A Fine-Grained Annotated Corpus for Target-Based Opinion Analysis of Economic and Financial Narratives

Jiahui HU*§ and **Patrick Paroubek†§**
*Natixis CIB, Paris, France
†CNRS, France
§Paris-Saclay University, France
§LISN, Bât 507, Rue du Belvedère, 91400 Orsay, France
⋆jiahui.hu@student-cs.fr
†pap@limsi.fr

Abstract

In this paper about aspect-based sentiment analysis (ABSA), we present the first version of a fine-grained annotated corpus for target-based opinion analysis (TBOA) to analyze economic activities or financial markets. We have annotated, at an intra-sentential level, a corpus of sentences extracted from documents representative of financial analysts' most-read materials by considering how financial actors communicate about the evolution of event trends and analyze related publications (news, official communications, etc.). Since we focus on identifying the expressions of opinions related to the economy and financial markets, we annotated the sentences that contain at least one subjective expression about a domain-specific term. Candidate sentences for annotations were randomly chosen from texts of specialized press and professional information channels over a period ranging from 1986 to 2021.

Our annotation scheme relies on various linguistic markers like domain-specific vocabulary, syntactic structures, and rhetorical relations to explicitly describe the author's subjective stance. We investigated and evaluated the recourse to automatic pre-annotation with existing natural language processing technologies to alleviate the annotation workload. Our aim is to propose a corpus usable on the one hand as training material for the automatic detection of the opinions expressed on an extensive range of domain-specific aspects and on the other hand as a gold standard for evaluation TBOA.

In this paper, we present our pre-annotation models and evaluations of their performance, introduce our annotation scheme and report on the main characteristics of our corpus.

1 Introduction

Financial markets are places where the exchange and the processing of information are crucial in determining prices. Thus, the knowledge about the market participants' opinions is an essential driver of markets. Understanding how these opinions evolve is therefore valuable for financial participants and regulators.

The interaction of market participants determines market dynamics, and how market participants act is largely determined by their beliefs on the outlook. The importance of beliefs in shaping market dynamics reflects an element of a self-fulfilling prophecy: beliefs guide decisions that then validate the underlying belief. One of the classic models of this socio-psychological phenomenon is the Diamond and Dybvig's model (1983) about bank runs. The model demonstrates that even healthy banks can go bankrupt because of depositors' expectations about the behavior of others. If a large number of depositors expect the others to withdraw their funds, the only rational option for them is to be the first one to withdraw the money. If all the depositors decide to withdraw their money, the bank will go bankrupt despite of its initial financial health. In other words, unlike natural science, market dynamics are also determined by its participants' beliefs about the future. An investor buys or sells an asset based on his view about its expected future price. A recent study, for example, finds some evidence that investors' beliefs can be reflected in respondents' multi-asset allocation strategy (Giglio et al., 2021).

People's beliefs are formed as a result of changes in business activities, economic data, and financial outlook. Specialized press and communication of corporates are essential inputs in belief formation. Economists have been using sentiment indices obtained from surveys, such as the Purchasing Manager Index, to gauge the underlying economic activity. One limitation of sentiment indices is that information is restricted by the questionnaire design and update frequency. Consequently, these surveys provide an indication for the general direction of the economy but do not provide much additional

Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 1–12
Punta Cana and online, November 11.

color in terms of "why" things are changing or what the underlying narrative is. One approach to get a better sense of the underlying factors is to analyze news articles. Numerous studies on news articles have been conducted in the past, though without using any algorithms ((Tetlock, 2007), (Baker et al., 2015)) or designing algorithms dependent on sentiment lexicons. For example, (Consoli et al., 2021) use the Loughran McDonald dictionary (Loughran and Mcdonald, 2011) to extract negative emotion from financial newspapers as explainable variable to predict sovereign debt spread. Thanks to recent progress in natural language understanding, it becomes possible to monitor economic narratives and opinions expressed in written texts. There are many approaches to address the problem, e.g., opinion classification (classifying opinions into positive, negative, and neutral) (Malo et al., 2013), topic modeling (probability models for discovering clusters based on the frequency of words in a collection of texts) (Azqueta-Gavaldon et al., 2020).

Although model accuracy of opinion categorization is crucial for its downstream economic and financial analysis tasks, performing fine-grained level target-based opinion analysis on texts written by domain experts is still a relatively unexploited field. In the literature about text mining applications in finance, one encounters most often dictionary-based or unsupervised learning approaches, and the performance of the language model is rarely discussed. One of the reasons for this state of affairs could be the lack of relevant annotated corpora.

To fill this gap, we introduce an annotated corpus that includes both information labeled by human and by algorithms. The corpus consists of texts from different reliable sources; its contents cover corporate news, macroeconomic statements, and comments relevant to financial markets. Each sentence in our corpus is annotated with the following information: (1) domain-specific concepts (2) span of words that potentially indicate the author's opinion (3) named entities (4) negative patterns. If the targets of opinion can be identified, we annotated the pair (target, polarity). In addition to these labels and relations, we also propose to consider specific rhetorical modes like domain experts do.

