CrudeOilNews: An Annotated Crude Oil News Corpus for Event Extraction

Meisin Lee, Lay-Ki Soon, Eu-Gene Siew, Ly Fie Sugianto
Monash University
Malaysia Campus, Jalan Lagoon Selatan, 47500 Selangor, Malaysia
{mei.lee, soon.layki, siew.eu-gene, lyfie.sugianto}@monash.edu

Abstract

In this paper, we present CrudeOilNews, a corpus 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. This paper describes the data collection process, the annotation methodology, and the event typology used in producing the corpus. Firstly, a seed set of 175 news articles were manually annotated, of which a subset of 25 news was used as the adjudicated reference test set for inter-annotator and system evaluation. The inter-annotator agreement was generally substantial, and annotator performance was adequate, indicating that the annotation scheme produces consistent event annotations of high quality. Subsequently, the dataset is expanded through (1) data augmentation and (2) Human-in-the-loop active learning. The resulting corpus has 425 news articles with approximately 11k events annotated. As part of the active learning process, the corpus was used to train basic event extraction models for machine labeling; the resulting models also serve as a validation or as a pilot study demonstrating the use of the corpus in machine learning purposes. The annotated corpus is made available for academic research purpose at [https://github.com/meisin/CrudeOilNews-Corpus](https://github.com/meisin/CrudeOilNews-Corpus).

Keywords: Crude Oil News, Annotated Dataset, Event Extraction, Financial Information Extraction, English corpus

1. Introduction

Financial markets are sensitive to breaking news on economic events. Specifically for crude oil markets, it is observed in [Brandt and 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 et al., 2021]. Therefore, accurate and timely automatic identification of events in news items is crucial for making timely trading decisions. Commodity news typically contains these few key information: (i) analysis of recent commodity price movements (up, down or flat), (ii) a retrospective view of notable event(s) that led to such a movement, and (iii) forecast or forward-looking analysis of supply-demand situation as well as projected commodity price targets. Here is a snippet taken from a piece of crude oil news:

(1) 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% as Tuesday.

There is a small number of corpora in the Finance and Economics domain such as SENTIVENT in [Jacobs and Hoste, 2021], but all are focused on company-specific events and are used mainly for the purpose of stock price prediction. As acknowledged by [Jacobs and Hoste, 2021], due to the lack of annotated dataset in the Finance and Economics domain, only a handful of supervised approaches exist for Financial Information Extraction. To the best of our knowledge, there is no available annotated corpus for crude oil or any other commodities. We aim to contribute towards resource building and facilitate future research in this area by introducing CrudeOilNews corpus, an event extraction dataset focusing on Macro-economic, Geopolitical, and crude oil supply and demand events. The annotation schema used are aligned to ACE (Automatic Content Extraction) and ERE (Entities, Relations, and Events) standards, so event extraction systems developed for ACE/ERE can be used readily on this corpus.

The contributions of this work 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 corpus through data augmentation;
- Used Human-in-the-Loop Active Learning to expand the corpus with model inference while optimizing human annotation effort to focus on just less confident (and likely less accurate) predictions.
2. Related Work

2.1. Annotation Methodologies

The annotation methodologies presented here are conceived based on the annotation standards of ACE, ERE, and supervised training. An extensive comparison has been made in (Agúndez 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). We have strived to align to ACE/ERE programs as closely as possible, but have made minor adaptations to cater to unique characteristics found in crude oil news. Tense and Genericity defined in ACE2005 are dropped from our annotation scope while the new property - Intensity is introduced.

2.2. Finance and Economic Domain

In the domain of Finance and Economics, the majority of available datasets are on company-related events and are used mainly for extracting company-related events for company stock price prediction. A variety of methods are used in economic event detection, such as hand-crafted rule-sets, ontology knowledge-bases, and using techniques like distant, weakly or semi-supervised training. For a more targeted discussion, we focus only on manually annotated datasets suitable for supervised training.

