INTERNSHIP REPORT

Disclaimer:

This report will be sent to:

Marc Evard : mevrard.upsay@gmail.com
François Yvon : francois.yvon@limsi.fr

This report is supervised by:

Jérémy Lhour: jeremy.lhour@cfm.com
Sylvain Champonnois: sylvain.champonnois@cfm.com

PREPARED BY:

Yihan Zhong: yihan.zhong@universite-paris-saclay.fr

Enhancing Quantitative Hedge Fund Strategies with NLP-Powered Financial News Analysis CONFIDENTIAL – FOR THE EXCLUSIVE USE OF THE RECIPIENTS

Yihan ZHONG [1]

Abstract

The primary goal of a quantitative researcher in a hedge fund is to generate alpha signals, enabling data-driven trading strategies for superior financial performance. My contribution involves harnessing the power of NLP on financial news to extract additional value across various dimensions, including sentiment analysis, novelty detection, and examining the correlation between returns and significant events. I have addressed three key challenges in this work: Firstly, I've tackled the issue of multilingual news headline analysis, developing solutions for languages without in house sentiment model. Secondly, I've devised methods to detect news volume spikes by employing topic tags. This is vital for us to effectively prioritize news events within the extensive volume of news data. Lastly, I've undertaken the pretraining of an advanced BERT model with in-house data. This step was necessary as our current reliance on an open-source model may fail to adequately capture the nuances of the financial domain. The initial challenge was addressed through a combination of techniques, including multilingual embedding for news matching and ridge regression applied to sentiment scores extracted from English news. The second challenge yielded intriguing findings, uncovering notable correlations between returns and specific news events. For the third challenge, we successfully pretrained an in-house model; however, we were unable to demonstrate its performance on the Mask Language Modeling task within the time constraints.

[1]Université Paris-Saclay, Gif-sur-Yvette 91990, France. Correspondence to: Yihan ZHONG <yihan.zhong@universite-paris-saclay.fr>.

1. Background

In this section, we delve into the background of the three projects undertaken during the internship, each explored through three distinct dimensions: the context, review of the state of the art, and my contributions to the existing methodologies.

1.1. Context

Multilingual news headline analysis: The project aims to explore the efficacy of leveraging multilingual data for sentiment analysis and assess its potential in generating enhanced alpha signals for financial analysis through backtesting. The utilization of foreign language data represents a significant untapped resource within the research team. While the English language models have exhibited strong performance, the exploration of foreign language data sources holds promise for enhancing the prediction model and gaining novel insights. In particular, the investigation focuses on leveraging Chinese language data extracted from financial News to uncover new perspectives and valuable information that would otherwise remain inaccessible. In order to establish the validity and generalizability of our findings, it is imperative to benchmark the performance of our Chinese language model against a well-established English language model known for its high performance. Given that we are unable to directly evaluate the Chinese model in isolation, we undertake a matching process to align the Chinese language data with its English language counterpart using multilingual embedding. Subsequently, we compare the performance of our Chinese model against the existing English model. Through this bench marking procedure, we ascertain the effectiveness of our Chinese language model and enable the extrapolation of insights garnered from the project to encompass other languages and diverse data sources.

Classification of financial news volume spikes with topic tags: The second project aims to gain a deeper understanding of the events that occur during volume spikes in relation to news headlines. The main objective is to analyze these volume spikes and identify market-moving news events, which will be leveraged to guide strategic decision-making in financial contexts. Understanding volume spikes

holds significant importance, as it is widely anticipated that market-moving news often coincides with increased trading activity. By comprehending the dynamics of volume spikes, the project can effectively identify and interpret significant events that occur alongside earnings days. This understanding will provide valuable insights to inform strategic decision-making processes. To achieve the objective of identifying topics associated with volume spikes, the project follows a systematic approach. Firstly, data is sampled from the identified volume spikes to ensure representative information. Next, the collected data is categorized and organized effectively for further analysis. A manual method is then employed to identify specific topic tags present within the categorized data. Ultimately, the reclassification of volume spikes based on these topic tags helps to gain valuable insights into the correlation between market returns and the occurrence of specific market-moving events.

Pretrain a BERT model with downstream data sources: The third project focuses on the pretraining of a BERT model using an in-house dataset, aiming to investigate whether a pretrained model with downstream financial datasets can yield improved results for sentiment analysis tasks. The significance of this project lies in the uniqueness of the financial news corpus, which demands specific language representations. We will leverage the platform MosaicML, a cutting-edge startup solution, to streamline and automate the training process on multiple GPUs, ensuring efficiency and scalability in our exploration of the model's performance. To compare the performance of self-pretrained models against off-the-shelf ones, we are subjected to use a mask language task using our in-house dataset for evaluation.