2 Related Works

Opinion analysis is the task of natural language understanding for classifying texts into positive, negative and neutral polarities. It sees increasing importance in the banking industry among regulators and financial participants. Most existing approaches of opinion analysis applied to economics and finance, however, attempt to detect the overall polarity of a sentence (Malo et al., 2013), or text (Cortis et al., 2017), leaving aside opinion targets.

A study on Twitter sentiment classification showed that 40% of errors resulted from the ignorance of target (Jiang et al., 2011). To the best of our knowledge, FiQA[1] task 1 is the first corpus that labels opinion polarities together with its targets from sentences of microblog and headlines. Nevertheless, its size is relatively small (1,313 samples), and it contains only relatively short sentences[2]. In the texts in which we are interested, sentences written by financial experts are likely to be more complex; as detailed in Table 12, the average sentence length from different texts sources is about 30 tokens, and the most extended sentence is 258.

As of the time of writing, among the analysis tasks, usually known as Aspect-Based Sentiment Analysis (ABSA) ((Pontiki et al., 2014), (Nazir et al., 2020)), there is only one economic news article related study on Target-Based Opinion Analysis (TBOA) (Barbaglia et al., 2020), which focuses on six macroeconomic aggregates. However, the dataset is not publicly available, and details regarding the statistics of this corpus are not specified.

We have created our corpus to respond to the need for TBOA corpus in economy and finance by incorporating argumentative and conditional opinions and opinions expressed inside a consequence or explicit speculation about the future. We are interested in combining two subtasks of ABSA (Pontiki et al., 2014): aspect term extraction (SB 1) and aspect term polarity (SB 2). This corpus aims to provides training and evaluation data to solve the following research problem:

Definition 1. Target-Based Opinion Analysis (TBOA)
Given a sentence of interest with n words **w**= (w_1,...,w_n) and its corresponding word embedding **x**= ($\vec{x}_1$,...,$\vec{x}_n$), the goal of TBOA is to predict **all targets** $\hat{t}$= ($\vec{t}_1$,...,$\vec{t}_m$) which are communicated as central message and their **associated opinions** $\hat{y}$= ($\hat{y}_1$,...,$\hat{y}_m$) **simultaneously**.

[1] https://sites.google.com/view/fiqa/home
[2] The longest sample in FiQA task 1 is composed of 31 words.

3 How we address author's stance

In this work, we are interested in sentences containing at least one expression of opinion about evolution in the economic and financial landscape[3], called subjective sentences. Their particularity is that writers use language to monitor and judge the flow of events or assess their impacts. Authors may:

(1) monitor changes by using language to describe in which direction an event or a concept evolves [a],
(2) express a judgment about these dynamics by clarifying their preference; furthermore their expectations can be diversely grounded in a mix of rationality and/or emotions,
(3) and assess the intensity of these dynamics.

[a] in the DOWN & LOW category, *plummet* and *decrease* convey the notion of scaling *rapid* and *median*, respectively.

Figure 1: Our focus on specific aspects of texts written by financial experts

3.1 Appraisal Theory

As theoretical grounding for our work, we have chosen appraisal theory (Martin and White, 2005) because it provides meaning-making elements to analyze any text that conveys positive or negative assessment, the text's corresponding intensity, and the author's involvement. It deals both with the evaluative language that construes the experience of the world (i.e., development of economy and business activities) and the language that opinion holders use to reveal their (personal) judgments to enact interpersonal relations in a communicative context. Thus, this framework is appropriate for analyzing beliefs that market participants form as a result of changes in the business and economic circumstances.

Under the appraisal framework, the evaluative language resources are divided into three semantic domains, namely

- Attitude: positive and negative assessments. It encompasses values by which a writer reveals his value by emotional response (called *affect*), institutionalized norms or ethics (called *judgment*), aesthetic and social valuation (called *appreciation*).
- Graduation: force and focus. It is composed of meanings by which propositions are strengthened (called *force*), and their boundaries are sharpened or blurred (called *focus*).
- Engagement: resources for positioning a writer's stance with respect to propositions conveyed by a text.

Inspired by the study about appraisal in opinion expressions (Asher et al., 2009), we make three axes to regroup opinion expressions that are relevant to opinions about changes of economic and financial activities:

- **Variation axis** which corresponds to (1) in Figure 1
- **Attitude axis**, i.e. (2) in Figure 1
- **Graduation axis**, i.e. (3) in Figure 1. It is complementary to *Variation axis* and *Attitude axis*. It can describe (a) the intensity, speed or quantity of changes, (b) whether the change happens suddenly or not, and (c) the measurement of quantity, extent, or proximity in time and space (e.g, small & large, a few& many, near & far)

3.2 Rhetorical Modes

Certain rhetorical modes are non-negligible linguistic phenomena in economic narratives, because they give hints to identify central messages in texts written by/for financial professionals. The presence of discourse techniques makes the opinions of some targets more relevant than others, and we only annotate (targets, opinions) which are perceived as central messages. For example, in the following argumentative sentence, author comments both on "*asset x*" and "*interest rates*". We restrict our annotation to the positive opinion toward the target "*asset x*", because the opinion about "*interest rates*" is the premise to support the author's stance on "*asset x*. The performance of the downstream task TBOA will be attenuated if the algorithm fails to distinguish the central message.