In (Jacobs et al., 2018; Lefever and Hoste, 2016), 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. As a continuation of the work, the authors introduced SentiEVENT, a fine-grained ACE/ERE-like dataset in (Jacobs and Hoste, 2021). Just like the earlier work, their focus is mainly on company-related and financial events. Among the list of defined event topology, 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 the news. While they choose to remain at a broad level, our work complements theirs by defining key events in the Macroeconomic and Geo-political category at a detailed level. As part of the search for commodity news related resources, we came across RavenPack’s 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 only contains sentiment scores without any event annotations. However, Ravenpack’s event taxonomy on crude oil-related events prove to be a useful resource in helping us define our own event typology. Details of event typology are covered in Section 4.3.1.

3. Dataset Collection

First, we crawled crude oil news articles from investing.com a financial platform, and financial/business news aggregator and is considered one of the top three global financial websites in the world. We crawled news articles dating from Dec 2015 to Jan 2020 (50 months). From the pool of crude oil news, we uniformly sampled 175 pieces of news articles throughout the 50-month period 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 were selected out of the gold-standard dataset as the adjudicated set (ADJ).

4. Annotation Setup

The dataset is annotated using Brat rapid annotation tool (Stenetorp et al., 2012), a web-based tool for text annotation. An example of a sentence annotated using Brat is shown in Figure 1. The annotation process is designed to have 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 this criterion, we recruited two annotators 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.
- Layer 2: Annotate events by identifying event triggers.

RavenPack is an analytics provider for financial services. Among their services are finance and economic news sentiment analysis. More information here: https://www.ravenpack.com/

1. RavenPack is an analytics provider for financial services. Among their services are finance and economic news sentiment analysis. More information here: https://www.ravenpack.com/

2. See https://www.investing.com/commodities/crude-oil-news for more information.
4.1. Annotation Guidelines

This section describes our definition of events and principles for annotation of entity mentions and events. Annotation of events is further divided into (i) annotating entity mentions, (ii) annotating event triggers, (iii) linking entity mentions to their respective events and identifying the argument roles each entity plays, (iv) assigning the right property labels to Polarity, Modality and Intensity.

4.2. 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. There are 21 entity types identified and annotated in the dataset, see Appendix C for the full list. Nominal entities relating to Finance and Economics are annotated. Apart from crude oil-related terms, below 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

4.3. 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 Event. Details and rules for identifying event triggers and event Arguments are covered below:

Event Triggers Our annotation of event trigger is aligned to ERE where an event trigger (known as event nugget in the shared task in [Mitamura et al., 2015]) can be either a single word (main verb, noun, adjective, adverb) or a continuous multi-word phrase. Here are some examples found in the dataset:

- Verb: Houti rebels attacked Saudi Arabia.
- Noun: The government slapped sanctions against its petroleum.
- Adjective: A fast growing economy has...
- Multi-verb: The market bounced back...

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, 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 1, 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 1.

4.3.1. Event Typology

According to [Brandt and Gao, 2019], who analyzed Ravenpack’s sentiment score of each event type and oil price, events that move commodity prices are geopolitical, macro-economic, and commodity supply and demand in nature. Based on Ravenpack’s event taxonomy, we have defined a set of 18 oil-related event types as our annotation scope. Event types and the cor-
Table 1: List of Event Arguments of example in Figure

| Entity                  | Argument Role  |
|-------------------------|----------------|
| U.S.                    | SUPPLIER       |
| crude stockpiles        | ITEM           |
| 1,350 million barrels   | DIFFERENCE     |
| December                | REFERENCE_POINT_TIME |
| 200 million barrels     | INITIAL_VALUE  |
| 438.9 million barrels   | FINAL_VALUE    |

responding list of example trigger words and example key arguments are listed in Table 2. See Appendix D for event schema for all 18 event types.

4.3.2. Event Property: Event Polarity, Modality, and Intensity

After events are identified, they are also assigned a label each for the properties respectively.

POLARITY (POSITIVE and NEGATIVE)

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.