1.2. Review of the State-of-the-Art

The work of (Artetxe & Schwenk, 2018) presents a cutting-edge methodology for learning joint multilingual sentence representations across an impressive 93 languages, with tremendous potential for diverse applications in cross-lingual transfer and other related tasks. The approach utilizes a single BiLSTM encoder with a shared BPE vocabulary for all languages and is trained on publicly available parallel corpora. This enables the system to learn a classifier on top of the resulting embeddings using English annotated data only, and transfer it to any of the 93 languages without any modification. The approach establishes a new state-of-the-art for all language pairs except English-Chinese. However, it is worth noting that the approach still achieves a relatively high accuracy of 68.4% on the XNLI test set for this pair, which is only slightly lower than the previous state-of-the-art of 69.5%.

In the context of sentiment analysis on a multilingual corpus, the scarcity of labeled data poses a challenge. Conventional text classification methods heavily rely on labeled data to establish the connection between text and labels. To address this challenge, (Yin et al., 2019) introduced a novel approach based on pre-trained NLI models. This method treats the sequence to be classified as the NLI premise and generates hypotheses from candidate labels, effectively serving as a zero-shot sequence classifier. By leveraging this method, we tackle the issue of the absence of a Chinese sentiment analysis model.

Comprehending the emergence of new news articles holds significant importance for trading, as it provides crucial initial information. In recent research on news clustering (Linger & Hajaiej, 2020), sparse TF-IDF features are utilized to represent articles. These features consist of 9 TF-IDF weighted sub-vectors, and the article similarity is computed through a linear combination of cosine similarities among these sub-vectors. We attempted to correlate the cosine similarity of headline embeddings with volume spikes to explore potential associations. However, the results were inconclusive, leading us to further investigate research directions related to volume spikes.

In recent research, "Self-pretraining" has emerged, where a model undergoes pre-training using the same dataset that is later utilized for fine-tuning. This methodology has demonstrated remarkable performance gains in downstream tasks, comparable to the traditional transfer learning method, which involves pre-training on large external datasets (Krishna et al., 2023). Additionally, empirical evidence has established that deduplication plays a pivotal role in optimizing language models. By deduplicating training data, language models achieve improved text quality, enhanced training efficiency, and more accurate evaluation (Lee et al., 2021). Based on these findings, we decided to use a new Bert (Devlin et al., 2018) architecture Mosaic-bert-uncased, which has proved to have higher accuracy on pretraining and finetuning when benchmarked with hugging face bert-base-uncased model. Mosaic-bert model (Portes et al., 2023) was based on Flash attention(Dao et al., 2022), AliBi (Press et al., 2022), and Gated Linear Units (Shazeer, 2020). Flash Attention optimizes the attention process by dividing the matrix into smaller blocks and calculating attention scores independently for each block. This reduces memory reads and writes during training, leading to faster model training and improved quality with extended context modeling. Moreover, FlashAttention avoids storing the large attention matrix in the forward pass, efficiently recomputing it in the backward pass and reducing memory usage. Attention with Linear Biases (ALiBi), eliminates the need for position embeddings and biases query-key attention scores with a linearly decreasing penalty proportional to the distance between the relevant key and query. This approach eliminates position embeddings and enables efficient extrapolation. GLU (Gated Linear Units) represents a neural network

layer characterized by the element-wise multiplication of two linear projections. One of these projections is initially subjected to a sigmoid function, rendering GLU a viable substitute for the conventional activation function within the feedforward sublayers of the Transformer model. Based on these findings, we are prompted to undertake model pretraining using our in-house dataset using a Mosaic-bert model, aiming to assess the potential impact on enhancing the capabilities of language models for various applications.

1.3. Contribution

Throughout the internship, my contributions have been dedicated to addressing industrial challenges through the application of current methodologies. In the multilingual news headline analysis project, I conducted a matching process to align Chinese news headlines with their corresponding English headlines, utilizing various vector representations. Subsequently, I performed a comparative analysis to assess the accuracy of these models.

Building on the successful headline matching, I conducted a sentiment correlation analysis using diverse sentiment models, including zero-shot learning, the sentiment model, and rigid regression on embeddings.

In the context of the classification of financial news volume spikes, I achieved a significant advancement by replacing the manual reading and classification of volume spike events with a more efficient approach. Specifically, I integrated an Open AI API with a GPT-4 model, streamlining the classification process.