"*The asset x is going to be sought out, because I think interest rates ...*"

Argumentation

The author of any text can use a wide range of formulations and argumentative constructions to convince his audience to agree with his views. A common technique that Mario Draghi[4] used in his speeches from Q4 2011 to Q4 2013 is to establish shared premises to develop common ground with his audience, to persuade the world to believe

[3] We will use interchangeably *in the economic landscape* or *in the financial landscape* in the rest of this paper.

[4] President of European Central Bank from November 2011 to October 2019.

in the euro, and reassure that European Monetary Union will recover from the sovereign debt crisis (Jalo, 2021). Establishing shared premises is only a means to his goal: gaining adherence to his conclusion. Recent research in cognitive science defends the view that, in some circumstances, the purpose of reasoning is indifferent from assembling premises that support a conclusion favored by the author (Mercier and Sperber, 2011). The conclusion drawn by an investor may sometimes be more grounded in his desire to favor his own beliefs than in pure rationality.

Cause-effect relation

Similarly, the cause-effect relation represents 7% of the sentences in the speeches of central banks (see Table 10). A cause is an event precedent and contiguous to an effect (Richards, 1965). This succession order implies that the latter, i.e., new information, is more salient. Therefore we only extract opinions expressed inside the effect part of a causal relation.

Conditional Opinions

We also distinguish conditional opinions. In a conditional sentence, opinion expressions can be challenging to determine due to the condition clause (Narayanan et al., 2009). For example, in the sentence, *"If rents fail to keep pace with inflation, the requirement for higher yields will drive down real assets prices."*, the author is pessimistic about *"real asset prices"* under the stated hypothetical scenario, but does not express an opinion on *"rents"*. However, if we remove *"if"* in this sentence, the first clause becomes negative. It is crucial not to misinterpret the author's intent. To do so, we want to enable our algorithm to recognize conditional sentences. Therefore, in addition to annotating opinions inside the main clause, we label the span of words corresponding to the dependent clause.

Explicit speculations

Finally, it is common that an author indicates his level of conviction with respect to his attitudinal propositions. For example, in *"Within our mandate, the ECB is ready to do whatever it takes to preserve the euro. And believe me, it will be enough.*[5]*"*, Mario Draghi conveyed his strong conviction to save the euro; this strong degree of commitment plays an important role in appealing for adherence to his goal. Under the appraisal framework, the formulations related to the author's stance are grouped into the notion of *Engagement*. The two following example sentences[6] indicate respectively greater and lesser degrees of personal commitment from the author to defend his opinion. The level of commitment reflects the author's willingness to open up dialogic space for alternative viewpoints and his confidence in the stated attitudinal proposition.

- *The stock markets are in robust form.*
- *The stock markets are expected to be in a robust form.*

A low degree of belief is, in our standpoint, explicit speculation about the future. People are less inclined to make decisions based upon opinions with a low degree of conviction; thus, these opinions are less likely to be the main driver of market dynamics.

4 Methodology

Supervised learning approaches have showed their efficiency to predict target-oriented opinions in a similar task, i.e. SemEval-2014 Task 4 (Pontiki et al., 2014) which is a widely used benchmark for TBOA. However, to the best of our knowledge, no publicly available annotated corpus corresponds to our research problem. Therefore, we follow the standard scheme of corpus creation in natural language processing to create our corpus for TBOA specific to economy and finance.

The overall picture of our research is illustrated in Figure 2. It is mainly divided into three phases: Data Collection and Pre-annotation (i.e., machine-assisted annotation), Data Annotation, and Model Training. Only the first two phrases will be fully examined in this paper, especially our work in pre-annotation, which delivers satisfying accuracy and reduces laborious annotation workload. Future work will investigate appropriate neural network architecture to solve the task TBOA (see Definition 1) by using this corpus.

[5] source: ECB link

[6] These formulations were invented for demonstration.

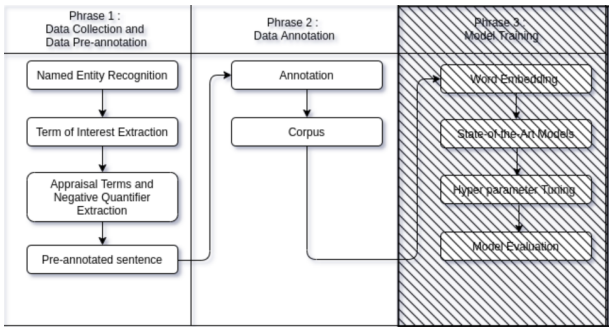

Figure 2: The structure of this study

Before discussing technical details, we want to specify why we use machine-assisted pre-annotations prior to applying manual annotations.

Motivations for adopting pre-annotation

Texts written by experts contain both factual and subjective sentences disproportionately. For our final task, i.e. TBOA, only subjective sentences with terminology terms in economics or finance are relevant candidates.

Definition 2 (Sentence of interest)**.** A sentence of interest is a sentence that includes at least one potential opinion indicator (i.e., appraisal terms) and one potential target (i.e.domain-specific terms, see 4.2.2).

Our pre-annotation experiments show that, except for earning call transcripts, sentences of interest are relatively rare compared to other types of sentences (see Table 6). This disproportionality will result in poor predictive performance on minority class (i.e., sentences of interest) in end-to-end leaning algorithms.

