GADFA: Generator-Assisted Decision-Focused Approach for Opinion Expressing Timing Identification

Chung-Chi Chen,[1] Hiroya Takamura,[1] Ichiro Kobayashi,[2] Yusuke Miyao,[3] Hsin-Hsi Chen[4]
[1] Artificial Intelligence Research Center, AIST, Japan
[2] Ochanomizu University, Japan
[3] University of Tokyo, Japan
[4] Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
c.c.chen@acm.org, takamura.hiroya@aist.go.jp,
koba@is.ocha.ac.jp, yusuke@is.s.u-tokyo.ac.jp, hhchen@ntu.edu.tw

Abstract

The advancement of text generation models has granted us the capability to produce coherent and convincing text on demand. Yet, in real-life circumstances, individuals do not continuously generate text or voice their opinions. For instance, consumers pen product reviews after weighing the merits and demerits of a product, and professional analysts issue reports following significant news releases. In essence, opinion expression is typically prompted by particular reasons or signals. Despite long-standing developments in opinion mining, the appropriate timing for expressing an opinion remains largely unexplored. To address this deficit, our study introduces an innovative task - the identification of news-triggered opinion expressing timing. We ground this task in the actions of professional stock analysts and develop a novel dataset for investigation. Our Generator-Assisted Decision-Focused Approach (GADFA) is decision-focused, leveraging text generation models to steer the classification model, thus enhancing overall performance. Our experimental findings demonstrate that the text generated by our model contributes fresh insights from various angles, effectively aiding in identifying the optimal timing for opinion expression.

1 Introduction

Opinion mining has been a popular topic for a long time (Liu, 2012). We are now able to perform well in sentiment analysis (Wu and Shi, 2022), aspect-based sentiment analysis (Ling et al., 2022; Chen et al., 2022), opinion helpfulness estimation (Diaz and Ng, 2018), and other opinion understanding tasks (Chen et al., 2021; Shi et al., 2022). With the development of large pre-trained language models, we have also made significant improvements in synthesizing opinions and arguments (Wachsmuth et al., 2018; Hua et al., 2019; Schiller et al., 2021; Chen and Takamura, 2024). Although there are many discussions on how opinions are generated and what should be generated, few studies pay attention to when we should generate opinions. Since models can generate text based on given input at any time, timing becomes an important issue. For example, there are many news articles every day, but professional analysts do not write comments on every article and only release reports a few times per year. Following this line of thought, this paper proposes a novel task: identifying the timing of expressing opinions. The aim of this task is to learn when is the proper timing to express opinions.

Professionals' opinions, attitudes, and behaviors are important for the masses' decision-making and the future of the field, market, country, and so on. For example, the opinions of the professionals in the Centers for Disease Control and Prevention (CDC) will influence people's attitudes toward COVID-19 in the past three years, and politicians' attitudes will influence the country-level and even global-level political situations. In the financial market, the opinions and behaviors of the professionals are shown to be related to different market properties (Hirst et al., 1995; Niehaus and Zhang, 2010; Kothari et al., 2016; Kim and Ryu, 2022). Given the importance of professionals' opinions, attitudes, and behaviors, this paper aims to forecast professional stock analysts' behaviors based on the latest publicly-available information, i.e., news.

When the latest news was published, the first decision that analysts will make is whether they should write a report to update their view and explain it to their customers. That is, if the news does not matter to the company's operation or stock price, analysts will not release reports to share their opinions. Because investors will react to the released reports (Hirst et al., 1995; Niehaus and Zhang, 2010), we argue that this is an essential task in analysts' behavior modeling, and it is also considered as the opinion expressing timing identification task in this paper. To the best of our

Proceedings of the 31st International Conference on Computational Linguistics, pages 10781–10794
January 19–24, 2025.

knowledge, this paper is the first work to explore the news-triggered opinion expressing timing identification task.

Training classification models with all news as input is an intuitive way to address the proposed task. However, given the length limitation of models' input, generating news summarization before training a classifier is the other rational direction for news-triggered classification tasks. Different from the above directions, this paper attempts to answer the question of whether we can improve the performances by adding an opinion generator in the loop. The rationale of our design for this approach is that we want to let models mimic the decision process of professionals. Instead of directly deciding whether to write a report after reading the news, professionals will first form their opinions toward the given news, and then further decide whether to share this opinion with their customers. Therefore, we propose a Generator-Assisted Decision-Focused Approach (GADFA), which first trains a generator to generate analysis based on each news, and further uses the generated opinions and the news as models' input. The important difference between the generated opinions and summarizations is that opinions contain subjective information which is not included in the original articles. In contrast, summarization is just to rewrite and shorten the given news.

In sum, this paper makes the following contributions: (1) We propose a novel task: news-triggered opinion expressing timing identification with a new dataset. (2) We design a decision-focused approach for enhancing the performances. (3) We provide in-depth discussions on the influence of using different generators and using cross-generators in the proposed approach, and also analyze the generated text from several aspects.