In the context of pretraining a mask language model on a downstream dataset, I conducted the pretraining process on the MosaicML platform, employing various advanced techniques such as gradient accumulation and modification of batch size to enhance training efficiency. Moreover, I leveraged state-of-the-art algorithms to ensure a highly effective and resilient model training process.

2. Material and Methods

In this section, I will delve into the datasets used for each project and elaborate on the machine learning and quantitative techniques employed not only for analysis but also to effectively address the objectives and solve the underlying challenges in each project.

2.1. Datasets

During the Multilingual news headline analysis project, the dataset employed consisted of Chinese financial news from 2011 to April 2023, comprising approximately 3.8 million news articles. The dataset's unique key is derived from the `suid` and `last update` columns. Additionally, the dataset includes valuable information such as the `tickers` column, indicating the traded companies, the `topic` column, and the `headline` and `body` contents.

To gain insights into the dataset's characteristics, a preliminary quantitative analysis was conducted, focusing on its seasonality and frequency. Subsequently, a representative sample of 2000 datapoints was extracted from the entire dataset to facilitate content analysis and deeper exploration. It was observed that the quantity of Chinese news can be cut into half since there are always Chinese traditional version and Chinese simplified version. Furthurmore, all the important headline starts with * and the news' body contains "Original title" / "By ... automation" have original English reference .

These observations shed light on the structure and composition of the dataset, which may have implications for subsequent analyses and models developed during the course of the project. Having gained an understanding that the Chinese dataset does not offer additional trading value, an additional idea emerged. The availability of English reference news in our data pool allows us to establish connections between Chinese and English language data. This aids in setting up a multilingual data analysis pipeline by benchmarking against existing English language models.

The matching process involved data from financial news in both English and Simplified Chinese on April 17, 2023. After eliminating nan values on tickers, the dataset contained 215 Chinese news and 2368 English news. Removing news that only had Chinese versions resulted in 183 remaining news articles for evaluation.

The dataset for the news volume spike classification projects consisted of English multistocks news from 2008 to 2022, comprising approximately 20 million news articles after deduplication for each single stock. Volume spikes were sampled from the top 10,000 z-scores.

The dataset employed for pretraining the off-the-shelf language model encompasses English multistock news spanning from 2008 to 2016. The dataset was deduplicated, without taking into consideration the single stock level information, resulting in a total of approximately 5 million news, results in a total of 92,491,236 tokens .

During the training phase, we introduced a smaller dataset specifically for model benchmarking. This smaller dataset is derived from news articles spanning the years 2008 to 2009, comprising approximately 700,000 news, results in a total of 12,283,296 tokens.

2.2. Methodology

Chinese-English Headline Matching: The matching process involved evaluating several methods. Firstly, it was cat-

egorized into matching with and without conditions. Matching was performed using a translation model to convert Chinese headlines into English headlines, followed by employing Sentence Transformer embeddings or direct use of multilingual embeddings. The matching culminated in a similarity search based on cosine similarity between the embeddings of news headlines. As shown in Figure 1, the figure presents presents a mind map illustrating the comparative analysis.

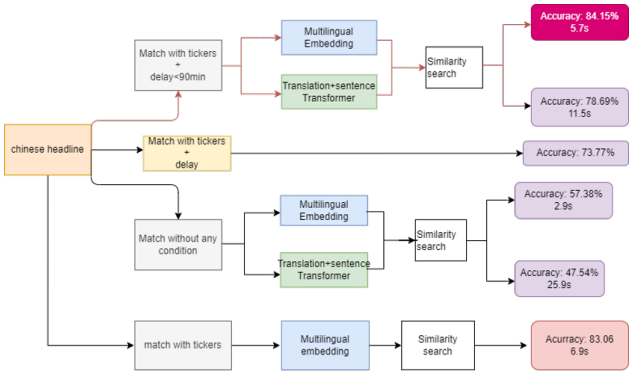


Figure 1. Comparative analysis of Chinese-English headline matching models

Sentiment analysis: Two distinct sentiment models were employed. The first model utilized a Lexicon-based method, incorporating the Loughran-McDonald sentiment lexicon (Loughran & McDonald, 2011). A corresponding Chinese dictionary was sourced from the referenced GitHub repository (Deng & Nan, 2022).

The second sentiment model used was a fine-tuned bert model on Financial news, which will output the probability for positive, negative and neutral news. Since there is no existing model that has been fine-tuned on Chinese headline which can output these three categories. Two different method has been used: a zero-shot learning inference model and a ridge regression on the Chinese news embeddings with the matched English news sentiment score as a three-dimensional label. The sentiment analysis employed a fine-tuned BERT model specifically trained on financial news data. This model will output probabilities indicating whether the news is positive, negative, or neutral. As no pre-existing model was available for Chinese headlines containing these three probabilities, two distinct approaches were adopted.