Furthermore, supervised learning models require a tremendous amount of labeled data to discover meaningful patterns in a dataset, even when disposing of a collection of sentences of interest. For a real-world application, creating a large dataset is costly and time-consuming. The primary motivation of automatically annotating certain information is to overcome these challenges by alleviating deep learning algorithm's difficulties in feature-finding and thus encouraging the algorithm to allocate more efforts in linking meaningful features to solve the assigned work. Side benefits are that these pre-annotations can ease the workload of the data annotator and can be represented in a structured way for data analysis from the first phase.

4.1 Data Collection

Our corpus is collected from a wide range of reputable sources of textual information which are complementary to each other, ranging from corporate finance to macroeconomics. While a company tends to attract investors by conveying positive news during earning calls, it usually adopts more prudential tones in the MD&A section of 10-K fillings (see below). News articles and Tweets provide outsider opinions on information communicated by central banks and corporates. Our raw dataset covers a period ranging from 1986 to 2021 (see Figure 5).

The size of raw data set of each data type is detailed in Table 1. Written texts of central banks are collected from their official websites[7]. Sentences of MD&A section (see further explanation below) are randomly selected from the dataset made available by (Ewens, 2019), and our 1,065 sentences of earning calls are randomly chosen from earning transcripts published between October 2017 and May 2021 by 8,912 public companies listed on NASDAQ, NYSE, NYSE MKT and TSX.

Central Banks	Earning Calls	MD&A	Tweets
22,259	1,065	4,793	2,628

Table 1 Number of sentences intended for pre-annotation

- **Central banks:**
  - ECB: Press conferences about ECB's monetary policy decision and speeches of ECB
  - FOMC: Meeting minutes of Federal Open Market Committee
- **MD&A**[8] (Ewens, 2019): Management Discussion and Analysis (item 7) is a mandatory and non-audited section of 10-K filings, an annual report that corporates submit to the U.S. Securities and Exchange Commission. MD&A represents the thoughts and opinions of the management of a public company[9] (called the C-suite) and provides a forecast of future operations. According to a survey of 140 sell-side analysts (Epstein and Palepu, 1999), the MD&A section is well-read and used.
- **Earning calls:** During earning call confer-

[7] source of ECB Speeches :link
source of ECB Press Conferences: link
source of FOMC minutes: link
[8] source of MD&A data: link
[9] A public company is a company whose shares are traded freely on a stock exchange.

ence, the C-suite, analysts, investors, and the media discuss the company's financial results and future plans. Participants can ask questions that may yield valuable information or ask for clarification on particular topics which were not addressed by the C-suite. They are considered as one of the key resources for investors. Financial analysts combine information obtained during the earning calls and from the MD&A section in fundamental analysis. A study on German companies finds some evidence that earning calls improve analysts' ability to forecast better future earnings of German companies (Bassemir et al., 2013).

- **News articles:** We use the Financial PhraseBank (FPB) dataset (Malo et al., 2013) as a composition of news article sentences. It is composed of 4,840 annotations at the sentence level. We randomly choose sentences from the corpus with 75% agreement (i.e. 3,453 sentences) among 5 - 8 annotators per sentence , and apply our annotations (see section 4.4.1). Note that we do not reuse any annotation of the FPB dataset and a small number of our polarity annotations differ from those of the original corpus.
- **Tweets**[10]**:** Social media messages of 19 domain experts.

4.2 Data pre-annotation pipeline

4.2.1 Named Entity Recognition

Named Entity Recognition (NER) is the task of detecting token-level instances of named entities from unstructured texts into pre-defined categories. Geographic locations (LOC), persons (PER), names of organizations or companies (ORG) are important hints for linking financial narratives to the corresponding entities. In *"Stock markets in [Europe$_{(B-LOC)}$] are expected to be in a robust form."*, the location [Europe$_{(B-LOC)}$] is an attribute of our term of interest *"Stock markets"* (see subsection 4.2.2). Its identification helps to target investors and analysts who follow the information flow of specific entities. It also enables joint analysis with structured numeric financial data organized by corporate name or geographic unit.

In order to choose an adapted model to label our dataset, we employ the CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003), one of the most famous corpora of NER, to evaluate different state-of-the-art models' performance for this task (see section 4.3). In additional to [LOC], [ORG], [PER] classes, [MISC] (miscellaneous) is the category of words derived from location (e.g., *Italian*), organization, person. Sentences of this dataset are taken from the Reuters Corpus[11], which is composed of Markets & Finance news of 90 categories[12]. 14,035 sentences were used for training, 3,250 for development and 3,453 for evaluation.

We have experimented with four NER systems, first three open-source libraries: SpaCy[13], Stanza (Qi et al., 2020)[14], and FLERT (Schweter and Akbik, 2020). In the FLERT model, firstly a transformer is fine-tuned on the NER task, and then the resulting features are provided to a BiLSTM-CRF (Huang et al., 2015) sequence labeling architecture. We also modeled a transformer with self-attention heads using BERT embeddings (Devlin et al., 2019) to predict named entities.