2 Related Work

As it is relatively easy to collect textual data, such as tweets or news articles, and align them with market prices, numerous datasets have been developed for predicting market information, including price movement prediction (Xu and Cohen, 2018; Li et al., 2020b) and volatility forecasting (Qin and Yang, 2019; Li et al., 2020a). In our belief, short-term price movement follows the random walk hypothesis (Fama, 1995), and several asset pricing models have adopted this concept to model asset price movement, such as the Black–Scholes model (Black and Scholes, 1973) for option pricing. Building upon this notion, we argue that learning to make professional decisions is a more tangible direction. Thus, this paper aligns news articles with professionals' behavior, specifically whether professional analysts will release reports after the news is published. The proposed task has several downstream applications. For example, choosing the timing to share opinions is an important task when constructing an AI analyst because we expect the AI analyst to only share important information instead of generating numerous unnecessary explanations. In other words, rather than sharing generated text continuously, the AI analyst also needs to select the timing of expressing opinions.

Analysts' behavior has been a long-standing topic in financial literature. Some studies focus on analyzing reports and market reactions. For instance, Devos et al. (2015) discuss the market response to changes in analysts' views, indicating that analysts' view changes are informative for investment, particularly for stocks with less transparency. Hsieh et al. (2016) examine the readability of reports and its impact on stock returns, finding that readability plays a significant role in eliciting positive reactions from the market. Others attempt cross-document analysis. Conrad et al. (2006) explore the relationship between analyst recommendations and major news. Keith and Stent (2019) model changes in analysts' views based on pragmatic and semantic features of earnings calls. To summarize our survey, there are existing discussions on analysts' view changes, but no previous study has focused on the timing of opinion expression. Furthermore, there is currently no publicly-available dataset for investigating the expressing timing of analysts' opinions. The dataset proposed in this paper is the first of its kind released for such a task.

3 Dataset

3.1 Task Design

While numerous new events occur daily in the financial market, originating from various information sources, most studies assume that news articles contain the most up-to-date information and also summarize information from other documents such as financial reports or company meetings. Additionally, news articles may report on popular discussion threads from social media platforms. Therefore, we have chosen news articles as the primary source of

	Release Report	Not Release Report
Train	2,717	2,717
Development	322	322
Test	325	325
Total	6,728	

Table 1: Statistics of the dataset.

	$T = 0$	$T = 5$
One News	3,632	3,051
More Than One News	3,096	3,677

Table 2: Statistics of the number of news.

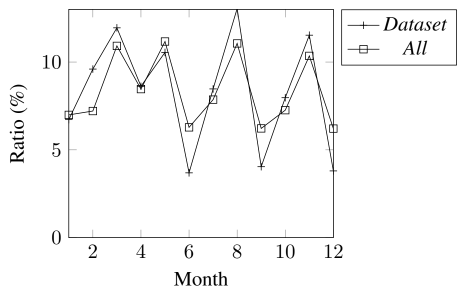

Figure 1: Distribution of report release date within year. Dataset and All denote the statistics of the experimental dataset and all reports from 2014 to 2020, respectively.

information for our models.

The proposed task of identifying the timing of opinion expression triggered by news is defined as follows. Given the news related to the target stock from day $t - T$ to t, our objective is to predict whether at least one professional analyst will release an analysis report on day $t + 1$. In this paper, we conduct experiments for two different values of T, namely $T = 0$ and $T = 5$, to discuss the timeliness of the news.

3.2 Dataset Creation Process

To construct the dataset for the proposed task, we follow a series of steps. Firstly, we download all analysis reports for the Taiwan stock market from Bloomberg Terminal.[1] Additionally, we obtain the Chinese news released by two major financial news vendors, Economic Daily News[2] and Commercial Times.[3] Secondly, we align these data based on their respective release times. It is important to note that this dataset covers the period from 2014 to 2020, comprising a total of 401,559 news articles and 40,205 reports. Thirdly, we filter the instances that have news on day t and at least one report released on day $t + 1$. This filtering process results in 3,364 positive instances labeled as "Release Report". Instead of randomly selecting negative instances labeled as “Not Release Report,” we control the target stock based on the positive instances. This means that the negative instances are selected from the same stock pools as the positive instances. The reason for controlling the target stock is that previous studies have shown potential bias towards the target stock in pre-trained models (Chuang and Yang, 2022) and managers’ gender (Sawhney et al., 2021). Using this approach, we identify negative instances that have news on day t but no report released on day $t + 1$.

Table 1 presents the number of instances in the training, development, and test sets. Additionally, Table 2 provides statistics on the number of news articles under different settings of T based on the opinion expressing timing detection dataset. We observe that when T is set to 0 and 5, 46.16% and 54.65% of instances, respectively, have multiple news articles. Furthermore, it indicates that 8.64% ((3,677-3,096)/6,728) of stocks are mentioned on different days during the five-day period. These results motivate us to explore the timeliness of information and the potential benefits of summarizing multiple news articles in the proposed tasks.

3.3 Distribution of Report Release Date

This section presents statistics on the distribution of report release dates. Figure 1 illustrates the statistics within the year. In addition to the statistics based on the experimental dataset, we include the statistics based on all reports from 2014 to 2020 for comparison. Firstly, the distributions based on the experimental dataset and all reports are similar. Secondly, we observe that analysts tend to release more reports in March, August, and November, while fewer reports are released in June, September, and December. One reason for the increased number of reports in March, August, and November is that companies are required to release their yearly, second-quarter, and third-quarter financial statements before the end of these months. The first-quarter financial statement is expected to be released before the end of May, resulting in the fourth-highest number of reports in May. Since professional analysts do not frequently change their views, the months following the financial statement release dates, i.e., June, September, and December, have fewer reports.

