.56	0.50	0.53
	9. LOCATION	0.86	0.81	0.83
	10. MONEY	0.91	0.89	0.90
	11. NATIONALITY	0.81	0.82	0.91
	12. NUMBER	0.95	0.92	0.93
	13. ORGANIZATION	0.79	0.86	0.79
	14. OTHERACTIVITY	<b>0.45</b>	<b>0.55</b>	<b>0.50</b>
	15. PERCENT	0.86	0.83	0.84
	16. PERSON	0.90	0.87	0.88
	17. PHENOMENON	<b>0.55</b>	<b>0.53</b>	<b>0.54</b>
	18. PRICEUNIT	0.86	0.79	0.82
	19. PRODUCTONUNIT	0.80	0.85	0.82
	20. QUANTITY	0.80	0.92	0.86
	21. STATEORPROVINCE	<b>0.79</b>	<b>0.86</b>	<b>0.82</b>

### D.3.2 ED Results Post-Transfer Learning

Table D.7 shows the detailed breakdown of results by event types of final model performance reported in Table 5.5.

Table D.7: Detailed results of Event Detection (ED) Post-Transfer Learning.

	Event Type	Precision	Recall	F1
Post-Transfer Learning	1. CAUSE_MOVEMENT_DOWN_LOSS	0.92	0.93	0.92
	2. CAUSE_MOVEMENT_UP_GAIN	0.87	0.89	0.88
	3. CIVIL_UNREST	1.00	0.92	0.96
	4. CRISES	1.00	1.00	1.00
	5. EMBARGO	0.97	1.00	0.98
	6. GEOPOLITICAL_TENSION	0.75	0.88	0.81
	7. GROW_STRONG	0.79	0.84	0.81
	8. MOVEMENT_DOWN_LOSS	0.92	0.93	0.92
	9. MOVEMENT_FLAT	1.00	0.86	0.92
	10. MOVEMENT_UP_GAIN	0.87	0.89	0.88
	11. NEGATIVE_SENTIMENT	0.93	0.94	0.93
	12. OVERSUPPLY	0.80	0.83	0.82
	13. POSITION_HIGH	0.91	1.00	0.96
	14. POSITION_LOW	0.91	0.97	0.94
	15. PROHIBITING	0.83	0.83	0.83
	16. SHORTAGE	0.91	1.00	0.95
	17. SLOW_WEAK	0.93	0.73	0.82
	18. TRADE_TENSIONS	0.75	0.88	0.81

### D.3.3 ARP Results Post-Transfer Learning

Table D.8 shows the detailed breakdown of results by event types of the final model performance reported in Table 5.5.

Table D.8: Detailed results of Argument Role Prediction (ARP) Post-transfer Learning. Model improvements are in bold.

	Argument Roles	Entity Type	P	R	F1
Post-Transfer Learning	NONE	-	0.89	0.95	0.92
	Attribute	FINANCIAL_ATTRIBUTE	0.84	0.80	0.82
	Item	ECONOMIC_ITEM	<b>0.84</b>	<b>0.95</b>	<b>0.89</b>
	Final_value /	MONEY / PRODUCTION UNIT /	0.81	0.75	0.78
	Initial_value /	PRICE UNIT / PERCENTAGE /	<b>0.78</b>	<b>0.80</b>	<b>0.79</b>
	Difference /		<b>0.79</b>	<b>0.85</b>	<b>0.82</b>
	Reference_point }	MONEY / QUANTITY	0.85	0.74	0.79
	Initial_reference_point }	DATE	<b>0.79</b>	<b>0.65</b>	<b>0.71</b>
	Contract_date }		0.82	0.81	0.80
	Duration	DURATION	0.79	0.87	0.83
	Type	LOCATION	<b>0.73</b>	<b>0.84</b>	<b>0.78</b>
	Imposer •	Country / State or province	0.73	0.91	0.81
	Imposee •	Country / State or province	<b>0.59</b>	<b>0.75</b>	<b>0.66</b>
	Place •	Country / State or province	0.88	0.64	0.74
	Supplier_consumer •	Country / State or province / Nationality / Group	0.71	0.87	0.78
	Impacted_countries •	Country	<b>0.75</b>	<b>0.79</b>	<b>0.77</b>
	Participating_countries •	Country	0.87	0.89	0.88
	Forecaster	ORGANIZATION / GROUP	<b>0.89</b>	<b>0.80</b>	<b>0.84</b>
	Forecast	FORECAST_TARGET	0.77	0.85	0.81
	Situation	PHENOMENON / OTHER ACTIVITIES	<b>0.69</b>	<b>0.65</b>	<b>0.67</b>