Model	[LOC]	[ORG]	[PER]	[MISC]	Time (min)
SpaCy	0.71	0.23	0.67	/	0.19
Stanza	0.92	0.74	0.89	0.85	7.55
BERT	0.93	0.87	**0.97**	**0.95**	8.07
FLERT	**0.98**	**0.88**	**0.97**	0.92	19.81

Table 2 F1-scores and computation time of the test set of CoNLL 2003 dataset

From Table 2, we conclude that classification precision is proportionate with computation cost. FLERT performs the best across all classes, its improvement on the class [LOC] with respect to BERT (the second-best model) are significant. Nonetheless, its computation costs is much higher than our BERT model. For this reason, we use the BERT model in our pipeline to automatically annotate named entities.

4.2.2 Term of Interest Extraction

The goal of terminology extraction is to locate relevant terms from unstructured texts. Terms of interest (TOI) are concepts in economics and finance in the form of their full name or abbreviation, including but not limited to:

- market-related terms: *FX, Forex markets, S&P 500 futures, traded commodities, etc.*
- accounting drivers: *free cash flow, ROA or return on assets, etc.*
- valuation divers: *EBITDA, EPS, NAV, etc.*
- macroeconomic aggregates: *growth rate, inflation, CPI, PMI, etc.*

[10]Obtained with Twitter API

[11]http://www.reuters.com/researchandstandards/

[12]for more details, see https://martin-thoma.com/nlp-reuters/

[13]Trained on *OntoNotes 5* dataset (18 classes) https://spacy.io/api/entityrecognizer

[14]https://stanfordnlp.github.io/stanza/ner.html

- risks factors: *VIX, volatility, CDS spread, etc.*
- others

To the best of our knowledge, there is no labeled dataset in economics & finance for terminology extraction. Therefore, we choose to identify these terms in an unsupervised manner: we firstly extract terminological candidates by identifying syntactically plausible noun phrases. Noun phrases that contain elements of a domain-specific thesaurus are labeled as terms of interest.

4.2.3 Appraisal Terms Extraction

The aim of appraisal terms extraction is to identify potential candidates in which an author expresses his opinion. More details about the composition of appraisal terms are provided in Table 13. Appraisal terms are pre-annotated with Algorithm 1.

Algorithm 1 key terms extractor

Input: sentence, predefined_list, tag
1: S' = spacy.nlp(sentence)
2: **for** $token$ in S'
3: **if** $token$ is in predefined_list **then do**
4: $res \leftarrow token$, start_char, end_char, tag
5: **end if**
6: **end for**
7: **return** res

4.2.4 Negation

Negative quantifiers are assigned to an appraisal expression that inverses the appraiser's perception towards a target. Negative quantifiers are also pre-annotated with Algorithm 1. The list of negative patterns is detailed in Table 14.

4.3 Data Pre-annotation Evaluation Measures

To evaluate the performance of the chosen models for NER (i.e., BERT) and TOI, the annotator is asked to correct wrongly pre-annotated tags or add missing ones. The performance of pre-annotation our corpus is measured with F1-score:

$$F1 = \frac{2*precision*recall}{precision+recall}$$

At the time of writing, performance evaluation is calculated based on 2,297 sentences annotated by one annotator familiar with the domain terminology (one of the authors).

Evaluation of NER

NER is labeled with the IOB (inside, outside, beginning) format:

"[The$_O$] [European$_{(B-ORG)}$] [Commission$_{(I-ORG)}$] [said$_O$] [on$_O$] [Thursday$_O$] [it$_O$] [did$_O$] [agreed$_O$] [with$_O$] [German$_{(B-MISC)}$] [advice$_O$] [to$_O$] [consumers$_O$] [to$_O$] [shun$_O$] [British$_{(B-MISC)}$] [lamb$_O$] [until$_O$] [scientists$_O$] [determine$_O$] [whether$_O$] [mad$_O$] [cow$_O$] [disease$_O$] [can$_O$] [be$_O$] [transmitted$_O$] [to$_O$] [sheep$_O$]." .

The model performance is evaluated on entity-level with Seqeval[15] (Nakayama, 2018). We use its default mode compatible with *conlleval*, an evaluation system developed for the CoNLL-2000 shared task. The prediction of *[European Commission]* is considered as *True Positive* when the model return [B-ORG, I-ORG]. When the model makes a boundary error, for example [I-ORG, I-MISC], the entity is counted as two errors (false negative and false positive).

Class	Precision	Recall	F1-score	Nb entities
LOC	96%	99%	97%	336
ORG	96%	93%	95%	888
PER	85%	98%	91%	150
MISC	88%	72%	79%	412

Table 3 The performance of our corpus using the BERT model

The performance of [LOC] and [ORG] is slightly better on our corpus than on the test set of the CoNLL 2003 dataset. This is related to how machine-assisted annotation are validated for the evaluation of our corpus. The annotator corrects the label of NER if necessary. For our intended usage, we are more permissive to boundary variations. For example, under the CoNLL 2003 annotation guideline, *[Finnish KCI Konecranes]* should be labeled [B-MISC, B-ORG, I-ORG]. But the labels [B-ORG, I-ORG, I-ORG] given by the algorithm are accepted by the annotator, because the information valuable to our intended usage, i.e., company name, is identified correctly. The type of boundary variations which are not accepted in the CoNLL 2003 annotation guideline [16] but accepted in our corpus are listed below:

- articles: a, an, the
- honorifics: Mr., Ms., etc.
- *[MISC]* that precedes an organization

Evaluation of TOI

We consider the identification of TOI as a task of sequence labeling. Under the IOB format, the TOI *Forex market* can be tagged with ['B-TOI', 'I-TOI']. Its evaluation is computed with the Seqeval

[15] A python framework for sequence labeling evaluation. https://github.com/chakki-works/seqeval

[16] https://www.clips.uantwerpen.be/conll2003/ner/annotation.txt

(Nakayama, 2018).