MODALITY (ASSERTED and OTHER)

Event modality determines whether the event represents a “real” occurrence. ASSERTED is assigned if the author or speaker refers to it as though it were a real occurrence, and OTHER otherwise. OTHER covers believed events, hypothetical events, commanded and requested event, threats, proposed events, discussed events, desired events, promised events, and other unclear construct.

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. Oftentimes, events reported in Crude Oil News is about the intensity of an existing event, whether the event is further intensified or eased.

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

(2) ...could hit Iraq’s output and deepen a supply shortfall. [INTENSIFIED]

(3) Libya’s civil strife has been eased by potential peace talks. [EASED]

The event strife (civil unrest) in sentence (2) 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’.

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

(4) OPEC cancelled a planned easing of output cuts. [NEGATIVE, OTHER, EASED]

(5) In order to end the global crisis, OPEC may hesitate to implement a planned loosening of output curbs. [NEGATIVE, OTHER, EASED]

(6) Oil prices rose to $110 a barrel on rumours of a renewed strife. [POSITIVE, OTHER, INTENSIFIED]

4.4. Inter-Annotator Agreement

Inter-annotator agreement (IAA) is a good indicator of how clear our annotation guidelines are, how uniformly annotators understand them, how robust are the event typology and overall how feasible the annotation task is. We evaluate IAA on each annotated category separately (see Table 9 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 annotations 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 and 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 pairwise F1 score as the measurement. Hence for the evaluation of entity spans and trigger spans, we report on both F1 as well as “token-level” kappa. Both scores were measured without taking into account the un-annotated tokens - labelled “O”.

As for the rest of the annotation category, we report only on Cohen’s Kappa as this is the standard measure of IAA for the classification task. We calculate the agreement by comparing annotation outcomes of the two annotators with each other, arbitrarily treating one as the ‘gold’ reference. We also scored each annotator 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 shows the average agreement scores for all annotation categories.

The event nugget scoring method introduced in Liu et al., 2015 was not used here because their assessment is inappropriate 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 pairwise F1 score as the measurement. Hence for the evaluation of entity spans and trigger spans, we report on both F1 as well as “token-level” kappa. Both scores were measured without taking into account the un-annotated tokens - labelled “O”.

Analysis We benchmark these IAA scores with the ‘strength of agreement’ of each Kappa ranges as set out by (Landis and Koch, 1977). Most annotation categories achieved a 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:
Table 2: 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 | cut, trim, reduce, disrupt, curb, squeeze, choked off | oil production, oil supplies, interest rate, growth forecast |
| 2. CAUSED-MOVEMENT-UP-GAIN  | boost, revive, ramp up, prop up, raise                  | oil production, oil supplies, growth forecast |
| 3. CIVIL-UNREST             | violence, turmoil, fighting, civil war, conflicts      | Libya, Iraq                                  |
| 4. CRISIS                   | crisis, crises                        | debt, financial                              |
| 5. EMBARGO                  | embargo, sanction                    | Iraq, Russia                                 |
| 6. GEOPOLITICAL-TENSION     | war, tensions, deteriorating relationship     | Iraq-Iran                                   |
| 7. GROW-STRONG              | grow, picking up, boom, recover, expand, strong, rosy, improve, solid | crude oil price, U.S. dollar, gross domestic product (GDP) growth |
| 8. MOVEMENT-DOWN-LOSS       | fell, down, less, drop, tumble, collapse, plunge, downturn, slump, slide, decline | crude oil price, U.S. dollar, gross domestic product (GDP) growth |
| 9. MOVEMENT-FLAT            | unchanged, flat, hold, maintained         | oil price                                   |
| 10. MOVEMENT-UP-GAIN        | up, gain, rise, surge, soar, swell, increase, rebound | oil price, U.S. employment data, gross domestic product (GDP) growth |
| 11. NEGATIVE-SENTIMENT      | worries, concern, fears                  |                                            |
| 12. OVERSUPPLY              | glut, bulging stock level, excess supplies |                                            |
| 13. POSITION-HIGH           | high, highest, peak, highs               |                                            |
| 14. POSITION-LOW            | low, lowest, lows, trough                |                                            |
| 15. PROHIBITION             | ban, bar, prohibit                      | exports, imports                            |
| 16. SHORTAGE                | shortfall, shortage, under-supplied      | oil supply                                  |
| 17. SLOW-WEAK               | slow, weak, tight, lackluster, falter, weaken, bearish, slowdown, crumbles | global economy, regional economy, economic outlook, crude oil demand |
| 18. TRADE-TENSIONS          | price war, trade war, trade dispute       | U.S.-China                                  |