# References

- Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., & Ellis, J. (2014). A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the second workshop on events: Definition, detection, coreference, and representation* (pp. 45–53).
- Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638–1649).
- Alyafeai, Z., AlShaibani, M. S., & Ahmad, I. (2020). A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.
- Arendarenko, E., & Kakkonen, T. (2012). Ontology-based information and event extraction for business intelligence. In *International conference on artificial intelligence: Methodology, systems, and applications* (pp. 89–102).
- Bai, Y., Li, X., Yu, H., & Jia, S. (2022). Crude oil price forecasting incorporating news text. *International Journal of Forecasting*, 38(1), 367–383.
- Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 3615–3620).
- Ben Ami, Z., & Feldman, R. (2017). Event-based trading: Building superior trading strategies with state-of-the-art information extraction tools. *Available at SSRN 2907600*.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th*

*annual meeting of the association of computational linguistics* (pp. 440–447).

- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1), 1–8.
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3), 992–1033.
- 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.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Chakraborty, S., Venkataraman, A., Jagabathula, S., & Subramanian, L. (2016). Predicting socio-economic indicators using news events. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1455–1464).
- Chan, S. W., & Chong, M. W. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, 94, 53–64.
- Chen, D., Ma, S., Harimoto, K., Bao, R., Su, Q., & Sun, X. (2019). Group, extract and aggregate: Summarizing a large amount of finance news for forex movement prediction. In *Proceedings of the second workshop on economics and natural language processing* (pp. 41–50).
- Chen, D., Zou, Y., Harimoto, K., Bao, R., Ren, X., & Sun, X. (2019, November). Incorporating fine-grained events in stock movement prediction. In *Proceedings of the second workshop on economics and natural language processing*. Hong Kong: Association for Computational Linguistics.
- Chen, Y. (2021). A transfer learning model with multi-source domains for biomedical event trigger extraction. *BMC Genomics* 22.
- Das, D., & Smith, N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1435–1444).
- Das, D., & Smith, N. A. (2012). Graph-based lexicon expansion with sparsity-inducing penalties. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 677–687).
- Del Corro, L., & Hoffart, J. (2021). From stock prediction to financial relevance: Repurposing attention weights to assess news relevance without manual annotations. In *Proceedings*

*of the third workshop on economics and natural language processing* (pp. 45–49).

- Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., et al. (2012). Building gold standard corpora for medical natural language processing tasks. In *Amia annual symposium proceedings* (Vol. 2012, p. 144).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/N19-1423> doi: 10.18653/v1/N19-1423
- Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1415–1425).
- Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.
- Dor, L. E., Gera, A., Toledo-Ronen, O., Halfon, A., Sznajder, B., Dankin, L., ... Slonim, N. (2019). Financial event extraction using wikipedia-based weak supervision. In *Proceedings of the second workshop on economics and natural language processing* (pp. 10–15).
- Duan, J., Zhang, Y., Ding, X., Chang, C. Y., & Liu, T. (2018). Learning target-specific representations of financial news documents for cumulative abnormal return prediction. In *Proceedings of the 27th international conference on computational linguistics* (pp. 2823–2833).
- Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2018). Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, 44(3), 408–421.
- Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. *Communications of the ACM*, 51(12), 68–74.
- Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, 38(1), 34–105.
- Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010). The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning{shared task}* (pp. 1–12).

- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4), 82–89.
- Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011). The stock sonar—sentiment analysis of stocks based on a hybrid approach. In *Twenty-third iaai conference*.
- Feuerriegel, S., & Neumann, D. (2013). News or noise? how news drives commodity prices. *ICIS 2013 Proceedings*.
- Forman, G. (2002). Choose your words carefully: An empirical study of feature selection metrics for text classification. In *European conference on principles of data mining and knowledge discovery* (pp. 150–162).
- Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., . . . Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. In *Proceedings of workshop for nlp open source software (nlp-oss)* (pp. 1–6).
- Glavaš, G., & Šnajder, J. (2014). Event graphs for information retrieval and multi-document summarization. *Expert systems with applications*, 41(15), 6904–6916.
- Gui, L., Xu, R., Lu, Q., Du, J., & Zhou, Y. (2018). Negative transfer detection in transductive transfer learning. *International Journal of Machine Learning and Cybernetics*, 9(2), 185–197.
- Gui, T., Zhang, Q., Gong, J., Peng, M., Liang, D., Ding, K., & Huang, X.-J. (2018). Transferring from formal newswire domain with hypernet for twitter pos tagging. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2540–2549).
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8342–8360).
- Han, S., Hao, X., & Huang, H. (2018). An event-extraction approach for business analysis from online chinese news. *Electronic Commerce Research and Applications*, 28, 244–260.
- Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication* (1973), 45(4), 363–407.
- Hogenboom, A., Hogenboom, F., Frasincar, F., Schouten, K., & Van Der Meer, O. (2013). Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1), 27–52.

- Hogenboom, F., de Winter, M., Frasincar, F., & Kaymak, U. (2015). A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10), 4667–4675.
- Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F., & Caron, E. (2016). A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85, 12–22.
- Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3), 296–298.
- Hu, Z., Liu, W., Bian, J., Liu, X., & Liu, T.-Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh acm international conference on web search and data mining* (pp. 261–269).
- Huang, L., Ji, H., Cho, K., Dagan, I., Riedel, S., & Voss, C. (2018, July). Zero-shot transfer learning for event extraction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*.
- Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).
- Jacobs, G., & Hoste, V. (2020). Extracting fine-grained economic events from business news. In *Proceedings of the 1st joint workshop on financial narrative processing and multilingual financial summarisation* (pp. 235–245).
- Jacobs, G., & Hoste, V. (2021). Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. *Language Resources and Evaluation*, 1–33.
- Jacobs, G., Lefever, E., & Hoste, V. (2018). Economic event detection in company-specific news text. In *The 56th annual meeting of the association for computational linguistics* (pp. 1–10).
- Jiang, N., & de Marneffe, M.-C. (2021). He thinks he knows better than the doctors: Bert for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9, 1081–1097.
- Jiménez-Zafra, S. M., Morante, R., Teresa Martín-Valdivia, M., & Ureña-López, L. A. (2020). Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1), 1–52.

- Judea, A., & Strube, M. (2016, December). Incremental global event extraction. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*.
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*.
- Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2), 155–190.
- Konstantinova, N., De Sousa, S. C., Díaz, N. P. C., López, M. J. M., Taboada, M., & Mitkov, R. (2012). A review corpus annotated for negation, speculation and their scope. In *Proceedings of the eighth international conference on language resources and evaluation (Irec'12)* (pp. 3190–3195).
- Konyushkova, K., Sznitman, R., & Fua, P. (2017). Learning active learning from data. *Advances in neural information processing systems*, 30.
- Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J. G., Smith, N. A., & Dyer, C. (2015). Frame-semantic role labeling with heterogeneous annotations. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (pp. 218–224).
- Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society*, 18(3), 502–520.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 260–270).
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
- Lefever, E., & Hoste, V. (2016). A classification-based approach to economic event detection in dutch news text. In *Proceedings of the tenth international conference on language*



*resources and evaluation (Irec'16)* (pp. 330–335).