Model	Precision	Recall	F1-score
our method	88%	90%	89%

Table 4 Model performance for identifying TOI in and out of the thesaurus

4.4 Annotation Process

After the pre-annotation phase, only sentences of interest are retained. The portion of sentences of interest varies from one type of data to another.

Type	CB	EC	MD&A	Tweets
Percentage of sentenced of interest	26%	63%	27%	15%
Randomly chosen sentences	22,259	1,065	4,793	2,628

Table 5 Percentage of sentences of interest from randomly chosen sentences

At the time of writing, we have annotated 2,297 sentences randomly selected from pre-annotated sentences of interest. The list of possible tags, labeled by algorithms and by human, can be found in Table 11. Our final corpus size is expected to be 4,000 sentences[17].

Type	CB	EC	News	MD&A	Tweets	Total
Expected Size	800	800	800	800	800	4, 000
Percentage annotated	87.5%	83.9%	93.8%	12.5%	9.5%	57.4%

Table 6 Size (number of sentences) of the corpus

4.4.1 Data Annotation

Our corpus is annotated with INCEpTION (Klie et al., 2018), an open-source annotation platform. As exemplified in Figure 3, pre-annotated tag [TOI] stands for terms of interest and [APPRAISAL] for appraisal terms. The annotator is expected to **identify all targets towards which opinions are expressed, as well as their polarities**. Following the evaluation campaign DEFT 2018 (Paroubek et al., 2018), the annotator

(i) selects minimal information about the *Opinion & Emotion Expression* (i.e. *"dysfunctional"*, tagged [OEE]),

(ii) selects the most complete information about the target (i.e. *Sovereign bond market* in Figure 3) and attributes polarity (tagged [-] in Figure 3) to it,

(iii) then draws an unlabeled arc from [OEE] toward its corresponding target.

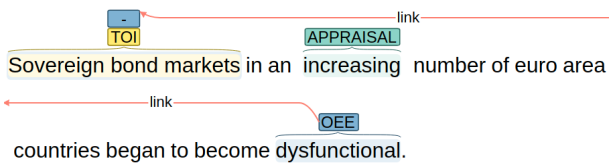

Figure 3: Example of a annotated sentence (explicit opinion)

While the opinion toward target *Sovereign bond market* displayed in Figure 3 is explicit, our corpus also contains implicit opinions based on knowledge of the world, such as Figure 4:

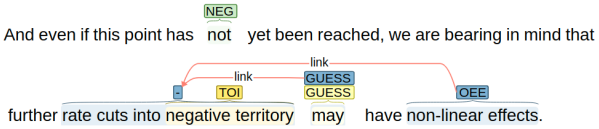

Figure 4: Example of a annotated sentence (implicit opinion)

We define four classes of polarities, namely positive, negative, noncommittal, and unknown; our corpus also contains sentences of interest for which no target can be identified. Polarities are attributed based on the following stance:

- the judgment related to the health of corporates and economic activities
- the judgment related to economic norms, which means *financial conditions* of an economy that contribute to financial stability (Arrigoni et al., 2020). The positive opinion of a target is associated with an easy access to finance. It means that economic agents (individuals, enterprises, or governments) can obtain adequate financial services, including credit, deposit, payment, insurance, and other risk management services.

As mentioned before, in the current stage the annotation was done by one of the authors.

4.4.2 Summary of the annotated corpus

At the time of writing, the annotated corpus contains 2,297 sentences labeled by human annotator.

Summary	CB	EC	News	MD&A	Tweets	Total
# sent. w. target(s)	433	438	540	61	50	1,522
# sent. w/o target	267	233	210	39	26	775
# sent.	700	671	750	100	76	2,297

Table 7 Statistics about annotated sentences with (w.) and without (w/o) target(s)

The following statistics were extracted from the 1,522 annotated *sentences w. target(s)*, called *subjective sentences* [18].

Aligned with our hypothesis, C-suites tend to convey positive information during earning calls, and central bankers are more likely to comment on both positive and negative sides of the economic environment (see Table 8). Targets of our corpus are composed of more tokens than the corpora of SemEval-2014 on Laptop and Restaurant Review (Pontiki et al., 2014) (see Table 9). When it comes to the usage of language, texts of central banks are more likely to express opinions with rhetorical modes, i.e. "argumentative opinions", "explicit

[17]We have not yet reached the desired size for this corpus; we plan to make the corpus available once kappa will have been measured and legal aspects have been assessed.

[18]Due to the small size of annotated sentences from MD&A and Tweets, we refrain from comments on their statistics.

expectations" and "opinions as consequences" (see bold-italic items in Table 10).