Table 3: IAA for all annotation categories. For categories involving spans (marked by *), both Cohen’s kappa (calculated at “token level”) and F1 score measurements are provided.

| Task              | Cohen’s Kappa (k) | 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              | -        |

(7) 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, we achieve 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 interpretation 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”.

From the cases where annotators disagree, we analyze and found that most of them stem from differences in interpreting special concepts, for example:

- The word outlook, should it be interpreted as forecast? Or should it be considered as a cue word for event modality?
- 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-straightforward 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 an agreed annotation.

5. 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.
### 5.1. Data Augmentation

The main purpose of introducing augmented data is to address the issue of serious class imbalance in Event Properties in the dataset. Table 4 shows event properties classification results. The pink-coloured cells show the model’s F1-score when trained on gold-standard dev dataset, F1-scores for minority classes are rather low. As a strategy to overcome class imbalance, we manually over-sample the minority classes for data augmentation and introduce them into the dataset. To this end, we carried out data augmentation through (i) trigger word replacement and (ii) event argument replacement.

- **Trigger word replacement**: FrameNet was utilized to augments available data and to generate both diverse and valid examples. Authors in [Aguilar et al., 2014](https://framenet.icsi.berkeley.edu) pointed out that all events, relations, and attributes that were represented by ACE/ERE can be mapped to FrameNet representations through some adjustments. In the selected sentences, we replaced the event trigger words with words (known as lexical units in FrameNet) of the same frame in FrameNet. The idea is to replace the existing trigger word with another valid trigger word while maintaining the same semantic meaning (in FrameNet’s term - maintaining the same frame). Through this exercise, we also introduced richer lexical variance in the dataset.

- **Event argument replacement**: Event argument replacement candidates were chosen from a pool of candidates of the same entity type and the same argument role within the pool of existing annotations, as illustrated below:

  1. **The benchmark for oil prices **advanced** 1.29% to $74.71.**
     - Candidates: [surged, rose, appreciated, climbed]
  2. **....after civil-unrest in Libya...**
     - Candidates: [Iraq, Nigeria, Ukraine]

After adding augmented data into the training, the green-coloured cells in Table 4 show improved F1-scores for minority classes across all three event properties. We add augmented data into the gold-standard dev set to form the new development set and use it to train baseline models for human-in-the-loop active learning.

### 5.2. 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 on unlabeled data. Predictions that are ‘uncertain’ (or of low confidence) are then given to human experts for verification. Verified labels are then added into 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 in. This way, we narrow down the scope and have human experts work specifically on these instances. Rather than blindly adding more training data incurring more cost and time, here we target instances that are near the model’s decision boundary, they are valuable when labeled correctly and added to the training data to improve model performance. The whole active learning process is shown in Figure 2.

**Least Confidence score**: Least confidence score, \( \phi_{LC} \) 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 \( y^* \) is the highest confidence softmax score:

\[
\phi_{LC}(x) = (1 - P(y^* | x)) \times \frac{n}{n - 1}
\]

The equation produces a Least Confidence (LC) scores to 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 divided by \( n - 1 \). Hence it can be used in binary classification as well as multi-class classification. Any model predictions with \( \phi_{LC} \) score above the
Figure 2: Human-in-the-loop active learning cycle: (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.

threshold is sampled as they are most likely to be classified wrongly and need to be relabeled by a human annotator.