In the first approach, a zero-shot learning inference model was utilized. This innovative technique allows the model to generalize its understanding of sentiment across languages without requiring explicit training data in Chinese.

In the second approach, a ridge regression technique was employed on the Chinese news embeddings. The ridge regression model was trained using the matched English news sentiment scores obtained by Finbert model (Araci, 2019) as a three-dimensional label. By aligning the sentiment scores of English news headlines with their corresponding Chinese embeddings, the model was able to map the sentiment analysis context from English to Chinese headlines. Figure 2 depicts the ridge regression technique applied to Chinese news embeddings.

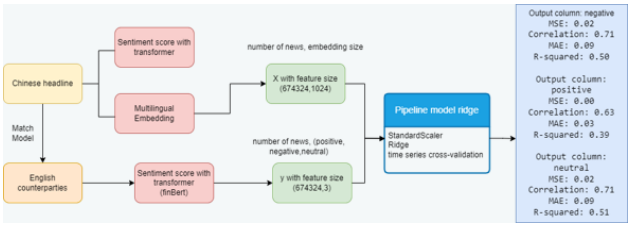


Figure 2. Flowchart of the regression process for reproducing Finbert sentiment score on Chinese headline

Preparing dataset using Generative AI: For confidentiality reason, the sampling of volume spike will not be mentioned in this report. Randomly sampling 500 observations from the top 10,000 z-scores for a specific company and date, the volume spike events in financial news were efficiently categorized and understood using an OpenAI API with a GPT model and a carefully edited prompt.

Example of the Prompt:

```
"Act as a financial analyst.
Create a category for the main event of
    {company}, on {date} based on the
   provided headline: {headline}.
The category should be concise and
   general, such as debt, rating, data
    breach, management changes, legal
   issues, or natural disasters, etc.
Show as a list."
```

During the study, another idea is to detect news articles that may indicate a potential scandal. However, a significant challenge prevalent in the industry is the scarcity of labeled data suitable for supervised learning. While few-shot learning remains an option, the process of assembling a small dataset is time-consuming and often lacks the desired level of generalization.

To address this issue, a workflow that involves leveraging a GPT model to enhance the generalization of a small dataset for few-shot learning is proposed. This workflow comprises the following steps: 1. Identifying potential scandals from

instances of sudden increases in news volume. 2. Extracting headlines related to specific companies on specific dates. 3. Employing the GPT model to generate summaries of two headlines that best encapsulate the underlying causes of the potential scandal.

Figure 3 is an illustrative example of the results obtained from this approach:

```
input_text

"PG&E: 100,000 IN NAPA, SONOMA WITHOUT ELECTRICITY AS OF MID-DAY. PG&E: ABOUT 30,000 IN NAPA, SONOMA WERE WI
THOUT GAS SERVICE. PG&E Says 100,000 Without Electricity From North Bay Wildfires. MOODY'S UPGRADES RATING O
N NCPA'S LODI ENERGY CENTER INDENTURE. PG&E Falls on Speculation Downed Power Lines Factor in Wildfires. Dea
dly California Fires Send PG&E Sliding by Most in 7 Years (1). California Reminds PG&E to Preserve Evidence
Related to Fires. California Regulator Looks at PG&E Activity in Area of Fires (1). PG&E Is Having Its Worst
Week in Seven Years as Wildfires Rage. PG&E Selloff Too Early to Call Overdone, Guggenheim Says. PG&E Tumble
s; RBC Cuts Price Target as 'Wildfire Risk Is Real'. PG&E Is Having Worst Week in Nine Years as Wildfires Ra
ge (1). PG&E: ROLE OF POWER LINES BEING PROBED IN CALIFORNIA FIRES. PG&E: UNKNOWN IF UTILITY WILL HAVE LIABI
LITY FOR FIRES. PG&E HAS ABOUT $800M IN LIABILITY INSURANCE. PG&E Says Power Lines' Role Being Probed in Cal
ifornia Fires. Probe Into California Wildfires Could Hang Over PG&E for Months. PG&E Loses $6.5b Over Two Da
ys As Mizuho Defends. PG&E Has Worst Week in Nine Years on California Wildfires."

response = openai.ChatCompletion.create(
model='gpt-3.5-turbo',
  messages=[
    {"role": "user", "content": prompt}],
max_tokens=500,
temperature=0,
)

print(response["choices"][0]["message"]["content"])

1. PG&E Falls on Speculation Downed Power Lines Factor in Wildfires.
2. PG&E: ROLE OF POWER LINES BEING PROBED IN CALIFORNIA FIRES.
```