[1] https://www.bloomberg.com/professional/products/bloomberg-terminal/
[2] https://money.udn.com/money/index
[3] https://ctee.com.tw/

Period	Dataset	All
Beginning of the month	29.04%	29.73%
Middle of the month	34.48%	35.41%
End of the month	36.47%	34.85%

Table 3: Distribution of release date within month.

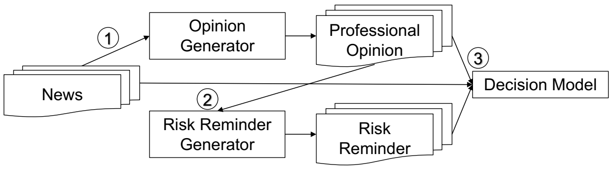

Figure 2: Illustration of the proposed GADFA.

Table 3 presents the distribution of report release dates within a month. It is observed that analysts release fewer reports at the beginning of the month, while the ratios at the middle and end of the month are similar. This trend can be attributed to the requirement for companies to disclose the previous month's revenue before the 10th of each month, with most companies releasing this information around that time. Consequently, professionals tend to release reports during the middle of the month to incorporate the latest information into their opinions. Nevertheless, there are still instances of reports being released at other times. The proposed task and dataset serve as a testbed for identifying the optimal timing of expressing an opinion based on news information. The provided approach and dataset offer avenues for future research, such as enhancing the performance of the proposed GADFA or developing a new end-to-end model to address the proposed task.

4 Method

Figure 2 illustrates the proposed GADFA. The design of GADFA is inspired by the notion of decision-focused approaches (Wilder et al., 2019; Hsu and Tan, 2021; Mandi et al., 2022), which involve making several predictions first and then using these predictions as input to an optimization algorithm for generating a decision. We believe that this pipeline approach can also be applied to natural language processing (NLP) applications, as the generated text can contain multiple information pieces from different aspects, such as sentiments and analysis. Considering the behavior model of professionals, we consider generating an opinion on the given news and then making further decisions as an ideal way to perform decision-focused learning for the proposed task.

News	Key takeaways from Quanta's 1Q22 analyst call included: 1) 1Q22 GMs dipped to an eight-quarter low due to inferior product mix and inefficient production; 2) weak 2Q22 earnings outlook with both PC demand weakness and supply constraints…
Professional Opinion	We believe that QCT could suffer from a potential slowdown in enterprise spending, while China hyperscalers' demand also seems a bit mellowed down in the near term. We forecast 23%/10% server revenue growth for Quanta in 2022/23.
Risk Reminder	Key downside risks include margin erosion in servers and a potential slowdown of PC demand post Covid-19.

Table 4: Examples of the news, professional opinion, and risk reminder.

Therefore, the proposed GADFA consists of three steps. Firstly, the opinion generator generates a professional opinion based on the given news. Table 4 presents examples of professional opinions, which include subjective views on whether investors should buy or sell the related stocks mentioned in the news. Secondly, the risk reminder generator generates a risk reminder based on the generated opinions. Table 4 provides examples of risk reminders, which indicate possible risks that may affect the accuracy of the generated opinion. Finally, the decision model combines the given news, generated opinions, and generated risk reminders to determine the appropriate timing for expressing opinions.

As shown in Figure 2, the proposed GADFA consists of two generators and one decision model. We first generate a professional opinion based on the given news and then generate a risk reminder based on the generated opinion. The decision model then fuses the news, professional opinion, and risk reminder to make the final decision on whether to release a report or not. For the generators, we select three well-performing pre-trained language models for comparison:

- Multilingual T5 (mT5) (Xue et al., 2021) covers 101 languages and was pre-trained on a Common Crawl-based dataset.

- Pegasus (Zhang et al., 2020)[4] is pre-trained with a masked sentence generation task and performs well in summarization tasks.

[4]The Chinese Pegasus is provided in UER toolkit (Zhao et al., 2019)

	Opinion Generator				Risk Reminder Generator			
	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score	ROUGE-1	ROUGE-2	ROUGE-L	BERT Score
mT5	0.1238	0.0475	0.1195	0.6471	0.6630	0.4878	0.6554	0.8581
Pegasus	0.2098	0.0951	0.1997	0.7034	**0.6883**	**0.5306**	**0.6801**	**0.8752**
Mengzi T5	**0.2397**	**0.1150**	**0.2250**	**0.7061**	0.6652	0.5130	0.6591	0.8745

Table 5: Experimental results of opinion generator and risk reminder generator.

- Mengzi T5 (Zhang et al., 2021) uses T5 (Raffel et al., 2020) as a backbone and is pre-trained for Chinese.

For the decision model, we use the standard BERT model (*bert-base-chinese*) (Devlin et al., 2019) in the proposed approach after comparing it with CPT (Shao et al., 2021), which is a tailor-made model for Chinese understanding and generation.

To train the opinion generator, we obtain 2,004 news-opinion pairs from a securities company. Since all opinions are written by professional analysts in the company, we use this data to fine-tune pre-trained language models for generating opinions on the given news. In this experiment, 1,603 (80%) instances are used for training, and the remaining instances are used for evaluation. To train the risk reminder generator, we extract 1,356 suggestion-reminder pairs from the reports collected from the Bloomberg Terminal. These reports are also written by professional analysts, and the risk reminder is used to indicate possible events that may affect the accuracy of the suggestion. We use this data to fine-tune pre-trained language models for generating risk reminders based on the given professional opinion. We use 80% of the instances for training and the rest for evaluation.