5.2.1. Baseline models
As the baseline for the first round of Active Learning, we trained a number of basic or ‘vanilla’ machine learning models, one for each sub-tasks using the new development set (described in Section 5.1) as training data and ADJ set as test data (See Table 5 for key statistics). These “vanilla” models also act as the pilot study for the use of this dataset in event extraction. The following section describes how these models are trained.

Entity Mention Detection Model: We formalize Entity Mention Detection task as a multi-class token classification task. Similar to the approach used in (Nguyen et al., 2016), we employ BIO annotation schema to assign entity type labels to each token in the sentences. For the model architecture, we use Huggingface’s BERTForTokenClassification to fine-tune on this task.

Event Extraction Model: We jointly train Event Detection together with Argument Role Prediction using JMEE (Joint Multiple Event Extraction), an event extraction solution proposed by (Liu et al., 2018). The original version of JMEE uses GloVe word embedding; for this work we used a modified version of JMEE that replaces GloVe with BERT (Devlin et al., 2019) contextualized word embeddings, and codes are available here.

Event Properties Classification: We use BERTForSequenceClassification model to fine-tune on this task. For every event identified in the earlier model, we extract the event ‘scope’ 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, we use \( n = 8 \). Using the example sentence presented in Figure 1, 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.

5.2.2. Experiments & Analysis
Least Confidence (LC) threshold: In order to find the optimum sample size for human relabeling, we need to determine the suitable LC threshold. We design the uncertainty sampling exercise as a Binary Classification task with two outcomes: sampled and not-sampled. We experimented with different threshold values 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. We checked the sampled and not-sampled instances against the ground-truth in ADJ, and were able to construct the confusion matrix and obtain Precision, Recall, and F1 scores. Ideally, we want 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). We experimented with different LC threshold value ranging from 0 to 1 in order to find the best threshold that produces sampled and not-sampled split with the best F1 score (the highest precision-recall pair). We carry out all iterations of active learning (described next) 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: We carried out five iterations of active learning, each iteration involves 50 unlabeled crude oil news being labeled through model prediction. Then we ran uncertainty sampling and arranged for two annotators to validate the samples and relabel them if needed. For sentence not sampled, they are deemed ‘confident’ and therefore being validated/checked by just a single annotator.

Analysis: Overall, we see improvements in model performance across all sub-tasks. As shown in Figure 3, model 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. 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 confused with other types. For example, the model erroneously classify trade tension as Geopolitical-Tension when the right class should be Trade-Tensions. As the word ‘tension’ exists 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 for minority classes, the model has significantly fewer data to learn from, leading the model to generate predictions that are less ‘confident’.

6. Corpus Statistics and Analysis

In total, we managed to produce a final dataset consisting of 425 documents, which consist of 7,059 sentences, 10,578 events, 22,267 arguments. The breakdown is shown in Table 5.

| Statistics of CrudeOilNews corpus. | Gold-standard | Aug | S-Iter |
|-----------------------------------|---------------|-----|-------|
| # 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|

6.1. Key Characteristics

We observe a few key characteristics of this corpus that are distinct from ACE2005 and ERE datasets. These need to be taken into consideration when adapting existing event extraction systems or building a new one for this corpus:

1. Obvious class imbalance in event properties distribution where the majority class outnumbers the minority classes by a large margin (see Table 4). We have attempted to minimize this margin by oversampling minority classes through data augmentation, but the margin is still quite substantial;
2. Homogenous entity types but play different argument roles (e.g., price - non-distinguishable from entity type MONEY or UNIT-PRICE, play different role such as opening price, closing price, and price difference).
3. Number intensity: Numbers (e.g., price, difference, percentage of change) and dates (including date of the opening price, dates of closing price) are abundant.