Figure 3. Selection of Headlines Generated by the GPT Model for a Potential Scandal

Correlation the stock return with detected topics:A dictionary was created manually to link detected volume spikes with relevant topic tags. Subsequently, a reverse classification was applied to sampled volume spikes using these topic tag dictionaries. Each category has a binary masking dataframe, which is multiplied with the Markowitz return dataframe to establish lead-lag relationships. As shown in Figure 4, the figure presents presents a mind map illustrating the correlation analysis.

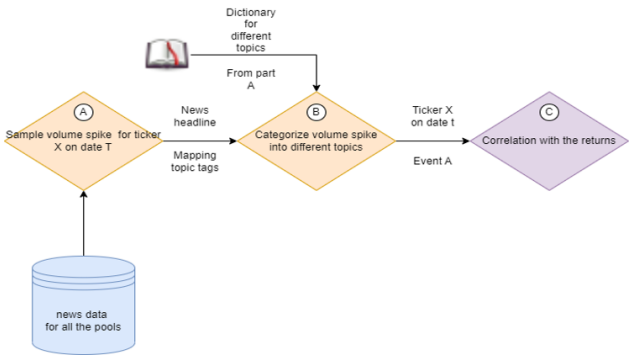


Figure 4. Correlation analysis between volume spike and return

Pretraining of a bert model: Deduplication of the training data has been demonstrated to significantly enhance the performance of the language model, particularly in the context of financial news data where an abundance of duplicates is inherent. Prior to model pretraining, a thorough deduplication process was executed using the text-dedup GitHub repository, as detailed in (Mou, 2023).

The establishment of the training environment encompassed efforts including package compatibility, the integration of a new tokenizer, and other considerations to ensure an optimal setup. While the specific technical configurations are beyond the scope of this discussion, it's noteworthy that the model was ultimately trained within AWS SageMaker Studio, employing a dedicated image tailored to Mosaic BERT, as per the official guidelines provided by the platform. The selected instance type was 'ml.g5.2xlarge,' equipped with one A10G GPU.

It is worth noting that, at the time of submission, the integration of the FlashAttention module (Dao et al., 2022) with OpenAI Triton implementation was not feasible, as it might necessitate the availability of GPUs of types A100 or H100. Consequently, we utilized an equivalent version of the PyTorch attention module. Nonetheless, future research may explore the potential efficiency gains achieved by the Triton kernel during training.

Regarding training methodology, a comparative analysis was conducted between the standard BERT model and the Mosaic BERT model on a small dataset. To optimize for efficiency and cost-effectiveness, the dataset was trained with identical settings as the Mosaic BERT implementation, omitting hyperparameter fine-tuning.

In our training configuration, it's worth to mention that several changes have been made. Notably, we've aligned the vocabulary size with our new in-house tokenizer, which is set at 16,384 tokens, differing from the original BERT uncased tokenizer with a size of 30,624 tokens. This adjustment in vocabulary size impacts the initial layer of the model and results in a reduction in the number of parameters. For instance, the BERT model initially had 109,510,000 parameters, which decreased to 98,642,000 with the new tokenizer.

Additionally, variation in the dropout rate for the attention layer is needed . For flash attention layer application with Triton, a dropout rate of 0.0 is obligatory, while for PyTorch-based attention layers, we've set the dropout rate to 0.1. Since we have a limit for Triton resource, we have set the dropout rate to 0.1

Selecting the appropriate hyperparameters, particularly the number of epochs, is an empirical process demanding careful consideration. In the initial training of MosaicBERT (Portes et al., 2023) on the "Colossal, Cleaned, Common Crawl" (C4) dataset in English, totaling 365 million sequences and 156 billion tokens,each with a length of 128 tokens. It was trained by a duration of 286 million se-

Table 1. Hyperparameters for pretraining Bert and Mosaic Bert

HYPERPARAM	BERT	MOSAIC BERT
NBR OF PARAMS	9.8642e+7	1.2652e+8
VOCABULARY SIZE	16384	16384
NBR OF LAYERS	12	12
HIDDEN SIZE	768	768
DROPOUT	0.1	0.1
ATTENTION DROPOUT	0.1	0.1
PEAK LEARNING RATE	3e−4	3e−4
BATCH SIZE	512	512
MICRO BATCH SIZE	128	128
LEARNING RATE DECAY	LINEAR	LINEAR
ADAM-ϵ	1e−6	1e−6
ADAM-β_1	0.9	0.9
ADAM-β_2	0.98	0.98
WARMUP RATIO	0.06	0.06
PRECISION	AMP_FP16	AMP_FP16