5 Experiment

5.1 Opinion and Risk Reminder Generators

To evaluate the results of the generation, we employ ROUGE (Lin, 2004) and BERT Score (Zhang et al., 2019) as evaluation metrics. Table 5 presents the outcomes of different pretrained language models in professional opinion generation and risk reminder generation tasks. In both tasks, the models demonstrate high BERT Scores. Among the three models, Mengzi T5 exhibits the best performance in the professional opinion generation task. We summarize our observations regarding the generated opinions as follows: Firstly, mT5 generates more repeated sentences compared to the other two models, which is the primary reason for its inferior performance. Secondly, Pegasus and Mengzi T5 are capable of producing fluent opinions, albeit with occasional hallucinations such as incorrect numerals. Thirdly, the generated opinions include forward-looking perspectives that are not present in the given news.

In the risk reminder generation task, Pegasus outperforms other models. We observe that both ROUGE and BERT Score yield high scores in the risk reminder generation task. Upon comparing the generated results with the ground truth, we find that this phenomenon can be attributed to the fact that certain companies within the same industry share identical risks. Consequently, professional analysts include very similar (sometimes identical) risk reminders in their reports. Furthermore, the same company may encounter the same risk in different reports. While the professional opinions may differ in such reports, the risk reminders remain the same. This discrepancy in the risk reminder generation task is the reason why the ROUGE scores and BERT Scores of the experimental models differ significantly from those in the professional opinion generation task.

5.2 Timing Identification

In this section, we evaluate the results of the proposed news-triggered opinion expressing timing identification task using macro-averaged F1-score. In addition to using vanilla CPT and BERT as baselines, we employ XL-Sum (Hasan et al., 2021) to summarize the given news, and then substitute the generated opinion and risk reminder with the summarization in the proposed approach. Table 6 illustrates the experimental results. Firstly, comparing the baseline results reveals that CPT performs worse than BERT. Additionally, including the abstract of the news articles leads to a decline in performance. Secondly, we observe that the proposed GADFA demonstrates superior performance in the proposed task regardless of the generator used. Thirdly, the proposed GADFA achieves the best performance when utilizing Mengzi T5 as the generator for generating both professional opinion and risk reminder. Lastly, the results also indicate the importance of timeliness analysis in the task. Incorporating more recent news ($T = 5$) proves to

Method	Input of Decision Model	$T = 0$	$T = 5$
CPT	News	74.31%	70.44%
BERT	News	75.64%	77.79%
BERT	News + XL-Sum Abstract	73.46%	77.25%
GADFA	News + mT5 Professional Opinion and Risk Reminder	77.80%	78.28%
GADFA	News + Pegasus Professional Opinion and Risk Reminder	77.75%	78.72%
GADFA	News + Mengzi T5 Professional Opinion and Risk Reminder	**78.12%**	**79.52%**

Table 6: Experimental results of news-triggered opinion expressing timing identification task.

Opinion Generator	Macro-F1
mT5	76.71%
Pegasus	78.36%
Mengzi T5	78.34%
mT5 + Pegasus + Mengzi T5	**78.54%**

Table 7: Ablation analysis — Remove risk reminder.

Opinion Generator	Macro-F1
mT5	67.45%
Pegasus	67.84%
Mengzi T5	**70.58%**

Table 8: Ablation analysis — Remove news and risk reminder.

be beneficial in the proposed task.

6 Discussion

6.1 Ablation Analysis

In Table 6, we have already demonstrated the effectiveness of adding generated opinion and risk reminder in the proposed task, and we have also shown that models perform better when using $T = 5$.

In this section, we present an ablation analysis of the proposed GADFA by removing the risk reminder from the decision model's input under the $T = 5$ setting. The upper part of Table 7 reveals that the performances deteriorate when using only news and generated opinion. Notably, among these results, we observe that using the professional opinions generated by Pegasus or Mengzi T5 yields better performance compared to using those generated by mT5.

Considering that generators may produce professional opinions from different perspectives, we further explore the scenario of using multiple professional opinions generated by different generators. Specifically, we treat each generator as an independent expert and incorporate the opinions of different experts into the decision-making process. The lower part of Table 7 displays the results of different combinations. We find that utilizing all opinions leads to slightly improved performance. Moreover, the performance achieved by using the generated opinions of Pegasus or Mengzi T5 in conjunction with mT5 surpasses that obtained by using only the generated opinions of mT5.

We further exclude the news from the decision model's input to address the following two research questions: (1) Does the decision model still require news for reference despite generating professional opinions based on the given news? and (2) Does the performance of the generated results in Table 5 reflect the performance in the proposed downstream task? Table 8 presents the ablation analysis for these two research questions. Firstly, we observe that the decision model still relies on news for reference as the performance experiences a significant drop when the news is removed. Secondly, we find a positive correlation between the generation performance presented in Table 5 and the performance of timing identification. This result suggests that there might be a positive correlation between the quality of generation and the performance of the downstream task.