- Li, J., Xu, Z., Xu, H., Tang, L., & Yu, L. (2017). Forecasting oil price trends with sentiment of online news articles. *Asia-Pacific Journal of Operational Research*, 34(02), 1740019.
- Li, Q., Ji, H., Hong, Y., & Li, S. (2014, October). Constructing information networks using one single model. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.
- Li, Q., Ji, H., & Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 73–82).
- Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 278, 826–840.
- 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.
- Liebmann, M., Hagenau, M., & Neumann, D. (2012). Information processing in electronic markets: Measuring subjective interpretation using sentiment analysis.
- Liu, J., & Huang, X. (2021). Forecasting crude oil price using event extraction. *IEEE Access*, 9, 149067–149076.
- Liu, Q., Cheng, X., Su, S., & Zhu, S. (2018). Hierarchical complementary attention network for predicting stock price movements with news. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 1603–1606).
- Liu, X., Huang, H.-Y., & Zhang, Y. (2019). Open domain event extraction using neural latent variable models. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2860–2871).
- Liu, X., Luo, Z., & Huang, H. (2018, October–November). Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1247–1256). Brussels, Belgium: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/D18-1156> doi: 10.18653/v1/D18-1156
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Z., Mitamura, T., & Hovy, E. (2015). Evaluation algorithms for event nugget detection: A pilot study. In *Proceedings of the the 3rd workshop on events: Definition, detection, coreference, and representation* (pp. 53–57).

- Lösch, U., & Nikitina, N. (2009). The newsevents ontology: an ontology for describing business events. In *Proceedings of the 2009 international conference on ontology patterns-volume 516* (pp. 187–193).
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1), 35–65.
- Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187–1230.
- Lyu, Q., Zhang, H., Sulem, E., & Roth, D. (2021). Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: Short papers)* (pp. 322–332).
- MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of economic literature*, 35(1), 13–39.
- Malik, H. H., Bhardwaj, V. S., & Fiorletta, H. (2011). Accurate information extraction for quantitative financial events. In *Proceedings of the 20th acm international conference on information and knowledge management* (pp. 2497–2500).
- Marasović, A., & Frank, A. (2016). Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st workshop on representation learning for nlp* (pp. 111–120).
- Marujo, L., Ribeiro, R., Gershman, A., de Matos, D. M., Neto, J. P., & Carbonell, J. (2017). Event-based summarization using a centrality-as-relevance model. *Knowledge and Information Systems*, 50(3), 945–968.
- Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The unified and holistic method gamma ( ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3), 437–479.
- Mausam, M. (2016). Open information extraction systems and downstream applications. In *Proceedings of the twenty- fth international joint conference on artificial intelligence* (pp. 4074–4077).
- Meftah, S., & Semmar, N. (2018). A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the eleventh international conference on language resources and evaluation (Irec 2018)*.
- Meftah, S., Semmar, N., Tamaazousti, Y., Essafi, H., & Sadat, F. (2021). On the hidden negative transfer in sequential transfer learning for domain adaptation from news to

- tweets. In *Proceedings of the second workshop on domain adaptation for nlp* (pp. 140–145).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., & Strassel, S. (2015). Event nugget annotation: Processes and issues. In *Proceedings of the the 3rd workshop on events: Definition, detection, coreference, and representation* (pp. 66–76).
- Morante, R., & Blanco, E. (2012). \* sem 2012 shared task: Resolving the scope and focus of negation. In \* sem 2012: *The first joint conference on lexical and computational semantics{volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012)}* (pp. 265–274).
- Morante, R., & Sporleder, C. (2012). Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2), 223–260.
- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
- Nguyen, T. H., Cho, K., & Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 300–309).
- Nguyen, T. H., & Grishman, R. (2018). Graph convolutional networks with argument-aware pooling for event detection. In *Aaai* (Vol. 18, pp. 5900–5907).
- Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1354–1364).
- Nguyen, T. M., & Nguyen, T. H. (2019). One for all: Neural joint modeling of entities and events. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 6851–6858).
- O’Gorman, T., Wright-Bettner, K., & Palmer, M. (2016, November). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd workshop on computing news storylines (CNS 2016)* (pp. 47–56). Austin, Texas: Association for Computational Linguistics.

- Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).
- Qian, Y., Deng, X., Ye, Q., Ma, B., & Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing & Management*, 56(6), 102086.
- Rönnqvist, S., & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264, 57–70.
- Rosenstein, M. T., Marx, Z., Kaelbling, L. P., & Dietterich, T. G. (2005). To transfer or not to transfer. In *In nips'05 workshop, inductive transfer: 10 years later*.
- Ruder, S. (2019). *Neural transfer learning for natural language processing* (Unpublished doctoral dissertation). NUI Galway.
- Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*, 2.
- Rudinger, R., White, A. S., & Van Durme, B. (2018). Neural models of factuality. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 731–744).
- Sadik, Z. A., Date, P. M., & Mitra, G. (2020). Forecasting crude oil futures prices using global macroeconomic news sentiment. *IMA Journal of Management Mathematics*, 31(2), 191–215.
- Saha, S., Pal, H., et al. (2017). Bootstrapping for numerical open ie. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 317–323).
- Sanh, V., Wolf, T., & Ruder, S. (2019). A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 6949–6956).
- Saurí, R., & Pustejovsky, J. (2009). Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3), 227–268.
- Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.
- Settles, B. (2009). Active learning literature survey.