Table 2. Dataset characters for training

HYPERPARAM	SMALL	LARGE1	LARGE2
NBR OF SAMPLES	$707k$	$5M$	$5M$
NBR OF TOKENS	$12M$	$92M$	$92M$
MAX DURATION	3EPS	3EPS	$286M$ SAMPLES
TOTAL STEPS	4K	129K	31K

quences, covering approximately 78.6% of the C4 dataset. While by employing this training duration , our results were suboptimal, this will be delve into in the section 3.

For our small dataset, we conducted training over three epochs, while for the larger dataset, we set a maximum duration for both 286 million sequences and three epochs. It's worth noting that the number of epochs remains an experimental parameter, and training can be extended, although we were unable to present the results within the submission date. Further details on the character about the dataset used for training can be found in the table 2.

3. Results

To demonstrate the potential of NLP in financial domain, this part will present various intermediate findings. These findings provide insight into how leveraging diverse NLP techniques can yield maximum insights from our dataset, unlock untapped potential within the dataset using the existing model, and enhance our baseline model.

3.1. Multilingual news headline analysis

The accuracy results for matching Chinese headlines with their English counterparts are depicted in Figure 1. The findings reveal that the most effective approach involves an initial timeline and tickers-based matching, followed by multilingual embedding-based matching using a manually defined threshold for cosine similarity. This combination demonstrates superior performance in terms of both accuracy and efficiency. As shown in Figure 5, the density plot illustrates the cosine similarity scores between matched Chinese headlines and their corresponding English headlines using various methods. Curve 1 represents the cosine similarity scores obtained using Laser embedding with specific conditions. Curve 2 corresponds to the cosine similarity scores from translation with the Helsinki-NLP/opus-mt-zh-en model (**?**), also with conditions. Curve 3 depicts cosine similarity scores obtained using Laser embedding, but without any specific conditions. Curve 4 shows cosine similarity scores obtained through translation, again without conditions.Curve 5, on the other hand, showcases cosine similarity scores achieved through translation using the alirezamsh/small100 (**?**) model, without specific conditions. This particular comparison aims to assess the performance of different translation models. The examination of cosine similarity scores across various matching approaches reveals distinct patterns. The Laser model consistently demonstrates limited variance, achieving relatively high scores between 0.65 and 0.8, regardless of ticker inclusion. Conversely, translation-based methods generally yield lower scores, potentially due to varying translation quality. Omitting tickers leads to reduced accuracy, attributed to concise headlines with shared terms. Notably, extending delay constraints fails to significantly enhance results, despite some correct matches spanning 500 minutes. These findings illuminate the behaviors and implications of different matching strategies within the context of cosine similarity analysis.

Figure 5. Density plot of the cosine similarity score for different kinds of matching

Regarding the correlation of the sentiment score obtained by Lexicon model, we observe a relatively low correlation between the sentiment scores, particularly when comparing Chinese headlines to their English counterparts from Figure 6. This discrepancy may be attributed to the discrepancy

in Lexicon quality; the Chinese sentiment lexicon, is comparatively less well-maintained than its English counterpart. However, it is noteworthy that there remains a reassuringly low likelihood of misclassifying potentially negative news as positive. This underscores the robustness of the model's sentiment classification, even in less-ideal conditions.

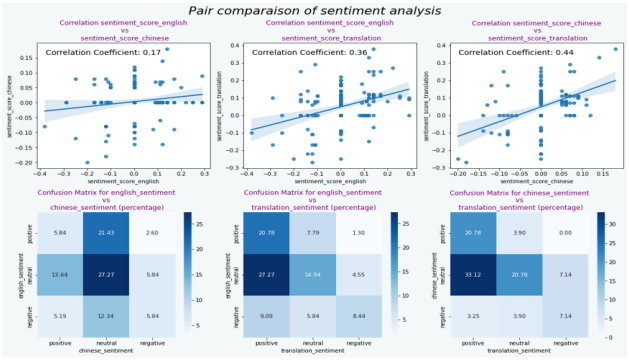


Figure 6. Correlation in Sentiment Scores between Chinese Headlines, Translated Chinese Headlines, and Matched English Headlines obtained by Lexicon based method

The results depicted in Figure 2 provide a comprehensive evaluation of ridge regression on accuracy concerning correlation, Mean Absolute Error (MAE), and R-squared metrics across the spectrum of sentiment probabilities: negative, positive, and neutral. Additionally, Figure 7 illustrates the variation in Mean Squared Error (MSE) corresponding to a range of alpha values. Notably, it reveals a sustained stability in accuracy when employing alpha values spanning from 10^{-2} to 10^{4}. The computed Mean Squared Errors consistently hover within the range of 0 to 0.02, affirming the robustness of the ridge regression model's predictive performance.