6.2 Sentiment and Topic

When analyzing opinions, two important aspects need to be considered: sentiment and topic. To evaluate the generated opinions and risk reminders in depth, we employ Stanza (Qi et al., 2020) and FR2KG (Wang et al., 2021) for sentiment estimation and financial entity extraction, respectively. Stanza is a linguistic analysis toolkit. We obtain the sentiment estimation of both the ground truth and the generated text using Stanza and calculate the ratio of generated text having the same sentiment as the ground truth. FR2KG is a knowledge graph constructed based on Chinese financial reports. We use the entity list in FR2KG to determine the extent to which the generated text contains the same financial entity as the ground truth. If two contents contain the same financial entity, it implies that they are discussing similar topics.

Table 9 presents the evaluation results. Firstly, we observe that Mengzi T5 performs the best when

	Professional Opinion		Risk Reminder	
	Sentiment	Entity	Sentiment	Entity
mT5	25.44%	0.47%	65.68%	72.05%
Pegasus	48.38%	39.96%	71.96%	72.35%
Mengzi T5	49.38%	40.06%	76.01%	79.50%

Table 9: Evaluation from sentiment and entity aspects.

	Professional Opinion		Risk Reminder	
	Sentiment	Entity	Sentiment	Entity
mT5 vs. Pegasus	24.94%	25.60%	63.47%	43.76%
mT5 vs. Mengzi T5	29.68%	27.63%	64.57%	70.69%
Pegasus vs. Mengzi T5	52.37%	25.05%	59.78%	47.77%

Table 10: Comparison among the generated text of different generators.

evaluating both sentiment and entity aspects in the professional opinion generation task. Secondly, in the risk reminder generation task, although Pegasus exhibits the best performance in Table 5, Mengzi T5 outperforms Pegasus in both sentiment and entity aspects. This suggests that one of the possible reasons why the proposed GADFA performs the best is when using Mengzi T5 as the generator for professional opinions and risk reminders. Thirdly, we notice that mT5 performs significantly worse in generating the same entity as the ground truth. Finally, the results align with the findings in Table 5: generating risk reminders appears to be much easier than generating professional opinions.

In addition to comparing the generated text with the ground truth, we provide cross-model comparisons in Table 10. Although the generators are trained on the same training set, the sentiment of the generated text varies greatly when given the same news, particularly in the professional opinion generation task. Additionally, Pegasus tends to generate different entities compared to the other two models. This phenomenon is evident in the risk reminder generation task.

6.3 Analysis of Professionals' and Generated Opinions

To provide a more in-depth analysis of the events that trigger professionals' behaviors, we employ pointwise mutual information (PMI) to calculate word-level scores. PMI is a widely used method for constructing sentiment dictionaries (Khan et al., 2016), and we believe it can also provide valuable insights for mining professionals' behaviors. Apart from considering the timing of report releases, we conduct additional analyses to comprehend the factors that prompt professional analysts to alter their perspective on a given stock. The PMI score of a

Timing - Release		View - Change	
lift rates	1.933	honeymoon	2.789
trade war	1.918	end	2.759
interfere	1.892	slow down	2.719
bulk order	1.836	surprise	2.567
exchange rate	1.827	gap	2.496

Table 11: Keywords selected based on PMI score.

word is computed as:

$$\log_2 \frac{p(w, condition)}{p(w)p(condition)} - \log_2 \frac{p(w, \neg condition)}{p(w)p(\neg condition)}, \quad (1)$$

Here, w represents the target word in the provided news, while $condition$ denotes analysts' behaviors, such as releasing reports or changing their view.

Table 11 presents the statistical results. Firstly, we observe a strong correlation between the timing of report releases and macroeconomic events such as lift rates, trade wars, exchange rates, as well as significant company news like bulk orders. Secondly, we find that changes in views are primarily influenced by changes in status (ending or slowing down) and unexpected events (surprises or gaps). It is important to note that "gap" refers to a substantial rise or fall in stock prices or earnings.

In Table 12, we provide an illustrative example demonstrating the models' proficiency in generating coherent and plausible investment opinions. Furthermore, both the generation models and professional analysts express neutral sentiments towards the given company. However, there exist some nuanced inaccuracies, which in turn raise additional research questions for future investigations. Firstly, the models may generate false numerical information, as evidenced by the provided news and Pegasus's result (5.032 billion vs. 99 billion). Secondly, the temporal information in the generated opinions may differ from that expressed by analysts, as highlighted by the analyst's opinion and Pegasus's result (2018/19 vs. 2019/20). Thirdly, estimations differ between analysts' opinions and generated opinions, as exemplified by the analyst's opinion and Pegasus's result (3.08/3.17 yuan vs. 3.44/3.68 yuan). These findings prompt the following research questions for future investigations:

- How can we generate accurate numerical information?
- Should the generator be considered faulty when the generated estimation deviates from that of professional analysts?