Polarity	CB	EC	News	MD&A	Tweets
Positive	45%	74%	68%	46%	45%
Negative	39%	17%	27%	50%	50%
Non-committal	16%	8%	5%	4%	4%
Unknown	0	1%	0	0	1%
Total opinions	710	706	626	102	74
Total nb sentences w. target(s)	433	438	540	61	50

Table 8 Statistics about polarities from subjective sentences

# tok.s	CB	EC	News	MD&A	Tweets	Laptop	Restau.
1	22%	31%	28%	23%	53%	62%	75%
2	38%	38%	49 %	31%	30%	29%	17%
≥ 3	40%	31%	23%	46%	17%	9%	8%
# targ.	710	706	626	102	74	2,966	4,728

Table 9 Statistics about number of tokens per target of our corpus v.s. the SemEval-2014 Task 4

Type	CB	EC	News	MD&A	Tweets
argumentative opin.	***5.76%***	4.35%	4.07%	6.56%	6%
conditional opin.	3.23%	1.14%	0%	3.28%	0%
explicit expect.	***7.16%***	2.51%	4.99%	6.56%	6%
opin. as conseq.	***6.93%***	0.91%	2.04%	11.48%	0%
Total	23.07%	8.91%	11.03%	27.87%	12%
#nb sent. w. target(s)	433	438	540	61	50

Table 10 Statistics about specific elements of our corpus

5 Conclusion

This paper defines an annotation scheme for identifying target-based opinions from sentences related to the economy and financial markets, by incorporating specific-domain use of language, namely formulations of argumentative and conditional opinions, opinions expressed in cause-effect relations, and opinions with a low level of commitment. The contribution of this paper can be summarized as follows:

1. Our annotation scheme considers rhetorical modes, which is, as far as we know, a novelty in opinion analysis applied to economic and financial narratives. It enables our corpus to describe faithfully how financial professionals communicate and analyze the evolution of events and changes in business data.
2. We designed a tailor-made pipeline by leveraging a set of techniques in Natural Language Processing (NLP) to pre-annotate informative features for identifying candidates relevant to TBOA and alleviating the workload of annotators. At this stage of our project, these pre-annotation models evaluation scores are significant enough to assume that most pertinent sentences are retained for TBOA.
3. Our manually annotated corpus responds to the need for high-quality training and evaluation data for the development of supervised learning algorithms in the research field of TBOA applied to economy and finance.

In the future, we want to develop neural models adapted to our corpus by incorporating fundamental approaches in NLP and domain-specific knowledge, in particular, we investigate the contribution of data augmentation approach to augment the size of our corpus.

Acknowledgment

We express our deepest thanks to Dr. Dirk Schumacher (Natixis Macroeconomic Research) for his valuable comments and support on this paper.

This work is supported by the grant CIFRE, a partnership between Natixis CIB Research and the LISN Laboratory (Interdisciplinary Laboratory of Digital Sciences).

References

Simone Arrigoni, Alina Bobasu, and Fabrizio Venditti. 2020. The simpler the better: measuring financial conditions for monetary policy and financial stability. Working Paper Series 2451, European Central Bank.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of Opinion Expressions in Discourse. *LI*, 32(2):279–292.

Andres Azqueta-Gavaldon, Dominik Hirschbühl, Luca Onorante, and Lorena Saiz. 2020. Economic policy uncertainty in the euro area: an unsupervised machine learning approach. Working Paper Series 2359, European Central Bank.

Scott Baker, Nicholas Bloom, and Steven Davis. 2015. Measuring economic policy uncertainty. NBER Working Papers 21633, National Bureau of Economic Research, Inc.

Luca Barbaglia, Sergio Consoli, and Sebastiano Manzan. 2020. Forecasting with economic news. *Available at SSRN 3698121*.

Moritz Bassemir, Zoltan Novotny-Farkas, and Julian Pachta. 2013. The effect of conference calls on analysts' forecasts – german evidence. *European Accounting Review*, 22(1):151–183.

Sergio Consoli, Luca Tiozzo Pezzoli, and Elisa Tosetti. 2021. Emotions in macroeconomic news and their impact on the european bond market. *Journal of International Money and Finance*, 118:102472.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Marc J. Epstein and Krishna G. Palepu. 1999. What financial analysts want.

Michael Ewens. 2019. Mda statements from public firms: 2002-2018.

Stefano Giglio, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus. 2021. Five facts about beliefs and portfolios. *American Economic Review*, 111(5):1481–1522.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jyri Jalo. 2021. JyriJalo-MastersThesis. page 82.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, page 151–160, USA. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Tim Loughran and Bill Mcdonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2013. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *arXiv:1307.5336 [cs, q-fin]*. ArXiv: 1307.5336.

J. Martin and Peter White. 2005. *The Language of Evaluation: Appraisal in English*.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *The Behavioral and brain sciences*, 34:57–74; discussion 74.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, page 180–189, USA. Association for Computational Linguistics.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, pages 1–1.