7. Conclusion and Future Work

Event extraction in the domain of finance and economics at the moment are limited to company-related events only. To contribute to the building of resources in this domain, we have presented CrudeOilNews corpus, an ACE/ERE-like corpus. This corpus contains 425 documents, with around 11,000 events annotated. We have also shared methodologies of how these information were annotated. The inter-annotator agreement is generally substantial, and annotator performance is adequate, indicating that the annotation scheme produces consistent event annotations of high quality.

There are a number of avenues for future work. The main area that can be further explored is to expand the annotation scope to cover more event types. Next, this work can also be expanded to cover event co-reference and event-event relations such as causal-relation, main-sub-event, event-sequence, and contradictory event relation. Besides that, the current sentence-level annotation can be extended to cater for event relations spanning multiple sentences, so that event extraction and relation extract can be done at the document level.
8. Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and 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, pages 45–53.

Brandt, M. W. and Gao, L. (2019). Macro fundamentals or geopolitical events? a textual analysis of news events for crude oil. Journal of Empirical Finance, 51:64–94.

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, volume 2012, page 144. American Medical Informatics Association.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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), pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association, 12(3):296–298.

Jacobs, G. and Hoste, V. (2021). Sentivent: enabling supervised information extraction of company-specific events in economic and financial news. Language Resources and Evaluation, pages 1–33.

Jacobs, G., Lefever, E., and Hoste, V. (2018). Economic event detection in company-specific news text. In The 56th Annual Meeting of the Association for Computational Linguistics, pages 1–10. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, pages 159–174.

Lefever, E. and 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 (LREC’16), pages 330–335.

Liu, Z., Mitamura, T., and 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, pages 53–57.

Liu, X., Luo, Z., and Huang, H. (2018). Jointly multiple events extraction via attention-based graph information aggregation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1247–1256, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (\(\gamma\)) for inter-annotator agreement measure and alignment. Computational Linguistics, 41(3):437–479.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., and Strassel, S. (2015). Event nugget annotation: Processes and issues. In Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 66–76.

Nguyen, T. H., Cho, K., and 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, pages 300–309.

O’Gorman, T., Wright-Bettner, K., and Palmer, M. (2016). Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016), pages 47–56, Austin, Texas, November. Association for Computational Linguistics.

Settles, B. (2009). Active learning literature survey.

Stenetorp, P., Pyysalo, S., Topič, G., Ohta, T., Ananiadou, S., and 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, pages 102–107.

Wu, B., Wang, L., Lv, S.-X., and Zeng, Y.-R. (2021). Effective crude oil price forecasting using new text-based and big-data-driven model. Measurement, 168:108468.
A. Detailed Corpus Statistics

Table 6: Event type distribution and sentence level counts

| Event type                          | Gold Annotation | Augmented | 5-Iter AL | Final Count | # Instance | Ratio    |
|-------------------------------------|-----------------|-----------|-----------|-------------|------------|----------|
|                                     | Dev             | Test/ADJ  |           |             |            |          |
| Cause-movement-down-loss            | 359             | 37        | 179       | 426         | 1,001      | 9.46%    |
| Cause-movement-up-gain              | 72              | 7         | 16        | 83          | 178        | 1.68%    |
| Civil-unrest                        | 57              | 3         | 43        | 47          | 150        | 1.42%    |
| Crisis                              | 19              | 4         | 11        | 34          | 68         | 0.64%    |
| Embargo                             | 115             | 7         | 44        | 44          | 210        | 1.99%    |
| Geopolitical-tension                | 42              | 10        | 25        | 125         | 202        | 1.91%    |
| Grow-strong                         | 167             | 16        | 71        | 280         | 534        | 5.05%    |
| Movement-down-loss                  | 697             | 149       | 213       | 1,881       | 2,940      | 27.70%   |
| Movement-flat                       | 47              | 2         | 14        | 42          | 105        | 0.99%    |
| Movement-up-gain                    | 683             | 178       | 203       | 1,637       | 2,701      | 25.53%   |
| Negative-sentiment                  | 116             | 43        | 73        | 307         | 539        | 0.99%    |
| Oversupply                          | 65              | 9         | 45        | 112         | 231        | 2.18%    |
| Position-high                       | 132             | 33        | 19        | 377         | 561        | 5.03%    |
| Position-low                        | 99              | 47        | 24        | 323         | 493        | 4.66%    |
| Prohibition                         | 39              | 1         | 3         | 6           | 49         | 0.46%    |
| Shortage                            | 31              | 1         | 10        | 5           | 47         | 0.44%    |
| Slow-weak                           | 164             | 27        | 52        | 262         | 505        | 4.77%    |
| Trade-tensions                      | 39              | 3         | 16        | 6           | 64         | 0.61%    |
| Total                               | 2,943           | 577       | 1,061     | 5,997       | 10,578     |          |