Given News	Uni-President announced its Q3 financial report last night. The net profit after tax in the third quarter was **5.032 billion yuan**, down 8.23% quarter-on-quarter and 11.92% year-on-year. The single-quarter EPS was 0.89 yuan, a relatively large decline.
Analyst	Third-quarter earnings fell short of expectations, mainly due to slower earnings growth at several key subsidiaries. We continue to view UPC's high-end product strategy positively, coupled with a 3% cash yield, which is usually a safe haven during market volatility. However, it is believed that the current stock price upside is limited, mainly due to the slow growth of the Chinese business entering the off-season and the slow growth of convenience stores due to e-commerce fees. It is recommended to be neutral, and the EPS in **2018/19** is estimated to be **3.08/3.17 yuan**.
mT5	The Center believes that the decline in UPC's net profit after tax in the fourth quarter is expected to increase quarter by quarter, because 1) the Chinese government intends to reduce the pressure through early repayment and lower interest rates. Stress is lessened, so we remain neutral on Uni-President Enterprises Corporation.
Pegasus	Uni-President's 3Q19 net profit after tax was **99 billion yuan**, lower than expected, mainly due to the group's deferred capital expenditures and shipments. Looking forward to the market outlook, we believe that Uni-President will continue to benefit from its high-value strategy, coupled with the strengthening of Uni-President's profit rate outlook, and Uni-President's same-store sales growth in the second half of 2019 is expected to improve, and we recommend paying close attention. Estimated EPS for **2019/20** are **3.44/3.68 yuan**.
Mengzi T5	Although Uni-President's revenue showed a quarterly decline, Uni-President's revenue showed a quarter-to-quarter decline that was slightly lower than expected. We continue to take a positive view on UPC's high-end product strategy and the recent purchases on revenue and profit. However, we suggest waiting for a better time, because: 1) China's business is facing slower consumption upgrades and intensified competition; 2) UPC has surpassed this year's Earnings outlook is flat, as e-commerce fee revenue growth momentum slows and reinvestment gains are modest.

Table 12: Examples of generated opinions.

		Professional Opinion	Risk Reminder
Analyst	Avg. # Words	93.29	44.56
	Unique Token	15,217	2,113
	Words from Input	15.33%	49.60%
	Positive	46.88%	20.66%
	Negative	33.17%	49.08%
	Neutral	19.95%	30.26%
mT5	Avg. # Words	66.23	37.77
	Unique Token	1,873	915
	Words from Input	33.68%	72.54%
	Positive	12.72%	21.40%
	Negative	3.24%	49.08%
	Neutral	84.04%	29.52%
Pegasus	Avg. # Words	63.53	40.20
	Unique Token	3,865	1,110
	Words from Input	32.91%	74.88%
	Positive	46.88%	17.71%
	Negative	30.92%	35.43%
	Neutral	22.19%	46.86%
Mengzi T5	Avg. # Words	68.08	38.82
	Unique Token	3,913	1,054
	Words from Input	31.76%	75.41%
	Positive	42.14%	19.56%
	Negative	30.42%	50.55%
	Neutral	27.43%	29.89%

Table 13: Statistics of generated results. The input of professional opinions is news, and the input of risk reminders is professional opinions.

- Does numerical information influence the performance of the decision model?

To further analyze the model properties, we present the statistics of the generated results in Table 13. Firstly, the models generate shorter texts than the ground truth for both investment generation and risk reminder generation tasks. Secondly, the models utilize fewer tokens compared to the ground truth, with mT5 employing the least tokens among the three models. Thirdly, approximately 50% of words in risk reminders can be directly copied from the input, which is a professional opinion. This is reasonable since the risk of the professional opinion, which is based on the expectation of sales growth, would be an event that slows down the sales growth. For instance, the risk reminder for the professional opinion "Multiple high-end new *drivers* can grow significantly on Silergy's improving product offerings and project wins" is "Slower progress in any of the key growth *drivers*".

Finally, we analyze the distributions of the sentiment using the same tool, Stanza, in Section 6.2. We observe that Pegasus's sentiment distribution in the professional opinion generation task is much closer to the ground truth's distribution, while Mengzi T5's sentiment distribution in the risk reminder generation task is also much closer to the ground truth's distribution. Unlike the other two models, mT5 generates a higher proportion of neutral professional opinions. By comparing the statistics of the ground truth in both tasks, we discover that analysts tend to write more positive reports than negative reports. Moreover, risk reminders are predominantly written in a negative tone rather than a positive tone.

6.4 Direction for Multimodal Research

In this paper, we posit that decision-making is influenced by news events, as demonstrated through experiments with news articles. Additionally, structured data such as financial statement tables, historical price data, and economic indices also play a significant role in decision-making. Future studies will incorporate these data types, introducing additional agents for comprehensive analysis. This section introduces a novel approach: chart pattern analysis generation. Chart pattern analysis, a prevalent method in technical analysis, is employed to

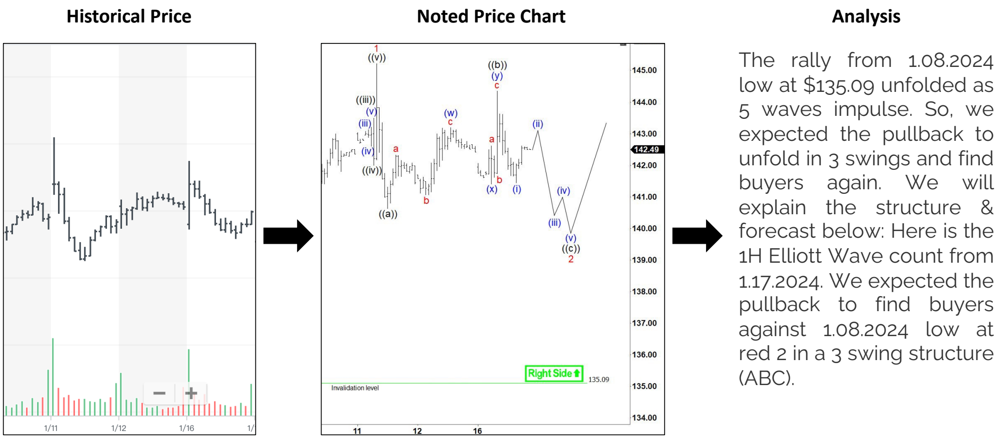

Figure 3: Example of chart pattern analysis generation.