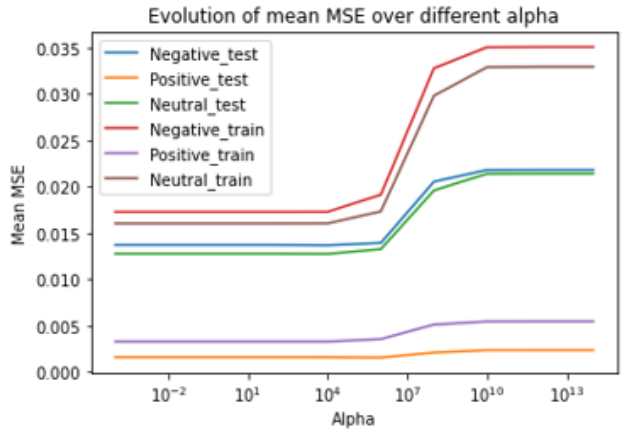
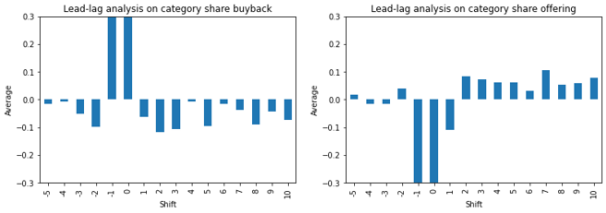


Figure 7. Ridge regression result with different alpha values

3.2. Classification of financial news volume spikes with topic tags

In this section, we will explore the relationship between a specific event and stock returns from two distinct perspectives: positive correlation and negative correlation. Given the quantitative nature of this analysis, we will provide a concise overview, reserving a more comprehensive discussion of how these findings can be effectively utilized by AI model for the conclusion chapter.

The result reveals that share buybacks exhibit a positive impact on stock returns on the first day of the event, followed by a gradual decline in the subsequent days. Conversely, in the case of share offerings, we observe a decrease in returns in the days leading up to and including the event, with subsequent days showing a rebound effect.



(a) Share Buyback (b) Share Offering

Figure 8. Lead-Lag analysis of the correlation between Volume Spikes and Stock Returns

3.3. Pretrain a BERT model with downstream data sources

The training results represent an empirical process designed for model comparison and initial performance assessment. In order to enhance model performance, an extended training process with a greater number of epochs is required.

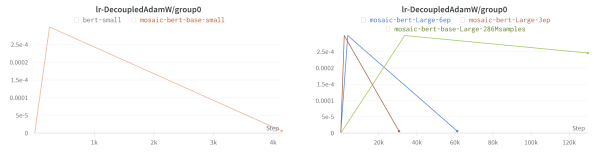
Figure 9 illustrates the learning rate schedule during training. Notably, in the context of a smaller dataset, the learning rate achieves its peak considerably faster than in the case of a larger dataset. This phenomenon also accounts for the initial higher accuracy observed on the evaluation set.

Figure 11 demonstrates that MosaicBert outperforms Bert in terms of both accuracy and language cross entropy. However, it's noteworthy that the training duration, as shown in Table 3, does not exhibit a significant difference. This observation may be attributed to the absence of the Triton application for flash attention.

In the context of training on a larger dataset, a prolonged training period spanning 128k steps does not exhibit superior performance when compared to a shorter 62k step training. This observation can be attributed to several factors, such as hyperparameter settings with batch size and learning rate

scheduling.

Specifically, in the case of longer training durations, the learning rate may not be helpful to the initial stages of training as it gradually reaches its peak at a slower pace. To optimize model convergence, it might be judicious to consider employing a larger learning rate in conjunction with an increased batch size. A larger batch size will reduce the stochasticity of the gradient updates thus a larger learning rate is more update.