B. Active Learning details

Table 7: The percentage of instances (not number of sentences) sampled through uncertainty sampling (ϕLC 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 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. | % of # tokens | % of # tokens | % of Trigger-Entity Pair | % of events | % of events | % of 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        |

Table 8: 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 4 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 |

Note: New development set the baseline model described in Section 5.2.1.
## C. Entity Mention Types

Table 9: List of Entity Types

| Entity Type       | Examples                                                                 |
|-------------------|--------------------------------------------------------------------------|
| 1. Commodity      | oil, crude oil, Brent, West Texas Intermediate (WTI), fuel, U.S Shale, light sweet crude, natural gas |
| 2. Country**      | Libya, China, U.S, Venezuela, Greece                                      |
| 3. Date**         | 1998, Wednesday, Jan. 30, the final quarter of 1991, the end of this year |
| 4. Duration**     | two years, three-week, 5-1/2-year, multiyear, another six months         |
| 5. Economic Item  | economy, economic growth, market, economic outlook, employment data, currency, commodity-oil |
| 6. Financial attribute | supply, demand, output, production, price, import, export                |
| 7. Forecast target| forecast, target, estimate, projection, bets                            |
| 8. Group          | global producers, oil producers, hedge funds, non-OECD, Gulf oil producers |
| 9. Location**     | global, world, domestic, Middle East, Europe                             |
| 10. Money**       | $60, USD 50                                                              |
| 11. Nationality** | Chinese, Russian, European, African                                      |
| 12. Number**      | (any numerical value that does not have a currency sign)                 |
| 13. Organization**| OPEC, Organization of Petroleum Exporting Countries, European Union, U.S Energy Information Administration, EIA |
| 14. Other activities | (free text)                                                                |
| 15. Percent**     | 25%, 1.4 percent                                                          |
| 16. Person**      | Trump, Putin (and other political figures)                                |
| 17. Phenomenon    | (free text)                                                               |
| 18. Price unit    | $100-a-barrel, $40 per barrel, USD58 per barrel                         |
| 19. Production Unit | 170,000 bpd, 400,000 barrels per day, 29 million barrels per day         |
| 20. Quantity      | 1.3500 million barrels, 1.8 million gallons, 18 million tonnes           |
| 21. State or province** | Washington, Moscow, Cushing, North America                              |
D. Event Schema

D.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            |               |
| 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              |               |
| Difference                  | Percentage, Number, Money, Production_unit, Quantity | $2.50         |
| Forecast                    | Forecast_target                  |               |
| Duration                    | Duration                         |               |
| Forecaster                  | Organization                     |               |

D.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            |               |
| Place                       | Country, Group, Organization, Location, State or province, Nationality | global, 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          |

D.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        |
### D.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                               |                     |

### D.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 exports |                     |
| Attribute                 | Financial_attribute                        |                     |
| Reference_point_time      | Date                                       | Friday              |
| Activity                  | Other_activities                           |                     |

### D.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 oil markets   |
| Attribute                 | Financial_attribute                        |                     |
| Difference                | Production_unit                            | forecasts           |
| Forecast                  | Forecast_target                            |                     |
D.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                                 |               |

D.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   |

D.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 administra-tion |
| 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.”
D.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    |

D.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                   | financial     |

D.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.