Patrick Paroubek, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, Amel Fraisse, Agata Jackiewicz, Jihen Karoui, Laura Monceaux, and Juan-Manuel Torres-Moreno. 2018. DEFT2018 : Recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de France. In *DEFT 2018 - 14ème atelier Défi Fouille de Texte*, volume 2 of *Actes de la conférence Traitement Automatique des Langues, TALN 2018*, pages 1–11, Rennes, France.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Thomas J. Richards. 1965. Hume's two definitions of 'cause'. *The Philosophical Quarterly (1950-)*, 15(60):247–253.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

A Appendix

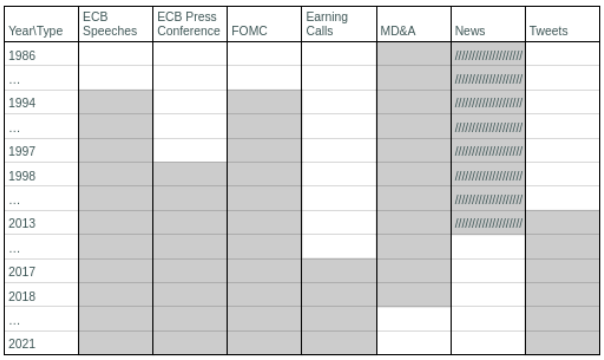

Figure 5: Shaded slashes of the column 'News' indicate that the time range of news sentences from the Financial PhraseBank dataset is incognito.

Labeled	Groups	Tags	Definition	Arrow
by algorithms	Named Entities	PERSON	people's name	None
		GEO	regions, countries	
		ORG	companies or institutions	
	Other Elements	NEGATION	absolute, partial or verbal negation	from [NEGATION] to its relevant [OEE]
		TOI	terms of interest in economics and finance	None
		APPRAISAL	opinion and emotion expressions	
by human	Opinion & Emotion Expressions	OEE	word of the span of words which indicated opinion holder's appraisal	from [OEE] toward its corresponding target
	Targets	-	target toward which a negative opinion is expressed	None
		+	target toward which a positive opinion is expressed	
		non-committal	target toward which opinion holder do not express their thoughts, attitude or intentions clearly.	
		unknown	annotators do not know which polarity to attribute for a particular concept/jargon, but **this tag should be used with caution**	
	Rhetorical Modes	RESULT	the span of words that declares: " something produces another thing as a result"; relation (cause, consequence) expressed by author	from [RESULT] to its relevant targets
		GUESS	explicit speculation about future	from [GUESS] to its relevant targets
		PREMISE	arguments used by opinion holder to support his opinion	None
		CONDITIONAL	expressed opinion are conditional on a specific clause	

Table 11 The list of possible tags labeled by algorithms and by human

Type	CB	EC	News	MD&A	Tweets	Laptop	Restaurant
min	6	8	6	13	10	3	7
max	77	80	53	258	27	83	79
mean	28.65	27.22	24.96	34.79	17.52	18.93	17.16
standard deviation	12.3	11.75	10.29	26.02	3.92	10.78	8.94

Table 12 The statistics about sentence length (i.e. number of tokens) by type of data source

Axis	Group	Definition	Example
Variation	UP/HIGH	gain in quantity or volume	rise, grow, upturn, recovery ...
	DOWN/LOW	loss in quantity or volume	escalate, downturn, plummet, ...
	STABLE	stable state	maintain, unchanged, hold, ...
	OTHERS	other expressions of variation	accumulation, surge ...
Attitude (rational)	APPRECIATE	recognition of value	good, well, sufficient ...
	NO_APPRECIATE	recognition of loss in value	inadequate, erosion, destructive..
	UNCERTAINTY	lack of visibility	risk, consolidation, speculative...
	ALARM	anxious awareness of undesirable outcome	pressure, tension, danger ...
	NON_SURPRISE	indicators to declare that observation are close to one's expectation.	in line with our expectation ...
	OTHERS	other expressions of attitude, such as multi-word expressions	surprise, surprisingly, surprised, ...
Attitude (emotional)	Over-confidence	intense feeling of excitement and strong desire to put ideas into practices	thrill, euphoria, greed,...
	FEAR	feeling of helpless, without any degree of control on the situation	anxiety, denial, panic...
	OTHERS	other expressions of over-pessimism or over-optimism, such as multi-word expressions	
Graduation	STRONG	high intensity	extremely, completely, rapid...
	WEAK	low intensity	gradually, slowly, steadily ...
	OTHERS	other expressions of intensity, strength and focus	

Table 13 Appraisal terms to monitor, judge the development an event or a concept, or to measure its impact.

	Words
Negative pattern	scarcely, scarce, fail, no more, nowhere, not, no, none, neither, little, deficiency, no longer, few, nowise, seldom, miscarry, in no manner, miscarry, nor, to no degree, hardly, zilch, nobody, cipher, by no means, lack, rarely, barely, nothing, deny, without, null, in no way, never, nada, cypher, naught, nix, zero, absence, nil, negate

Table 14 List of possible negative patterns

Type	CB	EC	News	MD&A	Tweets	Laptop	Restaurant
w. 1 target	61%	60%	86%	61%	60%	64%	51%
w. 2 targets	25%	26%	12 %	28%	32%	23%	29%
$\geq$ 3 targets	14%	14%	2%	11%	8%	13%	20%
Nb sentences w. target(s)	433	438	540	61	50	2,966	4,728

Table 15 Statistics about number of targets per sentence of our corpus v.s. the SemEval-2014 Task 4