analyze and predict price movements. Figure 3 illustrates an example. This method extends beyond mere reliance on numerical values of stock prices using historical data. Investors annotate charts with key price levels, as depicted in the annotated price chart of Figure 3. Subsequently, they perform analyses based on these annotated charts. Despite its popularity in the investment sector, chart pattern analysis generation has seldom been explored. This section initiates the discussion on this topic, presenting preliminary investigations.

We collect the dataset from the Elliott Wave Forecast Blog.[5] All posts on this blog analyze the price chart based on the same technical analysis method, Elliott Wave Theory. Elliott Wave Theory focuses on the patterns of historical prices and makes forecasts for the price movement based on the corrective price waves. Finally, we obtained 6,987 chart-argument pairs for the Blog. We experiment with the transformer-based encoder-decoder architecture, TrOCR (Li et al., 2021), which uses BEiT (Bao et al., 2021) as an encoder and RoBERTa (Liu et al., 2019) as a decoder. We further explore several representative image encoders, including DeiT (Touvron et al., 2021), Swin (Liu et al., 2021), and ViT (Dosovitskiy et al., 2021). We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and LinkBERT (Yasunaga et al., 2022) as decoders.

We adopt ROUGE metrics (Lin, 2004) and BERTScore (Zhang et al., 2019) for evaluating all results. Table 14 reports the results. Vit-RoBERTa performs the best among all permutations. Additionally, RoBERTa performs the best when using ViT and Swin encoders, and BERT performs the best when using DeiT. Surprisingly, LinkBERT, which performs well in several benchmarks and contains cross-document knowledge, does not outperform its ancestors.

Encoder	Decoder	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
TrOCR		0.2167	0.0958	0.1517	0.81015
DeiT	BERT	0.2410	0.1120	0.1646	0.82776
	RoBERTa	0.2333	0.1083	0.1615	0.82536
	LinkBERT	0.2373	0.1118	0.1604	0.83267
Swin	BERT	0.2323	0.1071	0.1588	0.82660
	RoBERTa	0.2440	0.1198	0.1685	0.83143
	LinkBERT	0.2431	0.1170	0.1665	**0.83514**
ViT	BERT	0.2336	0.1019	0.1554	0.82535
	RoBERTa	**0.2444**	**0.1213**	**0.1695**	0.83311
	LinkBERT	0.2361	0.1105	0.1598	0.83283

Table 14: Results of chart pattern analysis generation task.

Future studies could build upon our exploration to further adding multimodal information to the proposed approach. Generating analysis based on other documents, such as earning calls (Goldsack et al., 2024), could also be included.

7 Conclusion

We introduced the task of identifying the timing of expressing opinions triggered by news and proposed GADFA to address this task. We demonstrated that augmenting the decision model with risk reminders improves performance. Various discussions on model comparison, linguistic features, and analyst behavior based on temporal conditions provided a comprehensive understanding of GADFA and the task. Our approach is extendable to other professional behavior forecasting tasks, aligning with the steps involved in the decision-making process. In the future, we plan to explore GADFA's application in domains like clinical decision-making and forecasting trading behavior.

[5]https://elliottwave-forecast.com/

Limitations

The limitations of this paper are four-fold as follows:

1. Due to the data availability, this paper only explores the proposed task with Chinese data, and thus cannot claim that the proposed GADFA is a general approach to any languages. Future work can use the proposed approach in other languages' application scenarios.

2. As we mentioned, experts in other fields, such as CDC's experts and politicians, also need to identify the timing of expressing opinions. However, this paper only explores one application scenario, and cannot claim that the proposed approach is also useful in other scenarios. We believe that this paper provides a comprehensive analysis of analysts' opinion expression timing. Future works can follow our line of thought to explore the proposed task in other domains.

3. Since this paper focuses on the timing identification task, we mainly pay attention to the performance of the proposed task. Although we analyzed the generation results with ROUGE scores, BERT Scores, sentiment, and entity aspects, we did not perform the human evaluation on the generated results. Because the performances of the generator used in the experiments were already verified with human evaluation when they were proposed, we did not do human evaluation again. We focused on how these models influence the performance of the proposed task. If future work attempt to improve the performance by proposing a new generator, we think that it would be great to compare generators with human evaluation.

4. The decision to write a financial report can be influenced by the preferences of the financial analyst. Analysts may have personal preferences, prior experiences, or relationships with certain companies that could shape how they interpret and present financial data. This could lead to reports that selectively emphasize certain aspects of a company's financial situation. Although this paper primarily addresses the technical aspects of financial reporting, it is important to acknowledge that preferences could affect the objectivity of the reports and, consequently, the accuracy of the conclusions drawn. Future work should consider addressing how such preference could be minimized in the financial analysis process.