- Sha, L., Qian, F., Chang, B., & Sui, Z. (2018, Apr.). Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.
- Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 world wide web conference* (pp. 1105–1114).
- Søgaard, A., & Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 231–235).
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics* (pp. 102–107).
- Stone, P. (2002). General inquirer harvard-iv dictionary. <http://www.wjh.harvard.edu/~inquirer/>, updated on, 9(12), 2002.
- Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437–1467.
- Theil, C. K., Štajner, S., & Stuckenschmidt, H. (2018). Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the first workshop on economics and natural language processing* (pp. 32–37).
- Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242–264). IGI global.
- Upreti, B. R., Back, P. M., Malo, P., Ahlgren, O., & Sinha, A. (2019). Knowledge-driven approaches for financial news analytics. In *Network theory and agent-based modeling in economics and finance* (pp. 375–404). Springer.
- Vaidya, A., Mai, F., & Ning, Y. (2020). Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 683–693).
- Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11), 4999–5010.
- Veyseh, A. P. B., Nguyen, T. H., & Dou, D. (2019). Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the*

- 57th annual meeting of the association for computational linguistics* (pp. 4393–4399).
- Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019, November). Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*.
- Wang, J., Athanasopoulos, G., Hyndman, R. J., & Wang, S. (2018). Crude oil price forecasting based on internet concern using an extreme learning machine. *International Journal of Forecasting*, 34(4), 665–677.
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 1–40.
- Wex, F., Widder, N., Liebmann, M., & Neumann, D. (2013). Early warning of impending oil crises using the predictive power of online news stories. In *2013 46th hawaii international conference on system sciences* (pp. 1512–1521).
- Wu, B., Wang, L., Lv, S.-X., & Zeng, Y.-R. (2021). Effective crude oil price forecasting using new text-based and big-data-driven model. *Measurement*, 168, 108468.
- Wu, B., Wang, L., Wang, S., & Zeng, Y.-R. (2021). Forecasting the us oil markets based on social media information during the covid-19 pandemic. *Energy*, 226, 120403.
- Xiang, W., & Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, 7, 173111–173137.
- Xie, B., Passonneau, R., Wu, L., & Creamer, G. G. (2013). Semantic frames to predict stock price movement. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 873–883).
- Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1), 49–73.
- Yang, B., & Mitchell, T. M. (2016, June). Joint extraction of events and entities within a document context. In *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 289–299). San Diego, California: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N16-1033> doi: 10.18653/v1/N16-1033
- Yang, H., Chen, Y., Liu, K., Xiao, Y., & Zhao, J. (2018). Dcfec: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of acl 2018, system demonstrations* (pp. 50–55).
- Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th conference of*

*the association for computational linguistics* (pp. 5284–5294).

- Zhang, J., Qin, Y., Zhang, Y., Liu, M., & Ji, D. (2019). Extracting entities and events as a single task using a transition-based neural model. In *Ijcai* (pp. 5422–5428).
- Zhang, X., Qu, S., Huang, J., Fang, B., & Yu, P. (2018). Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6, 50720–50728.
- Zhang, Y., Qi, P., & Manning, C. D. (2018, October–November). Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2205–2215). Brussels, Belgium: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/D18-1244> doi: 10.18653/v1/D18-1244
- Zhao, L.-T., Liu, L.-N., Wang, Z.-J., & He, L.-Y. (2019). Forecasting oil price volatility in the era of big data: A text mining for var approach. *Sustainability*, 11(14), 3892.
- Zhao, L.-T., Zeng, G.-R., Wang, W.-J., & Zhang, Z.-G. (2019). Forecasting oil price using web-based sentiment analysis. *Energies*, 12(22), 4291.