(a) Small dataset (b) Large dataset

Figure 9. Learning Rate Schedule: Linear Warmup to Peak LR with Decoupled AdamW Optimizer

(a) Loss (b) MaskedAccuracy (c) Language-CrossEntropy

Figure 10. Model Training Metrics for small dataset

(a) LanguageCrossEntropy (b) MaskedAccuracy

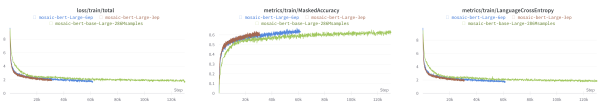
Figure 11. Model Evaluation Metrics for small dataset

4. Conclusion

In summary, this comprehensive study has yielded insights into the realm of financial news data analysis, driven by the powerful capabilities of Natural Language Processing (NLP). The findings extend beyond the mere exploration of data; they offer real-world implications for extracting maximum value from noisy financial datasets and the available information resources. The results of this study hold significant promise for enhancing strategy-making processes and paving the way for fine-tuning an industrial orientated Language model.

Table 3. Comparative Performance Analysis of Bert and Mosaic Bert Models on Datasets of Varied Sizes and Training Durations. Note that while the training duration and epoch settings can be further optimized, this table provides an initial assessment of model performance

MODEL	ACCURACY	DURATION
Small Dataset training for 3 epochs		
BERT	0.56	34 MINUTES
MOSAICBERT	0.55	34 MINUTES
Large Dataset training with different duration		
MOSAICBERT_3EP	0.62	5.7H
MOSAICBERT_6EP	0.64	11H
MOSAICBERT_286MSP	0.63	24H



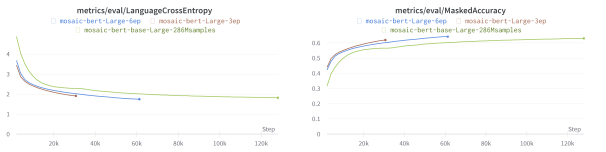
(a) Loss (b) MaskedAccuracy (c) Language-CrossEntropy

Figure 12. Model Training Metrics for large dataset

Moreover, our findings shed light on various key aspects, including evaluating the multilingual dataset by benchmarking it against an English dataset, efficiently sampling events through generative models, and highlighting the advantages of pretraining models with in-house data sources. These insights collectively contribute to a deeper understanding of how to harness NLP techniques for advanced financial data analysis and model development.

However, it is crucial to acknowledge the inherent limitations of this research. While we have demonstrated various methods and analyses across multiple tasks, certain key aspects deserve attention. Specifically, we have not presented back-test results that compare the performance of sentiment analysis on Chinese-language headlines to that of English-language headlines. Additionally, we have not proposed a method for weighting news headlines based on their association with the detected volume spike topics. Furthermore, it's essential to note that, as of the submission deadline, the BERT model has not been fully evaluated in the context of a sentiment analysis task applied to a financial corpus. These are avenues that remain to be explored and merit further investigation. Empirical testing to understand the relationships between different hyperparameters such as epochs, batch size, and learning rate is also an important aspect that requires attention and further exploration.

Thus, there are questions that demand further investigation. How can news articles be weighted according to their impact on financial returns? Is it possible to anticipate



(a) LanguageCrossEntropy (b) MaskedAccuracy

Figure 13. Model Evaluation Metrics for large dataset

market-moving news events using pre-trained models? Are there multilingual models that can replicate the effectiveness observed in English within the financial domain? And, importantly, to what degree can domain-specific language models contribute to the enhancement of market sentiment prediction?

These questions underscore the importance of delving deeper into the realm of leveraging large, diverse datasets and using NLP skills to create high-quality predictive signals for hedge fund.

References

Araci, D. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

Artetxe, M. and Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464, 2018. URL http://arxiv.org/abs/1812.10464.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

Deng, X. and Nan, P. cntext: a python tool for text mining, 9 2022. URL https://github.com/hiDaDeng/cntext.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Krishna, K., Garg, S., Bigham, J. P., and Lipton, Z. C. Downstream datasets make surprisingly good pretraining corpora, 2023.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. *CoRR*, abs/2107.06499, 2021. URL https://arxiv.org/abs/2107.06499.

Linger, M. and Hajaiej, M. Batch clustering for multilingual news streaming. *CoRR*, abs/2004.08123, 2020. URL https://arxiv.org/abs/2004.08123.

Loughran, T. and McDonald, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x.

Mou, C. text-dedup github repository. GitHub, 2023. URL https://github.com/ChenghaoMou/text-dedup.

Portes, J., Trott, A., King, D., and Havens, S. Mosaicbert: Pretraining bert from scratch for $20, 2023. URL https://www.mosaicml.com/blog/mosaicbert.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation, 2022.

Shazeer, N. Glu variants improve transformer, 2020.

Yin, W., Hay, J., and Roth, D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161, 2019. URL http://arxiv.org/abs/1909.00161.