Impact Statement

Although this paper can help go one more step further in opinion mining and automatic stock analysis assistance, some potential negative outcomes exist if developers use our method with bad intent (Solaiman et al., 2019). For example, the NLP-based trading algorithm would easily be influenced by changing one word (Xie et al., 2022). We think that the proposed timing identification task can also be used to influence the financial market. For example, the developer can select some timing to release generated opinions with false information, and it may lead to large market volatility. However, we believe that understanding the task and models' properties in-depth can also help us avoid this potential risk.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956. This work of Hsin-Hsi Chen was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-113L900901.

References

Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Fischer Black and Myron Scholes. 1973. The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3):637–654.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2051–2064, Dublin, Ireland. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From opinion mining to financial argument mining*. Springer Nature.

Chung-Chi Chen and Hiroya Takamura. 2024. Term-driven forward-looking claim synthesis in earnings calls. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15752–15760, Torino, Italia. ELRA and ICCL.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105, Dublin, Ireland. Association for Computational Linguistics.

Jennifer Conrad, Bradford Cornell, Wayne R Landsman, and Brian R Rountree. 2006. How do analyst recommendations respond to major news? *Journal of Financial and Quantitative Analysis*, 41(1):25–49.

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.

Erik Devos, Wei Hao, Andrew K Prevost, and Udomsak Wongchoti. 2015. Stock return synchronicity and the market response to analyst recommendation revisions. *Journal of Banking & Finance*, 58:376–389.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 698–708.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Eugene F Fama. 1995. Random walks in stock market prices. *Financial analysts journal*, 51(1):75–80.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2024. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. *arXiv preprint arXiv:2410.01039*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *ACL-IJCNLP*.

D Eric Hirst, Lisa Koonce, and Paul J Simko. 1995. Investor reactions to financial analysts' research reports. *Journal of Accounting Research*, 33(2):335–351.

Chia-Chun Hsieh, Kai Wai Hui, and Yao Zhang. 2016. Analyst report readability and stock returns. *Journal of Business Finance & Accounting*, 43(1-2):98–130.

Chao-Chun Hsu and Chenhao Tan. 2021. Decision-focused summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 117–132, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Farhan Hassan Khan, Usman Qamar, and Saba Bashir. 2016. Sentimi: Introducing point-wise mutual information with sentiwordnet to improve sentiment polarity detection. *Applied Soft Computing*, 39:140–153.

Yongsik Kim and Doojin Ryu. 2022. Firm-specific or market-wide information: How does analyst coverage influence stock price synchronicity? *Available at SSRN 4010314*.

Sagar P Kothari, Eric So, and Rodrigo Verdi. 2016. Analysts' forecasts and asset pricing. *Annual Review of Financial Economics*, 8:197–219.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020a. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3063–3070.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Wei Li, Ruihan Bao, Keiko Harimoto, Deli Chen, Jingjing Xu, and Qi Su. 2020b. Modeling the stock relation with graph network for overnight stock movement prediction. In *IJCAI*, volume 20, pages 4541–4547.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2149–2159, Dublin, Ireland. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Jayanta Mandi, Victor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. 2022. Decision-focused learning: Through the lens of learning to rank. In *International Conference on Machine Learning*, pages 14935–14947. PMLR.

Greg Niehaus and Donghang Zhang. 2010. The impact of sell-side analyst research coverage on an affiliated broker's market share of trading volume. *Journal of Banking & Finance*, 34(4):776–787.

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*, pages 101–108, Online. Association for Computational Linguistics.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. 2021. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757, Online. Association for Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4232–4241, Dublin, Ireland. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wenguang Wang, Yonglin Xu, Chunhui Du, Yunwen Chen, Yijie Wang, and Hui Wen. 2021. Data set and evaluation of automated construction of financial knowledge graph. *Data Intelligence*, 3(3):418–443.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. 2019. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2438–2447, Dublin, Ireland. Association for Computational Linguistics.

Yong Xie, Dakuo Wang, Pin-Yu Chen, Xiong Jinjun, Sijia Liu, and Oluwasanmi Koyejo. 2022. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1979, Melbourne, Australia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

A Implement Detail

In our experiment, we use the Hugging Face transformers package (Wolf et al., 2019).[6] Intel Xeon Gold CPU and Nvidia Tesla V100 w/32GB are the CPU and GPU we used. Table 15 provides the links to the models we used in the experiments. We release these datasets for academic use under the CC BY-NC-SA 4.0 license.[7]

[6] https://huggingface.co/docs/transformers/index

[7] http://fintime.nlpfin.com/

	URL
BERT (Devlin et al., 2019)	https://huggingface.co/bert-base-chinese
CPT (Shao et al., 2021)	https://huggingface.co/fnlp/cpt-base
XL-Sum (Hasan et al., 2021)	https://huggingface.co/spaces/krrishD/csebuetnlp_mT5_multilingual_XLSum
mT5 (Raffel et al., 2020)	https://huggingface.co/google/mt5-base
Pegasus (Zhang et al., 2020)	https://huggingface.co/uer/pegasus-base-chinese-cluecorpussmall
Mengzi T5 (Zhang et al., 2021)	https://huggingface.co/Langboat/mengzi-t5-base

Table 15: Reference for the models in our experiments.